OPINION ARTICLE

BioHackathon series in 2013 and 2014: improvements of semantic interoperability in life science data and services

[version 1; peer review: 2 approved with reservations]

Toshiaki Katayama¹, Shuichi Kawashima¹, Gos Micklem², Shin Kawano¹, Jin-Dong Kim¹, Simon Kocbek¹, Shinobu Okamoto¹, Yue Wang¹, Hongyan Wu³, Atsuko Yamaguchi¹, Yasunori Yamamoto¹, Erick Antezana⁴, Kiyoko F. Aoki-Kinoshita⁵, Kazuharu Arakawa⁶, Masaki Banno⁷, Joachim Baran⁸, Jerven T. Bolleman⁹, Raoul J. P. Bonnal¹⁰, Hidemasa Bono¹, Jesualdo T. Fernández-Breis¹¹, Robert Buels¹², Matthew P. Campbell¹³, Hirokazu Chiba¹⁴, Peter J. A. Cock¹⁵, Kevin B. Cohen¹⁶, Michel Dumontier¹⁷, Takatomo Fujisawa¹⁸, Toyofumi Fujiwara¹, Leyla Garcia¹⁹, Pascale Gaudet⁹, Emi Hattori²⁰, Robert Hoehndorf²¹, Kotone Itaya⁶, Maori Ito²², Daniel Jamieson²³, Simon Jupp¹⁹, Nick Juty¹⁹, Alex Kalderimis², Fumihiko Kato²⁴, Hideya Kawaji²⁵, Takeshi Kawashima¹⁸, Akira R. Kinjo²⁶, Yusuke Komiyama²⁷, Masaaki Kotera²⁸, Tatsuya Kushida²⁹, James Malone³⁰, Masaaki Matsubara³¹, Satoshi Mizuno³², Sayaka Mizutani²⁸, Hiroshi Mori³³, Yuki Moriya¹, Katsuhiko Murakami³⁴, Takeru Nakazato¹, Hiroyo Nishide¹⁴, Yosuke Nishimura²⁸, Soichi Ogishima³², Tazro Ohta¹, Shujiro Okuda³⁵, Hiromasa Ono¹, Yasset Perez-Riverol¹⁹, Daisuke Shinmachi⁵, Andrea Splendiani³⁶, Francesco Strozzi³⁷, Shinya Suzuki²⁸, Junichi Takehara²⁸, Mark Thompson³⁸, Toshiaki Tokimatsu³⁹, Ikuo Uchiyama¹⁴, Karin Verspoor⁴⁰, Mark D. Wilkinson⁴¹, Sarala Wimalaratne¹⁹, Issaku Yamada³¹, Nozomi Yamamoto²⁸, Masayuki Yarimizu⁷, Shoko Kawamoto¹⁸, Toshihisa Takagi^{29,42}

¹Database Center for Life Science, Kashiwa, Japan

²Department of Genetics, University of Cambridge, Cambridge, UK

³Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

⁴Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

⁵Faculty of Science and Engineering, Soka University, Hachioji, Japan

⁶Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan

⁷Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo, Japan

⁸CODAMONO, Toronto, Canada

⁹Swiss Institute of Bioinformatics, Geneva, Switzerland

¹⁰Istituto Nazionale Genetica Molecolare, Romeo ed Enrica Invernizzi, Milan, Italy

¹¹Universidad de Murcia, IMIB-Arrixaca-UMU, Murcia, Spain

¹²Department of Bioengineering, University of California Berkeley, Berkeley, USA

¹³Institute for Glycomics, Griffith University, Southport, Australia

- ¹⁴National Institute for Basic Biology, National Institutes of Natural Sciences, Okazaki, Japan
- ¹⁵The James Hutton Institute, Dundee, UK
- ¹⁶Biomedical Text Mining Group, Computational Bioscience Program, University of Colorado School of Medicine, Denver, USA
- ¹⁷Institute of Data Science, Maastricht University, Maastricht, The Netherlands
- ¹⁸National Institute of Genetics, Mishima, Japan
- ¹⁹European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK
- ²⁰Division of Rare Cancer Research, National Cancer Center Research Institute, Chuo, Japan
- ²¹Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
- ²²Pharmaceuticals and Medical Devices Agency, Chiyoda, Japan
- ²³Biorelate, Manchester, UK
- ²⁴National Institute of Informatics, Chiyoda, Japan
- ²⁵Preventive Medicine and Applied Genomics Unit, RIKEN Advanced Center for Computing and Communication, Yokohama, Japan
- ²⁶Institute for Protein Research, Osaka University, Suita, Japan
- ²⁷The Institute of Medical Science, The University of Tokyo, Minato, Japan
- ²⁸School of Life Science and Technology, Tokyo Institute of Technology, Meguro, Japan
- ²⁹National Bioscience Database Center, Japan Science and Technology Agency, Chiyoda, Japan
- ³⁰FactBio, Cambridge, UK
- ³¹The Noguchi Institute, Itabashi, Japan
- ³²Department of Bioclinical Informatics, Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan
- ³³Center for Information Biology, National Institute of Genetics, Mishima, Japan
- ³⁴Tokyo University of Technology, Hachioji, Japan
- ³⁵Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan
- ³⁶Intellileaf Ltd, Cambridge, UK
- ³⁷Enterome Bioscience SA, Paris, France
- ³⁸Leiden University Medical Center, Leiden, The Netherlands
- ³⁹DBJ Center, National Institute of Genetics, Mishima, Japan
- ⁴⁰School of Computing and Information Systems and the Health and Biomedical Informatics Centre, The University of Melbourne, Melbourne, Australia
- ⁴¹Escuela Técnica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas, Universidad Politécnica de Madrid, Madrid, Spain
- ⁴²Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo, Japan

v1 First published: 23 Sep 2019, 8:1677
<https://doi.org/10.12688/f1000research.18238.1>
 Latest published: 23 Sep 2019, 8:1677
<https://doi.org/10.12688/f1000research.18238.1>

Abstract

Publishing databases in the Resource Description Framework (RDF) model is becoming widely accepted to maximize the syntactic and semantic interoperability of open data in life sciences. Here we report advancements made in the 6th and 7th annual BioHackathons which were held in Tokyo and Miyagi respectively. This review consists of two major sections covering: 1) improvement and utilization of RDF data in various domains of the life sciences and 2) meta-data about these RDF data, the resources that store them, and the service quality of SPARQL Protocol and RDF Query Language (SPARQL) endpoints. The first section describes how we developed RDF data, ontologies and tools in genomics, proteomics, metabolomics, glycomics and by literature text mining. The second section describes how we defined descriptions of datasets, the provenance of data, and quality assessment of services and service discovery. By enhancing the harmonization of these two layers of machine-readable data and

Open Peer Review

Approval Status **??**

	1	2
version 1	?	?
23 Sep 2019	view	view
1. Todd J. Vision , University of North Carolina at Chapel Hill, Chapel Hill, USA 2. João Moreira , University of Twente, Enschede, The Netherlands		
Any reports and responses or comments on the article can be found at the end of the article.		

knowledge, we improve the way community wide resources are developed and published. Moreover, we outline best practices for the future, and prepare ourselves for an exciting and unanticipatable variety of real world applications in coming years.

Keywords

BioHackathon, Bioinformatics, Semantic Web, Web services, Ontology, Databases, Semantic interoperability, Data models, Data sharing, Data integration

This article is included in the **Hackathons** collection.

Corresponding author: Toshiaki Katayama (ktym@dbcls.jp)

Author roles: **Katayama T:** Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Kawashima S:** Conceptualization, Project Administration, Resources, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Micklem G:** Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Kawano S:** Project Administration, Software, Writing – Original Draft Preparation; **Kim JD:** Project Administration, Software, Writing – Original Draft Preparation; **Kocbek S:** Project Administration, Software, Writing – Original Draft Preparation; **Okamoto S:** Project Administration, Software, Writing – Original Draft Preparation; **Wang Y:** Project Administration, Software, Writing – Original Draft Preparation; **Wu H:** Project Administration, Software, Writing – Original Draft Preparation; **Yamaguchi A:** Project Administration, Software, Writing – Original Draft Preparation; **Yamamoto Y:** Project Administration, Software, Writing – Original Draft Preparation; **Antezana E:** Software, Writing – Original Draft Preparation; **Aoki-Kinoshita KF:** Software, Writing – Original Draft Preparation; **Arakawa K:** Software, Writing – Original Draft Preparation; **Banno M:** Software, Writing – Original Draft Preparation; **Baran J:** Software, Writing – Original Draft Preparation; **Bolleman JT:** Software, Writing – Original Draft Preparation; **Bonnal RJP:** Software, Writing – Original Draft Preparation; **Bono H:** Software, Writing – Original Draft Preparation; **Fernández-Breis JT:** Software, Writing – Original Draft Preparation; **Buels R:** Software, Writing – Original Draft Preparation; **Campbell MP:** Software, Writing – Original Draft Preparation; **Chiba H:** Software, Writing – Original Draft Preparation; **Cock PJA:** Software, Writing – Original Draft Preparation; **Cohen KB:** Software, Writing – Original Draft Preparation; **Dumontier M:** Software, Writing – Original Draft Preparation; **Fujisawa T:** Software, Writing – Original Draft Preparation; **Fujiwara T:** Software, Writing – Original Draft Preparation; **Garcia L:** Software, Writing – Original Draft Preparation; **Gaudet P:** Software, Writing – Original Draft Preparation; **Hattori E:** Software, Writing – Original Draft Preparation; **Hoehndorf R:** Software, Writing – Original Draft Preparation; **Itaya K:** Software, Writing – Original Draft Preparation; **Ito M:** Software, Writing – Original Draft Preparation; **Jamieson D:** Software, Writing – Original Draft Preparation; **Jupp S:** Software, Writing – Original Draft Preparation; **Juty N:** Software, Writing – Original Draft Preparation; **Kalderimis A:** Software, Writing – Original Draft Preparation; **Kato F:** Software, Writing – Original Draft Preparation; **Kawaji H:** Software, Writing – Original Draft Preparation; **Kawashima T:** Software, Writing – Original Draft Preparation; **Kinjo AR:** Software, Writing – Original Draft Preparation; **Komiyama Y:** Software, Writing – Original Draft Preparation; **Kotera M:** Software, Writing – Original Draft Preparation; **Kushida T:** Software, Writing – Original Draft Preparation; **Malone J:** Software, Writing – Original Draft Preparation; **Matsubara M:** Software, Writing – Original Draft Preparation; **Mizuno S:** Software, Writing – Original Draft Preparation; **Mizutani S:** Software, Writing – Original Draft Preparation; **Mori H:** Software, Writing – Original Draft Preparation; **Moriya Y:** Software, Writing – Original Draft Preparation; **Murakami K:** Software, Writing – Original Draft Preparation; **Nakazato T:** Software, Writing – Original Draft Preparation; **Nishide H:** Software, Writing – Original Draft Preparation; **Nishimura Y:** Software, Writing – Original Draft Preparation; **Ogishima S:** Software, Writing – Original Draft Preparation; **Ohta T:** Software, Writing – Original Draft Preparation; **Okuda S:** Software, Writing – Original Draft Preparation; **Ono H:** Software, Writing – Original Draft Preparation; **Perez-Riverol Y:** Software, Writing – Original Draft Preparation; **Shinmachi D:** Software, Writing – Original Draft Preparation; **Splendiani A:** Software, Writing – Original Draft Preparation; **Strozzi F:** Software, Writing – Original Draft Preparation; **Suzuki S:** Software, Writing – Original Draft Preparation; **Takehara J:** Software, Writing – Original Draft Preparation; **Thompson M:** Software, Writing – Original Draft Preparation; **Tokimatsu T:** Software, Writing – Original Draft Preparation; **Uchiyama I:** Software, Writing – Original Draft Preparation; **Verspoor K:** Software, Writing – Original Draft Preparation; **Wilkinson MD:** Software, Writing – Original Draft Preparation; **Wimalaratne S:** Software, Writing – Original Draft Preparation; **Yamada I:** Software, Writing – Original Draft Preparation; **Yamamoto N:** Software, Writing – Original Draft Preparation; **Yarimizu M:** Software, Writing – Original Draft Preparation; **Kawamoto S:** Conceptualization, Project Administration; **Takagi T:** Conceptualization, Funding Acquisition, Project Administration, Supervision

Competing interests: No competing interests were disclosed.

