ISSN 0147-6882, Scientific and Technical Information Processing, 2013, Vol. 40, No. 4, pp. 253–263. © Allerton Press, Inc., 2013.
Original Russian Text © A.O. Erkimbaev, V.Yu. Zitserman, G.A. Kobzev, V.A. Serebrjakov, K.B. Teymurazov, 2013, published in Nauchno-Technicheskaya Informatsiya, Seriya 1, 2013, No. 12, pp. 1–11.

Publishing Scientific Data as Linked Open Data

A. O. Erkimbaev[a], V. Yu. Zitserman[a], G. A. Kobzev[a], V. A. Serebrjakov[b], and K. B. Teymurazov[b]

[a] *Joint Institute for High Temperatures, Russian Academy of Sciences, Moscow, Russia*
[b] *Dorodnicyn Computing Center, Russian Academy of Sciences*
e-mail: adilbek@ihed.ras.ru, vz1941@mail.ru, gkbz@mail.ru, serebr@ultimeta.ru, kbt@intring.ru

Received July 29, 2013

Abstract—The concept and computer methods for the implementation of new form-linked data technologies for scientific communication are observed. This concept is being advanced by the Internet creator Tim Berners–Lee in the framework of the general idea of the Semantic WEB with the introduction in internet pages of descriptions that are understood by computers. An overview of the technologies and tools that enable online publishing of open data so that they are automatic linked with thematically related resources is given. Many advantages of the new form of publications in the field of science, viz., the integration of heterogeneous data, access from publications to the original arrays of raw data, and to software, as well as standardization of terms and concepts of the application domain through an appeal to ontologies and vocabularies posted on the web are shown.

Keywords: data integration, linked data, database, Semantic Web
DOI: 10.3103/S014768821304014X

INTRODUCTION

In recent decades the accelerated development of information technologies has had an extremely strong impact on the world of scientific information, especially on the storage and dissemination of data and knowledge. The determination of computer science as an independent direction in a number of disciplines (for example, geo-and bioinformatics), the universal use of bibliographic and factual databases (DBs), the transition from print to electronic forms of publication, the active use of the Internet to disseminate information, and many other issues may be noted in this context. The creation of the special term *e*-Science, which means the predominant role of information and its processing in scientific research, reflects new trends. Up to a point the computer expansion was mostly technical in nature, which led to a large increase in the flow of information and simplified the methods of data distribution and data-access methods. Moreover, the general transition to electronic resources has generated new problems of so-called *interoperability*, i.e., of ensuring the data format and structure conformity in the sphere of heterogeneous sources.

The prospects of a quantum leap emerged with the advent in 2001 of a principally new concept, viz., the *Semantic WEB*, which was suggested by the Internet creator Tim Berners–Lee [1]. This concept means a view of the network`s future where a certain structure that allows software agents to understand the meanings of pages and to carry out instruction of users bring in the contents of these pages. In interacting in the network, the agents would have to have a formal representation of the knowledge for each resource. The introductory role of the general, explicit, and formal specification of knowledge is given by the authors of [1] to ontologies. Ontologies, which are regarded as the genuine core of the Semantic WEB, are a system of concepts in the application domain, which is represented as a set of entities connected by various relationships. An ontology represents knowledge in the form of a formal structure that is available for computer processing. In 2004, a universal standard for network exchange of ontological information, viz., Web Ontology Language (OWL) was proposed by the World Wide Web Consortium (W3C).[1]

With OWL domain-application experts and application developers can create, modify, and combine different ontologies.

[1] W3C is an international consortium that was formed in 1994 as part of CERN. The purpose of its creation was the development of common protocols that enhance the interoperability of WWW resources, as well as a guide to WEB evolution. The consortium is developing recommendations on new technologies, as well as specifications on the status of the standard; it supports a vast repository of documents about developed and accepted standards and prototypes of tools and of applications that demonstrate the use of new technologies. A detailed description of the W3C in Russian can be found in the reference book by M.R. Kogalovsky [2].

Although the conceived idea of the Semantic WEB refers to the resources of any subject (business, art, politics, etc.), it is exactly natural sciences that form the most suitable "platform" for the development of new concepts. The conceptual foundation of such sciences as physics, chemistry, and astronomy is initially formalized enough to be put into the basis of an ontological description. This opens up the theoretical possibility of scientific data integration by specifying of the contents of heterogeneous sources. Automated text processing, high-precision information searching, and logical inference tools are among other features of the Semantic WEB. Concrete examples of data integration based on the physical properties of materials using a new concept were considered in our paper [3]. This points to the high potential of this concept. At the same time, despite some successes, we cannot say that the idea of the Semantic WEB has become widely used at present.

In 2006, Tim Berners–Lee [4] proposed an idea that shifts the focus to the publication of so-called "linked open data," which is a product and a development of the Semantic WEB concept. The movement for "open data," particularly in policies, administration, finance, and so on [5], which arose around the same time, is an another source of this idea. The governments of a series of countries, The World Bank, United Nations, the Partnership in the field of renewable energy sources (REEEP), and many other organizations (for a detailed review, see [6]) are among the organizations that make a practice of open publishing of their data.

Data are recognized as "open" if they are available for society where they meet the following principles: completeness, freedom of distribution without any restriction in the form of copyrights, patents, or other control mechanisms. However, open data achieves its full potential when data are converted into the linked open data by presenting them in a special RDF format (Resource Description Framework, see below) to identify the elements that they contain. The path from being "open data" to "open linked data" has been described by Tim Berners–Lee [7] in the form of a "five-star construction," where new options for data presentation are opened on each floor:

★ information is available on the WEB in any format under an open license;

★★ information is available in the form of structured data;

★★★ the use of open formats (for example, CSV instead of Excel) is allowed;

★★★★ URI identifiers that allow one to display them through a browser as individual data are used for all objects;

★★★★★ data is connected with other data forming a single context.

