Quantifying the content of Biomedical Semantic Resources as a core for Drug Discovery Platforms

Ali Hasnain[1] and Dietrich Rebholz-Schuhmann[1]

Insight Centre for Data Analytics, National University of Ireland, Galway firstname.lastname@insight-centre.org

Abstract. The biomedical research community is providing large-scale data sources to enable knowledge discovery from the data alone, or from novel scientific experiments in combination with the existing knowledge. Increasingly semantic Web technologies are being developed and used including ontologies, triple stores and combinations thereof. The amount of data is constantly increasing as well as the complexity of data. Since the data sources are publicly available, the amount of content can be measured giving an overview on the accessible content but also on the state of the data representation in comparison to the existing content. For a better understanding of the existing data resources, i.e. judgements on the distribution of data triples across concepts, data types and primary providers, we have performed a comprehensive analysis which delivers an overview on the accessible content for semantic Web solutions (from publicly accessible data servers). It can be derived that the information related to genes, proteins and chemical entities form the core, whereas the content related to diseases and pathways forms a smaller portion. As a result, any approach for drug discovery would profit from the data on molecular entities, but would lack content from data resources that represent disease pathomechanisms.

Keywords: Biomedical Ontologies and Databases, Life Sciences Linked Open Data (LSLOD)

1 Introduction

The deluge of biomedical data in the last few years, partially caused by the advent of high-throughput gene sequencing technologies, has been a primary motivation for efforts related to curating, integrating, publishing, querying and visualising biomedical data [6,11]. The biomedical research domain encompasses a wide range of spatial and temporal scales, from genes to organism through protein, cell, tissue, and organ, as well as from molecular events to human lifetime through cell signalling, diffusion, motility, mitosis and protein turnover. Information available at those different scales is organised in data resources where each data resource mainly specialises in a particular type of data [18]. The result is a large number of established online datasets that describe human biology [23].

Nevertheless an efficient and comprehensive search activity across these datasets can become quite problematic since similar data is located in many distributed datasets and is usually available in different data models and formats[9,8,10,27]. As a result an individual scientist could perform manual search in several databases, take the results returned, change their format and paste them to the next database in search for an answer. Such a procedure would be very cumbersome and does not contribute to efficient scientific workflows [15].

The semantic connectivity between biomedical data constitutes a critical issue of biomedical scientific research and has been successfully exploited in a number of research projects for translational medicine and drug discovery [16]. Moreover the adoption of linked data technologies will allow the integration of biomedical datasets provided by different and heterogeneous data sources (i.e. research groups, libraries, databases), as well as the provision of an aggregated view of the biomedical data in a machine-readable and semantically-enriched way that will facilitate reuse [21].

At the schema level, these resources mainly consist of both domain ontologies and terminological resources [13,25]. Jimeno-Yepes et al. [14], propose a loose coupling between the domain ontologies and lexicon. Considering both have different purposes, they cannot be treated with the same techniques nor simply merged into a common resource [20]. Term vocabularies, Dictionaries and Lexicon are used interchangeably and consist of a compendium of words enriched with information of its usage [12]. Whereas a domain ontology is an explicit specialisation of a conceptualisation. Jimeno-Yepes et al. [14] propose a

spectrum that classifies ontologies and lexicon based on their semantic expressiveness. Formal lexical resources are placed towards the left hand side of the spectrum, e.g. Biolexicon[1], which contain terminology from several resources with some linguistic relevant information. UMLS[2] has been categorised as a lexicon that has been used within several applications in Natural Language Processing and text mining. Several resources have been classified between a lexicon and an ontology (e.g. UNIPROT KB Taxonomy). More complex formally-defined and semantically enriched resources lie in between the definition of ontology and lexicon such as NCI Metathesaurus[3], MeSH, the UMLS Metathesaurus and the OBO ontologies [24]. The extreme right-hand side of the spectrum represents formal ontologies such as FMA or Galen with richer expressivity.

In a recent study, the scope and the size of the terminological resources have been estimated taking into consideration the semantic domain covered by a specific resource [22]. This analysis – for the first time – quantified the "Lexeome", i.e. the full range of terms provided from the terminological (and ontological) resources to give an upper estimate of relevant entities captured in semantic resources.

In this paper the focus lies on introducing biomedical resources especially ontologies, repositories, and other data resources relevant in the context of *Drug Discovery* and *Cancer Chemoprevention*. We monitor the transformation of content into the triple representation and quantify the available content. The analysis gives an overview of which resources have to be considered, what amount of data requires integration and provides the opportunity to tailor semantic solutions to specific needs in terms of size and performance.

2 Biomedical Ontologies

There are several initiatives that address the need to standardise biomedical data. The first standard terminology, namely the International Classification of Diseases (ICD), was created in 1893 [4]. Since then several terminologies have been created. However, emphasis was given only to ensure that there are enough terms to cover the domain of focus. Over the period of time, terminologies have advanced from simple lists and hierarchies of terms to formal representations of concepts in a semantically standardised structure. As a result, terminologies that use formal representations and usable by computers are often called "ontologies" [5], [17].

In contrast to manually-created hierarchical organisations of terms (referred to as taxonomies), ontologies make use of formal structures, relations and definitions to provide a conceptualisation of domain knowledge. A large collection of biomedical ontologies or bio-ontologies are available nowadays through services e.g Bioportal[5] and OBO foundry[6]. These have mostly been developed as joint efforts by various communities to enable easy integration of biomedical data from both the literature and publicly-available biomedical databases. This section highlight the most well-studied and prominent ontologies applicable to biomedical research and especially relevant for Drug Discovery and other scenarios. Furthermore, several general ontologies used for medical and clinical terms are also investigated in order to provide insights into how data can be represented.

These ontologies can fall into three main categories, namely 1) biomedical Ontologies, 2) drugs and chemical compound ontologies and 3) upper level ontologies. The biomedical ontologies are mainly used by biomedical applications and define the basic biological structures (e.g. genes, pathways etc). The Drugs and Chemical Compound Ontologies are related to the clinical drugs and their active ingredients. Finally, the upper level ontologies describe general concepts that many biomedical ontologies share.

[1] `http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html` (retr.10/02/2017)
[2] `https://www.nlm.nih.gov/research/umls/` (retr.10/02/2017)
[3] `https://ncim.nci.nih.gov/ncimbrowser/` (retr.10/02/2017)
[4] `http://www.who.int/classifications/icd/en/HistoryOfICD.pdf` (retr.10/02/2017)
[5] `https://bioportal.bioontology.org/` (retr.10/04/2017)
[6] `http://www.obofoundry.org/` (retr.10/04/2017)

Biomedical Ontologies cover (amongst others): (1) Advancing Clinico-Genomic Trials on Cancer (ACGT) Master Ontology (MO)[7], (2) Biological Pathway Exchange (BioPAX)[8], (3) Experimental Factor Ontology (EFO)[9], (4) Gene Ontology (GO)[10], (5) Medical Subject Headings (MeSH)[11], (6) Microarray Gene Expression Data Ontology (MGED)[12], (7) National Cancer Institute (NCI) Thesaurus[13], (8) Ontology for biomedical Investigations (OBI)[14], (9) Unified Medical Language System (UMLS)[15].
Drugs and Chemical Compound Ontologies would mainly comprise RxNorm[16], and Generic and Upper Ontologies would consider: (1) Basic Formal Ontology (BFO)[17], (2) OBO Relation Ontology (RO)[18], (3) Provenance Ontology (PROVO)[19].