Grant information: Funding for the BioHackathon was supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2019 Katayama T *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Katayama T, Kawashima S, Micklem G *et al.* **BioHackathon series in 2013 and 2014: improvements of semantic interoperability in life science data and services [version 1; peer review: 2 approved with reservations]** F1000Research 2019, 8:1677 <https://doi.org/10.12688/f1000research.18238.1>

First published: 23 Sep 2019, 8:1677 <https://doi.org/10.12688/f1000research.18238.1>

Introduction

Big data in the life sciences - especially from 'omics' technologies - is challenging researchers with scalability concerns in terms of computational and storage needs, while at the same time, there is also a stronger drive towards the promotion of open data including the sharing of analyses and their outputs. Consistent with this, the “[Open Data Charter](#)” issued by the 2013 G8 summit meeting states that the release of high-value open data is important for improving democracies and encouraging innovative reuse of data. Experimental results including genome data, as well as research and educational activities, are recognized as of high value in the *Science and Research* category of the Charter. To fully utilize open data in life sciences, semantic interoperability and standardization of data are required to allow innovative development of applications.

During the 6th and 7th NBDC/DBCLS BioHackathons in 2013 and 2014, which were hosted by the National Bioscience Database Center (NBDC) and the Database Center for Life Science (DBCLS) in Japan, we focused on the improvement of Resource Description Framework (RDF) data for practical use in biomedical applications by developing guidelines, ontologies and tools especially for the genome, proteome, interactome and chemical domains. Also, to host these data effectively, we explored best practices for representing dataset metadata, as well as assessing the capabilities of triple stores and the quality of service of endpoints. The [BioHackathon 2013](#) was held in Tokyo and [BioHackathon 2014](#) was held in Miyagi. Both were sponsored by the NBDC and the DBCLS in the series of NBDC/DBCLS BioHackathons¹⁻⁴, which bring together database providers and bioinformatics software developers to make their resources integrable in effective ways.

Improvement and utilization of RDF data in life sciences

Publishing data based on the RDF model and its serialization formats (e.g. [Turtle](#)), along with relevant biomedical ontologies, is becoming widely accepted within the bioinformatics community⁵⁻⁹ as a way of serving semantically annotated data. In this section, we describe recent developments in RDF standardization for the genomics, proteomics, glycomics, chemoinformatics and text-mining domains.

Genomic information

Genome data is a key component in modern life sciences as it serves as a hub for data integration. In the previous BioHackathons, we have developed ontologies, such as the [Feature Annotation Location Description Ontology](#) (FALDO)¹⁰ and the [Genomic Feature and Variation Ontology](#) (GFVO)¹¹, and produced RDF data from heterogeneous datasets for integrated databases and applications. In this section, we describe how we modeled genomic annotations and related resources in RDF and ontologies.

Ontology for locations on biological sequences

During the [BioHackathon 2012](#)⁴, it was recognized that a common schema ontology was desirable for the Semantic Web integration of sequence annotation across multiple databases. In

depth group discussions including bioinformatics software developers and major database representatives identified common core needs in defining locations on biological sequences (both nucleic acids and proteins). This produced a draft specification for the Feature Annotation Location Description Ontology (FALDO), and proof of principle data conversion tools. This work continued at the BioHackathon 2013, with a specific focus on ensuring that all the existing annotations in the International Nucleotide Sequence Database Collaboration (INSDC)¹² feature tables could be converted into RDF triples using FALDO, as well as standardizing the coordinate system, and making sure that the starts of features are biologically sensible i.e. the start value is numerically higher than the end for genes located on the reverse strand. Subsequently, in May 2014, DBCLS organized a closed meeting, the [RDF Summit](#), where a small group of developers from DBCLS, DNA Data Bank of Japan (DDBJ), Swiss Institute of Bioinformatics (SIB), European Bioinformatics Institute (EBI) and Stanford gathered to standardize the RDF representation of genomic annotations. The group agreed to use the FALDO ontology (see the section below) for annotating the coordinates of genomic annotations and to represent gene/transcript/exons in RDF. As a result, the RDF model of DDBJ, Ensembl¹³ and TogoGenome⁹ are now aligned such that common SPARQL queries can retrieve sequence annotations from these distinct data sources interoperably.

Human genome and variation

After defining a common RDF model to represent the INSDC feature tables, one of the major remaining needs was to standardize the RDF representation of genome variations, which was discussed during the BioHackathon 2014.

A group from EBI, DBCLS and Tohoku University surveyed existing databases that represent clinical annotation of variants. National Center for Biotechnology Information (NCBI) ClinVar¹⁴ provides information on the relationships between human genetic variation and phenotypes along with supporting evidence; [Online Mendelian Inheritance in Man](#) (OMIM)¹⁵ provides relationships between genes and disease; [Leiden Open Variant Database](#) (LOVD)¹⁶ provides gene variants related to colon cancer; [Human Gene Mutation Database](#) (HGMD)¹⁷ is commercial but widely used; Thomson Reuters Gene Variant Database (GVDB) is also a commercial database. Tohoku Medical Megabank had a license to jointly develop the RDF version of the GVDB with Thomson Reuters and they completed the initial version to test queries like “find shared variations among diseases” and “find related variations from a specific disease”. In parallel, the EBI group started to convert Ensembl variation data into RDF in which an “allele” is related to “gene_variant”, “sequence_alteration” and “regulatory_region_variant” instances in the sequence ontology (SO), and its location is represented by means of a FALDO region [[Figure 1](#)].

The H-invitational database (H-InvDB)^{18,19} group developed RDF data and an ontology for their database covering ncRNA annotations. During the Biohackathon 2013, the RDF version of the H-InvDB was expanded and its ontology was published including recent advancement in understanding of non-coding

Figure 1. Proposed schema for the Ensembl variation.

RNA (ncRNA) function. To improve descriptions of the functional relationships between coding transcripts and ncRNA, links between transcripts in H-InvDB and two major RNA databases, Rfam²⁰ and miRBase²¹, were added. For miRBase, interactions between miRNA and transcripts were predicted using TargetScan²². For both of these databases, new classes were defined in the ontology to describe interaction events, such as binding between a transcript and a miRNA. At the BioHackathon 2014, the group tried to incorporate variant information into the RDF data.

Identifiers for sequences and annotations

There was discussion of how to represent gene names and chromosome Uniform Resource Identifiers (URIs). For gene names, it is recommended to use `rdfs:label` and `dc:identifier` for primary gene IDs and use `skos:altLabel` for gene synonyms. However, it is not mandatory because gene IDs are not always available, depending on the source of information. As for chromosome URIs, it would be useful if the bioinformatics community could agree to a common URI for each chromosome and version (e.g. human chromosome 19 in the GRCh38 assembly). However, we could not reach an agreement at the BioHackathon as it seemed to be impractical to cover every sequence assembly of all species, individuals, cells and samples in an unified manner as drafted in the [RDF summit](#). In this section, we describe the current situation and proposals relating to this issue.

Universal Biological Sequence ID (UBSID). An essential step in the merging of datasets is relating primary identifiers i.e. any

data can be joined if they contain the same identifiers. Therefore, all databases can be joined as fully connected Linked Data if appropriate universal identifiers are consistently used. To date, molecular biology has mainly developed around the Central Dogma concept in which higher levels of annotation (transcripts, proteins) are related to the underlying genomic sequence. Genes, as well as protein binding motifs and other features such as SNPs, can be related to DNA sequences, as can the transcriptome and proteome. Therefore, much of modern molecular biology data can in principle be related if the underlying nucleotide sequences are used as the basis for identifiers. However, the use of sequences *per se* as identifiers has several problems: for example, a sequence can be extremely long (e.g. human chromosome I), or very short (e.g. the location of a SNP), there can be multiple sequences that are highly similar or identical as in multi-copy paralogs, and a sequence feature can be on the sense or antisense strand. In order to overcome these problems, a universal sequence-based identifier scheme should incorporate position information, reference sequence information, the actual sequence (when there are differences, such as mutations, from the reference), strand information, and in addition, it would be ideal if all of such information is expressed as a short, human-comprehensible identifier. By using reference-based compression of DNA sequences based on offset and run-length encoding, the sequence can be expressed just by the mismatching positions and this can form the basis for an identifier system. Therefore, the G-language group proposed a Universal Biological Sequence ID (UBSID) to enable this encoding. For example, the human APOE mRNA

sequence is encoded as <http://rest.g-language.org/ubsid/ubsid-2seq/hg19-chr19:045409882+A42:=43-1092=193-580=718:>> as a URI in the [G-language REST service](#).

Identifiers used in the DDBJ and TogoGenome RDF. After the BioHackathon 2012, a group from DBCLS and DDBJ developed an [ontology which can capture semantically the data model of the INSDC](#), such as the records from GenBank, DDBJ and ENA, with restrictions on terms used in the feature table and qualifier key-values. A [converter](#) for INSDC and RefSeq²³ entries to RDF was developed based on the ontology, and the RDFized data is used in the TogoGenome application. TogoGenome integrates information on genes, proteins, organisms, phenotypes and environments. Because the genes in TogoGenome are currently extracted from INSDC and RefSeq records, the URI for each annotation is constructed as a fragment using Identifiers.org URIs in the form: [http://identifiers.org/\[insd c or refseq\]/\[entry_id\]#\[fragment\]>](http://identifiers.org/[insd c or refseq]/[entry_id]#[fragment]>). For example, the human APOE gene on chromosome 19 in the RefSeq record NC_000019.10 is internally represented in TogoGenome as http://identifiers.org/refseq/NC_000019.10#feature:44905782-44909393:1:gene.1424> and the information can be accessed at <http://togogenome.org/gene/9606:APOE> where 9606 is the taxon ID corresponding to human in the NCBI Taxonomy database and APOE is the gene name used in the record. This approach is slightly different from the proposed UBSID model which encodes sequence alignment with comments but can distinguish the source of information and feature types annotated in the INSDC/RefSeq record. The location of each gene and exons in TogoGenome RDF are described by the FALDO ontology.

Identifiers used in the Ensembl RDF. Ensembl generates their own IDs for genes, transcripts and exons in their database. For example, the human APOE gene is given an ID of ENSG00000130203, which encodes five transcripts (one of them is ENST00000252486) and one of the exons of this transcript is ENSE000003577086. It is natural to use these IDs when constructing URIs for the RDF dataset. In the 2014 development version of the Ensembl RDF, the human APOE gene is indicated as <http://rdf.ebi.ac.uk/resource/ensembl/ENSG00000130203> within a graph identified as <http://rdf.ebi.ac.uk/dataset/ensembl/77/9606> for the human genome dataset in the Ensembl release 77. The location of this gene on human chromosome 19 is designated by <http://rdf.ebi.ac.uk/resource/ensembl/77/chromosome:GRCh38:19:44905754-44909393:1>>. The strategy to generate unique URIs for each annotation in Ensembl is different from that employed by DDBJ/INSDC and TogoGenome, which all share both the same RDF model and use the FALDO ontology to describe the actual coordinates of annotations (e.g. genes and exons) on a chromosome. Thus at present further work is needed before all these providers are completely consistent and interchangeable.

Data integration beyond organisms

To facilitate more accurate and deeper integration of data, it is important to standardize metadata accompanying DNA sequences, orthologous gene relationships among organisms, phenotypic properties of organisms, inter-species and organism-environment interactions including host-pathogen relationships. We describe some of these efforts now.

Metadata on samples. DDBJ, EBI and NCBI are jointly hosting the BioSample database as an international collaboration. In this resource, metadata are accumulated on the samples from which DNA sequence in the INSDC database was collected and/or on which other research projects were conducted. The metadata includes species, type of samples (cell types etc.) and phenotypic or environmental information, and therefore it is valuable for data integration if the metadata is available as RDF. A group from DDBJ generated an RDF version of BioSample metadata during the BioHackathon 2014, using as a starting point 14,362 entries stored in the [DDBJ BioSample database](#) in XML format.