Publication in the sphere of "linked data" provides publishers and information consumers significantly more capabilities in comparison with the simple data placement in the traditional **Web of Documents**. Unlike hypertexts, where links connect single documents written in HTML, linked data technology provides communication between random items that are distinguished in the document by URI, which could identify any object, person or concept. This medium, which extends the capabilities of the **Web of Documents**, is called the **Web of Data** or **Web 3.0**. Combining structured documents in a standard way, this medium represents, from the point of view of the user, a huge database with the same efficient search capabilities for relevant information as in a usual database.

The principles of "linked data," which were first proposed in [4], provided the guidelines for publishers, who have started to master the new technology. Technical documents have been created that regulate publishing practice in the medium of "linked data" and special tools, such as browsers and search engines, provide the same features of navigation in the ordinary **Web of Documents** during work with the **Web of Data**. Even today, a large variety of linked data is available in the network. So-called LOD (Linked Open Data) clouds cover more than 50 billion entries from a variety of areas such as geography, mass media, biology, chemistry, economy, and energy. The www.reegle.info portal is one of the best and most accessible examples of the use of new technology in the field of science. This portal provides automatic partitioning of documents that relating to renewable energy, energy efficiency, and climate-change problems. In addition to supporting the thematic cloud, the portal allows a concise study of linked data technology, especially the RDF model description, dictionaries and ontology references, rules for binding documents, and so on. Next, we will make a detailed examination of the main principles and standards that are used when linking data in the network and we will give the most impressive examples of the new technological possibilities in the distribution and integration of natural science data.

BASIC ELEMENTS: FORMATS, STANDARDS, AND ONTOLOGIES

T. Berners–Lee, who proposed the idea of Linked Open Data [4], proposed four basic principles on which basis the data must be created and distributed in the network:

(1) for each entity that is included in data, such as a person, document, abstract concept, etc. a unique identifier called a URI (Uniform Resource Identifier) must be specified and used;

(2) for access to this entity in the network it is sufficient to use an appeal to HTTP URI (HTTP-Hypertext Transfer Protocol);

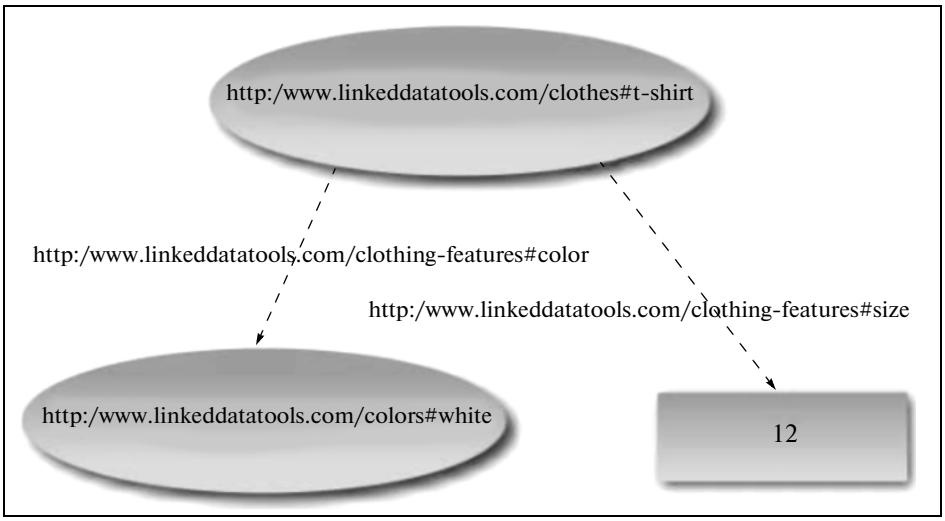

Fig. 1. A typical display of triplets [9].

(3) useful information by means of the RDF and SPARQL standards is provided by applying to the URI;

(4) the inclusion of links to other URIs allows one to find additional information that was not contained in the original document, such as information about the belonging of a person to a specific organization or information about the availability of related information on the subject of the original document.

These principles require the identification of each resource via an HTTP scheme based on a URI, which not only serves as its identifier but provides a representation of structured data. Hyperlinks between entities in different LOD documents are set by identifiers. Thus, binding of data in the network space takes place, which allows LOD applications to discover new sources of data, e.g., unknown authors as of the date of publication.

Thus, the LOD concept is based on three technologies, each of which is supported by the standards of W3C: HTTP, URI, and RDF. The third of these [4], RDF, which is also supported by the W3C standard, is proposed as a unified model of linked data. In fact, RDF is a definite model for the presentation of data and metadata, consisting of statements that are suitable for machine use. Each of these statements has the formal form "subject–predicate–object" and is called a *triplet*. Two examples taken from the *reegle* site guide [8] illustrate the meaning and the rules for writing a triplet:

Examples of RDF triplets

Subject	Predicate	Object
http://reegle.info/actors/2354	http://xmlns.com/foaf/0.1/name	"REEP"
actors:2354	foaf:name	"REEP"

In both examples some *performer* (actor) unambiguously defined by a URI is understood under the subject. The corresponding URI is also attributed to a predicate, i.e., to a property of the subject. In the given examples the *name* is this property. Finally, a so-called *literal* (text line) is used to indicate the object in these examples. The assertion in both examples is that performer 2354 has the name "REEP" (renewable energy and energy efficiency partnership). Abbreviation in the bottom row of the table by using the prefixes allows the subject and the predicate to use the short-cut notation of a triplet. The general rules for notation in the RDF format allow one to use an object as a literal and an identifier. In this example, the corresponding URI (www.reeep.org) could serve instead of "REEP."