Table 1 provides implementation details and quantitative overview of ontologies that are listed in Section 2, year of release (as per listed at Bioportal), the visibility (public/private) and implementation details (language and type of data) of different ontologies. Size and coverage of these ontologies in terms of total triples, number of entries/entities, dependency/or reuse of any ontology on others, sub-classification and brief description are also presented in the table. We also present the quantitative comparison of different ontologies in terms of *total number of classes*, *total number of properties*, *total number of individuals* and *maximum depth*.

3 Public Data Repositories for Drug Discovery

In this section, we analyse a comprehensive list of biomedical libraries and databases closely related to drug discovery that have been provided from the biomedical community. Since drug discovery has a focus to a specific disease domain, we have chosen to focus on cancer chemoprevention as a use case and thus list data resources relevant for this domain.

The databases are separated into the following categories: (i) *Gene, Gene Expression and Protein Databases* for gene and protein annotations as well as the expression levels and related clinical data, (ii) *Pathway databases* denoting the protein interactions and the overall functional outcomes, (iii) *Chemical and Structure Databases including Biological Activities* for the information related to drugs and other chemicals including also toxicity observations and clinical trials, (iv) *Disease Specific Databases for Prevention* which deliver content specific to the prevention of cancer, and the (v) *Literature databases*.

Table 2 provides implementation details and quantitative overview of the Life Sciences related databases presented in Section 3. In addition, it lays out information regarding the year of release, accessibility (public, private) and implementation details (language and type of data) of different databases. Size and coverage of these databases in terms of total triples, number of entries / entities, sub-classification and brief description are also presented in the table.

[7] `http://bioportal.bioontology.org/ontologies/ACGT-MO` (retr.10/02/2017)
[8] `http://www.biopax.org/` (retr.10/02/2017)
[9] `http://www.ebi.ac.uk/efo/` (retr.10/02/2017)
[10] `http://www.geneontology.org/` (retr.10/02/2017)
[11] `http://www.nlm.nih.gov/mesh/` (retr.10/02/2017)
[12] `http://bioportal.bioontology.org/ontologies/MO` (retr.10/02/2017)
[13] `http://ncit.nci.nih.gov` (retr.10/02/2017)
[14] `http://obi-ontology.org/page/Main_Page` (retr. 31/01/2017)
[15] `http://www.nlm.nih.gov/research/umls/about_umls.html` (retr. 10/02/2017)
[16] `http://www.nlm.nih.gov/research/umls/rxnorm` (retr. 22/02/2017)
[17] `http://ontology.buffalo.edu/bfo/` (retr. 10/03/2017)
[18] `http://obo.sourceforge.net/relationship/` (retr. 10/03/2017)
[19] `http://bioportal.bioontology.org/ontologies/PROVO/` (retr. 25/01/2017)

Table 1: Quantitative overview of implementation details of public ontologies (selected). (**BO**:biomedical Ontologies, **DCCO**:Drugs and Chemical Compound Ontologies, **GUO**:Generic and Upper Ontologies, **T/C**:Type/Category, **Y**:Year (acc. to Bioportal), **I**ndividuals, **C**lasses/Concepts, **P**roperties, **D**epth **Pub**lic, "–"=N/A.)

Ontology	T/C	Y	Topic	Implementation	Dependency	C	P	I	D	Sub classification/ Description
ACGT-MO	BO	2008	Cancer	OWL,CSV,RDF/XML, Diff	BFO/OBO	1'769	260	61	18	data exchange in oncology, integration of clinical and molecular data
BioPAX	BO	2010	Pathways	OWL,CSV,RDF/XML	–	68	96	0	4	metabolic, biochemical, transcription regulation, protein synthesis and signal transduction pathways
EFO	BO	2015	Modelling Experimental Factors	OWL,CSV,RDF/XML,Diff	–	18'596	35	0	14	enhance and promote consistent annotation, facilitate automatic annotation to integrate external data
GO	BO	2016	Genomic and Proteomic	OWL,CSV,RDF/XML,Diff	MOD	44'195	9	0	16	for describing biological processes, molecular functions and cellular components of gene products
MeSH	BO	2009	Health	RDF/TTL,CSV	GO	252'375	38	0	15	hierarchical structure for indexing, cataloguing, and searching for biomedical and health-related information
MGED	BO	2009	microarray experiment	OWL,CSV,RDF/XML	–	233	121	698	8	describes the biological sample, the treatment regarding sample and the micro-array chip technology used in the experiment
NCIT	BO	2007	Clinical care, translational research	OWL,CSV,RDF/XML,Diff	–	118'167	173	45'715	16	integrates molecular and clinical cancer-related information enabling researchers to integrate, retrieve and relate relevant concepts
OBI	BO	2008	integration of experimental data	OWL,CSV,RDF/XML,Diff	–	2'932	106	178	16	designs, protocols, instrumentation, materials, processes, data and types of analysis in biological and biomedical investigations
UMLS	BO	1993	Medical terms & concepts	RRF	–	3'221'702	–	–	–	meta-terminology that summarise the terminologies about biomedical and health related concepts in enable interoperability
RxNorm	DCCO	1993	clinical drugs	CSV,RDF/XML,Diff	–	118'555	46	0	0	contains standard names for clinical drugs (active drug ingredient, dosage strength, physical form) and links
BFO	GUO	2003	Genuine upper level ontology	OWL,CSV,RDF/XML,Diff	–	35	0	0	5	formalise entities such as 3D enduring objects and comprehending processes conceived as extended through time
RO	GUO	2005	Core relations for OBO ontologies	OWL,CSV,RDF/XML,Diff	–	–	–	–	–	provides methodology for providing formal definitions of the basic relations that cross-cut the biomedical domain
PROVO	GUO	2012	Provenance Data Model	OWL,CSV,RDF/XML,Diff	–	30	50	4	3	provides classes, properties and restrictions to represent provenance information

3.1 Gene, Gene Expression and Protein Databases

For the complete understanding of the molecular processes, e.g., in cancer, it is highly relevant to be able to analyse the molecular processes. Such processes leads into the need to decompose functional processes into molecular processes and to predict the outcomes of such processes from the genetic background. Although cancer genomics tends to be complex due to the fact that cancer cells deviate from regular process, the genomics information – in particular the data with regards to the function of genes, their expression and transformation into proteins – is a major source for the understanding of molecular processes. The following data sources have to be considered for a complete and coherent representation of such molecular processes.

