International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:12, 2016

A Survey of Semantic Integration Approaches in Bioinformatics

Chaimaa Messaoudi, Rachida Fissoune, Hassan Badir

***Abstract*—**Technological advances of computer science and data analysis are helping to provide continuously huge volumes of biological data, which are available on the web. Such advances involve and require powerful techniques for data integration to extract pertinent knowledge and information for a specific question. Biomedical exploration of these big data often requires the use of complex queries across multiple autonomous, heterogeneous and distributed data sources. Semantic integration is an active area of research in several disciplines, such as databases, information-integration, and ontology. We provide a survey of some approaches and techniques for integrating biological data, we focus on those developed in the ontology community.

***Keywords*—**Semantic data integration, biological ontology, linked data, semantic web, OWL, RDF.

I. Introduction

NOWADAYS, new technologies have emerged and revolutionized biological and biomedical research as advances in sequencing and mass spectrometry techniques. The emergence of these methods able to generate large amounts of data qualified as raw data that aims to obtain them faster with consequence, the exponential growth of data generated. All these data were rapidly stored in banks (or sources) of data. Several data sources have been developed to allow researchers to share and reuse data in the life sciences. Researchers often need to query various data sources to solve complex biological problems. This can be difficult; different data sources may assign the same name to distinct high-level concepts. These data sources are both distributed and heterogeneous: Each source has its own data format and its own structure, and it is common that the scientific terms used to describe the data differ from one source to another. And these semantic incompatibilities may create opportunities for the propagation of misinformation.

Semantic Web technologies have been proposed as a solution to data integration problems because they present formally defined semantics, make it possible to track data provenance, and support semantically rich knowledge representations. The World Wide Web Consortium (W3C) provides a set of standards to facilitate the representation, publication, linking, querying and discovery of heterogeneous knowledge using web infrastructure [1], including Extensible Markup Language (XML), Resource Description Framework (RDF) and RDF Schema (RDFS), the Web Ontology Language (OWL). The W3C proposes RDF as the standard model for data interchange on the Web. Recently, many research groups have endeavored to integrate data effectively from multiple resources in diverse specific domains such as immunology [2], Maritime [3], agronomic [4] and cyber security [5], using Semantic Web technologies and ontology. Also authors in [6], [7], explain and state ontology-based Approaches in general.

The goal of this paper is to provide an overview of semantic data integration approaches. We consider the challenges of information integration in biology from the prospective of researchers using information technology as an integral part of their discovery process. Specifically, Semantic-based technologies, such as ontologies that offer a proven method to exploit expert-based knowledge in the analysis of large datasets. We will start by explaining all necessary vocabulary related to Semantic-based technologies in Section II and III presents a review of biological data integration systems. Finally, in Section IV, we will present remarks with a discussion regarding these systems.

C. Messaoudi, R. Fissoune and H. Badir are with the Abdelmalek Essaadi University, National School of Applied Sciences, LABTIC, Tangier, 90 000, Morocco (e-mail: messaoudi.chaimaa@gmail.com, ensat.fissoune@gmail.com, hbadir@ensat.ac.ma).

II. Ontologies and Semantic Web

A. Ontology Basics

The concept of ontology is used in very different areas such as philosophy, linguistics or artificial intelligence. In philosophy, the ontology is a fundamental branch of metaphysics that deals with the notion of existence, the fundamental categories of existing and studies the most general properties of being. The first definition of ontology concept in computer science is proposed by Gruber [8] as a formal, explicit specification of a shared conceptualization. In this definition **conceptualization** refers to an abstract model of some domain knowledge in the world, which identifies the relevant concept in the domain, a model is a way of describing the important aspects of a domain while simplifying or omitting less important or irrelevant aspects. Models can be tools for communication, for analyzing or explaining observations, for predicting future developments, and can provide a framework for integrating data from different sources. **Shared** indicates that an ontology captures consensual knowledge; that is accepted by a group. **Explicit** means that the type of concepts in an ontology and the constraints on these concepts are explicitly defined. Finally, **formal** means that the ontology should be machine understandable. Authors in [9] considered an ontology to be an area of knowledge that is formalized, such that the individual terms (or concepts) are defined by a set of assertions that connect them to other terms.