In the RDF data model directed arc (predicate) connects the nodes relevant to subject and object. The graph in Fig. 1 comprises two triplets (*with statements about the color and size of a t-shirt*) [9], where the URI identifies the subject and both predicates, and both URI, and a literal are used for objects. With common identifiers in different triplets the computer connects

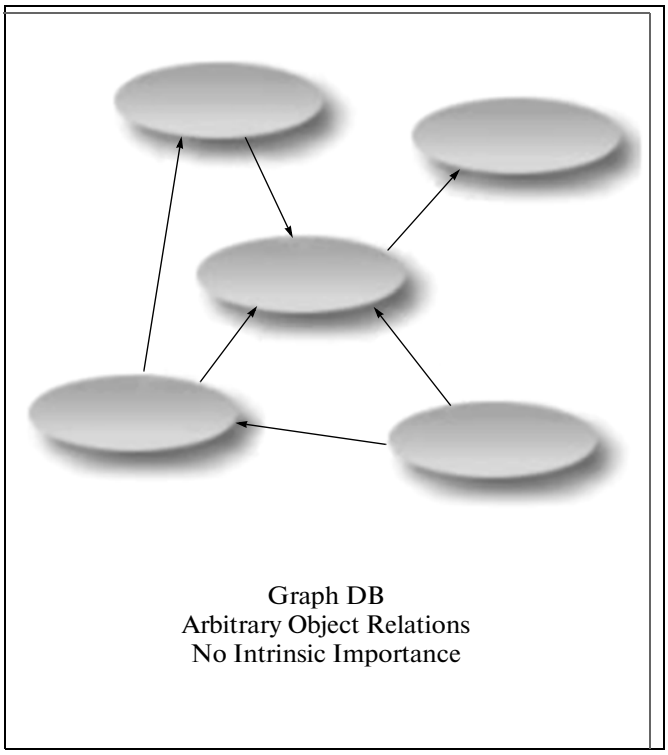

Fig. 2. A graphical presentation of a RDF model [9].

them into a single graph (see Fig. 1) doing this automatically, i.e., without the creator of the original document. Binding of data (an RDF link) is done via the triplet, whose subject and object refer to different sources.

As can be seen from Fig. 2, in an RDF presented as a graphic data pattern [9] binding of linked resources occurs without the explicit allocation of any of them as one of the initial ones (baseline). Each of the RDF triples is an element of a global medium of linked data and can serve as a reference point for its viewing. Thus, **RDF** is data model, which serves for the description of WEB resources that is designed to provide the perception of this data by a computer. In this case, RDF only provides a means to build the model, but in no way discloses the semantics (meaning) of the statements about resources. The interpretation of the resources content defined by a URI or by a *literal* string is possible due to the references to available online dictionaries and ontologies, which contain summaries of subject-oriented terms. Ontology, in addition, reserves the terms to identify concepts and their relationships. Among ontologies a wide field of following applications may be noted: **FOAF** (Friend of a Friend) for describing of current agents (persons or organizations), **SKOS** (Simple Knowledge Organization System) for category allocation, and **DBPedia properties** for different attributes. The term **Dbpedia file record**: **document** means that the concept of a "document" is interpreted in the wide **dbpedia** ontology, which was created to describe the structured information from Wikipedia. Within a specific subject area local schemes handle, for example, http://reegle.info/schema.rdf for renewable resources, so that the syntax for concept recording (for example, "the result of the project") has the form: **reegle: ProjectOutput**.

Thus, the compound of data model graphs provided by RDF or with specially created vocabularies and ontologies provides the foundation for all of the linked data publication technology. Figure 3 from [10] provides a simple view of such binding based on the example of two scientific articles indexed in the database **agris** under the conventional number **CH…179** and **CN…389**. Triplets disclose the data of each article indicating the author, title, etc. The same theme (subject) of the two publications is identified in the **agrovoc** dictionary under the same ID c_4416, which leads to the automatic binding of the two resources, although initially this was not intended and they were written in different languages. This technology is successfully applied in dealing with distributed knowledge through the automatic linking of RDF files placed in the network by any author, followed by the opportunity to find information in the assembled document that did not exist in any of its parts.

As noted, RDF allows the building of data models without touching the semantics itself and referring to the interpretation of the meaning of the data to network dictionaries and ontologies. The real mechanism that allows the use of dictionaries and attachment of RDF-data semantics consists in semantic annotation of metadata through a special syntax called RDF Schema or RDF Vocabulary Definition Language. The descriptions of relevant dictionaries are recorded themselves on RDF, representing the dictionary determination in the form of dictionary graphs and permitting their publication in the medium of Linked Open Data. In general, RDF Schema is no more than a semantic generalization of RDF that provides a framework for dictionary description, i.e., for object-oriented classes and features. To reflect the semantics, classes and properties are inserted in the same way as it is usually done in a language of object-oriented programming, such as Java. The only difference is that instead of a class determination in terms of the properties of its specimen, RDF Schema describes features in terms of classes of resources, to which this schema is applied. The language defined by the specification of RDF Schema is composed of a collection of RDF resources that can be used to describe the properties of other RDF resources in object-oriented dictionaries. The relevant document has the status of recommendations prepared by the W3C group in 2004. [11].