GenBank[20] is an open-access annotated collection of all publicly available nucleotide sequences and their protein translations. GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 380'000 distinct organisms.

ArrayExpress[21] archive is a database of functional genomics experiments including gene expression where one can query and download data collected to Minimum Information about a Microarray Experiment (MIAME) and Minimum Information about a high-throughput SeQuencing Experiment (MINSEQE).

Gene Expression Omnibus (GEO)[22] is a public repository that archives and freely distributes microarray, next-generation sequencing and other forms of high-throughput functional genomic data submitted by the scientific community.

Cancer Gene Expression Database (CGED)[23] is a database of gene expression profile and accompanying clinical information. This database offers graphical presentation of expression and clinical data with similarity search and sorting functions. CGED includes data on breast (prognosis and docetaxel datasets), colorectal, hepatocellular, esophageal, thyroid, and gastric cancers [4].

Universal Protein Resource (UniProt)[24] is a comprehensive resource for protein sequence and annotation data. The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate consistent and rich annotation [4]. This includes widely accepted biological ontologies, classifications and cross-references, as well as clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.

Protein Data Bank (PDB)[25] is a repository for the 3D structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography or NMR (Nuclear Magnetic Resonance) spectroscopy and submitted by biologists and biochemists from around the world, is freely accessible on the Internet. Most major scientific journals and some funding agencies require scientists to submit their structure data to the PDB[4].

Protein Database[26] is a collection of sequences from several sources, including translations from annotated coding regions in GenBank and TPA (Tissue plasminogen activator) as well as records from SwissProt, Protein Information Resource (PIR), Protein Research Foundation (PRF), UniProt and PDB. Protein sequences are the fundamental determinants of biological structure and function.

3.2 Pathway databases

As pointed out in the previous section, the modelling of pathways provides the crucial information to understand functional states in the cells. Different sources are available which partially overlap. The richest source is KEGG with about 50 M triples provided.

[20] `http://www.ncbi.nlm.nih.gov/genbank/` (retr. 10/01/2017)
[21] `http://www.ebi.ac.uk/arrayexpress/` (retr. 12/01/2017)
[22] `http://www.ncbi.nlm.nih.gov/geo/` (retr. 12/01/2017)
[23] `http://lifesciencedb.jp/cged/` (retr. 12/01/2017)
[24] `http://www.uniprot.org/` (retr. biomedical researchers can utilise cPath)
[25] `http://www.pdb.org` (retr. 20/08/2015)
[26] `http://www.hprd.org/` (retr. 20/08/2015)

Kyoto Encyclopedia of Genes and Genomes (KEGG) [27] is a database resource that integrates genomic, chemical, and systemic functional information. In particular, gene catalogues are linked to higher-level systemic functions of the cell, the organism, and the ecosystem. KEGG is further expanded towards more practical applications with molecular network-based views of diseases, drugs, and environmental compounds [4].

Reactome[28] is an open-source, open access, manually curated and peer-reviewed pathway database. The rationale behind Reactome is to convey the rich information in the visual representations of biological pathways familiar from textbooks and articles in a detailed, computationally accessible format. Entities (nucleic acids, proteins, complexes and small molecules), participating in reactions form a network of biological interactions, are grouped into pathways. Examples of biological pathways in Reactome include signalling, innate and acquired immune function, transcriptional regulation, translation, apoptosis and classical intermediary metabolism [4].

Wikipathways [19] is an open, collaborative platform dedicated to the curation of biological pathways. WikiPathways thus presents a model for pathway databases that enhance and complement ongoing efforts, such as KEGG, Reactome and Pathway Commons.

cPath: Pathway Database Software[29] is a software platform for collecting/querying biological pathways. It can serve as the core data handling component in information systems for pathway visualisation, analysis and modelling. cPath can be used for content aggregation, query and analysis. More specifically, its main features include: (i) Aggregate pathway data from multiple sources (e.g. BioCyc, KEGG, Reactome), (ii) Import/Export support with different formats PSI-MI (Proteomics Standards Initiative Molecular Interaction) and BioPAX, (iii) Data visualisation using Cytoscape and iv) Simple web service.

3.3 Chemical and Structure Databases including Biological Activities

The treatment of any disease and cancer in particular is based on chemical entities with a defined biological activity. Several data sources provide information on the chemical compound, on its relevance to specific treatments and the side effects that they may induce. The amount of data (i.e. triples) with regards to the different data sources is large and data integration is an ongoing difficult task (see OpenPhacs project). The following data sources are publicly available.

Chemical Compounds Database (Chembase)[30] collects and provides information on chemical compounds and their physical and chemical properties, NMR (Nuclear Magnetic Resonance) spectra, mass spectra, UV/Vis (Ultra-violet-Visible Spectroscopy) absorption and IR data. **Sigma-Aldrich**[31] product database includes datasheets for commercially available compounds including solubility **ChemDB**[32] is a public database of small molecules available on the Web. The database contains approximately 4.1 million commercially available compounds and 8.2 million counting isomers. The database includes a user-friendly graphical interface, chemical reactions capabilities as well as unique search capabilities.

Chemical Entities of Biological Interest (ChEBI)[33] is a database and ontology of small molecular entities. The term *'molecular entity'* refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer etc. that is identifiable as a separately distinguishable entity. Molecules directly encoded by the genome, such as nucleic acids, proteins and peptides derived from proteins by proteolysis cleavage, are not included.

DrugBank database [26] is a bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. se-

[27] `http://www.genome.jp/kegg/` (retr. 12/01/2017)
[28] `http://www.reactome.org` (retr. 12/01/2017)
[29] `http://cbio.mskcc.org/software/cpath/` (retr. 12/01/2017)
[30] `http://urlm.co/www.chembase.com#web` (retr. 12/01/2017)
[31] `https://www.sigmaaldrich.com/catalog/` (retr. 12/01/2017)
[32] `http://cdb.ics.uci.edu/` (retr. 12/01/2017)
[33] `http://www.ebi.ac.uk/chebi//` (retr. 12/01/2017)

quence, structure, and pathway) information. The database contains 6826 drug entries including 1431 Food and Drug Administration (FDA)-approved small molecule drugs, 133 FDA-approved biotech (protein/peptide) drugs, 83 nutraceuticals and 5211 experimental drugs. Additionally 4435 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries.

PubChem[34] provides information on the biological activities of small molecules including substance information, compound structures, and BioActivity data in three primary databases. PubChem is integrated with Entrez, NCBI's (National Center for Biotechnology Information) primary search engine, and also provides compound neighbouring, sub/superstructure, similarity structure, BioActivity data, and other searching features [4]. PubChem contains substance descriptions and small molecules with fewer than 1000 atoms and 1000 bonds.

