F20BD/F21BD Big Data Management

Assignment Two – Big Data vs Semantic Web

Deadline

- Edinburgh: Sunday 5 April 2015 (before midnight)
- Dubai: see separate announcement by Stephen Gill.

Overview

Big Data Management and Semantic Web Technologies are two key areas in computing today. They are different, but related. In this coursework you are asked to write a report which initially briefly outlines these two fields individually in terms of their motivation, underpinning technologies and sample applications. You should then conduct a study into the relationship between them and describe how they potentially can complement each other. Finally, you (F21BD only) are to write a one page review of the attached paper on "Semantic Web technologies for the big data in life sciences".

Details

- Introduction to the Big Data Management (approx. 1.5 pages):
  - What is the objective of Big Data Management?
  - What are the key technologies in Big Data Management?
  - Briefly describe a typical example of Big Data.
- Introduction to Semantic Web Technologies (approx. 1.5 pages):
  - What is the objective of the Semantic Web?
  - What are the key technologies for the Semantic Web?
  - Briefly describe a typical Semantic Web application example.
- What is the relationship between Big Data and the Semantic Web? (2-3 pages) You should carry out a search (e.g. via the online library facility and/or Google search) to identify a number of articles covering this topic and then summarise their points in your report. Make sure you appropriately reference the relevant articles.
- Provide a one page summary of the key points described in the attached paper by Wu and Yamaguchi. (1 page only. This part is for students on F21BD only.)

Reports should be submitted electronically through Vision. Use Ariel 11pt font and single line spacing. You are encouraged to make use of figures and diagrams where they aid the understanding of your report. The recommended length of sections above is for the text part only, i.e. if you use diagrams, they are in addition to the recommended text length.

Please clearly state your name, user id, degree programme, and for undergraduates your year of study4 in your submission.

A marking sheet will be issued separately for this assignment.

Review

DOI: 10.5582/bst.2014.01048

Semantic Web technologies for the big data in life sciences

Hongyan Wu*, Atsuko Yamaguchi*

Database Center for Life Science, Research Organization of Information and Systems, Japan.

Summary **The life sciences field is entering an era of big data with the breakthroughs of science and technology. More and more big data-related projects and activities are being performed in the world. Life sciences data generated by new technologies are continuing to grow in not only size but also variety and complexity, with great speed. To ensure that big data has a major influence in the life sciences, comprehensive data analysis across multiple data sources and even across disciplines is indispensable. The increasing volume of data and the heterogeneous, complex varieties of data are two principal issues mainly discussed in life science informatics. The ever-evolving next-generation Web, characterized as the Semantic Web, is an extension of the current Web, aiming to provide information for not only humans but also computers to semantically process large-scale data. The paper presents a survey of big data in life sciences, big data related projects and Semantic Web technologies. The paper introduces the main Semantic Web technologies and their current situation, and provides a detailed analysis of how Semantic Web technologies address the heterogeneous variety of life sciences big data. The paper helps to understand the role of Semantic Web technologies in the big data era and how they provide a promising solution for the big data in life sciences.**

Keywords: Big data, Semantic Web technologies, life-science databases

1. Introduction

Data deluge in the life sciences The life sciences field is entering an era of big data with the breakthroughs of science and technology. Moore's law shows that computers double in speed and halve in size every 18 months (*1*). A similar trend is observed for hard disks (*2*) and networks (*3*). The exponential growth of scientific instruments has resulted in an exponentially growing amount of scientific data (*4*). Until recent years, Moore's law kept outpacing the generation of biological sequence data by its growth in storage and processing capacity. This trend has remained true for approximately 40 years and was not broken until the completion of the Human Genome Project in 2003. From 2005, the sequencing output doubling rate decreased to 5 months because of the development of Next-Generation Sequencing technologies (NGS) (*5*). Since 2008, genomics data are outpacing Moore's Law by a factor of 4 (*6*). The 1,000 Genomes Project (*7*), which involves sequencing and cataloging human genetic variations, has deposited 2 times more raw data into GenBank (*http://www.ncbi.nlm.nih.gov/genbank/*) at National Center for Biotechnology Information (NCBI) during its first 6 months than all the previous sequences deposited in the last 30 years (*5*). In the last five years, more scientific data have been generated than in the entire history of mankind (*8*). Figure 1 illustrates the GenBank and Whole Genome Shotgun (WGS) statistics up to February 2014. Human DNA comprises approximately 3 billion base pairs with a personal genome representing approximately 100 gigabyte (GB) of data (*6*). Two nanopore sequencing platforms (GridION™ and MinION™) (*9*), reported in February 2012, are capable of delivering ultra-long sequencing reads (~100 kb) with additionally higher throughput and much lower cost. Sequencing a human genome has decreased in cost from $10,000 in 2007 to $1,000 in 2012 (*10*) and is likely to drop below $100 per genome in the next decade (*11*). In the third decade of the 21st century, it has been estimated that 1 billion people will be sequenced and that approximately 3,000 petabyte (PB) (1 PB is approximately equivalent to 10^6 GB) of storage will be needed. Grossman *et al*. predicated that

**Address correspondence to:*
Dr. Hongyan Wu and Dr. Atsuko Yamaguchi, DBCLS, Univ. of Tokyo Kashiwa-no-ha Campus Station Satellite 6F, 178-4-4 Wakashiba, Kashiwa-shi, Chiba 277-0871, Japan.
E-mail: wu@dbcls.rois.ac.jp (Wu HY)
atsuko@dbcls.rois.ac.jp (Yamaguchi A)

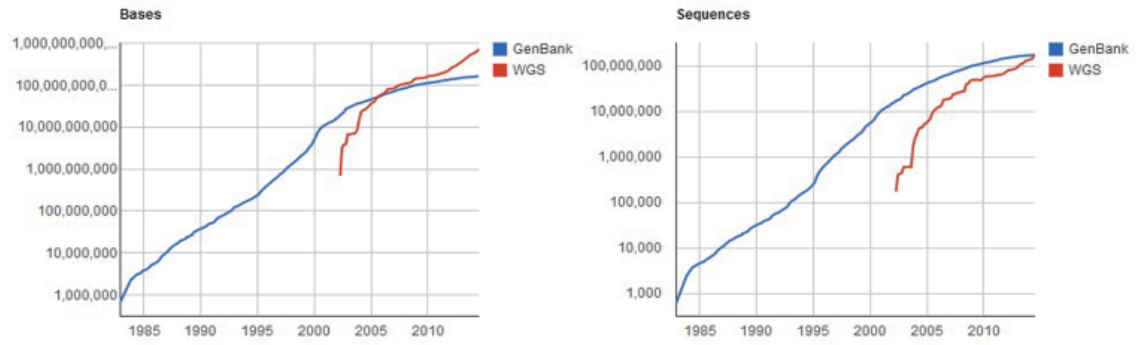

Figure 1. Growth of GenBank and WGS. NCBI GenBank ([Internet] [cited July 18, 2014]. Available from *http://www.ncbi.nlm.nih.gov/genbank/statistics*).

we would be in an era of ubiquitous sequencing within a few years, in which genome sequencing would become routine for both research and clinical applications (*11*).