The SPARQL query language (www.w3.org/TR/rdf-sparql-query) is one of the fundamental W3C standards. This designation is a recursive acronym from the English **SPARQL** Protocol and RDF Query

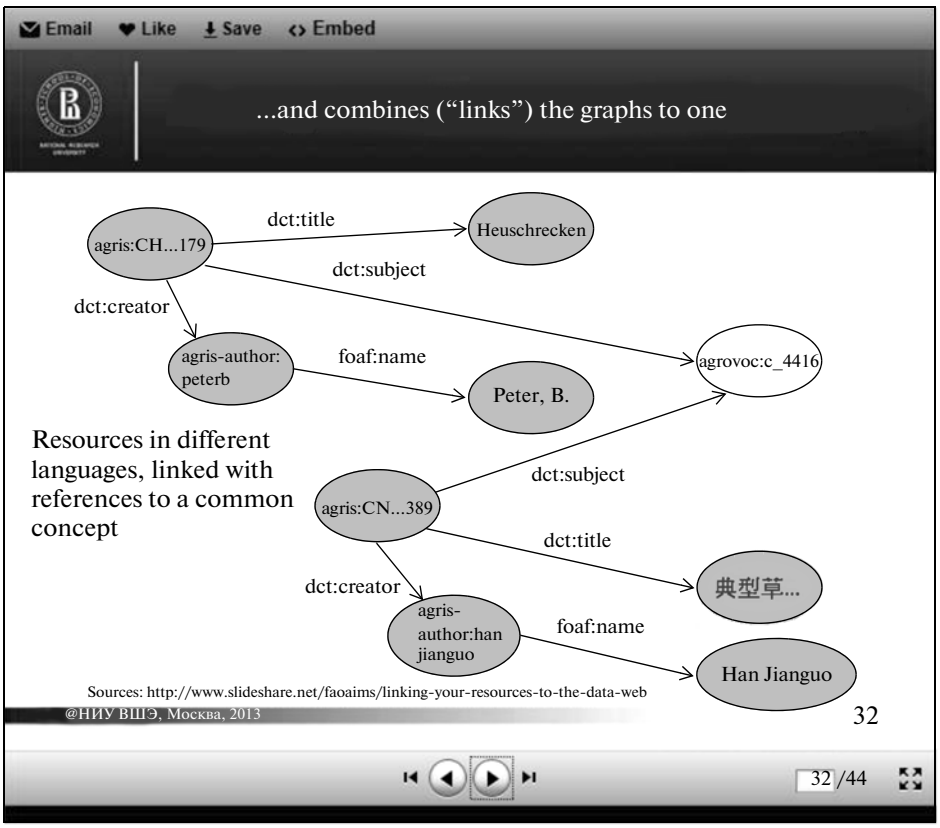

Fig. 3. An example of documents binding in the LOD medium [10].

Language. **SPARQL** is a query language for data represented by the RDF model, just as the SQL language (Structured Query Language) provides queries to relational database tables, as well as a protocol for the transmission of these queries and responses to them. SPARQL is a W3C recommendation and one of the semantic web technologies. Provision of a **SPARQL endpoint** is a recommended practice for the publication of data on the Web.

Practical recipes for publication of documents as linked data can be found in [12]. As well as techniques for providing this form of publication, several general principles are determined. LOD publication should correspond to these principles in order to display of advantages and opportunities of linked data in full. They are set out in the recommendations of the W3C [13] called the "seven best rules for the preparation of linked data":

1. Model the data.

2. Give URIs to concepts.

3. Whenever possible, use a dictionary

4. Provide data self-description in a form that is accessible to both people and computers.

5. Convert data into the RDF format.

6. Provide the publication with a free license and legal data dissemination.

7. Provide the published data with *hosting* and take measures for the wide advertisement of this data.

The first three rules include a preliminary informal analysis of the intended publication for clear separation of data and metadata, as well as for elimination of irrelevant information from the publication, which blocks "raw" data that is retrieved from a variety of sources: databases, XML files, CSV tables[2] etc. The author should then form a model of data representation in RDF format, what should be created for the dictionaries and thesaurus of concepts, and a URI identifier should be given to each of the entities. Search and selection of existing dictionaries, for example, using the catalog http://lov.okfn.org/dataset/lov, is an important aspect of advance preparation. Rule 4 calls for the author to present data as "self-describing" where the information about encoding the information is given in the RDF files themselves. The availability of this information, both for

[2] XML, eXtensible Markup Language, CSV, Comma-Separated Values

humans and for computers, ensures the implementation of the declared objectives of Linked Open Data.

The availability of a license for free and legal data dissemination (Rule 6) is an important requirement in the dissemination of open data. While the principle of open data means that they are available as "public domain" without copyright restrictions, however "exemption from restrictions" is provided through public licenses. Without the exact details of license conditions data do not acquire the status of open, even if they are presented in a machine-readable format on the network. The licenses that are issued by the nonprofit organization *Creative Commons*, whose aim is the legal spreading and use of knowledge and the results of creativity, are the most suitable licenses for this purpose.

Finally, advertising of data in order to make it well known to a wider range of network users is an important aspect of data publication in LOD. The addition of sets on the so-called *LOD cloud* (http://richard.cyganiak.de/2007/10/lod) with the visual representation of linked sets with meta-information that is maintained and updated on a certain node (http://thedatahub.org) is one of the best methods of advertising.

An extensive list of guidelines at different levels, with which we can explore the technology of linked publications and of linked data research, is presented on the http://linkeddata.org site. Recommendations on selecting of a variety of tools, viz., publishing platforms, editors, and control instruments ("validators") of RDF files, as well as of specialized browsers and search engines, are also given there. Several authors have attempted to create simplified guides that can provide a "quick start" in mastering the elements of the new technology, initially avoiding the problems that are associated with studying this technology in detail. In particular, we can point to [8] for a collection and submission of renewable energy information produced by LOD technology, as well as a short course [9] of five lessons (tutorials) to explore key elements of the technology: RDF, RDFS, semantic modeling, etc.