In addition, existing terminologies and ontologies for geological, archeological and morphological data were explored during the 2014 BioHackathon. For example, there are several resources for geolocations such as [W3C Geospatial Ontologies](#), [GeoRSS](#), [GeoNames](#) and [Global Biodiversity Information Facility](#) (GBIF). The National Aeronautics and Space Administration (NASA) has developed the [Global Change Master Directory](#) (GCMD) and the [Semantic Web for Earth and Environmental Terminology](#) (SWEET) which can be used to describe archaeological time scales. For morphology, the [Foundational Model of Anatomy](#) (FMA)²⁴, [Anatomy Reference Ontology](#) (AEO)²⁵, [Vertebrate Skeletal Anatomy Ontology](#) (VSAO)²⁶ and other domain specific ontologies²⁷ were surveyed. These ontologies are essential for encoding RDF data in environmental biology, such as biodiversity and biomolecular archeology. As a case study, a group developed a [semantic resource with information about corals](#) by integrating taxonomic, genomic, environmental, disease and coral bleaching information.

Ontologies for integration of microbial data. Within the field of microbiology, genomic and metagenomic data are expanding rapidly due to advances in next generation sequencing technologies. To effectively analyze these huge amounts of data, it is necessary to integrate various microbial data resources available on the Internet. Orthology can play an important role in summarizing such data by grouping corresponding genes across different organisms, and by annotating genes by transferring knowledge from highly curated model organism to newly sequenced genomes. Therefore RDF models were developed for representing the orthology data stored in the Microbial Genome Database for Comparative Analysis (MBGD)⁸, and these were used to construct an RDF version of MBGD. This also required the development of the OrthO ontology²⁸ for representing orthology and aligning concepts with the existing OGO ontology²⁹, with additional definitions mapped from OrthoXML³⁰. Orthology RDF data can now be linked with other databases published as RDF such as UniProt³¹, allowing the integrated dataset to be queried using SPARQL. When searching these data, ontologies can be utilized to specify complex search conditions. To assist making such precise queries, the [Microbial Phenotype Ontology](#) (MPO) was developed for describing microbial phenotypes such as microbial morphology, growth conditions, biochemical or physiological properties. During the hackathon, the ontology was updated to comply with a better classification of the hierarchical (is-a) and partonomical (part-of) structure. In addition, the [Pathogenic Disease Ontology](#) (PDO) was developed to describe pathogenic microbes that cause diseases in their

hosts. An RDF dataset that describes pathogenic information relating to each microbial genome sequence was created using the PDO. Since the genes within these genomes are connected to the ortholog information in the MBGD ortholog database, it is possible to calculate the set of orthologous gene groups that is enriched in the disease related microbes.

Knowledge extraction of factors related to diseases. Information and knowledge of the relationships between genes/ mutations/ lifestyle/ environment and diseases is required in order to predict the risk of a disease and for prognosis after the onset of a disease. In practice, it will also be necessary to collect individual lifestyle and environmental profiles as well as personal genetic data such as genome sequences to allow such predictions for individual people. The necessary underlying relationships are often described in the literature, but are not yet systematically collected in a database. To extract these relationships from the literature, there are two key steps that need to be addressed. First, entities must be annotated automatically using text mining software and, second, these annotations must be represented in a curation interface to allow confirmation that the information has been extracted accurately. Genes, genetic variants, diseases, environmental factors and lifestyle factors are the entity types that need to be annotated on the corpus. Existing software for extracting genes (e.g. GNAT³², GenNorm³³ etc.), mutations (e.g. tmVar³⁴, MutationFinder³⁵ etc.) and diseases (e.g. BANNER³⁶ with disease model) are openly available, along with existing datasets such as BioContext³⁷ and EVEX DB³⁸. Before environmental factors and lifestyle factors can be extracted systematically it is necessary to decide on a controlled vocabulary (whether existing or not) to represent them. Pregnancy Induced Hypertension (PIH) was chosen as a case study and 86 relevant open access PubMed Central articles identified. It was possible to extract genes from 32 of these articles using the BioContext dataset, while the other 54 articles were published more recently than BioContext. Attempts were made to extract mutations from 86 articles. For lifestyle and environmental factors, controlled vocabularies were collected in preparation for entity recognition. After obtaining all the entities in the 86 articles, they were curated using interfaces such as PubAnnotation³⁹, and the curated relationships represented as an RDF graph.

Tools for semantic genome data

Genome annotations have historically been represented and distributed in non-standard domain-specific data formats (e.g., INSDC, GFF3, GTF). The data formats themselves often include implicit semantics, making automatic interpretation and integration of the data with other resources challenging. Therefore, tools to convert those data into RDF and ontologies to support semantic representation of data need to be developed. BioInterchange is a tool to convert those file formats into RDF and was originally developed in the BioHackathon 2012⁴, with its functionalities and ontologies being enhanced over successive hackathons. Other tools for high-throughput data processing of Sequence Alignment/Map format (SAM), Binary SAM format (BAM)⁴⁰, Variant Call Format (VCF)⁴¹, Genome Variation Format (GVF)⁴², Header-Dictionary-Triples (HDT)⁴³ files have

also been developed and a middleware to enable SPARQL queries directly against these huge files on-the-fly for scalability was explored and results were incorporated into integrated semantic genome databases such as TogoGenome and MicrobeDB.jp.

Utilization of domain specific data formats in semantic web.

In BioHackathon 2013, VCF2RDF was developed and subsequently published as a Ruby program to convert VCF files into RDF, which represent positions in FALDO and alleles in its own *ad hoc* ontology terms. The resulting RDF data was loaded into Fuseki and queries were tested in the Jena framework, taking three minutes for the cow genome on a laptop to plot quality scores of variant calls for a million base pairs. During BioHackathon 2014, a group developed middleware to interpret SPARQL queries against SAM/BAM/VCF files on the fly. The first implementation was prototyped in JRuby so that the Java library for samtools can be used in a Ruby program. The resulting application, VCFotf, is packaged as a Docker image that serves a query interface on the Web page. Also, another implementation (sparql-vcf) was developed with Jena for improving query execution time, in which Jena property functions are used to introduce a “special predicate” which accelerates search performance; however, this ‘boutique’ query violates the SPARQL standard.

Use of compressed RDF for large scale genomic data.

BioInterchange was used in a Genomic HDT project as a feasibility study to convert a variety of genomic data files (e.g. GVF) containing coordinate-annotated genomic features into an ontology-annotated RDF representation. The RDF data file is then processed into an RDF/HDT file, which is a compressed, indexed, and queryable data archive. Using Ensembl’s human somatic variation data (81MB, 9MB gzipped), it was found that the RDF/HDT archive is only 20MB (1.5M triples; 15MB data + 5MB index), which is a significant reduction from the 313MB RDF N-triples representation. A JSON RESTful API was made available using Sinatra to provide access to the RDF/HDT file, and this allowed a demonstration of genome-based browsing of the RDF/HDT data file using the JBrowse genome browser.

Integrated semantic genome databases. TogoGenome⁹ was developed to integrate heterogeneous biomedical data using Semantic Web technologies. This utilizes the representation of genomic data in the standard RDF format, enabling interoperability with any other Linked Open Data (LOD) around the world. To support these efforts we have collaborated with DDBJ, UniProt, and the EBI RDF group to develop ontologies for representing locations and annotations of genome sequences and used these developments for all prokaryotic genomes and, later, eukaryotic genomes. To complement the above work we developed ontologies for taxonomies, phenotypes, environments and diseases related to organisms, so enabling faceted browsing of the entire datasets. Every TogoGenome report page is made up of modular components called TogoStanza, which is a generic framework to generate Web components querying SPARQL endpoints and rendering them as HTML elements.

Stanzas are re-usable modules which can be shared and embedded easily into other databases, and which have been developed in collaborations with MicrobeDB.jp, MBGD⁸ and CyanoBase⁴⁴, resulting in over 100 TogoStanzas being available so far.

Visualization of semantic annotations in JBrowse. JBrowse⁴⁵ was used by several projects within the BioHackathon as a demonstration platform. JBrowse running on top of the SPARQL endpoints, e.g. TogoGenome or a prototype InterMine⁴⁶

endpoint, or from indexed files produced by GenomicHDT, were comparable in performance with typical RDB back-ended settings. In addition, an unusual use of JBrowse was to view text instead of DNA sequence, with the annotation viewed being the output of natural language processing.

TogoGenome: JBrowse was extended to support the TogoGenome's SPARQL endpoint as a data source to retrieve and visualize genes on a chromosomal track (Figure 2). This enhancement

Figure 2. SPARQL back-ended JBrowse is integrated into the TogoGenome database.

is already merged into the official JBrowse release since version 1.10 in 2013. SPARQL queries are customizable in the JBrowse configuration file as long as they return *start*, *end*, *strand*, *type* (label), *uniqueID* and *parentUniqueID* of the annotation objects in a given range within a sequence. When scrolling to neighboring regions, the performance is good enough for browsing.

InterMine: Representatives from the [InterMine](#) project⁴⁶ produced proof-of-concept demonstrations of semantic extensions to InterMine data-warehouses. These included on the one hand a draft of how to [model InterMine data](#) as linked data, producing both an ontology of relationships and triples that conform to that ontology, and on the other hand a draft of a very limited SPARQL engine capable of operating on an InterMine data source directly. Together these investigations indicate that given some development effort, it is likely that significant progress can be made to integrating InterMine into the semantic web. An area that needs work, and is receiving attention, is the production of stable URIs for InterMine entities. In addition to this, work was done to implement a simple adaptor allowing, as described above, JBrowse to request data directly from InterMine RESTful web services.

Text-mining: In the community of BioHackathon, text mining resources were developed around [PubAnnotation](#), a public repository of literature annotation data sets. Usually text mining requires its own set of tools, e.g. viewers or editors. However, an interesting experiment was carried out during BioHackathon 2013 and 2014 to use JBrowse as a viewer of text annotation data. The idea behind the experiment was that both genomic data and text data are represented as character sequences, and that annotations of both type of data are attached to specific regions on the sequences. A simple script was developed to convert annotations in PubAnnotation to JBrowse format, and it was observed that text annotations can be nicely viewed in JBrowse. The result raises the possibility of further interoperability between tools for genomics and text mining.

Proteomics, metabolomics and glycomics information

In addition to genomic information, advancements in developing ontologies and RDF datasets for proteins, metabolites, and glycans were made during the hackathons. It took several years to design standard data models as a community agreement and to convert existing resources into RDF by adding semantics, and the BioHackathons have successfully facilitated the efforts of domain experts.

Protein structures, interactions and expressions

The European Bioinformatics Institute's (EBI) SIFTS "Structure Integration with Function, Taxonomy and Sequences" resource provides regularly updated residue-level mappings between UniProt and PDB entries⁴⁷. SIFTS has been distributed in Comma Separated Values (CSV) and Extensible Markup Language (XML) formats. Like many other proteome-related databases, SIFTS uses the classical protein chain ID specified by the author. However, in 2016, the Worldwide Protein Data Bank

(wwPDB) will abolish the conventional PDB format and instead will distribute RDF/XML based on the PDB exchange dictionary / macromolecular Crystallographic Information Format (PDBx/mmCIF) [PDBx/mmCIF]. At the same time wwPDB will start assigning protein chain identifiers, which will also be encoded as URIs in the wwPDB/RDF.

During the BioHackathon, an RDF version of SIFTS (RDF-SIFTS) was designed and implemented to provide residue-to-residue correspondence between PDB and UniProt entries in RDF⁴⁸. RDF-SIFTS links both the protein chain ID assigned by authors and the one assigned by wwPDB to SIFTS. RDF-SIFTS uses existing ontologies of PDB, UniProt, [EMBRACE Data and Methods](#) (EDAM)⁴⁹ as well as FALDO, and resources are linked to Identifiers.org⁵⁰ URIs.

The University of Tokyo Proteins (UTProt)⁵¹ is a project that is collecting and building RDF to support interactome linked data. During the BioHackathon, the UTProt group extended RDF-SIFTS to cover intermolecular interactions, and this resulted in six billion triples including, for each pair of residues in the interacting surfaces, their separation distance. This resource will be useful for analysis of structure and sequence in proteomics and interactomics. Serialization Ruby code, [RDF-SIFTS maker](#), is available through GitHub as open source software which can be used to convert new release of SIFTS data from EBI.