Many other kinds of life science big data are being produced at high speed as well as genomics data. Functional magnetic resonance imaging or functional MRI (fMRI) is a functional neuroimaging procedure using MRI technology that measures brain activity by detecting associated changes in blood flow. This technique generates complex data sets: ~100,000 locations, measured simultaneously hundreds of times, resulting in billions of pairwise relations, collected in multiple experimental conditions, and from dozens of participants per study (*12*). Other data, including Computerized Tomography (CT) Scan data, epidemic data, Electronic Health Records (EHR) system data, patient behavior and sentiment data *etc.*, are also being generated and gathered at a fast pace.

Big data-related projects and activities More and more big data-related projects and activities are being performed in the world. The Genome 10K project (*http://www.genome10k.org*) aims to assemble a genomic zoo, which will be a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. The 1001 Genomes Project (*http://www.1001genomes.org*), launched at the beginning of 2008, has the goal of discovering the whole-genome sequence variation in 1,001 strains of the reference plant Arabidopsis thaliana. The 1K Insect Transcriptome Evolution (1KITE) Project (*http://www.1kite.org*) aims to study the transcriptomes of more than 1,000 insect species encompassing all recognized insect orders. The ENCyclopedia of DNA Elements (ENCODE) project (*http://www.genome.gov/10005107*) aims to identify all functional elements in the human genome sequence. ENCODE generated more than 15 terabyte (TB) of raw data, and the data analysis consumed the equivalent of more than 300 years of computing time. The Cancer Genome Atlas (TCGA) (*http://cancergenome.nih.gov/*) began as a three-year pilot in 2006 with an investment of $50 million each from the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), confirming that an atlas of changes could be created for specific cancer types. The European life-science infrastructure for biological information (ELIXIR) (*http://www.elixir-europe.org/*) unites Europe's leading life science organizations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. ELIXIR aims to provide the facilities necessary for life science researchers, from bench biologists to cheminformaticians, to make the most of the rapidly growing store of information about living systems. Tohoku University Tohoku Medical Megabank Organization (*http://www.megabank.tohoku.ac.jp/english/*) was founded to establish an advanced medical system to foster the reconstruction from the Great East Japan Earthquake. The organization will develop a biobank that combines medical and genome information during the process of rebuilding the community medical system and supporting health and welfare in the Tohoku area. Approximately 60 PB of data representing 1.5 million genomic and medical pieces of information is predicted to be acquired. In China, Beijing Genomics Institute (BGI) and their publishing partner BioMed Central, a leader in scientific data sharing, announced the launch of a new journal, GigaScience, which publishes large-scale biological research in a unique format (*http://www.genomics.cn/en/news/show_news?nid=99134*).

Big data issues Life sciences data are continuing to grow in not only size but also variety and complexity with great speed. The role of genome sequencing in the life sciences is the tip of the iceberg. To investigate the complex systematic effects of drugs and other chemical compounds on biological systems and to validate a hypothesis in drug discovery, we require the data on diseases, compounds, genes, targets, side effects, and metabolic pathways, as well as from the clinic and other sources (*13*). These data reside in a number of different data sources, such as GenBank (*14*), Genome Sequence DataBase (GSDB) (*15*), SWISS-PROT (*16*), European Molecular Biology Laboratory (EMBL), Online Mendelian Inheritance in Man (OMIM) (*17*), and many others (*18*). Data sources can store different data types in different formats (*19*); for example, flat file (*e.g.*, tab-delimited file), sequence file (*e.g.*, FASTA), structure file

(*e.g.*, Protein Structure File (PSF)), Extensible Markup Language (XML) file (*e.g.*, KGML- Kyoto Encyclopedia of Genes and Genomes (KEGG) Markup Language for describing graph objects), and database management systems (DBMSs). Even for the same data type, data formats in different sources are often incompatible. In addition, new data formats are being invented along with the development of new technologies (*20*), such as Sequence Alignment/MAP (SAM) (21) and Genome Variation Format (GVF) (*22*).

To ensure that big data has a major influence in the life sciences, comprehensive data analysis across multiple data sources and even across disciplines is indispensable. For example, research on the neurodegenerative disease Alzheimer's disease (AD) spans the disciplines of psychiatry, neurology, microscopic anatomy, neuronal physiology, biochemistry, genetics, molecular biology, and bioinformatics (*23*). A series of combination and integration problems such as data, terminologies, knowledge, and service integration must be solved first (*24*). Eliminating the inconsistency of data and terms as well as finding and meaningfully combining information in the vast majority of data all require knowing the exact semantics of the data (*25*).

The increasing volume of data generated by new technologies at an unprecedented rate and the heterogeneous complex varieties of data are two principal issues mainly discussed in life science informatics (*26*). In the remainder of this paper, we provide insight into how the Semantic Web technologies address the heterogeneous variety of life sciences big data. We also present a survey of the state-of-the-art development of every technology and some related projects. Finally, we summarize the challenges and problems that we have to face now and in the future.

2. Semantic Web technologies

The ever-evolving next-generation Web, characterized as the Semantic Web (*27*), is an extension of the current Web, aiming to provide information for not only humans but also computers to semantically process data. Berners-Lee *et al*. (*27*) believed that this form of Web content that was meaningful to computers would unleash a revolution of new possibilities. The following introduces a series of the Semantic Web technologies.