INTEGRATION OF NATURAL-SCIENCE DATA: OPPORTUNITIES OF LINKED OPEN DATA

With regard to science, a number of the LOD-technology aspects that meet the long-standing needs of the scientific community, which include the rapid spread of knowledge, standardization of concepts and terms, and integration of text and structured data, have been noted. The market has repeatedly offered tools that are able to meet some of these needs to varying degrees. The facilities offered by the Web of Science international database could be given as an example. Starting from a certain publication they allow one to find a cluster of thematically related records by binding of bibliographic data and by creation of a citation map with two generations of forward and backward citations. For a number of years, hopes for a deeper level of integration were associated with possible domain-specific versions of XML, providing a clear view of metadata to describe the structure and semantics of the information resources on the Web. A new concept of scientific publications in the form of XML documents containing ordinary text as well structured data with relational or hierarchical structure was even proposed[14]. The imperfections of XML-related technologies when compared with the capabilities of the Semantic WEB were gradually realized. XML has very limited means for integrating resources that cover a range of subject areas. Syntax is more strongly supported in an XML document than semantics, as tags structure the data within a document, but they are poorly related to the content of the data. In this regard the Semantic WEB technology, including LOD, opens greater opportunities in data introduction and integration by taking both their structure and semantics into account. All the problems with integration are usually pinned on the data user, in this case they are placed on the publisher, who uses a rich toolset in the form of RDFS, OWL, etc. This form provides wide *dissemination* when the publisher offers open access through standard interfaces such as SPARQL or URI; he also offers the integration of data by supporting a list of *links* between different RDF data, providing access to them through the user's requests; normalization through the use of RDF data in a common set of vocabularies and ontologies.

Semantic WEB Technologies in the Life Sciences

To date, there is already some experience in publishing of linked natural-science data. For several reasons, the greatest activity in the development of these technologies is observed in the field of life sciences. The creation of a special structure in W3C Group, viz., *Health Care and Life Sciences Interest Group* (*HCLS IG*, www.w3.org/blog/hcls) is an indication of the particular attention of W3C to this field. The activities of this group are focused on the widespread implementation of integration technologies in three fields: biology, **traditional medicine**, and application of methods in a relatively new field: **translational medicine**.

In biology the main purpose of using LOD is to cope with the huge volume of new data received during research. Such studies are conducted on the broadest range of scales: molecules, cells, cell structures, tissues, organs, organisms, populations, and ecosystems. Many experimental procedures, instruments, and reagents are used. We can note a number of directions, for example, studies of gene expression, of phenotypes, and chemical screening, that deliver a particularly large amount of data, on the basis of which conclusions are made and new hypothesis are proposed. The vast majority of these data are concentrated on a variety of heterogeneous databases, which requires a

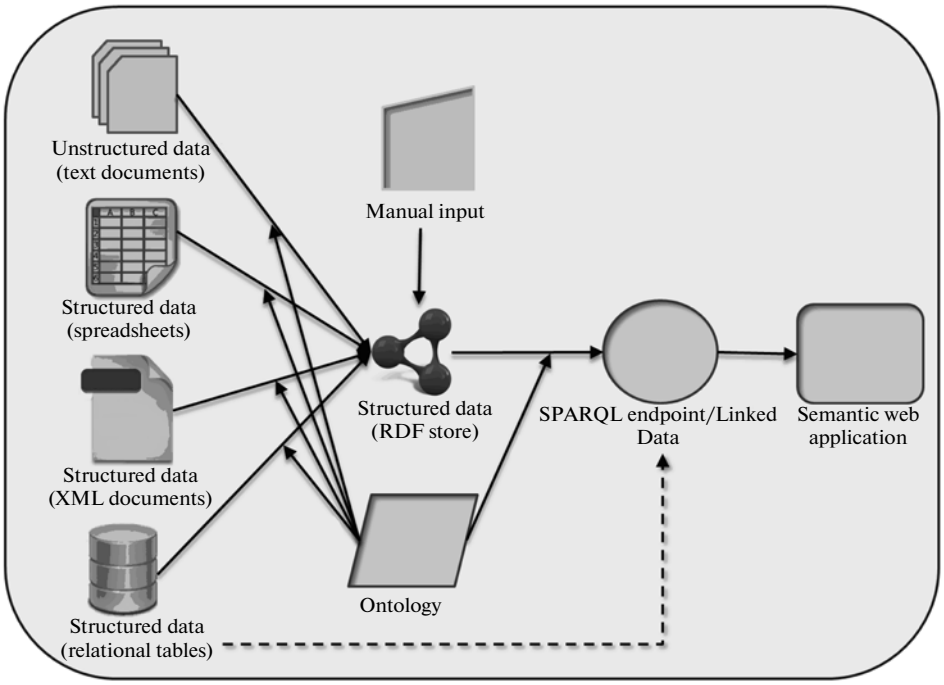

Fig. 4. A schematization of a working process during representation of unstructured and structured data in RDF format [15].

great deal of effort for their integration through the creation of integrating data models [2, 3].

The **HCLS IG** Group has prepared a set of technical advice on the conversion of raw data into the RDF format. In particular, the guidelines in [15] give detailed and well-documented answers to researchers to such questions as the methods and tools of relational database reflection in RDF; conversion of unstructured data (text, graphics, etc.) in RDF; opportunities for people and software agents to find and to use published RDF; licensing of published LOD data. The guide includes a total of 14 detailed recommendations on all aspects of the presentation of linked data; their application area overruns the biomedical theme.