"Omics" technologies are primarily aimed at the universal detection of genes (genomics), mRNAs (transcriptomics), proteins (proteomics) and metabolites (metabolomics) in a specific biological sample. Proteomics and metabolomics in particular have gained a lot of attention in recent years due the possibility of studying reactions, post-translational modifications, and pathways⁵². The proteomics community has been working for more than ten years in the standardization of file formats and proteomics data⁵³. Different XML-based file formats and open-source libraries have been released to handle proteomics data from spectra to quantitation results⁵⁴⁻⁵⁶.

In contrast, metabolomics is a relatively new "omics" field where the standardization of exchange formats is difficult, due to the variety of measurement methodologies ranging from nuclear magnetic resonance (NMR) spectroscopy to a variety of mass spectrometers (MS). Moreover, currently no single system can provide enough resolution to measure the entire set of small molecules within a biological sample; instead, data from multiple systems are combined to gain more comprehensive coverage, for instance combining Liquid Chromatography (LC), Gas Chromatography (GC), and Capillary Electrophoresis (CE) separation prior to analysis in a mass spectrometer. Recently, the mzTab data exchange format was introduced by the Human Proteome Organization (HUPO) Proteomics Standards Initiative, as a standardized format to report both qualitative and quantitative metabolomics and proteomics experiments in a simple tabular format⁵⁷. In BioHackathon 2014, a [Perl library](#) was developed to standardize the metabolomics data obtained from MasterHands software⁵⁸. MasterHands is a proprietary software for the analysis of CE-MS-based metabolomics used in the Institute

for Advanced Biosciences, Keio University, and at Human Metabolome Technologies Inc. The library allows the annotation of KEGG compound information using the KEGG REST API, and also allows the annotation of Reactome and MetaCyc information.

In the age of systems biology and data integration, proteomics data represent a crucial component to understand the “whole picture” of life. In this context, well-established databases for proteomics data include the Global Proteome Machine Database (GPMDB), PeptideAtlas, ProteomicsDB, and the Proteomics Identification (PRIDE) database among others⁵⁹. In addition, at BioHackathon 2014, the “omics” group worked on the standardization to RDF of different web services and APIs for proteomics and protein expression data. The GPMDB2RDF and PRIDE2RDF library allow the export of expression data from the GPMDB Database⁶⁰ and PRIDE Database⁶¹ respectively. The development of a standard interface for providing protein expression data will allow, in the future, exchange and proper reuse of public proteomics data. To this end, the “omics” group in the BioHackathon 2014 made the first steps towards the development of the [ProteomeXchange Interface](#) (PROXI) for protein expression data exchange⁵⁹.

Glycoinformatics

The glycoscience group participated in a satellite BioHackathon in Dalian, China, in parallel to the GLYCO 22 Meeting held June 23–28, 2013. Although a preliminary RDF format was developed at the previous BioHackathon in 2012⁶², there was a need to address not only glycan structures (sequences) but also supporting experimental data, the biological source of the sample analyzed, and publication information. Therefore, during BioHackathon 2013, a formal ontology to represent these features, as well as the glycan structures to which they relate, was discussed. The aim of the GlycoRDF group was to define a standard RDF representation, in the form of an ontology by integrating features from existing ontologies where possible and creating new classes and relationships where needed.

As it would be impossible, in a week, to create an ontology that could cover the full spectrum of glycomics information and experimental data, it was decided that the group would limit the first version to the data that currently exists in glycomics databases. On the other hand, the developers also attempted to define the ontology so that it could be easily extended with additional predicates and classes if needed, in case more data or more glyco-related databases utilize the proposed RDF format. As a result, by the end of BioHackathon 2013, the first version of the GlycoRDF ontology was agreed upon and is currently available at the GlycoRDF repository at [63](#). In 2014, work progressed to the point where all glyco-scientists who attended previous BioHackathons had now generated GlycoRDF-formatted versions of their databases. The updated list of these databases are listed and documented on the [GlycoRDF repository](#).

Enzymatic reaction ontology

Entities can be classified based on a variety of features, such as their function(s), structures/sub-structures, or chemical

properties. For example, genes and proteins are independently classified based on their functions, role, and cellular location, organized by the Gene Ontology (GO)⁶⁴. At the same time, gene and proteins are also classified based on their conserved partial substructures, such as protein domains in Pfam. ChEBI⁶⁵ classifies chemical substances by their overall functions (ChEBI role ontology) and by their partial structures (ChEBI molecular structure ontology). For enzymes, their overall functions are classified by the Enzyme List of International Union of Biochemistry and Molecular Biology, often referred to as the Enzyme Commission (EC) numbers⁶⁶. To date, however, there have been no standard ways to classify enzymes based on the partial structures of their enzymatic reactions. Therefore, during BioHackathon 2013 we discussed the development of an ontology that deals with the partial structures of enzymatic reactions, i.e. substrate-product pairs derived from reaction equations. This led to the [Enzyme Reaction Ontology for annotating Partial Information of biochemical transformation](#) (PIERO) being published in 2014⁶⁷. In BioHackathon 2014, we had further discussions to refine the PIERO data to establish the PIERO Ver0.3 Schema. This ontology was later used in *de novo* metabolic pathway reconstruction analysis⁶⁸ and for ortholog predictions⁶⁹.

Text mining and question-answering

In contrast to molecular resources, extraction and utilization of knowledge represented in the literature is still in progress. As an infrastructure, it is proposed to have a common open platform for sharing text annotations resulting from manual curation and various natural language processing (NLP) techniques. NLP methods are also applied to derive a SPARQL query from natural language.

Modeling text annotations on the Semantic Web

Text mining is becoming an increasingly common component of biological curation pipelines and biological data analysis, and as such there is increasing demand for both text that has been automatically annotated with natural language processing tools, and annotated document resources that can be used in development and evaluation of those tools. This demand in turn leads to a need for standard, interoperable representations for annotations over documents. Several proposals for general linguistic annotation representations have been made⁷⁰, including ones specifically for biomedical text annotation representations^{71,72}, as well as data models underpinning standard modular architectures such as Unstructured Information Management Architecture (UIMA)⁷³. However, these approaches have not been adapted to the Semantic Web. Recently, the Open Annotation Core Data Model has been proposed to enable interoperable annotations on the web⁷⁴. This project explored the application of the Open Annotation Model to the use case of capturing text mining output, by harmonizing the data models of the existing proposals.

The existing RDF-based representation of the PubAnnotation tool³⁹ was used as a starting point, and adapted for compatibility with the Open Annotation model. The Open Annotation model provides an annotation class that relates a web resource to information that is about that resource; this representational choice

is different from other models yet critically allows separation of metadata (e.g. provenance information) about the annotation itself, from meta-data about the content of the annotation⁷⁵. Several core requirements for text-based annotations were identified: (1) representation of document spans as annotation targets; (2) representation of “simple” associations, e.g. between a span of text and a concept such as an ontology identifier; (3) representation of “complex” associations, e.g. between several spans of text and a relation or event. In addition, the overall structure of a document corpus, which can consist of several documents, must be modeled in such a way as to allow those documents to have internal structure such as chapters, sections, passages or sentences. PubAnnotation models text spans relative to these internal structural elements, while **BioC** and **UIMA** have adopted absolute character offsets across a complete document. The model developed here allows for both, by allowing the target of annotation to be either a full document, or a document element as appropriate. It is hoped that the proposals made for web-based document annotation representations will enable interoperability with other Open Annotation-based data and tools, while also addressing the need to move linguistic annotation into the web.

During BioHackathon 2014, the integration of literature annotation resources was pursued with actual data sets. Colorado Richly Annotated Full-Text (CRAFT)⁷⁶ is a recent important achievement of biomedical text mining, which included 67 full papers with rich annotation based on 7 biomedical ontologies.

The GRO corpus⁷⁷ is a richly annotated corpus based on the **Gene Regulation Ontology**⁷⁸. Allie is an acronym-annotated collection of all PubMed titles and abstracts⁷⁹. They were all converted into PubAnnotation-compatible format, and submitted to PubAnnotation. The whole-PubMed-scale dataset, Allie, triggered the issue of scalability. However, integration of the two corpora, CRAFT and GRO, into PubAnnotation, demonstrated significantly improved utility.

Natural language query

SPARQL is a standard language for querying triple stores. However, SPARQL queries can be difficult to write, even for experts. Usability studies have shown natural language interfaces to SPARQL to be the preferred method of SPARQL query formation assistance⁸⁰. For this reason, software developers are encouraged to create applications that allow users to ask biomedical questions against triple stores using natural (i.e. human) language.

Building on the work in BioHackathon 2012 on querying **Systematized Nomenclature of Medicine-Clinical Terms** (SNOMED-CT), during BioHackathon 2013 effort was focused on the Online Mendelian Inheritance in Man (OMIM) SPARQL endpoint, with a simultaneous focus on building an evaluation data set. Social networking was used to obtain use cases from biologists and informaticians, and it was quickly discovered that the system had an issue with differentiating between broad semantic types and specific instances. For example, “heart disease” was correctly mapped to a specific entity, but

the word “genes” was incorrectly mapped to one specific gene. For this reason, dealing with the issue of recognizing broad semantic classes was the major focus of the development work, and testing semantic class recognition was the main focus of the testing effort. OMIM uses Type Unique Identifiers (TUIs), in the Unified Medical Language System (UMLS)⁸¹, to semantically type subjects and objects in its triple store, so we approached the problem of recognizing broad semantic classes as recognizing mentions of TUIs. Accordingly, a TUI concept recognizer was implemented into the open source **LODQA** system for automatic generation of SPARQL queries from natural language queries.

Efforts to develop a natural language interface were continued in BioHackathon 2014, during which the LODQA system was configured for two large scale RDF datasets, Bio2RDF and BioGateway. In this way, it was demonstrated that technology like LODQA can answer a question like, “Which genes are involved in calcium binding?”, based on RDF data sets like Bio2RDF. However, it also revealed remaining performance issues.

Metadata about RDF data resources

Because there is so far no solid guideline on publication of RDF data available, it is not clear for a researcher who wants to develop and release RDF data, how to create the associated metadata, how to describe the provenance of the data and how to assess the quality of the data/service. Also, understanding a dataset is not easy for users of data because there are so many classes, relations and possibilities. To resolve these issues, minimum requirements to represent statistics and characteristics of RDF data and services, including SPARQL endpoints, were discussed.

Dataset metadata

The **International Society for Biocuration** (ISB), in collaboration with the **BioSharing** forum, developed the BioDBCore⁸² which is a community-defined, uniform, generic description of the core attributes of biological databases. However, when it comes to the RDF datasets, one of the difficulties reported by users is that they find it difficult to figure out what data are in a dataset and how things are connected. Vocabulary of Interlinked Datasets (VoID) is a small vocabulary to describe key schemata style information about a dataset. It also includes key metadata such as when a dataset has been updated and under which license it falls. In this section, we propose a guideline for database providers, to provide useful extended VoID files for their users.

Access to consistent, high-quality metadata is critical to finding, understanding, and reusing scientific data - this is the core of the FAIR Data Principles⁸³. However, while there are many relevant vocabularies for the annotation of a dataset, none sufficiently capture all the necessary metadata. Towards providing guidance for producing a high-quality description of biomedical datasets, we identified RDF vocabularies that could be used to specify common metadata elements and their value sets. The resulting guidelines, finalized under the auspices of the W3C Semantic Web for Health Care and the Life

Sciences Interest Group (HCLSIG), cover elements of description, identification, versioning, attribution, provenance, and content summarization. This guideline reuses existing vocabularies, and is expected to meet key functional requirements including discovery, exchange, query, and retrieval.

Big data presents an exciting opportunity to pursue large-scale analyses over collections of data in order to uncover valuable insights across a myriad of fields and disciplines. Yet, as more and more data are made available, researchers are finding it increasingly difficult to discover and reuse these data. One problem is that data are insufficiently described to understand what they are or how they were produced. A second issue is that no single vocabulary provides all key metadata fields required to support basic scientific use cases. For instance, the [Data Catalog Vocabulary](#) (DCAT) is used to describe datasets in catalogs, but does not deal with the issue of dataset evolution and versioning. A third issue is that data catalogs and data repositories all use different metadata standards, if they use any standard at all, and this prevents easy search, aggregation, and exchange of data descriptions. Thus, there is a need to combine these vocabularies in a comprehensive manner that meets the needs of data registries, data producers, and data consumers.