2.1. *Resource Description Framework (RDF)*

The RDF (*28*) is a model for representing information about resources on the World Wide Web. The RDF model identifies items with Web identifiers (called Uniform Resource Identifiers, or URIs) and encodes data in the form of subject, predicate, and object (with the whole usually referred to as a "triple"). The subject is a URI or blank node. The object is a URI or string literal. The predicate specifies the relationship between the subject and object and is also represented by a URI. For example, in the KEGG database the breast cancer gene hsa:675 encodes the *Homo sapiens protein* with the number 119395734 in NCBI Protein database. This gene is the same as gene ENSG00000139618 in Ensemble database. This could be expressed as two triples: "<hsa:675> <encodes> <protein:119395734>. <hsa:675> <owl:sameAs> <ENSG00000139618>". The relationship among these three resources: hsa:675, protein:119395734, and ENSG00000139618, in three databases is established. Similar to how any document expressed in HyperText Markup Language (HTML) can be linked to any other document expressed in HTML, the information expressed in RDF can be connected to any other information expressed in RDF (*26*). However, with respect to HTML, a linked resource must be a whole document, whereas with RDF, any information defined as a resource can be linked together.

RDF is expressive with the simple triple format. The Semantic Web integrates not only resources that are themselves built or represented using RDF but also those resources that can be mapped to RDF (*29*).

2.2. *SPARQL Protocol and RDF Query Language (SPARQL)*

SPARQL is an RDF query language (*30*). A SPARQL endpoint is a conformant SPARQL protocol service as defined in the SPROT (SPARQL Protocol for RDF) specification. A SPARQL endpoint enables users (human or other) to query a knowledge base *via* the SPARQL language. SPARQL 1.1 specification, produced by the SPARQL Working Group on 21 March 2013, defines the syntax and semantics of the SERVICE extension, which allows a query author to direct a portion of a query to a particular SPARQL endpoint. The results are returned to the federated query processor and are combined with results from the rest of the query (*31*). The growing number of SPARQL query services offer data consumers an opportunity to merge data distributed across the Web. However, SPARQL query is still in its infancy, and its service provider tends to change its endpoint in the development stage. The site (*32*) monitors the availability of some SPARQL endpoints. Table 1 summarizes the main current available SPARQL endpoints in the life sciences.

2.3. *Ontology*

Semantic heterogeneities arise at the entry level where different terms are used for the same things or the same terms are used for different things. *Ontology* describes the types of entities in the world and how they are related. The RDF model enables a link between two resources. Ontology strengthens and implements the link by specifying the semantics of terminology systems in a well-defined and unambiguous manner (*33,*

Table 1. List of some available biomedical SPARQL endpoints

- Allie: http://allie.dbcls.jp/
- Bio2RDF:
  - HGNC: http://hgnc.bio2rdf.org/sparql
  - GO: http://go.bio2rdf.org/sparql
  - PharmGKB: http://cu.pharmgkb.bio2rdf.org/sparql
  - Pubmed: http://pubmed.bio2rdf.org/sparql
- BioGateway: http://www.semantic-systems-biology.org/biogateway/querying
- Cell Cycle Ontology: http://www.semantic-systems-biology.org/cco/queryingcco/sparql
- HDP: http://healthdata.tw.rpi.edu/sparql
- Linked Food: http://www.linkedfood.org:8890/sparql/
- Linked Life Data: http://linkedlifedata.com/sparql
- myExperiment: http://rdf.myexperiment.org/sparql
- NCBO: http://sparql.bioontology.org/
- Neuroscience Information Framework: http://rdf-stage.neuinfo.org/
- The EBI RDF platform:
  - BioModels :http://www.ebi.ac.uk/rdf/services/biomodels/sparql
  - BioSamples: http://www.ebi.ac.uk/rdf/services/biosamples/sparql
  - ChEMBL: http://www.ebi.ac.uk/rdf/services/chembl/sparql
  - Expression Atlas: http://www.ebi.ac.uk/rdf/services/atlas/sparql
  - Reactome: http://www.ebi.ac.uk/rdf/services/reactome/sparql
  - UniProt: http://beta.sparql.uniprot.org/

34). Ontology provides a shared understanding of data, services and processes and has thus far played a role in the semantic integration of databases (*35*).

The OWL Web Ontology Language (*OWL*) (*36*) is designed for use by applications that need to process the content of information instead of just presenting information to humans. By providing additional vocabulary along with formal semantics, OWL facilitates a greater machine interpretability of Web content than that supported by XML and RDF. Consider the following simple example (*37*): (*i*) frog and Amphibian are two classes, and both have an *is-a* property; (*ii*) there is a restriction in which Frog is a subclass of Amphibian; and (*iii*) Herry is one example of a Frog class. We can simplify the model as "<Frog> <rdfs:subClassOf> <Amphibian>" and "<Herry> <is-a> <Frog>", and then we can infer that "<Herry> <is-a> <Amphibian>". By including descriptions of classes, properties and their examples, the OWL formal semantics specifies how to derive its logical consequences, *i.e.*, facts not literally present in the ontology but *entailed* by the semantics. These entailments may be based on a single document or multiple distributed documents that have been combined using defined OWL mechanisms (*http://www.w3.org/TR/owl-guide/*). In this way, RDF enables the data publisher to explicitly state the nature of the connection (*38*). In contrast, HTML links typically only indicate that two documents are related in some way without showing the nature of the relationship. Together with RDF Schema (*39*), which provides a data-modeling vocabulary for RDF data, OWL offers a standard, machine-processable means of describing relationships between RDF statements, *e.g.*, that one property is an rdfs:subPropertyOf of another.

The life sciences are flourishing with ontologies to enable the data in distributed sources to be shared and analyzed. The Open Biological and Biomedical Ontologies (OBO) Foundry (*http://www.obofoundry.org/)* is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development, with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The ontologies developed by them include biological process, cellular component, chemical entities of biological interest, molecular function ontology and so on. The Ontology Working Group of the Health Care and Life Sciences (HCLS) is to facilitate creation, evaluation and maintenance of "core vocabularies and ontologies to support cross-community data integration and collaborative efforts". The Gene Ontology (GO) project (*http://www.geneontology.org/*) is a collaborative effort of Gene Ontology Consortium, to address the need for consistent descriptions of gene products in different databases. The Micro Array Gene Expression Data (MGED) Ontology (*40*) describes Microarray data and experiments. Biological Pathway Exchange (BioPAX) is an ontology for biological pathway data. National Center for Biomedical Ontology's (NCBO) BioPortal (*https://bioportal.bioontology.org/*) contains URIs for concepts from almost 300 biomedical ontologies and reference terminologies. BioPortal is a convenient tool that can be used to identify public ontologies that best map to the entities in biomedical and clinical data sets. Ontobee (*http://www.ontobee.org/*) aims to facilitate ontology data sharing, visualization, query, integration, and analysis. Several web services have been developed to efficiently use the existing ontologies. The Ontology Lookup Service (OLS, *http://www.ebi.ac.uk/ontology-lookup/*) provides a web service interface to query multiple ontologies from a single location with a unified output format. To support ontology production based on existing resources, the OntoFinder/OntoFactory system (*http://ontofinder.dbcls.jp/*) aims to provide