The second area that is supervised by the *HCLS IG* group where the effectiveness of LOD methods are quite high, is *traditional clinical medicine*. The extent of the application area of these methods is determined by the needs of a powerful information infrastructure that combines clinical data with the data of genomic research, bioinformatics, chemical informatics, and environmental data. One the main tasks that is undertaken by the group in the implementation of LOD technologies, consists in combining them with the applicable western standards on the collection and sharing of health data, such as **The Clinical Data Interchange Standards Consortium (CDISC)** and **Health Level Seven (HL7)**. Another task involves the creation of formal ontologies for clinical medicine and research using tools such as OWL and RDFS.

The methods of linked data are especially in demand in so-called *translational medicine* in a relatively new industry that is actively using preclinical data in daily diagnosis and treatment, as well as connecting the development of drugs directly with the patient's response to correct the selection of the drug dose and time of reception. Naturally, an effective procedure for binding data that is obtained in the laboratory and clinical trials is the key issue in fundamental and applied medicine.

Based on the experience with linked data-representation associated data, the *HCLS IG* group developed and recommended to users a detailed description on transformation of raw data into RDF triplet and linking these with other sources (Fig. 4.). In [16] several specific examples of this process were examined: RDF data binding, which was obtained with DNA microarray technology for the analysis of changes in gene expression; linkage of numerous data on the properties and testing of drugs; generation of an RDF concept index (which is connected to the database of side effects caused by the action of drugs) in unstructured clinical reports.

One of the problems encountered by experts during the linking of biomedical data is that initially these data were meaningfully collected in a relational database. The general opinion that prevails in the community that uses the Semantic WEB is that the data

"should be left where they are" and to generate RDF synchronized, always leaving the possibility of access to updated sources of data open. Many professionals who use the recommendations of the *HCLS IG*, choose the relational data base (RDB) format, sometimes importing initially unstructured text in the database, anticipating all the procedures of their conversion to RDF format. For this reason, the entire direction of works in order to build reflections of the RDB → RDF (RDB2RDF) type was formed.

Compared with the relational model, the RDF structure is more expressive and data recorded in RDF may be interpreted and processed by software agents. Apparently, the first time the idea of such a mapping was expressed in 1998 by T. Berners–Lee, who considered the similarities and differences in the RDF models and "entity–relationship" [17]. Since 2009, the dedicated RDB2RDF Working Group Charter group (www.w3.org/ 2001/sw/rdb2rdf/) has been working on this subject. The D2R Server (http://d2rq.org/d2r-server) is the most common tool in relational database presentation as Linked Data. Using declarative language, the publisher sets a display between the relational database schema and the target RDF dictionary. As a result, the server publishes linked data, which allows the client to request data from the database using the SPARQL protocol. Using a server involves several steps: (1) downloading and installation of server software; (2) automatic generation of the D2RQ display of the database schema; (3) manual setting of the display by replacing automatically generated terms with more appropriate ones selected from well-known and public RDF dictionaries; (4) installation of RDF links from external data sources; and (5) installation of several RDF references from the existing LOD data on the resources of a new set corresponding to the database, so that search engine spiders, which go through web pages, can discover new information. The D2R Server supports the most popular databases: Oracle, MySQL, PostgreSQL, SQL Server, HSQLDB, and Interbase/Firebird.

The meaning of the conversion operation RDB → RDF consists not only of format unification but also of a certain content enrichment due to explicit modeling of the relationships between entities, which were not significant in relational databases, and, of course, due to the inclusion of object-oriented semantics in the data. This conversion generates the "semantic view" of database content, which initially has no RDF representation. Thus, Fig. 4 shows that, in addition to generating an RDF database version through a declarative representation, a D2R server can directly provide an endpoint entry in the database for the SPARQL protocol. The recommendations of *HCLS IG* discuss all the specifics that are associated with the choice of display languages and tools for database-content implementation in the Linked Open Data medium in detail.

Semantic WEB Technologies in Chemistry

In chemistry, the idea of the Semantic WEB met a long-prepared basis. Since 1998 there has been a discipline called *Chemoinformatics*. The distribution and integration of multiple references and experimental data on the structure and properties of the compounds, chemical constants of reactions, etc, make up one of the tasks of this discipline. At the same time, chemistry is different from other disciplines in its conservative attitude to the concept of open data. As the author of the review [18] remarked, in the chemistry the dissemination of knowledge is sub-delegated to commercial publishers, which prevents the automatic release of data from journal articles, more strongly than in other areas of science. This begins to slow scientific progress, especially in the neighboring areas (in bioinformatics, genomics, and pharmaceuticals) and the implementation of integration technologies is gradually beginning in chemistry.

In 2010 the American Chemical Society (ACS) held a 2-day seminar on the use of the RDF model, where existing products that have provided communication within chemistry, and especially in conjunction with the problems of bioinformatics, were considered. According to the materials of this seminar, the *Journal of Cheminformatics* publishes the topical series "RDF technologies in chemistry" (www.jcheminf.com/ series/acsrdf2010) with a detailed review [19]. At that time, the use of RDF was sporadic, demonstrating success for only some projects. Apparently, the first proposal to use RDF for representation of chemical structures was made in 2004 in [20], who previously proposed the CML language (Chemical Markup Language) for data exchange. The SPARQL search language became applicable in 2007 for annotated systems of crystalline structures. Such projects as Bio2RDF, Chem2Bio2RDF and OpenTox were illustrated in the presentations at the workshop. In these projects knowledge from genomics, chemistry, and pharmaceutics was loaded onto the Semantic WEB using RDF. These projects were designed for the creation of a database with the chemical information available from a central point, combining the individual data sets. As well, smaller-scale projects using RDF, for example, the Open Notebook Science Solubility dataset, were implemented.