We developed a specification for the description of a dataset that meets key functional requirements (dataset description, linking, exchange, change, content summary), reuses 18 existing vocabularies, and is expressed using RDF. The specification covers 61 metadata elements pertaining to data description, identification, licensing, attribution, conformance, versioning, provenance, and content summary. Each metadata element includes a description and an example of use. The specification presents a three component model for modular description depending

on whether specific files and versions are known ([Figure 3](#)). The *summary level description* focuses on release-independent information that mirrors the one captured by dataset registries; the *distribution level description* focuses on specific data files, their formats and downloadable location; and the *version level description* links summary descriptions with distribution descriptions. Each description level is bound to a different set of meta-data requirements – mandatory, recommended, optional. A full worked example using the [ChEMBL](#) dataset is provided. The group is currently evaluating the specification with implementations for dataset registries such as [Identifiers.org](#)⁵⁰ and [IntegBio Database Catalog](#), as well as Linked Data repositories such as [Bio2RDF](#)⁸⁴. The specification is available from the [W3C site](#)⁸⁵.

VoID for InterMine and UniProt

As VoID is a vocabulary for describing datasets that can be used to generate documentation and assist users in finding key knowledge on how to write analytical data queries, the [InterMine](#) group worked on automatically generating VoID files for InterMine-based Model Organism Databases, while the UniProt group worked on the same for showing classes and predicates used in the named graphs on the [UniProt SPARQL endpoint](#).

InterMine⁸⁶ is an open source graph-based data warehouse system built on top of PostgreSQL. Through a collaboration (the InterMOD consortium⁸⁷), with most of the main animal Model Organism Databases (MODs) there are now InterMine databases available for budding yeast (SGD)⁸⁸, rat (RGD⁸⁹), zebrafish (ZFIN⁹⁰), mouse (MGI⁹¹), nematode (WormBase, unpublished), Fly (InterMine group)⁹², and [Arabidopsis](#)⁹³ with further MOD InterMine instances expected. Extensive data from the modENCODE project⁹⁴ are also available through modMine⁹⁵. As a step towards exposing these rich data as RDF, code was developed

Figure 3. Three component model for dataset description.

that uses existing InterMine RESTful web services to interrogate the FlyMine database and to generate a VoID description of the database. Further work is required to adjust the core InterMine data model to include additional database metadata items. This will then allow the automatic generation of VoID descriptions for any InterMine database. Further work is also required to ensure that appropriate standards are adhered to, especially for RDF predicates. In addition to the above developments, progress was made in creating a [Sesame](#)-based SPARQL endpoint for InterMine databases to complement the existing web application and web services. At the moment the endpoint only supports a small range of simple queries. It is hoped that in the future such endpoints will make available the rich data assembled and curated by the world wide Model Organism Database community. In the process this should provide opportunities for interoperation and also a mechanism for federation across the different resources.

UniProt³¹ is available as RDF and can be queried via [SPARQL](#) and [REST services](#). UniProt is a large and complicated database, that is difficult to explore due to its size. During the hackathon we implemented a procedure to generate a VoID file to describe UniProt data. [The VoID file](#), now available on FTP and via [the uniprot.org SPARQL Service Description](#) (application/rdf+xml), is updated every release in synchrony with our production, and show users what types of data (and how much) are available in the UniProt datasets. We also document how many links to other databases UniProt provides, demonstrating the hub effect of UniProt.org in the life science domain. For the UniProt SPARQL endpoint this VoID description is used as a key part of the user documentation describing the schema of the UniProt data.

Schema.org and RDFa for biological databases

[Schema.org](#) is a collection of extensible schemas that webmasters can use to mark up structured data on their web pages with the aim of improving search engine performance and enabling the creation of other applications. The initiative was founded by Google, Bing and Yahoo! as a collaboration to improve the web and their search results by using such structured data. More than 700 item types have been listed in [schema.org](#), some of which have been supported by these search engines. If webmasters mark up their content in [an acceptable markup format](#) (e.g. Microdata, microformats or RDFa), then web crawler programs can detect these structured data and they can be rendered as rich snippets in the search results.

During the BioHackathon, the members of this working group proposed two item types for a schema extension: [“BiologicalDatabase”](#) and [“BiologicalDatabaseEntry”](#). We discussed what item properties would be suitable for our purposes and how to label them in markup. Finally, we decided to use the [Microdata format](#) to mark up web pages and proposed five original properties: “entryID”, “isEntryOf”, “taxon”, “seeAlso” and “reference”. Work in this area is now being carried forward by the bioschemas.org project.

We also publicized our proposal and encouraged BioHackathon members to mark up their databases. A Microdata crawler was

created to extract these structured data. We modified “Sagace”⁹⁶, a web-based search engine for biomedical data and resources in Japan, developed at the NIBIOHN in collaboration with NBDC. We confirmed that marked-up data showed up as rich snippets in search results. [Ten databases have been marked up](#) with our new proposal and so can help improve the readability of search results. This service is freely available at <http://sagace.nibiohn.go.jp>.

Provenance of data

Several models for associating provenance for an assertion have been proposed, but there has been inadequate evaluation to determine how accurately they are able to represent the myriad of provenance details required to support citation and reuse. The approach taken at BioHackathon 2014 was to survey and document assertional provenance methods, develop tools to populate these models, develop evaluation metrics to compare them, and assess this comparison. We describe a selection of these activities below.

Nanopublication

A [nanopublication](#) is defined as the smallest unit of publishable information that represents a finely-grained, but complete idea. Nanopublications are composed of such fine-grained assertions coupled with provenance metadata about the assertion, such as the methods used to create it and personal and institutional attributions, and finally additional metadata about the nanopublication itself, such as who or what created it, and when. The aim is to make a formal, predictable, and transparent relationship between data and its provenance. Nanopublications will be discussed here with respect to their application to FANTOM5^{97,98} data, to track DBCLS literature curation, and within the Semantic Automated Discovery and Integration (SADI) framework⁹⁹.

The FANTOM5 project monitored transcription initiation at single base-pair resolution in mammalian genomes by Cap Analysis Gene Expression (CAGE) coupled with single molecule sequencing^{97,98}. Promoters were defined as upstream of CAGE peaks (transcription start site clusters) and their activities were quantified based on their read counts. The FANTOM5 promoters and their activities were described in nanopublications¹⁰⁰ to facilitate their open and interoperable exchange. Three classes of nanopublications, having the following assertions¹⁰¹, were generated: 1) A CAGE peak is defined in a specific region of the genome, 2) The CAGE peak is a transcription start site (TSS) region, which is part of a gene, 3) The CAGE peak is active at a certain level in a specified sample. Class 1 nanopublications (CAGE peaks) provide minimum information based on a model on genomic coordinates. They can be exported to genome browsers. Class 2 nanopublications (gene associations) are served as supplemental data to allow biological searches. This class of nanopublications may be re-released when a new data processing workflow is available or when different parameters or gene definitions are used. Class 3 nanopublications (activity levels of transcription in individual samples) are used only if the details of expression are relevant in a given biological search. By dissecting the whole data set into three classes of nanopublications with different granularities, its reusability is increased. These nanopublications are available at <http://rdf.biosemantics.org>, and they have been reported also in an article related to FANTOM5¹⁰¹.

The DBCLS has developed a web-based gene annotation tool, **TogoAnnotation**, has provides an easy way of accessing and adding annotations. Likewise *Gene Indexing* was developed as a simple named-entity recognition (NER) task in order to make connections between genomic loci and the literature. Gene Indexing generates micro-annotations by manually extracting gene and protein symbols from the text, tables and figures of full papers and connecting them to both PubMed IDs and genome location. A total of 10 curators cooperated over a five year period to manually annotate over 5,000 full papers relating to microbes.

In this way over 200,000 gene/protein micro-annotations were generated.

Based on the above data, during the BioHackathon 2014, a Nanopublication model was developed for these literature curation data, as well as a converter to make any annotation in the TogoAnnotation system representable as a Nanopublication RDF (Figure 4). It is intended that the curation data be integrated into the TogoGenome system and be expanded as a standard distributed annotation platform in the future.

```
@prefix : <http://togo.annotation.jp/nanopublication/nanopub_id> .
@prefix np: <http://www.nanopub.org/nschema#> .
@prefix cito: <http://purl.org/spar/cito/> .
@prefix faldo: <http://biohackathon.org/resource/faldo#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix so: <http://purl.org/obo/owl/SO#> .
@prefix tga: <http://togo.annotation.jp/vocabularies#> .
@prefix up: <http://purl.uniprot.org/core/> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix pav: <http://swan.mindinformatics.org/ontologies/1.2/pav/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

: {
  : a np:Nanopublication ;
    np:hasAssertion :assertion ;
    np:hasProvenance :provenance ;
    np:hasPublicationInfo :publicationInfo .
}

:assertion {
  :biological_entity a so:SO_0000104 ; # polypeptide
    rdfs:label "gene_name" ;
    faldo:location :location ;
    cito:citesAsAuthority || up:citation :citation ;
    tga:isTaggedBy <http://togo.annotation.jp/tags/pname:gene_name> .

  :location a faldo:Region ;
    faldo:begin <position_uri> ;
    faldo:end <position_uri> ;
    faldo:reference <sequence_uri> .

  :citation a up:Journal_Publication ;
    tga:isTaggedBy <http://togo.annotation.jp/tags/pmid:pubmed_id> ;
    tga:mentionedIn <http://togo.annotation.jp/tags/tag_id> . # e.g., Fig 1
}

:provenance {
  :assertion prov:wasDerivedFrom tga:gene_indexing_version2 .
  :citation prov:wasDerivedFrom <http://www.ncbi.nlm.nih.gov/pubmed/pubmed_id> .
  :location prov:wasDerivedFrom <http://genome.microbedb.jp/cyanobase/genes/gene_id,
    <http://togogenome.org/annotation/sequence_id/begin..end> .
}

:publicationInfo {
  : dc:created "date_time"^^xsd:dateTime ;
    dc:rights <http://creativecommons.org/licenses/by/3.0/> ;
    dc:rightsHolder <http://www.nig.ac.jp/section/yn/yn-e.html> ;
    pav:createdBy "curator_name1", "curator_name2" ;
    pav:importedFrom <http://togo.annotation.jp/bookmarks/bookmark_id> .
}
```

Figure 4. Proposed nanopublication data model for TogoAnnotation data.

The SADI Semantic Web services project also has a need to represent rich provenance data regarding how its services create their output. Given the rapid growth and notable success of the OpenPHACTS¹⁰² and NanoPublications¹⁰³ projects, it seems desirable that analytical Services - those following the SADI Semantic Web Service design patterns in particular - should output semantic data that follows the same NanoPublication paradigm. This would allow SADI services to publish new biomedical knowledge directly into the vast integrated NanoPublications space, and take advantage of their integration tools.

Extensions to the existing Perl SADI::Simple codebase in Comprehensive Perl Archive Network (CPAN) were undertaken at the hackathon. A key consideration was to ensure that the code could support distinct metadata for each triple, since SADI services are specifically designed to support multiplexed inputs potentially spread over a large number of processors for analysis, before being reassembled into an output message. As such, it is potentially the case that each triple has slightly distinct provenance information. The implemented solution guarantees globally unique identification of each of these nanopublications, for each execution, even over multiple iterations of the same input data.

NanoPublications are created when, through HTTP content negotiation, the client requests n-quads. The service responds with an RDF structure that follows the structure of the (proposed) [NanoPublication Collection](#).

Requesting quads from a 'legacy' SADI Service that does not support NanoPublications will result in a HTTP 406 (Not Acceptable) response, with an output body in application/rdf+xml, as is allowed by HTTP 1.1.