non-computer experts with an easy interface to assist ontology selection and term selection from BioPortal, until a user produces his/her seed ontology. Ontologies have been widely applied and play an important role in the life science field. Gene Set Enrichment Analysis (GSEA) (*41*) utilizes GO to determine whether an a priori defined set of genes shows significant, concordant differences between two biological states. Great efforts have been paid on how to design a good ontology system and use it in data integration and automated reasoning. The pathway data sharing in the BioPAX community standard (*42*) and cross-product extensions of the Gene Ontology (*43*) use ontology to verify the consistency of a data model. Multiple ontologies have been integrated (*44*). Formal approaches to ontology research and their potential impact on biomedical applications and analyses have been summarized (*45*).

2.4. *Linked data*

Similar to how the idea to add, search, and automatically discover documents in the world stimulated the Web's explosive growth, the same principles of linking, and therefore ease of discovery, can be applied to data on the Web (*38*). Different from a key word search, linked data make automated reasoning about data possible by using semantic technologies. We are moving from the era of "data on the web" to an era of "web of data (linked data)" (*46*). Linked data try to create the Web into a giant global database. The term *Linked Data* refers to a set of best practices for publishing and interlinking structured data on the Web. Tim Berners-Lee in his Design Issues introduced four *Linked Data* principles (*47*): (*i*) Use URIs as names for things. (*ii*) Use HTTP URIs to allow people to look up those names. (*iii*) When an individual looks up a URI, provide useful information using recommended standards (*e.g.*, RDF and SPARQL). (*iv*) Include links to other URIs so that more things can be discovered.

Hypertext Transfer Protocol (HTTP) URIs provide a simple way to create globally unique names and a means to access information describing the identified entity. The RDF model enables the establishment of RDF links between data. A SPARQL query facilitates the retrieval of the data of interest across the distributed sources.

Linked Data has gained significant uptake in the life sciences. The HCLS group works on the Linking Open Drug Data (LODD) project (*http://www.w3.org/wiki/HCLSIG/LODD*), which provides linked RDF data exported from several data sources such as ClinicialTrials.gov, DrugBank (*http://www.drugbank.ca/*), and DailyMed (*http://dailymed.nlm.nih.gov/dailymed/about.cfm*). In particular, the Bio2RDF project has interlinked more than 30 widely used data sets (*48*), including the Universal Protein Resource (UniProt), KEGG, the Chemical Abstracts Service (CAS), PubMed, and Gene Ontology. Linking Open Data (*http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData*), a W3C Semantic Web Education and Outreach (SWEO) community project, aims to publish existing open license datasets as Linked Data on the Web to interlink things between different data sources. In Figure 2 the pink corner shows the life science data of the Linking Open Data (LOD) Project Cloud Diagram.

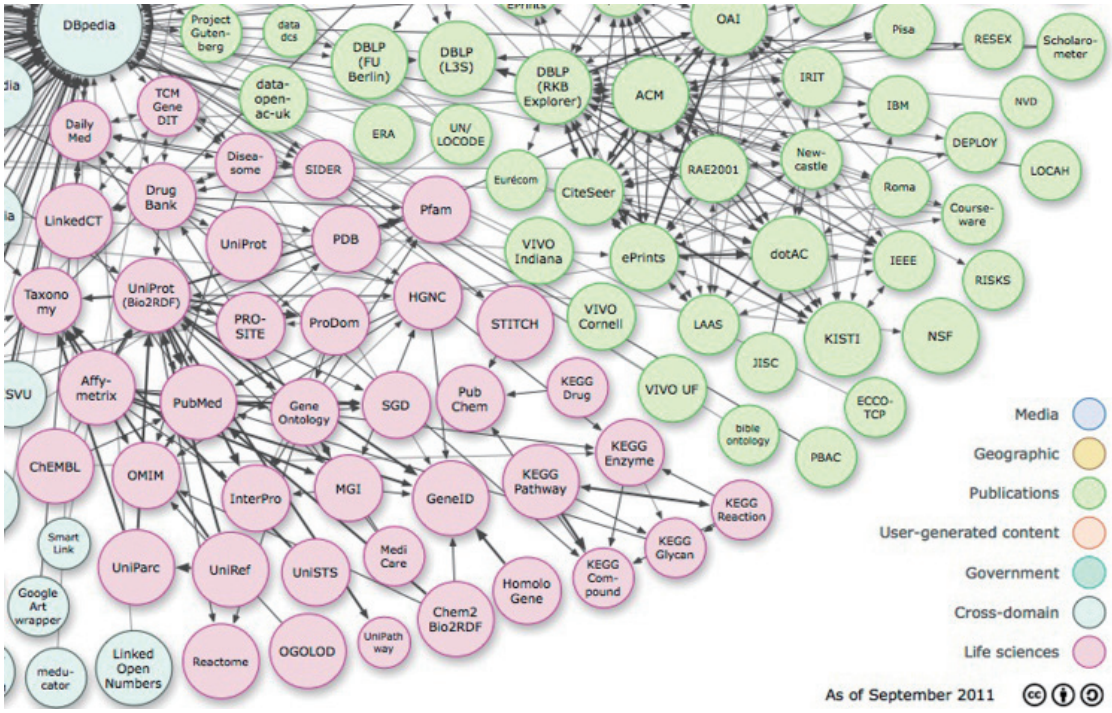

Figure 2. The lower right corner of the Linking Open Data cloud diagram ([Internet] [cited July 18, 2014], Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. Available from *http://lod-cloud.net/*). The pink part illustrates the life sciences data.