Aggregated Computational Toxicology Resource (ACToR)[35] is an online warehouse of all publicly available chemical toxicity data and can be used to find data about potential chemical risks to human health and the environment. ACToR aggregates data from over 500 public sources on over 500'000 environmental chemicals searchable by chemical name, other identifiers and by chemical structure [4]. It allows users to search and query data from chemical toxicity databases including: (1) ToxRefDB for animal toxicity studies, (2) ToxCastDB covering data from 1'000 chemicals in over 500 high-throughput assays, (3) ExpoCastDB consolidating human exposure and exposure factor data, and (4) Distributed Structure-Searchable Toxicity (DSSTox) for high quality chemical structures and annotations.

ClinicalTrials[36] is an up-to-date registry and results database of federally and privately supported clinical trials conducted in the United States and around the world [4].

TOXicology Data NETwork (TOXNET)[37] provides access to full-text and bibliographic databases oriented to toxicology, hazardous chemicals, environmental health and related areas.

3.4 Disease Specific Databases for Prevention

More of such databases will arise, once the specific data becomes available but currently it is limited to a smaller number of data resources with limited data contained.

Colon Chemoprevention Agents Database (CCAD) [3] contains results from a systematic review of the literature of Colon Chemoprevention in human, rats and mice. Target cancers are colorectal adenoma and adenocarcinoma, aberrant crypt foci (ACF) (a preneoplasic lesion), and Min mice polyp (adenomas in Apc+/- mutant mice). The Chemopreventive agents are ranked by efficacy (potency against carcinogenesis).

Dietary Supplements Labels Database[38] offers information on label ingredients in more than 5'000 selected brands of dietary supplements that enables users to compare label ingredients in different brands. Information is also provided on the "structure/function" claims made by manufacturers and can therefore be used to narrow down active ingredients in different types of food which may be applicable as Chemoprevention agents. Ingredients of dietary supplements in this database are linked to other databases such as MedlinePlus and PubMed [4].

REPAIRtoire Database[39] is a database resource for systems biology of DNA damage/repair. It collects and organises the information including: (i) DNA damage linked to environmental mutagenic and cytotoxic agents, (ii) pathways comprising individual processes and enzymatic reactions involved in the removal of damage, (iii) proteins participating in DNA repair and (iv) diseases correlated with mutations in genes encoding DNA repair proteins. It also provides links to publications and external databases. REPAIRtoire can be queried by the name of pathway, protein, enzymatic complex, damage and disease.

[34] `http://pubchem.ncbi.nlm.nih.gov/` (retr. 12/01/2017)
[35] `http://actor.epa.gov/actor/faces/ACToRHome.jsp` (retr. 12/01/2017)
[36] `http://clinicaltrials.gov/` (retr. 10/01/2017)
[37] `http://toxnet.nlm.nih.gov/` (retr. 12/01/2017)
[38] `http://www.dsld.nlm.nih.gov/dsld/` (retr. 20/03/2017)
[39] `http://repairtoire.genesilico.pl/` (retr. 14/01/2017)

3.5 Literature Databases

The scientific literature is still one of the most comprehensive data sources for experimental findings. The content is provided in an unstructured way and some of its content is delivered through data curation into the data sources above. The most relevant data sources are listed below.

Pubmed[40] is the most widely used source for biomedical literature. PubMed provides access to citations from the MEDLINE database and additional Life Science journals including links to many full-text articles at journal Web sites and other related Web resources. PubMed was first released in January 1996. The knowledge regarding Chemoprevention agents available as publications makes Pubmed a primary source of biomedical information [4].

PubMed Dietary Supplement Subset[41] is designed to limit search results to citations from a broad spectrum of dietary supplement literature including vitamin, mineral, phytochemical, ergogenic, botanical and herbal supplements in human nutrition and animal models. It retrieves citations on topics including: chemical composition; biochemical role and function - both in vitro and in vivo; clinical trials; health and adverse effects; fortification; traditional Chinese medicine and other folk/ethnic supplement practices. [4].

4 Linked Data

In March 2007 the W3C Semantic Web Education and Outreach (SWEO) Interest Group announced a new Community Project called *"Interlinking Open Data"*[42] that was subsequently shortened to *"Linking Open Data" (LOD)*. The goal of the Linked Open Data project is twofold: (i) to bootstrap the Semantic Web by creating, publishing and interlinking RDF exports from these open datasets, and, (ii) introduce the benefits of RDF and Semantic Web technologies to the broader Open Data community [2]. Linked Data aims to make data available on the Web in an interoperable format so that agents can discover, access, combine and consume content from different sources with higher levels of automation than would otherwise be possible. The envisaged result is a *"Web of Data"*, a Web of structured data with rich semantic links where agents can query in a unified manner, across sources, using standard languages and protocols. Over the past few years, hundreds of knowledge-bases with billions of facts have been published according to the Semantic Web standards (using RDF as a data model and RDFS and OWL to provide explicit semantics) following the Linked Data principles.

[40] `http://www.ncbi.nlm.nih.gov/pubmed` (retr. 22/02/2017)
[41] `http://ods.od.nih.gov/research/PubMed_Dietary_Supplement_Subset.aspx` (retr. 12/03/2017)
[42] `http://www.w3.org/blog/SWEO/page-2` retr. 05/02/2017

Table 2: Quantitative overview of implementation details of public libraries and databases (selected) (**LD**:Literature Databases, **T/C**:Type/ Category, **Pub**lic, **V**isibility, "–"=N/A)