B. Semantic Web Languages

Semantic web, proposed by Tim Berners Lee in 2001, is broadly accepted in biological research. The Semantic web

uses **Resource Description Framework (RDF)** a graph-based language in which resources are identified through their internationalized Resource Identifier (IRI) and statements take the form of triples (subject-predicate-object). Therefore, a set of RDF statements forms a labeled directed graph. RDF also comes with a predefined vocabulary that can be used to state the type of a resource (e.g. a class, or a literal) or represent relations between resources (e.g. labels of resources, subclass relations between resources). For example, (Book, name, Bioinformatics Concepts) describes a resource Book whose name is Bioinformatics Concepts. In the meantime, OWL is further expressive than RDF by additionally enabling reasoning and inference in a domain of interest [10].

Web Ontology Language (OWL) [11] is a language based on description logic and has a formal, model-theoretic semantics. Several sub-languages of OWL have been developed, including OWL-DL, OWL-EL, OWL-RL, OWL-QL and OWL Full, which support different language constructs, have different properties regarding decidability and complexity of reasoning tasks, and therefore different areas of application.

SPARQL Protocol and RDF Query Language [12] is a standardized NoSQL query language, which can be used to query RDF databases and supports query federation (i.e. querying data distributed across multiple databases). SPARQL can also be used to query other kinds of data, including relational databases and flat files.

Linked Data [13] represents a method of publishing and sharing data on the web. When publishing Linked Data sets, data items are identified through a URI, and links to other data items are included in the data set by explicitly referring to the URI that denotes the other items. The URIs used to denote data items should be dereferencable, i.e. it should be possible to obtain additional information about the item through the URI (depending on the method used to access the URI, the information could be presented as HTML, RDF, JavaScript Object Notation or similar).

The OBO (Open Biomedical Ontologies) Flatfile Format [14] is a graph-based knowledge representation language widely used for biological and biomedical ontologies. The majority of language constructs are compatible with OWL, and bi-directional transformations between the OBO Flatfile Format and OWL have been implemented.

C. Ontology and Bio-Ontologies

Ontologies are used in several contexts such as e-commerce, and World Wide Web (WWW) in order to organize, analyze, search or integrate data. But what makes bio-ontologies so special? These days, the term 'bio-ontologies' allows integration and exploration of scientific data. Firstly, biomedical ontologies describe biological research or medical data, and one of the most important bio-ontology languages, OBO, was designed specifically for the needs of biological research. The bio-ontologies are tools for annotation and integration of data that allow a large number of researcher using a common vocabulary to describe and communicate their results and give the bioinformatics tools for functional analysis of microarrays data, mass spectrometry data, semantic similarity for biological analysis and clinical diagnostics, as well as many other applications. The incorporation of bio-ontologies in data annotation systems enables the semantic integration of complex scientific data [15], facilitates the exchange of information between heterogeneous information systems and supports the consistency of data curation. The main features provided by ontologies to support the biological and biomedical research are [16]: Classes and relations, Domain vocabulary, Metadata and descriptions, Axioms and formal definitions. Combining the four main features of ontologies facilitates semantic integration of heterogeneous, multimodal data within and across domains, and enables novel data mining methods that span traditional boundaries between domains and data types . The use of standard identifiers for classes and relations in ontologies is what enables data integration across multiple databases because the same identifiers can be used across multiple, disconnected databases, files, or web sites. Each term in the ontologies that are associated with the OBO has an ID that has two components: A letter code that specifies the ontology type and a number. For example, PR:000025257 represents a heat shock protein 105 kDa that is encoded in the genome of mouse in the PRotein Ontology (OBO): the ontology type is defined by the prefix PR and the number represents a unique entity in the PR ontology. IDs can be used in two ways: to link a biological database to ontologies and to connect different biological databases (interoperability).
The most important ontologies that can be used to report proteomics experiments are listed in Table I. They are used by the XML-based proteomics standards defined by the HUPO PSI working groups and some of them can of course be used in other biological disciplines.