Table 2. Popular triple stores

Name	Language	Cluster	Inference	Available at
4store	C	Yes	No	http://www.4store.org/
Bigdata	Java	Yes	RDFS and limited OWL inference	http://www.bigdata.com/
Mulgara	Java	Yes	RDFS and limited OWL	http://www.mulgara.org/
OWLIM	Java	Yes	RDFS, OWL 2 RL and OWL 2 QL	http://www.ontotext.com/owlim
Virtuoso	C	Yes	limited RDFS and OWL	http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/
AllegroGraph	Common Lisp	Yes	RDFS and limited OWL	http://franz.com/agraph/allegrograph/
Apache Jena	Java	Yes	RDFS, OWL	http://jena.apache.org/
RDF-3X	C++	No	No	https://www.mpi-inf.mpg.de/~neumann/rdf3x/
Sesame	Java	Yes	RDFS	http://www.openrdf.org/

Life science has 40 datasets with more than 3 billion triples, accounting for 9.60% of all data.

2.5. *Triple store*

Life and health science communities have made remarkable progress as early adopters of Semantic Web technologies. A triple store is a database for the storage and retrieval of triples. The UniProt knowledge base (*49*) connects more than 150 molecular biology and chemoinformatics databases and integrates, interprets, and standardizes data from numerous resources to achieve the most comprehensive catalog of protein sequences and functional annotations. As another example, the Protein Data Bank Japan (PDBj) (*50*) accepts and processes PDB entries that are deposited mainly from Asian and Oceanic researchers and maintains a centralized archive of macromolecular structures in collaboration with other Worldwide Protein Data Bank (wwPDB) members, including the Research Collaboratory for Structural Bioinformatics (RCSB) PDB (*51*), the Biological Magnetic Resonance Bank (BMRB) (*52*) in the US, and the Protein Data Bank Europe (PDBe) (*53*) in Europe. DNA Data Bank of Japan (DDBJ) (*54*) contains approximately 8 billion triples, a number that will likely increase. Whether RDF stores can meet the needs of a biological database provider, such as loading, querying, and scaling the data efficiently, will be a major concern.

The triple store benchmark is a benchmark for evaluating the performance of storage systems, such as load cost, query performance and scalability. The Benchmark can be classified into a synthetic data benchmark and a real data benchmark. The Lehigh University Benchmark (LUBM) (*55*) and the Berlin SPARQL Benchmark (BSBM) (*56*) are two often-used general benchmarks, and they use a data generator to produce synthetic e-commerce knowledge data. Cell Cycle Ontology (*57*) and BioBenchmark Toyama 2012 (*58*) uses real biological data. BioBenchmark Toyama evaluated five triple stores, 4store (*59*), Bigdata (*60*), Mulgara (*61*), Virtuoso (*62*), and OWLIM-SE (*63*), with five biological data sets, Cell Cycle Ontology, Abbreviation/Long Form Search in Life Sciences (Allie), PDBj, UniProt, and DDBJ, ranging in size from approximately 10 million to 8 billion triples. Table 2 lists some popular triple stores according to their implemented language, inference ability and the presence of support for running in clusters. 4store was used in cell cycle ontology. Mulgara was used as an internal triple store in DDBJ. OWLIM-SE has been applied as a UniProt triple store. Virtuoso shows good performance in BSBM and DBpedia SPARQL Benchmark. Bigdata is a complete free open source triple that performs averagely well in BSBM, supports most inference functions and runs in both single node and cluster modes and could be a potentially good candidate to customize one's own triple store.

2.6. *Triple store in the cloud*

To address such large-scale data management and analysis, Semantic Web services necessitate the adoption of advances in high performance computing (*64*), such as cloud computing (*65,66*). Cloud computing has been proposed as a promising technology to solve both the economic and efficiency problems caused by the data explosion. Users do not need to purchase and install their own local expensive servers, and cloud computing vendors prepare all the computing resources and infrastructures as on-demand services. Users need only to pay the rental fee for the resources they have used in the cloud, which saves much money as the users pay by use instead of provisioning for peak (high-end sources purchased only for dealing with tough but few tasks). The most important benefit is that cloud computing greatly facilitates the sharing of analysis pipelines and data between researchers. According to the level of resources to be shared, cloud computing can be divided into four categories (*67*): Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). DaaS provides on-demand access to up-to-date public data that can be accessed and used through the Internet. SaaS provides online software services in publicly accessible servers. PaaS provides a platform that enables users to develop, test and deploy their applications in the cloud. IaaS provides virtualized resources, including hardware and software, through the Internet. Cloud computing

provides big data in the life sciences field with good storage space, web services and development platforms.

The ability to address big data studies on cloud-based triple stores is drawing more attention. Apache Cassandra (*68*) is a cloud database with linear scalability. CumulusRDF (*69*) is an RDF store on a cloud-based architecture, licensed under the GNU Affero General Public License. The current version uses Apache Cassandra as a storage backend. CumulusRDF supports a SPARQL1.1 endpoint and allows for fast queries of 1 billion triples on 16 nodes (*70*). Apache HBase is an open source, horizontally scalable, row consistent, low latency, random access data store. HBase has a proven track-record for scaling out to clusters containing approximately 1,000 nodes. It has been implemented as two versions: Jena-HBase (*71*), using Jena as the SPARQL query engine, and Hive+HBase, a SQL-like data warehousing tool that allows for querying using MapReduce (*72*). MapReduce is a programming model and an associated implementation for processing and generating large data sets. MapReduce is highly fault tolerant and scalable and can run on clusters with thousands of machines, facilitating its wide use as a cloud programming framework in bioinformatics (*67*). The projects (*72-76*) focused on developing large-scale RDF stores using the MapReduce paradigm. Fensel *et al*. (*77*) focused on web-scale data analysis and reasoning. Stratustore (*78*) is an RDF store that uses Amazon's SimpleDB as an RDF store back-end in combination with Jena's API. It is an open source project. The results show that its performance is not competitive with other RDF stores such as Virtuoso when using 20 simultaneous Stratustore instances. The throughput of the system also increases as the number of Stratustore instances grows. Bugiotti *et al*. used SimpleDB to store RDF files in the Amazon Simple Storage Service (S3) and used Amazon SimpleDB to store the index (*79*). Dydra (*80*) relied on the Amazon EC2 infrastructure, providing a SPARQL endpoint to query the data stored. SHARD (*81*), a Berkeley Software Distribution (BSD) licensed open source project, is a proof-of-concept high-performance, low-cost distributed computing technology to develop a highly scalable triple-store built on Hadoop and Hadoop Distributed File System (HDFS). Accumulo (*82*) is an open-source, distributed, column-oriented store model. Rya (*83*) uses Accumulo as a storage backend. The evaluation (*83*) showed that, in most cases, Rya outperforms existing distributed RDF solutions.