The most promising technology in the family of RDF (SPARQL protocol) was successfully used to solve problems of chemogenomics, multi-disciplinary field determining the correspondence between ligands and targets in biological objects. A search for biologically active compounds was carried out based on three databases (PubChem, Uniprot, and DrugBank), which are available on the Chem2Bio2RDF server. The server provides access to relational databases, transforming SPARQL queries into conventional SQL.

RDF technology connects data on chemical concepts (object, concept) with resources that provide

details related to these resources and the SPARQL Protocol provides the means for data searching and aggregation. The following standard of the Semantic WEB, viz., the Web Ontology Language (OWL) links the RDF technology with many ontologies. Like a controlled vocabulary or thesaurus, ontologies describe the meaning of concepts, binding terms with definitions, as perceived by people. At the same time, content presentation in explicit terms allows both people and computers to organize formal reasoning, and perhaps to find source errors.

To date, there are not very many ontologies in chemistry, especially those written in the OWL language. For example, Konyk et al. used OWL to connect the three major Data Warehouse (PubChem, Drug-Bank, and DBPedia), providing new ways of acquiring knowledge [21]. There is a special collection of biomedical ontologies (OBO Foundry ontologies, www.obofoundry.org), some of which include chemical problems, such as ChEBI (Chemical Entities of Biological Interest), CHEMINF (Chemical Information Ontology), CO (A chemical ontology for the identification of functional groups and semantic comparison of small molecules). For ontologies stored in the OBO format, there are means of conversion in OWL, which provides the application of this format as a general one for the integration of chemical data.

The CHESS system (Chemical Entity Semantic Specification) for the submission of polyatomic molecules and their components using the Semantic WEB [22] was developed on the basis of the CHEMINF ontology. CHESS specification includes three broad categories: (1) chemical entities, i.e., reactions, complexes, molecules, functional groups, bonds and atoms (possibly including electrons and macromolecules); (2) chemical descriptors; and (3) a chemical "configuration" that reflects the conditions under which the data were obtained, as well as the data source. The main requirement involves the ability to represent chemistry concepts in a manner that does not depend on the starting database, on software and on the particular branch of science. A harmonized coding system of atoms, bonds and functional groups, using the recommended IUPAC identifier InChI keys, which is a linear notation in the form of a character string (www.iupac.org/inchi), has been implemented for this purpose. Another feature of the system lies in its flexibility, which is manifested in the ability to perceive data and notations of different formats that define structural information (along with InChI, for example, the Simplified Molecular Input Line Entry System notation (SMILES)). CHESS also provides data storage with clear indications of the conditions in which these data is derived, as well as of related data. Finally, it is important that along with traditional search and aggregation services CHESS supports standard chemoinformatics tools for drug searches, chemical analogues, and model reactions via pattern matching.

Analysis of the abilities of CHESS [22] shows that, despite the above-noted conservatism, a gradual transition to the integration of data with sufficient universalism of this procedure, i.e., with the independence of the data structure and format, is observed in chemistry. Naturally, the project in [22] is not unique; the overall picture as of 2013, which was presented in a review in [23], concerns linking technology of documents as well, while in the reports of the ACS seminar [19], the focus was on the integration of databases. In particular, the review considered a form of so-called *Semantic Publications* based on the example of the publications of the Royal Society of Chemistry (RSC). In the structure of this society, the RSC Semantic publishing project (www.rsc.org/Publishing/Journals/Project-Prospect/ index.asp) was implemented, in which framework articles from magazines in 2008–2010 are linked with the ChemSpider open database (www.chemspider.ru). Manuscripts submitted to the RSC, are semantically marked to identify the important chemical data, particularly data according to structures. Marking provides "links" of texts with additional sources of data on properties. This allows search engines to use marking, in particular, to identify the date associated with a particular structure. The approach implemented by RSC Project demonstrates the benefits of publication in formats that are compatible with the Semantic WEB. The corresponding functionality associated with the formats of RDF, has been added to the interface of the giant ChemSpider open database.

CONCLUSIONS

The above discussion shows the untapped potential of the Semantic WEB and specifically, of Linked data technology in the field of natural sciences. Although the examples that were given in the article are borrowed from the field of Life sciences and chemistry, similar applications have been developed for other fields, for example for crystallography, thermodynamics, earth sciences, etc. The most significant issue is that semantic technologies revolutionize the process of publication for a scientist. A scientific publication ceases to be an isolated unit, which is reflected only by abstracting services, and becomes a part of a global database. The conversion of its content in RDF provides a binding with thematically related publications and database. The binding itself occurs without the participation of the author, based on the use of online dictionaries and ontologies. The publication of the Royal Chemical Society (Great Britain) is a perfect example. This publication is uses semantic markup and contains numerous references to relevant sources that contain more information on the structures and properties of these substances in the original publication.

The idea of such publication enrichment with additional data goes back to the early 2000s [14, 24], where instead of the traditional forms of publication, a structure that allows readers to have direct access to an original numerical file, which is obtained by experiment or numerical simulation, was proposed. Such a structure, which received the special name *datument* (a neologism signifying the synthesis of data and documents), was constructed in the form of the XML document, which should replace conventional forms of electronic publication such as PDF files according to the authors concept. In this concept marking up of texts with references can lead to a repository where the initial data is stored or to the corresponding software. Moreover, a concept is widely discussed where, along with the openness of a *datument*, it is possible to allow interaction of authors and readers during scientific communication without the traditional role of the publisher, limiting their function to the organization of peer reviews and support of impact factors. Free distribution of tables and graphic information, molecular structures, and mathematical constructs in the form of the elements of MathML are the main advantages of a *datument*.

Despite the great enthusiasm that was shown to similar forms of publication, support for XML was apparently a deterrent. We have already noted that the XML-related technologies are much weaker during the integration of many heterogeneous resources, as it supports syntax, but not semantics. At the same the LOD technology, which is not limited to the standardization of data exchange, provides ***truly semantic*** data integration by automatic linkage of a number of related resources in coordination with the terminology and concepts with common vocabularies and ontologies.