Database	T/C	Year	Topic	Implementation	Content
GenBank	GDD	1982	nucleotide sequences & translations	WebBased	more than 193'739'511 sequences over 65 B nucleotide bases in more than 61 M sequences
UniProt	PD	2002	protein sequence and annotation data	WebBased/LOD	63'686'057 sequences, 21'364'768'379 amino acids classifications, cross-references, annotation of proteins
PDB	PD	1971	3D structural data of proteins	WebBased/LOD	118280 Biological Structures evidence of experimentally validated protein structures
PDBs	PD	2009	3D Protein sequences	WebBased	30'047Protein Entries, 41'327PPIs translated coding regions from GenBank, TPA, SwissProt, PIR, PRF, UniProt and PDB.
GEO	GED	2011	microarray, NGS	WebBased	3'848 datasets gene expression for specific studies
ArrayExpress	GED	2011	genomics experiments	WebBased	65060 experiments 1'973'776 assays annotated data for gene expression from biological experiments
CGED	GED	–	gene expressions & clinical data	–	– cancer of breast, colorectal, hepatocellular, esophageal, thyroid, gastric cancers
KEGG	PAD	1995	genomic, chemical, systemic info	WebBased/ LOD	432'883PathwayMaps, 153'776hierarchies genome sequencing and high-throughput experimental technologies
Reactome	PAD	2003	pathways	WebBased	9'386 Proteins pathway data for signalling, transcriptional regulation, translation, apoptosis
Wikipathways	PAD	2007	biological pathways	WebBased	2'475 pathways pathway database complementing e g. KEGG, Reactome, Pathway Commons
cPath	PAD	2005	biological pathways	Desktop/ WebBased	31'698 pathways, 1'151'476 interactions pathway visualisation, analysis and modelling
Chembase	CPSD	–	chemical compounds	WebBased	150'000 pages compounds, their physical and chemical properties, mass spectra
Sigma-Aldrich	CPSD	1975	compounds datasheets	WebBased	200'00+ products, 500+ services data for commercial compounds
ChemDB	CPSD	1989	small molecules	WebBased	more than 4.1 M compounds 4.1 M commercial compounds, 8.2 M counting isomers
CCAD	DSCD	2002	colon chemoprevention	WebBased	1,137 agents literature data for colon chemoprevention in human, rats, mice
DrugBank	BACD	2008	drug data	WebBased/LOD	8,261 drugs, 4,164 targets, 243 Enzymes, 118 Transporters, drug (chemical, pharmaceutical), drug target (sequence, structure, pathway)
ChEBI	BACD	–	small molecules	WebBased/LOD	48'296 compounds natural and synthetic atom, molecule, ion, radical, conformer
PubChem	BACD	2004	compound structures, bioactivity data	WebBased/LOD	89'124'401 Compounds compound neighbouring, sub/superstructure, bioactivity data
TOXNET	TED	1987	toxicology database	WebBased	– toxicology, hazardous chemicals, environmental health and related areas
ACToR	BACD	2008	chemical toxicity data	WebBased	more than 500 public source environmental chemicals searchable by name and structure
REPAIRtoire	BACD	2019	DNA damage & repair	WebBased	– DNA damage links, pathways, proteins for DNA re-pair, diseases related to mutations
ClinicalTrials	TED	2000	clinical trial data	WebBased	213'868 studies offers information for locating clinical trials for diseases and conditions
Pubmed	LD	1996	biomedical Literature	WebBased/Excel	11 M + journal citations Primary source of information for bio-medical researchers
PDSS	LD	1999	citations of dietary supplement	WebBased	– dietary supplement literature including vitamin, mineral, botanical/herbal supplements
DSLD	NSCAD	2013	label ingredients of dietary supplements	WebBased	more than 5'000 selected brands Ingredients of dietary supplements linked to MedlinePlus and PubMed

4.1 Life Sciences Linked Open Data Cloud

This section reviews the linked biomedical datasets relevant in a Cancer Chemoprevention and drug discovery scenario, three significant providers are as follow: (1) Linked Open Drug Data (LODD), (2) Bio2RDF, and (3) LinkedLifeData.

Linked Open Drug Data (LODD) [43] is a set of linked datasets relevant to Drug Discovery. It includes data from several datasets including Drugbank, LinkedCT, DailyMed, Diseasome, SIDER, STITCH, Medicare, RxNorm, ClinicalTrials.gov, NCBI Entrez Gene and OMIM. The LODD datasets have been crawled by the Semantic Web Search Engine (SWSE)[44] that can be accessed via a faceted browsing interface.

Bio2RDF[45] constitutes a project that contains multiple linked biological databases including pathways databases such as KEGG, PDB and several NCBIs databases [1]. Bio2RDF is an open-source project that uses Semantic Web technologies to build and provide the largest network of Linked Data for the Life Sciences. Bio2RDF defines a set of simple conventions to create RDF(s) compatible Linked Data from a diverse set of heterogeneously formatted sources obtained from multiple data providers.

As of July 2014, Bio2RDF Release 3 contains[46] about 11 billion triples across 35 datasets (based on Virtuoso 7.1.0 as the SPARQL 1.1 endpoint). The new types of data have been included for example from OrphaNet, PubMed, SIDER, GenDR, and LSR. Further local endpoints have been integrated: Chembl, LinkedSPL, PathwayCommons, and Reactome. In the current version, every URI is an instance of an `owl:Class`, `owl:ObjectProperty`, or `owl:DatatypeProperty`.

LinkedLifeData (LLD)[47] is a semantic data integration platform for the biomedical domain containing 5 billion RDF statements from various sources including UniProt, PubMed, EntrezGene and 20 more. LDD allows writing complex data analytical queries, answering complex bioinformatics questions, helps navigate through the information or export results subsets. LDD offers two different access levels: (1) LLD Public – completely free anonymous access; and (2) LLD Enterprise – premium service access with extra features.

4.2 Quantitative overview of Datasets

Table 3 provides implementation details and quantitative overview of dataset listed in Section 4.1, but also information regarding the year of release (as per reported at `http://www.datahub.io`, `http://www.bio2rdf.org`, `http://www.linkedlifedata.com`), the visibility (public/ private) and the implementation details (language and type of data) provided by different datasets. Size and coverage of these datasets in terms of total triples, number of entries/entities, link of SPARQL endpoint, sub-classification and brief description is also presented in the table. Quantitative comparison of different datasets in terms of combination of information including *total number of classes*, *total number of properties*, *total number of Instances*, *total number of triples* and *total number of entities* are also presented.

As can be expected (see Tbl. 3), the largest triple store collections (2 to 10 B triples, top section of the table) have been retrieved from triple stores that reference genes or proteins and branch out to the reference information after data integration. These triple stores will serve as a reference data resource, since the data integration is performed by providers of several of the integrated databases.

[43] `http://www.w3.org/wiki/HCLSIG/LODD` (retr: 05/02/2017)
[44] `http://swse.deri.org/(retr.27-04-2016)`
[45] `http://bio2rdf.org` (retr: 05/02/2017)
[46] `https://github.com/bio2rdf/bio2rdf-scripts/wiki` (retr: 05/02/2017)
[47] `http://linkedlifedata.com` (retr: 05/02/2017)

Table 3: Counts of triples (**:T**) and entities **:E** across the most relevant datasets across LSLOD, Bio2RDF and LLD (**T/C**:Type/Category, **Y/D**:Year/Date, **E/F**:Environmental Factors, **SPLs**:Structured Product Labels, **DIKB**:Drug Interaction Knowledge Base, **LLD**:Linked Life Data) "–" = N/A)