III. Biological Data Integration

A. The Data Integration Problems

Due to the wide variety of sources, query them and exploit the wealth of information they contain is a complex task because it is facing enormous constraints. We can group these problems into three types of conflicts, including technological, syntactic and semantic.

1) Technological Level: The problems considered at the technical level are related to the interconnection of systems as diverse and complex and to the various formats of data exchange.

- **Diversity of data access:** The data access protocols are varied CGI/HTTP, FTP.
- **Variety of services and tools:** The sources propose tools able to search some data properties (often, these tools are used to return a source of data that is similar to experimental data presented to the input). A high diversity is present through these tools: each source has one or more variants of the same tool; Furthermore, the user has very rarely a complete description of the handled tool.

2) Syntactic Level: Syntactical conflicts are related to diversity and the multiplicity of models (structured, semi-structured, unstructured) and data formats.

TABLE I
IMPORTANT ONTOLOGIES USED IN THE PROTEOMICS FIELD

Ontology	Function	Reference	Website (accessed 4/2016)
Gene ontology(GO)	An ontology for describing the function of genes and gene products	[17]	http://purl.obolibrary.org/obo/go.owl
PSI-Molecular Interactions(MI)	A structured controlled vocabulary for the annotation of experiments concerned with protein-protein interactions	[18]	http://purl.obolibrary.org/obo/mi.owl
Chemical entities of biological(CHEBI)	A structured classification of molecular entities of biological interest focusing on 'small' chemical compounds	[19]	http://purl.obolibrary.org/obo/chebi.owl
PSI-Protein modifications(MOD)	An ontology consisting of terms that describe protein chemical modifications	[20]	http://purl.obolibrary.org/obo/mod.owl
Ontology for Biomedical Investigations(OBI)	An integrated ontology for the description of life-science and clinical investigations	[21]	http://purl.obolibrary.org/obo/obi.owl
Brenda tissue(BTO)	A structured controlled vocabulary for the source of an enzyme comprising tissues, cell lines, cell types and cell cultures	[22]	http://purl.obolibrary.org/obo/bto.owl
PRotein Ontology(PRO)	An ontological representation of protein-related entities in three major areas: proteins related by evolution; proteins produced from a given gene; and protein-containing complexes	[23]	http://purl.obolibrary.org/obo/pr.owl
Phenotypic qualities (properties)(PATO)	An ontology of phenotypic qualities (properties, attributes or characteristics)	-	http://purl.obolibrary.org/obo/pato.owl
Units of measurement(UO)	Metrical units for use in conjunction with PATO comprising tissues, cell lines, cell types and cell cultures	[24]	http://purl.obolibrary.org/obo/uo.owl
PSI-Mass Spectrometry(MS)	A structured controlled vocabulary for the annotation of experiments concerned with proteomics mass spectrometry	-	http://purl.obolibrary.org/obo/ms.owl
PSI-Sample Processing and Separations(SEP)	A structured controlled vocabulary for the annotation of sample processing and separation techniques in scientific experiments (gel electrophoresis, column chromatography...)	-	http://purl.obolibrary.org/obo/sep.owl
Cigarette Smoke Exposure Ontology(CSE)	A structured controlled vocabulary for systems toxicology	[25]	http://purl.bioontology.org/ontology/CSEO
eNanoMapper(ENM)	The eNanoMapper ontology covers the full scope of terminology needed to support research into nanomaterial safety	[26]	http://purl.bioontology.org/ontology/ENM

- **Diversity of data syntax (models and data formats):** Data models are different depending on the source: we find the relational model (in ensEMBL), the object model (often encountered in the case of warehouses as GEDAW [27], semi-structured XML models (in UniProt).
- **Diversity of query languages:** It follows from the preceding paragraph that the sources have different query languages. The query language of a database (such as PubMed / Medline, GenBank ...) is often a simple combination of words to search in the texts while relational databases can for example, be queried in SQL.

3) Semantics Level: Semantic conflicts are due to the presence of data from several sources, subject to different interpretations depending on the local context used (application domain). They manifest themselves in the way of denominate information causing terminological conflict (synonyms, homonyms, polysemy ...) and therefore causing misunderstandings between applications using taxonomies or using different data patterns.