3. Challenge

Semantic Web technologies were not born for big data. As the basis for Semantic Web technologies, RDF was originally designed as a metadata data model in 1997, providing interoperability between applications that exchange machine-understandable information on the Web. Six joint documents (Primer, Concepts, Syntax, Semantics, Vocabulary, and Test Cases) superseded the W3C RDF Recommendation and described updates to the syntax and a more detailed model in 2004. In 2014 RDF Schema 1.1, as well as more representation formats such as JSON-LD, was introduced. The introduction of vocabulary, semantics, formal syntax, and rich representation formats made RDF evolve into a general-purpose language for representing information on the Web. The introduction of vocabulary and semantics (such as RDF Schema (RDFS), OWL, *etc*.) laid a foundation for dealing with the *variety* problem of big data in life sciences. Take wwPDB as an example.

wwPDB is a collection of the experimentally determined 3D structures of biopolymers and their complexes. Metadata such as *Functional Keywords*, *Biological source and Total molecular weight* of an entry are encoded into RDF data directly, while the corresponding detailed structure information of the entry is encoded into URI links as a resource. Therefore the detailed information, such as atom model, can be retrieved from the linked file, "<PDBo:link_to_pdbml_extatom rdf:resource="ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/XML-extatom/1gof-extatom.xml.gz"/>" for the entry 1GOF in the PDBj database. Compared with relational database systems, RDF is more flexible for defining metadata with the current vocabulary. In the following statements, owl:DatatypeProperty defines a data type property instance metadata "datablockName". rdfs:domain indicates that the subjects of such property must belong to a "datablock" class, and the property itself should be a "string" Class.

```
<owl:DatatypeProperty xmlns:xsd="http://www.w3.org/2001/
XMLSchema" rdf:ID="datablockName">
<rdfs:domain rdf:resource="#datablock"/>
<rdfs:range rdf:resource="http://www.w3.org/2001/
XMLSchema#string"/>
</owl:DatatypeProperty>
```

According to the priority or importance of the data, one can choose to encode the information into the RDF model to do further analysis or act as search tags, or only include the detailed information into a linked file. By temporarily omitting the data file and concentrating on the metadata, the search or analysis can be reduced to a more effective space. Likewise Semantic web technologies can effectively manage the metadata of various kinds of data, such as videos and images, thus providing a good solution for the famous *variety* problem of big data.

On the other hand, the other two Vs, velocity and volume, are still posing a big challenge for Semantic Web technologies. SPARQL 1.1, proposed in 2013, facilitates the distributed RDF data query, and is promising for enabling a global big database. However, some kinds of practical problems are hindering the

query efficiency, such as some SPARQL endpoints do not support SPARQL 1.1 yet; no Vocabulary of Interlinked Datasets (VoID) for Semantic Web Integrator and Query Engine (SemWIQ) (*84*), Web of Data Query Analyzer (WoDQA) (*85*), and SPARQL Endpoint Federation Exploiting VOID Descriptions (Splendid) framework (*86*); no service description for DARQ system (*87*) and so on. Specifications and guidance about what artifacts a SPARQL endpoint is obliged to offer are needed to make a federated query responsive. A lot of effort has also been put into the research of triple store. Wu *et al*. in the BioBenchmark (58) show that a single Virtuoso 6.4 and OWLIM-SE 5.1 node can deal with 8 billion triples well. BSBM V3.1 also proves that Virtuoso 7 can handle 150 billion triples with 8 machines. At the same time distributed RDF systems based on Hadoop and other cloud platforms, as mentioned in Section 2.6, are also being developed rapidly. It still needs a great effort to effectively manage petabyte and even bigger data.

Data integration is a typical application of Semantic Web technologies in life sciences. TogoTable uses database identifiers (IDs) in the table as a query key for searching. Because TogoTable (*88*) uses RDF, it can integrate annotations from not only the reference database to which the IDs originally belong, but also externally linked databases *via* the LOD network. TogoGenome (*http://togogenome.org/*) is a Semantic Web-based genome database collection. Neurocommons project (*89*) uses Semantic Web technology for assembling and querying biomedical knowledge from multiple sources and disciplines.

However, the concept of big data, especially big data in the life sciences, is still in its infancy. Personal data, such as a personal genome, personalized medicine, and clinical data (*e.g.*, electronic health records), are mostly still in an embryonic stage and located in the local data warehouses of their specific organizations. Effective-enough data processing platforms are needed to provide enough incentives for biological organizations to publish and share their data. More importantly, the platform or systems must ensure data security in a collaborative environment and not risk medical privacy (*24*). S3QL provides a permission control mechanism that allows the users to protect their data by specifying contextual minutia (*90*). Cloud security solutions include the use of better security systems with advanced encryption algorithms and proper signing of Service level agreements (*91*). Private and hybrid clouds are being built to ensure data safety (*92*). Similar problems exist in the LOD project. In addition, maintaining semantics links in a dynamic big data era is another difficult problem for LOD.

Big data in the life sciences requires a high level of knowledge of both biology and computer science. Big data technologies, such as cloud-based applications, are based on parallel processing. Until now, few bioinformatics tools have been designed to run in parallel (*6*), a process that requires a high level of computational know-how. By contrast, ontology design, data analysis, hypothesis building and validation, and many other problems need specialized biological knowledge.

Despite these difficulties the flexibility on metadata, the development of distributed and cloud-based triple store systems, and the improvement of federated query systems facilitate Semantic Web technologies as a promising solution for the big data in life sciences, with great efforts and collaboration from the computer and biological community.

4. Conclusions

Big data poses a great challenge for the life sciences. To address the heterogeneous variety of life scientific big data, a series of Semantic Web technologies provides a promising solution. RDF, SPARQL, triple store and ontology facilitate the integration and analysis of heterogeneous multi-disciplinary data. Linked data turns the Web into a giant global database. Triple store in the cloud takes full advantage of cloud services to address the exponential growth of biological data. Although still in its infancy, the whole scientific community is making efforts to develop new technologies and tools to ensure that big data is accessible, analyzable and applicable to the field of life sciences.