Dataset	T/C	Y/D	Topic	Size/ Coverage	Description
LLD	LLD	2014-06-04	Drugs, Chromosomes etc	10'192'641'644**:T**, 15'536'206'391**:E**	25 public biomedical databases with access to complex bioinformatics
iProClass	Bio2RDF	2014-06-09	Proteins, pathways, genes	3'306'107'223**:T**, 364'255'265**:E**	UniProtKB and UniParc proteins, with links to biological databases
NCBI Gene	Bio2RDF	2014-09-20	Genes	2'010'283'833**:T**, 189'594'629**:E**	nomenclature, RefSeqs, maps, pathways, variations, phenotypes, locus-specific
PubMed	Bio2RDF	2014-06-27	Citations	500'5343'905**:T**, 412'593'720**:E**	citations from MEDLINE and LS journals for biomedical articles after 1950s
ChEMBL	Bio2RDF	–	bioactive compounds,bioactivities	409'942'525**:T**, 50'061'452**:E**	bioactive compounds, quantitative properties and bioactivities)
CTD	Bio2RDF	2014-06-09	Chemical-gene/protein interactions	326'720'894**:T**, 19'768'641**:E**	cross-species chemical-gene/protein interactions, chemical-gene-disease
PharmGKB	Bio2RDF	2014-06-27	genotype/phenotype	278'049'209**:T**, 25'325'504**:E**	genotype/phenotype data, gene variants, gene-drug-disease relationships
ClinicalTrials	Bio2RDF	2014-09-25	Clinical Trials	98'835'804**:T**, 7'337'123**:E**	publicly and privately supported clinical studies
GOA	Bio2RDF	2014-06-05	Gene Ontology Annotations	97'520'151**:T**, 5'950'074**:E**	Gene Ontology(GO) annotations to proteins in UniProtKB and IPI
Affymetrix	Bio2RDF	2014-08-01	Microarrays	86'942'371**:T**, 6'679'943**:E**	probesets used in the Affymetrix microarrays
KEGG	Bio2RDF	2014-08-13	Genes	50'197'150**:T**, 6'533'307**:E**	16 databases of biological, genomic, and chemical information
iRefIndex	Bio2RDF	2014-06-22	Proteins, pathways, genes	48'781'511**:T**, 3'110'993**:E**	protein interactions in BIND/BioGRID/DIP/HPRD/MPPI/OPHID
WormBase	Bio2RDF	2014-06-04	Genome	22'682'002**:T**, 1'840'311**:E**	genome of the Caenorhabditis elegans
Taxonomy	Bio2RDF	2014-05-27	Taxonomy	21'310'356**:T**, 1'147'211**:E**	organisms in the genetic databases with one nucleotide or protein sequence
BioPortal	Bio2RDF	2014-07-20	Biological/biomedical ontologies	19'920'395**:T**, 2'199'594**:E**	an open repository of biomedical ontologies
SIDER	Bio2RDF	2014-07-22	Drugs	17'627'864**:T**, 1'222'429**:E**	medicines and adverse drug reactions, side effect frequency/ classifications
SGD	Bio2RDF	2014-08-07	Biochemical reactions	12'494'945**:T**, 957'558**:E**	molecular biology and genetics of the yeast Saccharomyces cerevisiae
Reactome	Bio2RDF	–	Pathways	12'487'446**:T**, 2'461'010**:E**	core pathways and reactions in human biology
dbSNP	Bio2RDF	2014-07-15	Nucleotide substitutions	8'801'487**:T**, 530'538**:E**	single nucleotide substitutions, deletion insertion polymorphisms
OMIM	Bio2RDF	2014-09-19	Mendelian disorders, Genes	8'750'774**:T**, 1'013'389**:E**	human genes and genetic phenotypes
MGI	Bio2RDF	2014-06-05	Genes	8'206'813**:T**, 924'257**:E**	gene, nomenclature, mapping, homologies, sequence links, phenotypes, allelic's
MeSH	Bio2RDF	2014-05-27	terms and terminologies	7'323'864**:T**, 305'401**:E**	naming descriptors in hierarchical structure for searching specificity
HomoloGene	Bio2RDF	2014-07-04	Annotated Gene	7'189'769**:T**, 869'985**:E**	automated detection of homologs among the annotated genes
NDC	Bio2RDF	2014-08-02	Drugs Identifies	6'199'488**:T**, 488'146**:E**	UPI used in the USA for drugs intended for human use
PC	Bio2RDF	–	Pathways	5'700'724**:T**, 1'024'572**:E**	biological pathway information collected from public pathway databases
DrugBank	Bio2RDF	2014-07-25	Drugs	3'672'531**:T**, 316'950**:E**	detailed drug data with comprehensive drug target
HGNC	Bio2RDF	2014-07-04	Human Gene	3'628'205**:T**, 372'136**:E**	gives unique and meaningful names to every human gene
SABIO-RK	Bio2RDF	2014-06-05	Biochemical reactions	2'716'421**:T**, 448'248**:E**	biochemical reactions, kinetic equations, parameters, conditions
BioModels	Bio2RDF	2014-06-05	Biological/mathematical models	2'380'009**:T**, 188'380**:E**	store, search, retrieve published mathematical models of biological interests
InterPro	Bio2RDF	2014-06-02	Proteins and Genomes	2'323'345**:T**, 176'579**:E**	predictive protein "signatures"/ annotation of proteins and genomes
LinkedSPL	Bio2RDF	–	Drugs	2'174'579**:T**, 59'776**:E**	Linked Data version of DailyMed
WikiPathways	Bio2RDF	–	Pathway maps	514'397**:T**, 71'879**:E**	open and public collection of pathway maps
Orphanet	Bio2RDF	2014-06-02	Rare diseases/Orphan drugs	377'947**:T**, 28'871**:E**	rare diseases and orphan drugs. Diagnosis, care and treatment of rare diseases
GenAge	Bio2RDF	2014-06-03	Genes	73'048**:T**, 6'995**:E**	human and model organism genes related to longevity and ageing
LSR	Bio2RDF	2014-07-16	LS terminologies	55'914**:T**, 5'032**:E**	datasets and terminologies used in the Life Sciences
GenDR	Bio2RDF	2014-06-03	Genes	11'663**:T**, 1'129**:E**	genes associated with dietary restriction (DR)

The next collection of triple stores (200 to 500 M triples; PubMed, ChEMBL, CTD, PharmGKB) are primary data resources that cover individual observations, where a scientific publication is categorized similarly. All these data resources are growing at a rate that is linked to ongoing research in this domain, in contrast to a data resource that would report on scientific entities that can only be discovered once, e.g. a specific protein in a given species.

The following two fields of data resources (50 to 100 M triples; 12 to 50 M triples) contain different types of resources. The data in the resource from the first group correlates with experiments that are performed according to discovery needs and may lose relevance over time (see Affymetrix data). The second group contains reference data resources for species (Wormbase, SGD), pathways (KEGG, Reactome, iRefIndex), but also large-scale resources with a very specific purpose, such as Taxonomy, BioPortal, and SIDER.

For the remaining resources, it can be expected that they will be developing into large-scale resources as seen above (MGI, dbSNP, BioModels) whereas others by the nature of their content, would show only very limited growth, such as HGNC, DrugBank, Orphanet, and also possibly InterPro. Further resources have been considered (ref. tbl. 4), but could not be analysed to the degree of detail as for the data resources given in Tbl. 3.