B. The Data Integration Systems

The integration of data sources with complex structures and semantics, has become a very important area of research because of the explosion in the number and heterogeneity of data sources. In [28], the authors conducted five processes for semantic data integration that solve seven core problem. These processes include making explicit the differences between biomedical concepts and database records, aggregating sets of identifiers denoting the same biomedical concepts across data sources, and using declaratively represented forward-chaining rules to take information that is variably represented in source databases and integrating it into a consistent biomedical representation and demonstrate these processes and solutions by presenting KaBOB (the Knowledge Base Of Biomedicine), a knowledge base of semantically integrated data from 18 prominent biomedical databases using common representations grounded in Open Biomedical Ontologies. The importance and utility of use of RDF knowledge bases (KBs) in biomedicine have also been demonstrated.

GPKB [29], software architecture to create and maintain a Genomic and Proteomic Knowledge Base , which integrates several of the most relevant sources of such dispersed information (including Entrez Gene, UniProt, IntAct, Expasy Enzyme, GO, GOA, BioCyc, KEGG, Reactome and OMIM). This solution is general, as it uses a flexible, modular and multilevel global data schema based on abstraction and generalization of integrated data features, and a set of automatic procedures for easing data integration and maintenance, also when the integrated data sources evolve in data content, structure and number. These procedures also assure consistency, quality and provenance tracking of all integrated data, and perform the semantic closure of the hierarchical relationships of the integrated biomedical ontologies.

An ontological foundation for the Bio2RDF linked data is provided in [30] for the life sciences project and is used for semantic integration and discovery for SADI-based semantic

web services. SIO is freely available to all users under a creative commons by attribution license. See website for further information: http://sio.semanticscience.org. Ontodog [31] a web-based system that can generate an ontology subset based on Excel input, and support generation of an ontology community view, which is defined as the whole or a subset of the source ontology with user-specified annotations including user preferred labels. Ontodog allows users to easily generate community views with minimal ontology knowledge and no programming skills or installation required, accessible in http://ontodog.hegroup.org/.

More recently, in [32], the authors explored the potential of Semantic Web Technologies as a means of integration and development of Big Data applications. Specifically, they proposed the leveraging of Industrial Ontologies for the purpose of supporting connections between disparate data sources. In [33] the authors, have developed a framework based on a semantic mediator for environmental data analysis, where a web server is hosted to receive the stSPARQL queries from a user in the form of a request in a Web browser. This framework consists of four parts: the data translation, temporal relation inference, triplestore bulk load, and data preparation and visualization.

NoSQL stores are emerging as an efficient alternative to relational database management systems in the big data context. The authors in [34], thought to apply this alternative in the context of ontology based data access (OBDA) and show that OBDA is even more needed in the NoSQL ecosystem cause it provides a semantic conceptual schema over a repository of data and, due to its logical formalism, it is likely to support formal analysis, optimization and reasoning. In order to illustrate their approach, they present a medical social application which stores and processes patient information concerning their diseases, allergies, and drug prescriptions. The architecture is composed of three layers: query, semantic and storage. The Storage layer is composed of standard NoSQL databases. The Semantic layer is the cornerstone of this research and is dealt with schema features and integrity constraints. In this architecture, an end-user writes a SPARQL query which is sent to the OBDA system. They also proposed a mapping solution between a relational schema and a set of Nosql stores/RDBMS in [35].

SoFIA [36], a framework for workflow-driven data integration with a focus on genomic annotation. SoFIA conceptualises workflow templates as comprehensive workflows that cover as many data integration operations as possible in a given domain. An Omics data integration framework for annotating high throughput data sets. Available in https://github.com/childsish/sofia/-releases/latest under the GNU General Public License. Semantic annotation of ncRNA data lag behind their identification, and there is a great need to effectively integrate discovery from relevant communities. Identification of non-coding RNAs (ncRNAs) has been significantly enhanced due to the rapid advancement in sequencing technologies. Also, in [37] the Non-Coding RNA Ontology (NCRO) is being developed to provide a precisely defined ncRNA controlled vocabulary, which can fill a specific and highly needed niche in unification of ncRNA biology.