Acknowledgements

This work has been supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

References

1. Moore GE. Cramming more components onto integrated circuits. Proc IEEE. 1998; 86:82-85.
2. Walter C. Kryder's law. Sci Am. 2005; 293:32-33.
3. Reynolds C. As we may communicate. ACM SIGCHI Bulletin. 1998; 30:40-44.
4. Szalay A, Gray J. 2020 computing: Science in an exponential world. Nature. 2006; 440:413-414.
5. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010; 11:207.
6. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. J Biomed Inform. 2013; 46:774-781.
7. Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 2002; 30:4103-4117.
8. Harvard School of Public Health. The promise of 'big data'. 2012. *http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/* (accessed August 4, 2014).
9. Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. Nat Biotech. 2012; 30:295-296.

10. Davies K. The $1,000 genome: the revolution in DNA sequencing and the new era of personalized medicine. Am J Hum Genet. 2010; 87:742.
11. Grossman RL, White KP. A vision for a biomedical cloud. J Intern Med. 2012; 271:122-130.
12. Turk-Browne NB. Functional interactions as big data in the human brain. Science. 2013; 342:580-584.
13. Wild DJ, Ding Y, Sheth AP, Harland L, Gifford EM, Lajiness MS. Systems chemical biology and the Semantic Web: what they mean for the future of drug discovery research. Drug Discov Today. 2012; 17:469-474.
14. NCBI. GenBank. *http://www.ncbi.nlm.nih.gov/genbank/* (accessed August 4, 2014).
15. Harger C, Skupski M, Bingham J, *et al*. The genome sequence database (GSDB): improving data quality and data access. Nucleic Acids Res. 1998; 26:21-26.
16. UniProt. *http://www.uniprot.org/* (accessed August 4, 2014).
17. OMIM-Online Mendelian Inheritance in Man. *http://www.omim.org/* (accessed August 4, 2014).
18. Davidson SB, Overton C, Buneman P. Challenges in integrating biological data sources. J Comput Biol. 1995; 2:557-572.
19. Li A. Facing the challenges of data integration in biosciences. Eng Lett. 2006; 13:327.
20. Zhang Z, Bajic VB, Yu J, Cheung KH, Townsend JP. Data integration in bioinformatics: current efforts and challenges. In: Bioinformatics – Trends and Methodologies. 2011. DOI: 10.5772/21654.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. Bioinformatics. 2009; 25:2078-2079.
22. Reese MG, Moore B, Batchelor C, Salas F, Cunningham F, Marth GT, Stein L, Flicek P, Yandell M, Eilbeck K. A standard variation file format for human genome sequences. Genome Biol. 2010; 11:R88.
23. Ruttenberg A, Clark T, Bug W *et al*. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007; 8(Suppl 3):S2.
24. Chen H, Yu T, Chen JY. Semantic web meets integrative biology: A survey. Brief Bioinform. 2013; 14:109-125.
25. Bizer C, Boncz P, Brodie ML, Erling O. The meaningful use of big data: four perspectives - four challenges. Sigmod Rec. 2011; 40:56-60.
26. Neumann EK, Miller E, Wilbanks J. What the semantic web could do for the life sciences. Drug Discov Today Biosilico. 2004; 2:228-236.
27. Berners-Lee T, Hendler J, Lassila O. The semantic web. Sci Am. 2001; 284:34-43.
28. W3C. RDF Primer. *http://www.w3.org/TR/2004/REC-rdf-primer-20040210/* (accessed August 4, 2014).
29. W3C. The Self-Describing Web. *http://www.w3.org/2001/tag/doc/selfDescribingDocuments#RDFSection* (accessed August 4, 2014).
30. W3C. SPARQL Query Language for RDF. *http://www.w3.org/TR/rdf-sparql-query/* (accessed August 4, 2014).
31. W3C. SPARQL 1.1 Federated Query. *http://www.w3.org/TR/sparql11-federated-query/* (accessed August 4, 2014).
32. AVAILABILITY. *http://sparqles.okfn.org/availability* (accessed August 4, 2014).
33. Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition. 1993; 5:199-220.
34. Guarino N. Formal Ontology and Information Systems. 1998. *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.1776* (accessed August 4, 2014).
35. Smith B, Kohler J, Kumar A. On the application of formal principles to life science data: A case study in the gene ontology. In: Data Integration in the Life Sciences (Lambrix P, Kemp G). Berlin Heidelberg, Springer, 2004; 2994:79-94.
36. W3C. OWL Web Ontology Language Overview. *http://www.w3.org/TR/owl-features/* (accessed August 4, 2014).
37. Suntisrivaraporn B. Empirical evaluation of reasoning in lightweight DLs on life science ontologies. 2009; *http://www.citeulike.org/user/cjm/article/6332859* (accessed August 4, 2014).
38. Bizer C, Heath T. Linked data: evolving the web into a global data space. 2011; *http://linkeddatabook.com/editions/1.0/* (accessed August 4, 2014).
39. W3C. RDF Schema 1.1. *http://www.w3.org/TR/rdf-schema/* (accessed August 4, 2014).
40. Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, Game L, Heiskanen M, Morrison N, Rocca-Serra P, Sansone SA, Taylor C, White J, Stoeckert CJ Jr. The MGED ontology: A resource for semantics-based description of microarray experiments. Bioinformatics. 2006; 22:866-873.
41. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Nat Acad Sc U S A. 2005; 102:15545-15550.
42. Demir E, Cary MP, Paley S, *et al*. The BioPAX community standard for pathway data sharing. Nature Biotechnol. 2010; 28:935-942.
43. Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, Harris MA, Hill DP, Lomax J. Cross-product extensions of the Gene Ontology. J Biomed Inform. 2011; 44:80-86.
44. Hoehndorf R, Dumontier M, Gennari JH, *et al*. Integrating systems biology models and biomedical ontologies. BMC Syst Biol. 2011; 5:124.
45. Hoehndorf R, Dumontier M, Gkoutos GV. Evaluation of research in biomedical ontologies. Brief Bioinform. 2013; 14:696-712.
46. Bauer F, Kaltenbock M. Linked open data: the essentials. 2011. *http://www.semantic-web.at/LOD-TheEssentials.pdf* (accessed August 4, 2014).
47. Bizer C, Health T, Berners-Lee T. Linked data - the story so far. IJSWIS. *http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf* (accessed August 4, 2014).
48. Bizer C. Interlinking scientific data on a global scale. Data Science Journal. 2013; 12.
49. Consortium U. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 2013; 41(D1):D43-D47.
50. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, Igarashi R, Kengaku Y, Cho H, Standley DM, Nakagawa A. Protein Data Bank Japan (PDBj): Maintaining a structural data archive and resource description framework format. Nucleic Acids Res. 2012; 40(D1):D453-D460.
51. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlić A, Quesada M. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res. 2013;