As a conclusion, the life science research community has to determine, which technological solutions allow the delivery of the large-scale semantic Web triple stores to the general public. Other data resources may well be replicated at different sites for local integration work.

5 Related Work

Zeginis et al. [28] proposed "meet-in-the-middle" approach to develop the semantic model relevant for cancer chemoprevention. Relevant data was analysed in a bottom-up fashion from analysing the domain whereas a top-down approach was considered to collect ontologies, vocabularies and data models. Hasnain et al. [7] proposed Linked Biomedical Dataspace with components namely: a) knowledge extraction, b) link creation, c) query execution and d) knowledge publishing, all of these access and use biomedical resources relevant for cancer chemoprevention.

6 Concluding Remarks

In this paper we introduce different tiers of biomedical Data relevant to *Cancer Chemoprevention* and *Drug Discovery* domain. This involves Ontologies, libraries and databases in Healthcare and the biomedical domain, Linked Data and Life Science Linked Open Data.

We classify ontologies into three main classes: i) biomedical Ontologies (e.g. EFO, OBI, GO etc), ii) Drugs and Chemical Compound Ontologies (e.g. RxNorm) and iii) Generic and Upper Ontologies (e.g. BFO, RO, PROV). Similarly we categorise libraries and databases in five categories that comprise (i) *Gene, Gene Expression and Protein Databases*, (ii) *Pathway databases*, (iii) *Chemical and Structure Databases including Biological Activities*, (iv) *Disease Specific Databases for Prevention*, and the (v) *Literature databases*.

Table 4: Quantitative overview of datasets involving LODD only without judgement on the number of entries versus triples. (**T/C**:Type/Category, **Y/D**:Year/Date, **E/F**:Environmental Factors, **SPLs**:Structured Product Labels, **DIKB**:Drug Interaction Knowledge Base, **LLD**:Linked Life Data), "–":N/A)

Dataset	T/C	Y/D	Topic	Size/Coverage	Description
DBpedia	LODD	2009	Drugs/ Diseases/ Proteins	218 M**:T**; 2'300 drugs, 2'200 proteins	2.49 M things extracted from Wikipedia
LinkedCT	LODD	–	Clinical Trials	25 M **:T**, 106'000 trials	Linked data source of trials from ClinicalTrials.gov
RxNorm	LODD	2011	Drugs	> 7.7 M **:T**	connects prescription drugs, ingredients and NDC through RXCUI
GHO	LODD	2011	Infectious Diseases/Demography/EF	3 M **:T**	infectious diseases at country, regional, global levels
DailyMed	LODD	2010	Drugs	1'604'893**:T**, 36'000+ product	All FDA-approved SPLs and NDF-RT
RDF-TCM	LODD	2009	Genes/Diseases/Medicine	117'643**:T**	Chinese medicine, gene, disease association and mapping to Extrez Gene IDs
Diseasome	LODD	2010	Diseases/ Genes	91'182**:T**; 2'600 genes	disorders and disease genes linked to disorder–gene associations
DIKB	LODD	2011	Drugs/ (DDIs)	> 41k **:T**	Drugs and DDIs Claims and Evidence for drug mechanisms and DDIs
UPNR	LODD	–	Drugs/Procedures/Diagnoses	38'664	800 full-text clinical notes from the University of Pittsburgh
Medicare	LODD	2010	Medicare Formulary	–	doctors, healthcare professionals, services

Access to the data repositories

Affymetrix (http://cu.affymetrix.bio2rdf.org/sparql), **BioModels** (http://cu.biomodels.bio2rdf.org/sparql), **BioPortal** (http://cu.bioportal.bio2rdf.org/sparql), **ChEMBL** (http://cu.chembl.bio2rdf.org/sparql, http://rdf.farmbio.uu.se/chembl/sparql), **ClinicalTrials** (http://cu.clinicaltrials.bio2rdf.org/sparql), **CTD** (http://cu.ctd.bio2rdf.org/sparql), **DailyMed** (http://purl.org/net/nlprepository/linkedSPLs), **DBpedia** (http://dbpedia.org/sparql), **dbSNP** (http://cu.dbsnp.bio2rdf.org/sparql), **DIKB** (http://dbmi-icode-01.dbmi.pitt.edu:2020/), **Diseasome** (http://www4.wiwiss.fu-berlin.de/diseasome/sparql), **DrugBank** (http://cu.drugbank.bio2rdf.org/sparql, http://www4.wiwiss.fu-berlin.de/drugbank/sparql), **GenAge** (http://cu.genage.bio2rdf.org/sparql), **GenDR** (http://cu.gendr.bio2rdf.org/sparql), **GHO** (http://gho.aksw.org), **GOA** (http://cu.goa.bio2rdf.org/sparql), **HGNC** (http://cu.hgnc.bio2rdf.org/sparql), **HomoloGene** (http://cu.homologene.bio2rdf.org/sparql), **InterPro** (http://cu.interpro.bio2rdf.org/sparql), **iProClass** (http://cu.iproclass.bio2rdf.org/sparql), **iRefIndex** (http://cu.irefindex.bio2rdf.org/sparql), **KEGG** (http://cu.kegg.bio2rdf.org/sparql), **LinkedCT** (http://data.linkedct.org/sparql), **LinkedLifeData** (http://linkedlifedata.com/sparql), **LinkedSPL** (http://cu.linkedspl.bio2rdf.org/sparql), **LSR** (http://cu.lsr.bio2rdf.org/sparql), **Medicare** (http://www4.wiwiss.fu-berlin.de/medicare/sparql), **MeSH** (http://cu.mesh.bio2rdf.org/sparql), **MGI** (http://cu.mgi.bio2rdf.org/sparql), **NCBI Gene** (http://cu.ncbigene.bio2rdf.org/sparql), **NDC** (http://cu.ndc.bio2rdf.org/sparql), **OMIM** (http://cu.omim.bio2rdf.org/sparql), **Orphanet** (http://cu.orphanet.bio2rdf.org/sparql), **PathwayCommons** (http://cu.pathwaycommons.bio2rdf.org/sparql), **PharmGKB** (http://cu.pharmgkb.bio2rdf.org/sparql), **PubMed** (http://cu.pharmgkb.bio2rdf.org/sparql), **RDF-TCM** (http://www.open-biomed.org.uk/sparql/endpoint/tcm), **Reactome** (http://cu.reactome.bio2rdf.org/sparql), **RxNorm** (http://link.informatics.stonybrook.edu/sparql/), **SABIO-RK** (http://cu.sabiork.bio2rdf.org/sparql), **SGD** (http://cu.sgd.bio2rdf.org/sparql), **SIDER** (http://cu.sider.bio2rdf.org/sparql, http://www4.wiwiss.fu-berlin.de/sider/sparql), **STITCH** (http://www4.wiwiss.fu-berlin.de/stitch/sparql), **Taxonomy** (http://cu.taxonomy.bio2rdf.org/sparql), **UPNR** (http://dbmi-icode-01.dbmi.pitt.edu:8080/sparql), **WikiPathways** (http://cu.wikipathways.bio2rdf.org/sparql), **WormBase** (http://cu.wormbase.bio2rdf.org/sparql).