Bio2RDF2SADI

Discovering and reusing data requires substantial expertise about where data are located and how to transform them into a more useable form for further analysis. While the Bio2RDF project transforms dozens of key bioinformatic resources into RDF, and is made available through public SPARQL endpoints, a key challenge still remained: how to identify which datasets contain the entities and relations that are of interest to solve a particular problem. To this end, Bio2RDF now generates and publishes summaries of the dataset contents in each of its SPARQL endpoints, thereby simplifying lookup, and reducing server load for expensive and common queries.

During the hackathon, an architecture was developed for an automated approach that utilizes the metadata from Bio2RDF's content summaries to automatically generate SADI Semantic Web Services that provide discoverable access to this Bio2RDF data¹⁰⁴. SADI Services use ontologies to formally describe their inputs and outputs, such that it is possible to find services of interest by querying their ontological descriptions via a global Service metadata registry. In the case of these Bio2RDF SADI Services, the input data-type is a simple Bio2RDF typed-URI (for example <http://bio2rdf.org/mesh:C025643> rdf:

type ctd:Chemical]) and the output is, as per the SADI specifications, the input node annotated with a Bio2RDF relation (for example <http://bio2rdf.org/mesh:C025643> sio:is-participant-in <http://bio2rdf.org/go:0008380>). Such metadata descriptions can be automatically generated from the Bio2RDF indexes, and moreover, the corresponding SPARQL queries that make up the business logic of the service can similarly be automatically constructed based on the information in these indexes. As such, both the service description, as well as the service itself, can be dynamically created to provide access to any Bio2RDF data of interest.

The advantage of exposing Bio2RDF as a set of SADI services is that the data in Bio2RDF becomes discoverable - software does not need to know, *a priori*, what data/relations exist in which Bio2RDF endpoint. Moreover, when exposed as SADI Services, Bio2RDF data can more easily be integrated into workflows using popular workflow editors such as Taverna¹⁰⁵ or as demonstrated by our use of these services within Galaxy workflows¹⁰⁶.

Quality assessment

A large amount of biomedical information is available via SPARQL endpoints, often in a redundant way. Life Sciences databases often integrate information from different sources to enrich the data they provide, and some information resources are pure aggregators whose value is in the harmonization of the information that they collect. As these resources publish their information on the Semantic Web, the result is that the same information is present in multiple endpoints. As a consequence, to decide which endpoint to use to access some particular data of interest is not a trivial task. Two hackathon activities addressed this issue. The development of a dataset descriptor is useful to know which data are present in an endpoint, with information on version, representation and update policies. But even if such a descriptor is provided, there is still an issue of the reliability of endpoints. It is also difficult to know which endpoints are actively maintained and which are not.

YummyData is a project that monitors endpoints by periodically running queries and performing a few tests. By collecting data over extended periods, it can provide a proxy for the reliability of an endpoint and the dynamism of the information it provides. More specifically, YummyData periodically queries datahub.io for datasets tagged as being of biomedical interest. It combines the result with a list of curated endpoints and, periodically, it runs a series of tests and queries and stores their results. YummyData performs some tests to determine whether the endpoint provides a VoID descriptor (see section above), as well as to measure response time. It also runs a series of queries that can be generic or endpoint-specific. Generic queries inspect aggregate information such as the number of statements, distinct resources, or properties. Specific queries are currently only implemented as a proof of concept, but they are intended to reveal aspects of the quality of the data provided by endpoints. For instance, a typical query would ask for the number of entities annotated via a given evidence code. Results over time are then aggregated in two types of rating: a SPARQL score that is a numeric value that results from a count of positive response

codes over time windows; a star rating that is intended to provide a more qualitative assessment of features (e.g. the availability of a valid VoID descriptor, or of a copyright notice, yields +1 star). At the time of writing, YummyData has collected data for about a year on a few tens of information resources. A subset of these data are accessible via the <http://yummydata.org> website.

Conclusion

To fulfil the mission of the DBCLS, which is to integrate life sciences databases, the annual BioHackathon series was started in 2008 to explore state-of-the-art technological solutions. The utilization of Semantic Web technologies as a means for database integration was introduced in BioHackathon 2010³. Since then, we have collaboratively worked as a community to promote the use of RDF and ontologies in life sciences. As one of the demonstration products, DBCLS released the first RDF-based genome database, TogoGenome, in 2013. Subsequently, the EBI RDF Platform was released by EMBL-EBI and PubChem RDF was published by NCBI, and these provide fundamental database resources in genomics to the wider biomedical research community as well as the pharmaceutical and biotechnology industries. The NBDC RDF portal launched in 2015 complements the above resources by adding other major domains such as protein structures and glycoscience resources. The 6th and 7th BioHackathons in 2013 and 2014 were held to develop and improve methods and best practices for creating and publishing these community wide resources. As a result, the field is becoming ready for testing in real world use cases such as dealing with human genome-scale biomedical data. Other domains (e.g. plants/crops) are less developed but gaining momentum (see for instance [AgroPortal](#)). At the same time we found another layer of demands for additional development in real world applications such as genotype-phenotype information to drug discovery, which define further challenges and will be addressed in the upcoming BioHackathons.

Data availability

Underlying data

No data are associated with this article.

Extended data

Records of the BioHackathon 2013 and 2014 meetings are aggregated at <https://github.com/dbcls/bh13/wiki> and <https://github.com/dbcls/bh14/wiki> respectively.

Zenodo: dbcls/bh13: Included repositories related to BH13. <http://doi.org/10.5281/zenodo.3271508>¹⁰⁷

This project contains the following extended data:

- dbcls/bh13-v1.0.0.zip (BioHackathon 2013 records)

Zenodo: dbcls/bh14: Included repositories related to BH14. <http://doi.org/10.5281/zenodo.3271509>¹⁰⁸

This project contains the following extended data:

- dbcls/bh14-v1.0.0.zip (BioHackathon 2014 records)

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

Grant information

Funding for the BioHackathon was supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

BioHackathon 2013 and 2014 are supported by the Integrated Database Project (Ministry of Education, Culture, Sports, Science and Technology of Japan) and hosted by the National Bioscience Database Center (NBDC) and the Database Center for Life Science (DBCLS). We thank Yuji Kohara, the director of DBCLS, for his supporting the BioHackathons.

References

1. Katayama T, Arakawa K, Nakao M, *et al.*: **The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows.** The DBCLS BioHackathon Consortium. *J Biomed Semantics*. 2010; 1(1): 8. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Katayama T, Wilkinson MD, Vos R, *et al.*: **The 2nd DBCLS BioHackathon: interoperable bioinformatics web services for integrated applications.** *J Biomed Semantics*. 2011; 2: 4. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Katayama T, Wilkinson MD, Micklem G, *et al.*: **The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies.** *J Biomed Semantics*. 2013; 4(1): 6. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Katayama T, Wilkinson MD, Aoki-Kinoshita KF, *et al.*: **BioHackathon series in 2011 and 2012: penetration of ontology and linked data in life science domains.** *J Biomed Semantics*. 2014; 5(1): 5. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Jupp S, Malone J, Bolleman J, *et al.*: **The EBI RDF platform: linked open data for the life sciences.** *Bioinformatics*. 2014; 30(9): 1338–1339. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Magrane M, UniProt Consortium: **UniProt Knowledgebase: a hub of integrated protein data.** *Database (Oxford)*. 2011; 2011: bar009. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Bolton EE, Wang Y, Thiessen PA, *et al.*: **PubChem: Integrated Platform of Small Molecules and Biological Activities.** *Annu Rep Comput Chem*. 2008; 4: 217–241. [Publisher Full Text](#)
8. Uchiyama I, Mihara M, Nishide H, *et al.*: **MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data.** *Nucleic Acids Res*. 2015; 43(Database issue): D270–D276. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

9. Katayama T, Kawashima S, Okamoto S, *et al.*: **TogoGenome/TogoStanza: modularized Semantic Web genome database.** *Database (Oxford)*. 2019; **2019**: bay132.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Bolleman JT, Mungall CJ, Strozzi F, *et al.*: **FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation.** *J Biomed Semantics*. 2016; **7**: 39.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Baran J, Durgahee BS, Eilbeck K, *et al.*: **GFVO: the Genomic Feature and Variation Ontology.** *PeerJ*. 2015; **3**: e933.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Cochrane G, Karsch-Mizrachi I, Takagi T, *et al.*: **The international nucleotide sequence database collaboration.** *Nucleic Acids Res*. 2016; **44**(D1): D48–D50.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
13. Aken BL, Achuthan P, Akanni W, *et al.*: **Ensembl 2017.** *Nucleic Acids Res*. 2017; **45**(D1): D635–D642.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Landrum MJ, Lee JM, Benson M, *et al.*: **ClinVar: public archive of interpretations of clinically relevant variants.** *Nucleic Acids Res*. 2016; **44**(D1): D862–D868.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Hamosh A, Scott AF, Amberger JS, *et al.*: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res*. 2005; **33**(Database issue): D514–517.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Fokkema IF, Taschner PE, Schaafsma GC, *et al.*: **LOVD v.2.0: the next generation in gene variant databases.** *Hum Mutat*. 2011; **32**(5): 557–563.
[PubMed Abstract](#) | [Publisher Full Text](#)
17. Stenson PD, Mort M, Ball EV, *et al.*: **The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine.** *Hum Genet*. 2014; **133**(1): 1–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Imanishi T, Itoh T, Suzuki Y, *et al.*: **Integrative annotation of 21,037 human genes validated by full-length cDNA clones.** *PLoS Biol*. 2004; **2**(6): e162.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
19. Takeda JI, Yamasaki C, Murakami K, *et al.*: **H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery.** *Nucleic Acids Res*. 2013; **41**(Database issue): D915–D919.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Burge SW, Daub J, Eberhardt R, *et al.*: **Rfam 11.0: 10 years of RNA families.** *Nucleic Acids Res*. 2013; **41**(Database issue): D226–32.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data.** *Nucleic Acids Res*. 2011; **39**(Database issue): D152–7.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.** *Cell*. 2005; **120**(1): 15–20.
[PubMed Abstract](#) | [Publisher Full Text](#)
23. O'Leary NA, Wright MW, Brister JR, *et al.*: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.** *Nucleic Acids Res*. 2016; **44**(D1): D733–D745.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Rosse C, Mejino JLV Jr: **The Foundational Model of Anatomy Ontology.** In *Anat Ontol Bioinforma Princ Pract. Computational Biol*. Edited by Burger A, Davidson D, Baldock R. Springer; 2008; **6**: 59–117.
[Publisher Full Text](#)
25. Bard JB: **The AEO, an Ontology of Anatomical Entities for Classifying Animal Tissues and Organs.** *Front Genet*. 2012; **3**(FEB): 18.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
26. Dahdul WM, Balhoff JP, Blackburn DC, *et al.*: **A unified anatomy ontology of the vertebrate skeletal system.** *PLoS One*. 2012; **7**(12): e51070.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Ramírez MJ, Coddington JA, Maddison WP, *et al.*: **Linking of digital images to phylogenetic data matrices using a morphological ontology.** *Syst Biol*. 2007; **56**(2): 283–294.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. Chiba H, Nishide H, Uchiyama I: **Construction of an ortholog database using the semantic web technology for integrative analysis of genomic data.** *PLoS One*. 2015; **10**(4): e0122802.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Miñarro-Gimenez JA, Madrid M, Fernandez-Breis JT: **OGO: an ontological approach for integrating knowledge about orthology.** *BMC Bioinformatics*. 2009; **10**(Suppl 10): S13.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Schmitt T, Messina DN, Schreiber F, *et al.*: **Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information.** *Brief Bioinform*. 2011; **12**(5): 485–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
31. UniProt Consortium: **UniProt: A hub for protein information.** *Nucleic Acids Res*. 2015; **43**(Database issue): D204–D212.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Hakenberg J, Plake C, Leaman R, *et al.*: **Inter-species normalization of gene mentions with GNAT.** *Bioinformatics*. 2008; **24**(16): i126–132.
[PubMed Abstract](#) | [Publisher Full Text](#)
33. Wei CH, Kao HY: **Cross-species gene normalization by species inference.** *BMC Bioinformatics*. 2011; **12**(Suppl 8): S5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Wei CH, Harris BR, Kao HY, *et al.*: **tmVar: a text mining approach for extracting sequence variants in biomedical literature.** *Bioinformatics*. 2013; **29**(11): 1433–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
35. Caporaso JG, Baumgartner WA Jr, Randolph DA, *et al.*: **MutationFinder: a high-performance system for extracting point mutation mentions from text.** *Bioinformatics*. 2007; **23**(14): 1862–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
36. Leaman R, Gonzalez G: **BANNER: an executable survey of advances in biomedical named entity recognition.** *Pac Symp Biocomput*. 2008; **13**: 652–663.
[PubMed Abstract](#) | [Publisher Full Text](#)
37. Gerner M, Sarafraz F, Bergman CM, *et al.*: **BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events.** *Bioinformatics*. 2012; **28**(16): 2154–61.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
38. Van Landeghem S, Björne J, Wei CH, *et al.*: **Large-scale event extraction from literature with multi-level gene normalization.** *PLoS One*. 2013; **8**(4): e55814.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Kim J, Wang Y: **PubAnnotation - a persistent and sharable corpus and annotation repository.** In *Proc 2012 Work Biomed Nat Lang Process*. 2012(BioNLP): 202–205.
[Reference Source](#)
40. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics*. 2009; **25**(16): 2078–2079.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
41. Danecek P, Auton A, Abecasis G, *et al.*: **The variant call format and VCFtools.** *Bioinformatics*. 2011; **27**(15): 2156–2158.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Reese MG, Moore B, Batchelor C, *et al.*: **A standard variation file format for human genome sequences.** *Genome Biol*. 2010; **11**(8): R88.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
43. Fernández JD, Martínez-Prieto MA, Gutiérrez C, *et al.*: **Binary RDF representation for publication and exchange (HDT).** *J Web Semant*. 2013; **19**: 22–41.
[Publisher Full Text](#)
44. Fujisawa T, Narikawa R, Maeda SI, *et al.*: **CyanoBase: a large-scale update on its 20th anniversary.** *Nucleic Acids Res*. 2017; **45**(D1): D551–D554.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
45. Buels R, Yao E, Diesh CM, *et al.*: **JBrowse: a dynamic web platform for genome visualization and analysis.** *Genome Biol*. 2016; **17**: 66.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
46. Kalderimis A, Lyne R, Butano D, *et al.*: **InterMine: extensive web services for modern biology.** *Nucleic Acids Res*. 2014; **42**(Web Server issue): W468–472.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
47. Velankar S, Dana JM, Jacobsen J, *et al.*: **SIFTS: Structure Integration with Function, Taxonomy and Sequences resource.** *Nucleic Acids Res*. 2013; **41**(Database issue): D483–9.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
48. Kinjo AR, Bekker GJ, Suzuki H, *et al.*: **Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures.** *Nucleic Acids Res*. 2017; **45**(D1): 282–288.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Ison J, Kalas M, Jonassen I, *et al.*: **EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats.** *Bioinformatics*. 2013; **29**(10): 1325–32.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
50. Juty N, Le Novère N, Laibe C: **Identifiers.org and MIRIAM Registry: community resources to provide persistent identification.** *Nucleic Acids Res*. 2012; **40**(Database issue): D580–586.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
51. Komiya Y, Masaki B, Saad G, *et al.*: **UTProt: Database Integration and Tool Development for Intracomplex Utilizing Biosemantics.** In *2013 Annu Conv Japanese Soc Bioinforma*. 2013; **81**: 81–82.
52. Perez-Riverol Y, Hermjakob H, Kohlbacher O, *et al.*: **Computational proteomics pitfalls and challenges: HavanaBioinfo 2012 Workshop report.** *J Proteomics*. 2013; **87**: 134–138.
[PubMed Abstract](#) | [Publisher Full Text](#)
53. Deutsch EW, Albar JP, Binz PA, *et al.*: **Development of data representation standards by the human proteome organization proteomics standards initiative.** *J Am Med Informatics Assoc*. 2015; **22**(3): 495–506.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
54. Perez-Riverol Y, Alpi E, Wang R, *et al.*: **Making proteomics data accessible and reusable: current state of proteomics databases and repositories.** *Proteomics*. 2015; **15**(5–6): 930–49.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
55. Perez-Riverol Y, Wang R, Hermjakob H, *et al.*: **Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective.** *Biochim Biophys Acta*. 2014; **1844**(1 Pt A): 63–76.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
56. Chervitz SA, Deutsch EW, Field D, *et al.*: **Data standards for Omics data: the**