The ability to present unstructured (text articles, documents, etc.) and rigidly structured data placed in a relational database at the same level (see, especially, Fig. 4 and the description of the D2R server in guidelines [15]) is another important advantage of the Semantic WEB technology.

Finally, in addition to the rich technological capabilities, the entire concept of related data raises a number of conceptual implications for those disciplines, where this concept shows its advantages. It fosters a transition towards open data, responding to ever-emerging initiatives such as Pubmedcentral, ePrints Initiative, Open Archives Initiative, Public Library of Science, etc. It creates communication space, having a noticeable advantage compared with traditional publications forms distributed by commercially oriented publishers. An increased attention to dictionaries and community taxonomies is another consequence of the new technology. Besides the fact that it is an appeal to the Common Terminology Systems provides a binding of documents, the need to work with ontologies suggests certain order in the subject area by standardizing the terminology, concepts, units, etc. Finally, the general rules for the publication of related documents (see the section "Basic elements ...") impose fairly strict requirements on a preliminary stage, such as a removal of contaminating information, a clear separation of data and metadata, an interpretation of the terms by references to public vocabularies or ontologies. Thus, already at the preparation stage the new publication technology encourages authors to provide a more rigorous and standardized form of information presentation.

ACKNOWLEDGMENTS

This work was supported by RFBR (the Russian Foundation for Basic Research)—a project 13-07-00218.

REFERENCES

1. Berners-Lee, T., Hendler, J., and Lassila, O., The semantic web, *Sci. Am.*, 2001, vol. 284, pp. 35–43.
2. Kogalovskii, M.R., *Entsiklopediya tekhnologii baz dannykh* (Encyclopedia of Data Base Technology), Moscow: Finansy Statistika, 2002.
3. Erkimbaev, A.O., Zitserman, V.Yu., Kobzev, G.A., Son, E.E., and Sotnikov, A.N., Integration of databases on substance properties: Approaches and technologies, *Autom. Docum. Mathem. Ling.*, 2012, vol. 46, pp. 170–176.
4. Berners-Lee, T., Design Issues: Linked Data. Online at. www.w3.org/DesignIssues/LinkedData.html.
5. Open Data – An Introduction "Today we find ourselves in the midst of an open data revolution". http://okfn.org/opendata
6. Bauer, F. and Kaltenböck, M., Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers, Vienna, 2012. www.semantic-web.at/LOD-TheEssentials.pdf
7. 5 Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers, Vienna, 2012. http://5★stardata.info/
8. Reegle LOD Developer Guide. http://data.reegle.info/developers/guide
9. Introducing Linked Data and The Semantic Web. http://www.linkeddatatools.com/semantic-web-basics
10. Radchenko, I.A., Introduction into Concept of Linked Open Data. Linked Open Data Webinar Cycle. AIMS, 2013. http://www.slideshare.net/iradche/linked-open-data-16524818
11. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, 2004. http://www.w3.org/TR/rdf-schema
12. Bizer, C., Cyganiak, R., and Heath, T., How to Publish Linked Data on the Web. http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial
13. Linked Data Cookbook. From Government Linked Data (GLD) Working Group Wiki. http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

14. Murray-Rust, P. and Rzepa, H.S., Scientific publications in XML – towards a global knowledge base, *Data Sci. J.*, 2002, no 1, pp. 84–98.
15. Health Care and Life Science (HCLS) Linked Data Guide. www.w3.org/2001/sw/hcls/notes/hcls-rdf-guide/
16. Marshall, M.S., Boyce, R., Deus, H.F., Zhao, J., Willighagen, E.L., Samwald, M., Pichler, E., Hajagos, J., Prud'hommeaux, E., and Stephens, S., Emerging practices for mapping and linking life sciences data using RDF – A case series, *Web semantics: Science, Services and Agents on the World Wide Web*, 2012, vol. 14, pp. 2–13.
17. Berners-Lee, T., Relational Databases on the Semantic Web, Design Issue Note, 1998–2009. www.w3.org/DesignIssues/RDB-RDF.html
18. Adams, N., Semantic Chemistry, *The Voice of Semantic Web Technology and Linked Data Businessm*. semanticweb.com, 2009.
19. Willighagen, E.L. and Brändle, M.P., Resource description framework technologies in chemistry, *J. Cheminformatics*, 2011, vol. 3, no. 15. www.jcheminf.com/content/3/1/15
20. Murray-Rust, P., Rzepa, H.S., Williamson, M., and Willighagen, E., Chemical Markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators, *J. Chem. Inf. Comput. Sci.*, 2004, vol. 44, 462–469.
21. Konyk, M., de Leon, A., and Dumontier, M., Chemical knowledge for the semantic Web, *Lect. Notes in Comp. Sci.*, 2008, vol. 5109, pp. 169–176.
22. Chepelev, L.L. and Dumontier, M., Chemical entity semantic specification: Knowledge representation for efficient semantic cheminformatics and facile data integration, *J. Cheminformatics*, 2011, vol. 3, no. 20.
23. Frey, J.G. and Bird, C.L., Cheminformatics and the semantic Web: Adding value with linked data and enhanced provenance, *WIREs Comput. Mol. Sci.*, 2013, vol. 3, pp. 465–481. http://onlinelibrary.wiley.com/doi/10.1002/wcms.1127/pdf
24. Murray-Rust, P. and Rzepa, H.S., XML for scientific publishing, *OCLC Syst. Serv.*, 2003, vol. 19, pp. 163–169.

Translated by O. Kupriyanova-Ashina