TABLE II
SOME BIOLOGICAL DATA INTEGRATION SYSTEMS

Category	Reference	Year
Data Warehouse	COLUMBA [44]	2005
	BioWarehouse [48]	2006
	[47]	2008
	GEDAW [40]	2008
	BioMart [43]	2011
	[41]	2016
	SoFIA [36]	2016
Linked open data and Semantic Web	[38]	2010
	[34]	2013
	Ontodog [31]	2014
	[30]	2014
	[45]	2015
	Kabob [28]	2015
	[33]	2015
	GPKB [29]	2016
	[32]	2016
Workflow	[50]	2005
	[49]	2010
	Taverna [51]	2013
	SoFIA [36]	2016

In [38], building a biological ontology recommender web service aimed at facilitating data integration by use of ontologies for data annotation. With the challenge to figure out best suitable annotation for specific datasets, it used word-based documented metadata in a domain and suggested ontologies that were most suitable for annotating data. Ontologies were decided on base of three criteria naming coverage (most terms covering input text), connectivity (ontologies mapped to other ontologies) and size (number of concepts). Scores are then assigned to ontologies based on these. The single most important consideration in selecting a bio-ontology is to understand requirements first before deciding to engage with a particular ontology or indeed before minting one's own ontology, authors in [39] also provided in its article ten rules to select a bio-ontologies for biologists and bioinformatics. Starting with specifing the ontology domain , and ending with the tenth rule that says, sometimes an Ontology is not needed at all. In the warehouse approach, [27], [40] developed a warehouse according to a relatively small object model and containing expression data of liver genes in various pathophysiological conditions. The aim of the warehouse is to offer the most precise and complete as possible annotation for the expression of liver genes. Therefore, the warehouse is characterized by the quality of the data stored on it. In GEDAW, data is semantically integrated both in schemes and instances.

In [41], the authors have developed an integrated system that combines all stages of cancer studies, from gathering of clinical data, through elaborate patient questionnaires and bioinformatics tools, to data warehousing and preparation of analysis reports. The central data warehouse developed in the SysCancer project is used to support multidimensional analysis

from local databases after data earmarked for public access has been exported from them. Additionally, it is used as a gateway for the computational cluster, responsible for performing complex analyses of data using advanced algorithms that are accessible through a self-explanatory simple interface.

IV. Remarks and Discussion

Data integration for the life sciences is by no means a new topic [38], [30], [40], [28]. In bioinformatics, the data integration regime has been studied in [42] (see Table II), present approaches can be broadly categorized into three classes: So Called data warehouses are relational databases that integrate a selected set of data into a common schema [43], [44]. The warehouse approach is increasingly used in the biological field because it is extremely well adapted to some needs of domain (confidentiality, treatment control, full data cleansing). However, difficulties associated with maintaining a warehouse are several (updating the data warehouse compared to the data sources). Accessing the data in a data warehouse requires either the ability to program complex queries (usually in SQL), or the usage of specific point-and-click user interfaces encapsulating such queries. The latter solution is the only option when programming expertise is lacking, but is inflexible and involves costly interface development. A second class of data integration systems are based on linked open data and Semantic Web standards [45], [28], [30], [38], [29], [32], [33]. These offer more flexibility in terms of data modelling, but require efforts comparable to data warehousing for building semantically integrated data sets [46]. Both approaches perform data integration prior to any concrete analysis, which implies that they usually try to be as comprehensive as possible to cover unforeseen applications. Creating or updating this large integrated data set is highly complex and time consuming, increasing the danger of using outdated data [40], [27], [47], [48], [41]. More recently, a third class of systems has emerged that are based on flexible integration workflows [49], [50], [51]. In these approaches, data integration is performed by starting a pipeline of steps that are defined in advance by a workflow developer. Results of these workflows are typically directly consumed by the user or by other tools and not meant to be materialised in a persistent, maintained manner. Accordingly, every analysis uses the most recent data available. To be fast, these workflows are specialised; a drawback when no available workflow exactly meets the user's requirements. Either a new workflow has to be developed, or multiple workflows with potentially overlapping subtasks have to be executed, yielding inflexibility and unnecessary computation. What is lacking is a data integration method that, based on a formalized understanding of an application domain, is able to automatically determine the minimal complete sequence of steps required to fulfil a given user request starting from a given set of input data. Contrary to that, using SoFIA [36], workflow designers specify comprehensive workflow templates covering as much of a given application domain as possible, much like defining the process used to populate a data warehouse. However, these templates are not intended to be executed in their entirety. Instead, they should be understood as a formalised knowledge base of processes transforming various types of input data into various types of annotations using background knowledge.