- basis of data sharing and reuse. In *Methods Mol Biol*. Edited by Mayer B. Springer; 2011; 719: 31–69.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
57. Griss J, Jones AR, Sachsenberg T, *et al.*: **The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience.** *Mol Cell Proteomics*. 2014; 13(10): 2765–2775.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 58. Sugimoto M, Kawakami M, Robert M, *et al.*: **Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis.** *Curr Bioinform*. 2012; 7(1): 96–108.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 59. Perez-Riverol Y, Uszkoreit J, Sanchez A, *et al.*: **ms-data-core-api: an open-source, metadata-oriented library for computational proteomics.** *Bioinformatics*. 2015; 31(17): 2903–2905.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 60. Craig R, Cortens JP, Beavis RC: **Open source system for analyzing, validating, and storing protein identification data.** *J Proteome Res*. 2004; 3(6): 1234–1242.
[PubMed Abstract](#) | [Publisher Full Text](#)
 61. Vizcaino JA, Côté RG, Csordas A, *et al.*: **The Proteomics Identifications (PRIDE) database and associated tools: Status in 2013.** *Nucleic Acids Res*. 2013; 41(Database issue): D1063–1069.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 62. Aoki-Kinoshita KF, Bolleman J, Campbell MP, *et al.*: **Introducing glycomics data into the Semantic Web.** *J Biomed Semantics*. 2013; 4(1): 39.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 63. Ranzinger R, Aoki-Kinoshita KF, Campbell MP, *et al.*: **GlycoRDF: an ontology to standardize glycomics data in RDF.** *Bioinformatics*. 2015; 31(6): 919–925.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 64. Ashburner M, Ball CA, Blake JA, *et al.*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet*. 2000; 25(1): 25–29.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 65. Degtyarenko K, de matos P, Ennis M, *et al.*: **ChEBI: a database and ontology for chemical entities of biological interest.** *Nucleic Acids Res*. 2008; 36(Database issue): D344–350.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 66. McDonald AG, Boyce S, Tipton KF: **ExplorEnz: the primary source of the IUBMB enzyme list.** *Nucleic Acids Res*. 2009; 37(Database issue): D593–597.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 67. Kotera M, Nishimura Y, Nakagawa Z, *et al.*: **PIERO ontology for analysis of biochemical transformations: effective implementation of reaction information in the IUBMB enzyme list.** *J Bioinform Comput Biol*. 2014; 12(6): 1442001.
[PubMed Abstract](#) | [Publisher Full Text](#)
 68. Yamanishi Y, Tabei Y, Kotera M: **Metabolome-scale de novo pathway reconstruction using regioisomer-sensitive graph alignments.** *Bioinformatics*. 2015; 31(12): i161–i170.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 69. Tabei Y, Yamanishi Y, Kotera M: **Simultaneous prediction of enzyme orthologs from chemical transformation patterns for de novo metabolic pathway reconstruction.** *Bioinformatics*. 2016; 32(12): i278–i287.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 70. Ide N, Suderman K: **GrAF: A Graph-based Format for Linguistic Annotations.** In *Proc Linguist Annot Work 2007*. 2007; 1–8.
[Reference Source](#)
 71. Comeau DC, Islamaj Doğan R, Ciccarese P, *et al.*: **BioC: a minimalist approach to interoperability for biomedical text processing.** *Database (Oxford)*. 2013; 2013: bat064.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 72. Ciccarese P, Ocana M, Garcia Castro LJ, *et al.*: **An open annotation ontology for science on web 3.0.** *J Biomed Semantics*. 2011; 2 Suppl 2: S4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 73. Ferrucci D, Lally A: **UIMA: an architectural approach to unstructured information processing in the corporate research environment.** *J Nat Lang Engineering*. 2004; 10: 327–348.
[Publisher Full Text](#)
 74. Sanderson R, Ciccarese P, Van de Sompel H: **Designing the W3C open annotation data model.** *Proceedings of the 5th Annual ACM Web Science Conference-WebSci '13*. 2013; 366–375.
[Publisher Full Text](#)
 75. Verspoor K, Ave E: **Towards Adaptation of Linguistic Annotations to Scholarly Annotation Formalisms on the Semantic Web.** In *Proc Sixth Linguist Annot Work*. 2012; 75–84.
[Reference Source](#)
 76. Bada M, Eckert M, Evans D, *et al.*: **Concept annotation in the CRAFT corpus.** *BMC Bioinformatics*. 2012; 13: 161.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 77. Kim J, Han X, Lee V, *et al.*: **GRO Task: Populating the Gene Regulation Ontology with events and relations.** *Work BioNLP Shar Task*. 2013; 50–57.
[Reference Source](#)
 78. Beisswanger E, Lee V, Kim JJ, *et al.*: **Gene Regulation Ontology (GRO): design principles and use cases.** *Stud Health Technol Inform*. 2008; 136: 9–14.
[PubMed Abstract](#)
 79. Yamamoto Y, Yamaguchi A, Bono H, *et al.*: **Allie: a database and a search service of abbreviations and long forms.** *Database (Oxford)*. 2011; 2011: bar013.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 80. Kaufmann E, Bernstein A: **How useful are natural language interfaces to the semantic Web for casual end-users?** *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2007; 4825 LNCS: 281–294.
[Publisher Full Text](#)
 81. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology.** *Nucleic Acids Res*. 2004; 32(Database issue): D267–D270.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 82. Gaudet P, Bairoch A, Field D, *et al.*: **Towards BioDBCore: A community-defined information specification for biological databases.** *Database (Oxford)*. 2011; 2011: baq027.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 83. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data*. 2016; 3: 160018.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 84. Callahan A, Cruz-Toledo J, Dumontier M: **Ontology-Based Querying with Bio2RDF's Linked Open Data.** *J Biomed Semantics*. 2013; 4 Suppl 1: S1.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 85. Dumontier M, Gray AJG, Marshall MS, *et al.*: **The health care and life sciences community profile for dataset descriptions.** *PeerJ*. 2016; 4: e2331.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 86. Smith RN, Aleksic J, Butano D, *et al.*: **InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data.** *Bioinformatics*. 2012; 28(23): 3163–5.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 87. Sullivan J, Karra K, Moxon SA, *et al.*: **InterMOD: integrated data and tools for the unification of model organism research.** *Sci Rep*. 2013; 3: 1802.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 88. Balakrishnan R, Park J, Karra K, *et al.*: **YeastMine—an integrated data warehouse for *Saccharomyces cerevisiae* data as a multipurpose tool-kit.** *Database (Oxford)*. 2012; 2012: bar062.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 89. Wang SJ, Laulederkind SJ, Hayman GT, *et al.*: **Analysis of disease-associated objects at the Rat Genome Database.** *Database (Oxford)*. 2013; 2013: bat046.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 90. Howe DG, Bradford YM, Conlin T, *et al.*: **ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics.** *Nucleic Acids Res*. 2013; 41(Database issue): D854–60.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 91. Motenko H, Neuhauser SB, O'Keefe M, *et al.*: **MouseMine: a new data warehouse for MGI.** *Mamm Genome*. 2015; 26(7–8): 325–330.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 92. Lyne R, Smith R, Rutherford K, *et al.*: **FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics.** *Genome Biol*. 2007; 8(7): R129.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 93. Krishnakumar V, Contrino S, Cheng CY, *et al.*: **ThaleMine: A Warehouse for Arabidopsis Data Integration and Discovery.** *Plant Cell Physiol*. 2017; 58(1): e4.
[PubMed Abstract](#) | [Publisher Full Text](#)
 94. Celniker SE, Dillon LA, Gerstein MB, *et al.*: **Unlocking the secrets of the genome.** *Nature*. 2009; 459(7249): 927–930.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 95. Contrino S, Smith RN, Butano D, *et al.*: **modMine: flexible access to modENCODE data.** *Nucleic Acids Res*. 2012; 40(Database issue): D1082–8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 96. Morita M, Igarashi Y, Ito M, *et al.*: **Sagace: a web-based search engine for biomedical databases in Japan.** *BMC Res Notes*. 2012; 5: 604.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 97. Forrest ARR, Kawaji H, Rehli M, *et al.*: **A promoter-level mammalian expression atlas.** *Nature*. 2014; 507(7493): 462–470.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 98. Andersson R, Gebhard C, Miguel-Escalada I, *et al.*: **An atlas of active enhancers across human cell types and tissues.** *Nature*. 2014; 507(7493): 455–461.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 99. Wilkinson MD, Vandervalk B, McCarthy L: **The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation.** *J Biomed Semantics*. 2011; 2(1): 8.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 100. Mons B, Velterop J: **Nano-Publication in the e-science era.** In *Proc Work Semant Web Appl Sci Discourse (SWASD 2009)*. 2009.
[Reference Source](#)
 101. Lizio M, Harshbarger J, Shimoji H, *et al.*: **Gateways to the FANTOM5 promoter level mammalian expression atlas.** *Genome Biol*. 2015; 16: 22.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
 102. Harland L: **Open PHACTS: A Semantic Knowledge Infrastructure for Public and Commercial Drug Discovery Research.** In *Knowl Eng Knowl Manag EKWQ 2012 Lect Notes Comput Sci*. Springer Berlin Heidelberg; 2012; 7603: 1–7.
[Publisher Full Text](#)
 103. Kuhn T, Barbano PE, Nagy ML, *et al.*: **Broadening the Scope of Nanopublications.** In *Semant Web Semant Big Data*. Edited by Cimianoe P, Corcho O, Presutti V, Hollink L, Rudolph S. Springer Berlin Heidelberg; 2013; 7882: 487–501.
[Publisher Full Text](#)