Acknowledgements

The work presented in this paper has been partly funded by EU FP7 GRANATUM project (project number 270139) and Science Foundation Ireland under Grant No. SFI/12/RC/2289.

References

1. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics 41(5), 706–716 (2008)
2. Berners-Lee, T., Bizer, C., Heath, T.: Linked data-the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
3. Corpet, D.E., Taché, S.: Most effective colon cancer chemopreventive agents in rats: a systematic review of aberrant crypt foci and tumor data, ranked by potency. Nutrition and cancer 43(1), 1–21 (2002)
4. Deus, K.T.W.P.C.N.T.B.C.G.C.K.H.F.: D1.1 – requirements analysis. Tech. rep., CERTH, NUIG-DERI, FIT, CYBION, UCY, and DKFZ (2011)
5. Greenes, R.A., McClure, R.C., Pattison-Gordon, E., Sato, L.: The findings–diagnosis continuum: implications for image descriptions and clinical databases. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. p. 383. American Medical Informatics Association (1992)
6. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences LOD Cloud. In: 1st International Workshop on Ontology Engineering in a Data-driven World collocated with EKAW12 (2012)
7. Hasnain, A., Kamdar, M.R., Hasapis, P., Zeginis, D., Warren Jr, C.N., et al.: Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery. In: International Semantic Web Conference (In-Use Track), October 2014 (2014)
8. Hasnain, A., Mehmood, Q., Sana e Zainab, S., Saleem, M., Warren, C., Zehra, D., Decker, S., Rebholz-Schuhmann, D.: Biofed: federated query processing over life sciences linked open data. Journal of Biomedical Semantics 8(1), 13 (2017), http://dx.doi.org/10.1186/s13326-017-0118-0
9. Hasnain, A., Mehmood, Q., e Zainab, S.S., Hogan, A.: Sportal: Profiling the content of public sparql endpoints. International Journal on Semantic Web and Information Systems (IJSWIS) 12(3), 134–163 (2016), http://www.igi-global.com/article/sportal/160175

10. Hasnain, A., Mehmood, Q., e Zainab, S.S., Hogan, A.: SPORTAL: searching for public SPARQL endpoints. In: Proceedings of the ISWC 2016 Posters & Demonstrations Track co-located with 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 19, 2016. (2016), http://ceur-ws.org/Vol-1690/paper78.pdf
11. Hasnain, A., e Zainab, S.S., Kamdar, M.R., Mehmood, Q., Warren Jr, C.N., Fatimah, Q.A., Deus, H.F., Mehdi, M., Decker, S.: A roadmap for navigating the life sciences linked open data cloud. In: Semantic Technology, pp. 97–112. Springer (2014)
12. Hirst, G.: Ontology and the lexicon. In: Handbook on ontologies, pp. 269–292. Springer (2009)
13. Hoehndorf, R., Dumontier, M., Gkoutos, G.V.: Evaluation of research in biomedical ontologies. Brief. Bioinformatics (Sep 2012)
14. Jimeno-Yepes, A., Jiménez-Ruiz, E., Berlanga, R., Rebholz-Schuhmann, D.: Use of shared lexical resources for efficient ontological engineering. In: Semantic Web Applications and Tools for Life Sciences Workshop (SWAT4LS). CEUR WS Proceedings. vol. 435, pp. 93–136 (2008)
15. Karim, R., Michel, A., Zappa, A., Baranav, P., Sahay, R., Rebholz-Schuhmann, D.: Data Workflow Systems in Life Science Research: Towards Cloud Infrastructures and use of Open Data. Brief. Bioinformatics (Aug 2017)
16. Machado, C.M., Rebholz-Schuhmann, D., Freitas, A.T., Couto, F.M.: The semantic web in translational medicine: current applications and future directions. Brief Bioinformatics p. bbt079 (2013)
17. Musen, M.A.: Dimensions of knowledge sharing and reuse. Computers and biomedical research 25(5), 435–467 (1992)
18. Neumann, E., Prusak, L.: Knowledge networks in the age of the Semantic Web. Brief. Bioinformatics 8, 141–149 (May 2007)
19. Pico, A.R., Kelder, T., Iersel, M.P., Hanspers, K., Conklin, B.R., Evelo, C.: Wikipathways: pathway editing for the people. PLoS biology 6(7) (2008)
20. Rebholz-Schuhmann, D., Oellrich, A., Hoehndorf, R.: Text-mining solutions for biomedical research: enabling integrative biology. Nat. Rev. Genet. 13(12), 829–839 (Dec 2012)
21. Rebholz-Schuhmann, D., Grabmuller, C., Kavaliauskas, S., Harrow, I., Kapushevsky, M., Westaway, M., Woollard, P., Wilkinson, N., Strutt, P., Braxtenthaler, M., Hoole, D., Wilson, J., o'Beirne, R., Kidd, R.R., Filsell, W., Marshall, C., Backofen, R., Clark, D.: Semantic integration of gene-disease associations for diabetes type ii from literature and biomedical data resources. Drug Discovery Today 19(7), 882–9 (2014)
22. Rebholz-Schuhmann, D., Kim, J.H., Yan, Y., Dixit, A., Friteyere, C., Backofen, R., Lewin, I.: Evaluation and cross-comparison of Lexical Entities of Biological Interest (LexEBI). PLoS One 8(10), e75185 (2013)
23. Ruttenberg, A., Rees, J.A., Samwald, M., Marshall, M.S.: Life sciences on the Semantic Web: the Neurocommons and beyond. Brief. Bioinformatics 10, 193–204 (Mar 2009)
24. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology 25(11), 1251–1255 (2007)
25. Splendiani, A., Gundel, M., Austyn, J.M., Cavalieri, D., Scognamiglio, C., Brandizi, M.: Knowledge sharing and collaboration in translational research, and the DC-THERA Directory. Brief. Bioinformatics 12(6), 562–575 (Nov 2011)
26. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucleic acids research 36(suppl 1), D901–D906 (2008)
27. e Zainab, S.S., Hasnain, A., Saleem, M., Mehmood, Q., Zehra, D., Decker, S.: Fedviz: A visual interface for sparql queries formulation and execution. In: Visualizations and User Interfaces for Ontologies and Linked Data (VOILA 2015), Bethlehem, Pennsylvania, USA. (2015)
28. Zeginis, D., et al.: A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web (2013)