V. Conclusion

In this paper, we have presented some recent systems that use biological ontologies to solve the problems involved in data integration. It intents help for the ontology-based data integration community giving different aspects of systems which have been used as a reference for further research. Many more have been left out: It was not feasible neither practical to include everything that has been done to date. Rather, we selected indicative examples that characterize a range of related works. But other several aspects have to be analyzed as the valuation of the architecture properties. The future applications of research comes down to Ontology-based semantic data integration that seems to be one of most promising approaches. The major challenge is to develop more automatic semantic data mining algorithms and systems by utilizing the full strength of formal ontology that has well defined representation language, formal semantics, and reasoning tools for logic inference and consistency checking.

References

[1] N. Shadbolt, T. Berners-Lee, and W. Hall, "The semantic web revisited," *IEEE intelligent systems*, vol. 21, no. 3, pp. 96–101, 2006.
[2] A. H. Asiaee, T. Minning, P. Doshi, and R. L. Tarleton, "A framework for ontology-based question answering with application to parasite immunology," *Journal of biomedical semantics*, vol. 6, no. 1, p. 1, 2015.
[3] G. Santipantakis, K. I. Kotis, and G. A. Vouros, "Ontology-based data integration for event recognition in the maritime domain," in *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*. ACM, 2015, p. 6.
[4] C. Jonquet, E. Dzalé-Yeumo, E. Arnaud, and P. Larmande, "Agroportal: a proposition for ontology-based services in the agronomic domain," in *IN-OVIVE'15: 3ème atelier INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement*, 2015.
[5] M. Iannacone, S. Bohn, G. Nakamura, J. Gerth, K. Huffer, R. Bridges, E. Ferragut, and J. Goodall, "Developing an ontology for cyber security knowledge graphs," in *Proceedings of the 10th Annual Cyber and Information Security Research Conference*. ACM, 2015, p. 12.
[6] H. Wache, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner, "Ontology-based integration of information-a survey of existing approaches," in *IJCAI-01 workshop: ontologies and information sharing*, vol. 2001. Citeseer, 2001, pp. 108–117.
[7] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," in *Semantic Computing (ICSC), 2015 IEEE International Conference on*. IEEE, 2015, pp. 244–251.
[8] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition, 5(2)*, pp. 199–220, 1993.
[9] B. L. Jonathan and S. Y. Rhee, "Ontologies in biology: Design application and future challenges." 2004.
[10] F. Manola and E. Miller, "Rdf primer," *World Wide Web Consortium*, 2004.
[11] B. Grau, I. Horrocks, and B. M. et al, "Owl 2: the next step for owl," *Web Semant*, vol. 6, pp. 309–322, 2008.
[12] A. Seaborne and E. Prud'hommeaux, "Sparql query language for rdf," *W3C Recommendation (W3C, 2008)*, 2008.
[13] C. Bizer, "Evolving the web into a global data space." in *BNCOD*, vol. 7051, 2011, p. 1.
[14] I. Horrocks, "Obo flat file format syntax and semantics and mapping to owl web ontology language," *University of Manchester*, 2007.
[15] J. Blake and C. Bult, "Beyond the data deluge: Data integration and bio-ontologies." *Journal of Biomedical Informatics*, pp. 314–320, 2006.

[16] R. Hoehndorf, P. Schofield, and G. Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective," *Brief. Bioinform, 16 (6)*, pp. 1069–1080, 2015.

[17] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K.Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology," *The Gene Ontology Consortium*, vol. Nat. Genet. 25, pp. 25–29, 2000.