41(D1):D475-D482.
52. Ulrich E L, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z. BioMagResBank. Nucleic Acids Res. 2008; 36(suppl 1):D402-D408.
53. Velankar S, Best C, Beuth B, *et al*. PDBe: Protein data bank in Europe. Nucleic Acids Res. 2010; 38(suppl 1):D308-D317.
54. Kodama Y, Mashima J, Kaminuma E, Gojobori T, Ogasawara O, Takagi T, Okubo K, Nakamura Y. The DNA DATA BANK of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. Nucleic Acids Res. 2012; 40(D1):D38-D42.
55. Guo Y, Pan Z, Heflin J. LUBM: A benchmark for OWL knowledge base systems. Web Semantics: Science, Services and Agents on the World Wide Web. 2005; 3: 158-182.
56. Bizer C, Schultz A. The Berlin sparql benchmark. Int J Semant Web Inf. 2009; 5:1-24.
57. Mironov V, Seethappan N, Blondé W, Antezana E, Lindi B, Kuiper M. Benchmarking triple stores with biological data. SWAT4LS. 2010.
58. Wu H, Fujiwara T, Yamamoto Y, Bolleman J, Yamaguchi A. BioBenchmark Toyama 2012: An evaluation of the performance of triple stores on biological data. J Biomed Semant. 2014; 5:32.
59. 4store. *http://4store.org/* (accessed August 4, 2014).
60. Bigdata. *http://www.systap.com/bigdata.htm* (accessed August 4, 2014).
61. Mulgara. SEMANTIC STORE. *http://www.mulgara.org/* (accessed August 4, 2014).
62. OPENLINK SOFTWARE. Virtuoso Universal Server. *http://virtuoso.openlinksw.com/* (accessed August 4, 2014).
63. ontotext. OWLIM. *http://www.ontotext.com/owlim* (accessed August 4, 2014).
64. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nat Rev Genet. 2010; 11:647-657.
65. Bateman A, Wood M. Cloud computing. Bioinformatics. 2009; 25:1475-1475.
66. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010; 11:207.
67. Zhou S, Liao R, Guan J. When cloud computing meets bioinformatics: A review. J Bioinform Comput Biol. 2013; 11:1330002.
68. Cassandra. *http://cassandra.apache.org/* (accessed August 4, 2014).
69. Harth G, Ladwig A. CumulusRDF: linked data management on nested key-value stores. ISWC 2011.
70. Cudré-Mauroux P, Enchev I, Fundatureanu S, Groth P, Haque A, Harth A, Keppmann F, Miranker D, Sequeda J, Wylot M. NoSQL databases for RDF: an empirical evaluation. ISWC 2013.
71. Vaibhav K, Hadilkar MK, Bhavani Thuraisingham, Castagna P. Jena-HBase: a distributed, scalable and efficient RDF triple store. ISWC 2012.
72. Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Comm ACM. 2008; 51:107-113.
73. Mika P, Tummarello G. Web semantics in the clouds. Ieee Intell Syst. 2008; 23:82-87.
74. Husain M, Khan L, Kantarcioglu M, Thuraisingham B. Data intensive query processing for large rdf graphs using cloud computing tools. IEEE CLOUD 2010.
75. Myung J, Yeon J, Lee Sg. SPARQL basic graph pattern processing with iterative MapReduce. MDAC 2010.
76. Schatzle A, Przyjaciel-Zablocki M, Lausen G. PigSPARQL: Mapping SPARQL to Pig Latin. SIGMOD/PODS 2011.
77. Fensel D, van Harmelen F, Andersson B, Brennan P, Cunningham H, Della Valle E, Fischer F, Huang Z, Kiryakov A, Lee TI. Towards LarKC: A platform for web-scale reasoning. ICSC 2008.
78. Stein R, Zacharias V. RDF on cloud number nine. SIGMOD/PODS 2012.
79. Bugiotti F, Goasdoue F, Kaoudi Z, Manolescu I. RDF data management in the Amazon cloud. EDBT/ICDT Workshops 2012.
80. DYDRA. *http://dydra.com/* (accessed August 4, 2014).
81. Rohloff K. Cloud computing for scalability: the SHARD triple-Store. DIDC 2011.
82. Accumulo. *http://wiki.apache.org/incubator/accumuloproposa* (accessed August 4, 2014).
83. Punnoose R, Crainiceanu A, Rapp D. Rya: A scalable RDF triple store for the clouds. Cloud-I 2012.
84. Langegger A, Wöß W, Blöchl M. SemWIQ-semantic web integrator and query engine. INFORMATIK 2008.
85. Akar Z, Halaç TG, Ekinci EE, Dikenelli O. Querying the web of interlinked datasets using void descriptions. LDOW 2012.
86. Görlitz O, Staab S. SPLENDID: SPARQL endpoint federation exploiting void descriptions. COLD 2011.
87. Quilitz B, Leser U. Querying distributed RDF data sources with SPARQL. ESWC 2008.
88. Kawano S, Watanabe T, Mizuguchi S, Araki N, Katayama T, Yamaguchi A. TogoTable: cross-database annotation system using the Resource Description Framework (RDF) data model. Nucleic Acids Res. 2014. doi: 10.1093/nar/gku403.
89. Ruttenberg A, Rees JA, Samwald M, Marshall MS. Life sciences on the Semantic Web: the Neurocommons and beyond. Brief bioinform. 2009:10:193-204.
90. Deus H, Correa M, Stanislaus R, Miragaia M, Maass W, Lencastre H, Fox R, Almeida J. S3QL: A distributed domain specific language for controlled semantic integration of life sciences data. BMC Bioinformatics. 2011; 12:1-15.
91. Nemade P, Kharche H. Big data in bioinformatics & the era of cloud computing. IOSR-JCE. 2013; 14:53-56.
92. Qiu J, Ekanayake J, Gunarathne T, Choi JY, Bae SH, Li H, Zhang B, Wu TL, Ruan Y, Ekanayake S. Hybrid cloud and cluster computing paradigms for life science applications. BMC bioinformatics. 2010; 11(Suppl 12):S3.

(Received April 7, 2014; Revised July 18, 2014; Re-revised August 4, 2014; Accepted August 12, 2014)