[18] S. Orchard, "Molecular interaction databases," *Proteomics*, vol. 12, pp. 1656–1662, 2012.

[19] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner, "Chebi: a database and ontology forchemical entities of biological interest," *Nucleic Acids Res*, pp. D344–D350, 2008.

[20] L. Montecchi-Palazzi, R. Beavis, P. Binz, R. Chalkley, J. Cottrell, D. Creasy, J. Shofstahl, S. Seymour, and J. Garavelli, "The psi-mod community standard for representation of protein modification data," *Nat. Biotechnol.*, vol. 26, pp. 864–866, 2008.

[21] R. Brinkman, M. Courtot, D. Derom, J. Fostel, Y. He, P. Lord, J. Malone, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, L. Soldatova, C. S. Jr., J. Turner, and J. Zheng, "O.b.i. consortium,modeling biomedical experimental processes with obi," *Biomed. Semant.*, vol. (Suppl. 1), 2010.

[22] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, and D. Schomburg, "The brenda tissue ontology (bto): the first all-integrating ontology of all organisms for enzyme sources," *Nucleic Acids Res*, pp. D507–D513, 2011.

[23] D. Natale, C. Arighi, W. Barker, J. Blake, C. Bult, M. Caudy, H. Drabkin, P. D'Eustachio, A. Evsikov, H. Huang, J. Nchoutmboube, N. Roberts, B. Smith, J. Zhang, and C. Wu, "The protein ontology: a structured representation of protein forms and complexes," *Nucleic Acids Res*, vol. 39, pp. D539–D545, 2011.

[24] G. Gkoutos, P. Schofield, and R. Hoehndorf, "The units ontology: a tool for integrating units of measurement in science," *Database (Oxford)*, vol. 6, pp. D539–D545, 2012.

[25] E. Younesi, S. Ansaril, M. Guendel, S. Ahmadi, C. Coggins, J. Hoeng, M. Hofmann-Apitius, and M. C. Peitsch, "Cseo - the cigarette smoke exposure ontology," *Journal of Biomedical Semantics*, 2014.

[26] E. Friederike, L. Rieswijk, C. Evelo, H. Sarimveis, P. Doganis, G. Drakakis, B. Fadeel, B. Hardy, J. Hastings, C. Helma, N. Jeliazkova, V. Jeliazkov, P. Kohonen, R. Grafstrom, P. Sopasakisa, G. Tsiliki, and E. Willighagen, "Ontology, database and tools for nanomaterial safety evaluation," *Journal of Biomedical Semantics*, 2015.

[27] E. Guérin, G. Marquet, A. Burgun, O. Loréal, L. Berti-Equille, U. Leser, and F. Moussouni, "Integrating and warehousing liver gene expression data and related biomedical resources in gedaw," in *International Workshop on Data Integration in the Life Sciences*. Springer, 2005, pp. 158–174.

[28] K. M. Livingston, M. Bada, W. A. Baumgartner, and L. E. Hunter, "Kabob: ontology-based semantic integration of biomedical databases," *BMC bioinformatics*, vol. 16, no. 1, p. 1, 2015.

[29] M. Masseroli, A. Canakoglu, and S. Ceri, "Integration and querying of genomic and proteomic semantic annotations for biomedical knowledge extraction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, pp. 209–219, 2016.

[30] M. Dumontier, C. J. Baker, J. Baran, A. Callahan, and L. C. et al., "The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery," *Biomed Semantics*, vol. vol. 5, p. p. 14, 2014.

[31] J. Zheng, Z. Xiang, C. J. Stoeckert, and Y. Hel, "Ontodog: a web-based ontology community view generation tool," *Bioinformatics*, vol. vol. 30, pp. pp. 1340–1342, 2014.

[32] D. Ostrowski, N. Rychtyckyj, P. MacNeille, and M. Kim, "Integration of big data using semantic web technologies," in *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*. IEEE, 2016, pp. 382–385.

[33] B.-H. Tran, C. Plumejeaud-Perreau, A. Bouju, and V. Bretagnolle, "A semantic mediator for handling heterogeneity of spatio-temporal environment data," in *Research Conference on Metadata and Semantics Research*. Springer, 2015, pp. 381–392.

[34] O. Curé, F. Kerdjoudj, D. Faye, C. Le Duc, and M. Lamolle, "On the potential integration of an ontology-based data access approach in nosql stores," *International Journal of Distributed Systems and Technologies (IJDST)*, vol. 4, no. 3, pp. 17–30, 2013.

[35] O. Curé, R. Hecht, C. Le Duc, and M. Lamolle, "Data integration over nosql stores using access path based mappings," in *International Conference on Database and Expert Systems Applications*. Springer, 2011, pp. 481–495.

[36] L. H. Childs, S. Mamlouk, J. Brandt, C. Sers, and U. Leser, "Sofia: a data integration framework for annotating high-throughput datasets," *Bioinformatics*, p. btw302, 2016.

[37] J. Huang, K. Eilbeck, J. A. Blake, D. Dou, D. A. Natale, A. Ruttenberg, B. Smith, M. T. Zimmermann, G. Jiang, Y. Lin *et al.*, "A domain ontology for the non-coding rna field," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 621–624.

[38] C. Jonquet, M. A.Musen, and N. H. Shah, "Building a biomedical ontology recommender web service," *Biomed Semantics*, pp. 1–18, 2010.

[39] J. Malone, R. Stevens, S. Jupp, T. Hancocks, H. Parkinson, and C. Brooksbank, "Ten simple rules for selecting a bio-ontology," *PLOS Comput Biol*, vol. vol. 30, pp. 12(2), e1 004 743, 2016.

[40] E. Guérin, F. Moussouni, B. Courselaud, and O. Loréal, "Modélisation d'un entrepôt de donnes dédié à l'analyse du transcriptome hépatique," *Actes des Journées Ouvertes Biologie Informatique Mathématiques (JOBIM)*, vol. vol. 30, pp. pp 319–324, 2008.

[41] W. Bensz, D. Borys, K. Fujarewicz, K. Herok, R. Jaksik, M. Krasucki, A. Kurczyk, K. Matusik, D. Mrozek, M. Ochab *et al.*, "Integrated system supporting research on environment related cancers," in *Recent Developments in Intelligent Information and Database Systems*. Springer, 2016, pp. 399–409.

[42] C. Goble and R. Stevens, "State of the nation in data integration for bioinformatics," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 687–693, 2008.

[43] A. Kasprzyk, "Biomart: driving a paradigm change in biological data management," *Database*, vol. 2011, p. bar049, 2011.

[44] S. Trißl, K. Rother, H. Müller, T. Steinke, I. Koch, R. Preissner, C. Frömmel, and U. Leser, "Columba: an integrated database of proteins, structures, and annotations," *BMC bioinformatics*, vol. 6, no. 1, p. 1, 2005.

[45] C. M. Machado, D. Rebholz-Schuhmann, A. T. Freitas, and F. M. Couto, "The semantic web in translational medicine: current applications and future directions," *Briefings in bioinformatics*, vol. 16, no. 1, pp. 89–103, 2015.

[46] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop *et al.*, "Why linked data is not enough for scientists," *Future Generation Computer Systems*, vol. 29, no. 2, pp. 599–611, 2013.

[47] T. Jörg and S. Deßloch, "Towards generating etl processes for incremental loading," in *Proceedings of the 2008 international symposium on Database engineering & applications*. ACM, 2008, pp. 101–110.

[48] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. W. Stringer-Calvert, J. D. Tenenbaum, and P. D. Karp, "Biowarehouse: a bioinformatics database warehouse toolkit," *BMC bioinformatics*, vol. 7, no. 1, p. 1, 2006.

[49] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the ensembl api and snp effect predictor," *Bioinformatics*, vol. 26, no. 16, pp. 2069–2070, 2010.

[50] B. Giardine, C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor *et al.*, "Galaxy: a platform for interactive large-scale genome analysis," *Genome research*, vol. 15, no. 10, pp. 1451–1455, 2005.

[51] K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher *et al.*, "The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud," *Nucleic acids research*, p. gkt328, 2013.