104. González AR, Callahan A, Cruz-toledo J, *et al.*: **Automatically exposing OpenLifeData via SADI semantic Web Services.** *J Biomed Semantics.* 2014; 5: 46.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
105. Wolstencroft K, Haines R, Fellows D, *et al.*: **The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud.** *Nucleic Acids Res.* 2013; 41(Web Server issue): W557–561.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
106. Aranguren ME, Wilkinson MD: **Enhanced reproducibility of SADI web service workflows with Galaxy and Docker.** *Gigascience.* 2015; 4(1): 59.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
107. Katayama T: **dbcls/bh13: Added CC-BY license as requested by the journal (Version 1.0.1).** *Zenodo.* 2019.
<http://www.doi.org/10.5281/zenodo.3271508>
108. Katayama T: **dbcls/bh14: Added CC-BY license as requested by the journal (Version 1.0.1).** *Zenodo.* 2019.
<http://www.doi.org/10.5281/zenodo.3271509>

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 28 August 2020

<https://doi.org/10.5256/f1000research.19950.r67760>

© 2020 Moreira J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? João Moreira

University of Twente, Enschede, The Netherlands

The paper presents an overview about experiences and produced work related to RDF from the 6th and 7th annual BioHackathons (2013 and 2014).

The paper is structured in two major sections about (1) RDF data management in life sciences; and (2) meta-data about the RDF datasets with emphasis on SPARQL endpoint resources. The paper provides an important contribution to the state-of-art in the context of RDF resources development and publication for life sciences, also presenting some discussion about good practices in these topics.

The paper is quite relevant and is well written, but there are some issues:

- My main concern regards how current (actual) the content presented is, since the paper describes the editions of 2013 and 2014, and 7 editions occurred after these. I know that many advances/improvements happened in these last editions, including for example the use of ontologies like the Ontology for Biomedical Investigations (OBI), the SemanticScience Integrated Ontology (SIO), the HCLS ontology profile, the SPAR ontologies (particularly FaBio), FHIR, among others, in different use cases and with more recent technologies (e.g., ShEx, SHACL, RDF*, GraphDB, JSON-LD, etc).
- Ideally, this survey should also incorporate these more recent BioHackathons (2015-2019) or at least provide a discussion about the main points where the most important advances/improvements happened. The current (recent) literature has many works produced in these last 7 editions.
- I missed a paragraph (or some sentences in the end of Introduction) explaining why the authors choose to structure the paper in these 2 major categories: (1) Improvement and utilization of RDF data in life sciences; and (2) Metadata about RDF data resources.
- The reasoning behind the paper structure is not so clear. For example, it is difficult to understand why the sub-sections of "Provenance of Data" were organized in Nanopublication, Bio2RDF2SADI and Quality assessment. The 2 first categories refer to

specific approaches, while the later refers to a generic data management activity. Similar issue happens on the other sections.

- The discussion on best practices, outlined in the abstract as a contribution of this paper, is quite difficult to track. One can find some recommendations spread in the text, but, in my point of view, it would be better to have a section only for Discussion and Recommendations. So, I suggest to create this section before the conclusions, moving all good practices' discussion to it.
- Furthermore, it would be valuable to have in this new section a summary of the main open issues regarding the 2 major categories, highlighting which of the presented approaches faced these issues, giving directions to solve them.
- In the Provenance of data section, I missed the reference to W3C PROV, which btw is used by the nanopublication approach and is a fundamental standard for reasoning over the data provenance.
- The Conclusion section can be improved by making explicit what were the main lessons learned, the main contributions of the paper and a summary of the open issues identified.

Is the topic of the opinion article discussed accurately in the context of the current literature?

Partly

Are all factual statements correct and adequately supported by citations?

Yes

Are arguments sufficiently supported by evidence from the published literature?

Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Ontology engineering, semantic interoperability, FAIR data principles, e-Health semantics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 18 November 2019

<https://doi.org/10.5256/f1000research.19950.r54207>

© 2019 Vision T. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Todd J. Vision

Department of Biology, School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

The authors report on the many activities conducted under the umbrella of the BioHackathons 2013 and 2014. While there are recurring intellectual threads, most notably a focus on support for RDF, the manuscript is really a collection of reports of the progress made by each of the different subgroups involved in these events. I admire the work that must have gone into fitting each of these individual efforts into a coherent narrative, and this is a very useful historical document of those efforts.

I do have some suggestions below where I think the manuscript could be improved - primarily to help developers and users put this work into context. Most of these are not critical flaws, however, and I would leave to the authors' discretion how they wish to respond to these comments and suggestions.

The one issue I think is critical for acceptance is the explanation of Figure 1 (see below).

One recurring issue is the ambiguity of the reference time frame, both for the starting and stopping points of the work described. This is important since the report covers work that took place 5-6 years in the past, a long time ago in the world of bioinformatics.

With respect to the starting point, the use of present tense makes it unclear in many passages whether the authors are describing the state of software as of 2013/2014, or today. To avoid confusion, I would recommend either fixing the relevant passages throughout (I list some of them below) or clarify upfront that the state of affairs of the software tools and resources described is as of 5-6 years ago, which in many cases will no longer be accurate.

Regarding the stopping point, it is not clear if the progress described reflects what was accomplished during the event, up until the next biohackathon, or some other time point.

This is perhaps a larger ask, but I wanted as a reader to know where these efforts stand today.

Putting the work in context of the progress since would take a weakness of the manuscript (ie. reporting on older work) and turn it into a virtue (evaluating the impact each of these efforts had on their fields with the benefits of a few years of hindsight). Perhaps this could be summarized by a "where-are-they-now" table.

In addition, there are a number of claims made, particularly regarding tool performance, without supporting evidence provided. In some cases, these might be from subjective assessments, which is fine as long as that is clear. But subjective or not, the lack of evidence makes the claims difficult to evaluate. I point some of these passages out below.

COMMENTS BY SECTION:

ABSTRACT

"an exciting and unanticipatable variety of real world applications in coming years" - this is an odd statement given that several years have already passed.

INTRODUCTION

"To fully utilize open data in life sciences, semantic interoperability and standardization of data are required to allow innovative development of applications" - while it may be true, this claim is not directly supported. A weaker claim might be more justified.

IMPROVEMENT AND UTILIZATION OF RDF DATA IN LIFE SCIENCES"

"RDF ... is becoming widely accepted" - can the other cite support for this claim?

GENOMIC INFORMATION

FIGURE 1 - needs further explanation via a caption and/or legend. What do the colors represent, and the different types of lines? The use of URLs deserves explanation - especially the one to a google doc.

IDENTIFIERS FOR SEQUENCES AND ANNOTATIONS

I couldn't quite figure out what I was supposed to understand about the conceptual differences among the identifier models.

TOOLS FOR SEMANTIC GENOME DATA

"To support these efforts we have collaborated with DDBJ, UniProt, and the EBI RDF group to develop ontologies for representing locations and annotations of genome sequences and used these developments for all prokaryotic genomes and, later, eukaryotic genomes" - this passage seems redundant with the descriptions above. Also, the use of "we" here and in next sentence appears to be legacy text referring to only the TogoGenome participants.

JBrowse on RDF was comparable in performance to typical RDB back-end - it would be good if the authors could support that claim with some of the data obtained.

The sections "Visualization of semantic annotations in JBrowse" and "TogoGenome" are somewhat redundant with each other. I recommend merging.

INTERMINE

"An area that needs work, and is receiving attention," - is this still true 5-6 years later?

"Implement a simple adaptor allowing, as described above, JBrowse to request data directly from InterMine RESTful web services" If this was actually described above, I missed it.

TEXT-MINING

I would personally find a screenshot of the text annotations in JBrowse much more interesting than the current Figure 2, which is exactly what you'd expect it to look like and doesn't provide evidence for the *performance* claim above.

PROTEIN STRUCTURES AND EXPRESSIONS

"the Worldwide Protein Data Bank (wwPDB) will abolish the conventional PDB format" and "wwPDB will start assigning protein chain identifiers, which will also be encoded as URIs in the wwPDB/RDF." 2016 is three years ago, so if this has already been accomplished, this should be in past tense.

"Recently, the mzTab data exchange format was introduced" - recently meaning 2014?

"the "omics" group in the BioHackathon 2014 made the first steps towards the development of the ProteomeXchange Interface (PROXI) for protein expression data exchange⁵⁹." - my impression is that this effort has evolved considerably since 2014. If true, it would be valuable to put this into context for how the technology is used currently by the ProteomeXchange consortium.

"integration of the two corpora, CRAFT and GRO, into PubAnnotation, demonstrated significantly improved utility" - is there supporting evidence for this claim?

"Access to consistent, high-quality metadata is critical to finding, understanding, and reusing scientific data - this is the core of the FAIR Data Principles⁸³." - it's odd to reference this manifesto of principles that postdates the hackathons by at least two years. This would be less awkward if there was a table or section devoted to developments *since* the 2014, i.e. "where are they now?"

DATASET METADATA

"The resulting guidelines, finalized under the auspices of the W3C Semantic Web for Health Care and the Life Sciences Interest Group (HCLSIG)" - there's no link or citation given here.

The 4th pgph redundantly includes the same coverage list: "..elements of description, identification, versioning, attribution, provenance, and content summarization."

"The group is currently evaluating the specification with implementations for dataset registries such as Identifiers.org⁵⁰ and IntegBio Database Catalog, as well as Linked Data repositories such as Bio2RDF⁸⁴." - update on progress since 2014?

W2C link appears to be dead, but a paper was published in 2016. I get the impression that there was further work on this post-hackathon not mentioned here.

VOID FOR INTERMINE AND UNIPROT

The relationship of prior work on VOiD (if any) to the outcome from the hackathon is not clear to me.

"there are now InterMine databases available" - is "now" as of writing, or as of the time of the hackathon?

The VoID file [for Uniprot] is updated every release in synchrony with our production" - replace use of "our" which I presume refers to the Uniprot team and not to the collaborative team of authors here.

SCHEMA>ORG AND RDFA FOR BIOLOGICAL DATABASES

what's "NIBIOHN"?

NANOPUBLICATIONS

A more explicit explanation of the relationship between NANOPUBLICATIONS and the overarching topic of PROVENANCE would be helpful.

"n-quads" - please explain or provide a reference.

"At the time of writing, YummyData has collected data for about a year" - what is the time of writing?

CONCLUSION

"The NBDC RDF portal launched in 2015 complements the above resources..." - again, this would be appropriate for a "where are they now?" section.

"will be addressed in the upcoming BioHackathons" - they have been held annually since 2014, so it is odd to not mention what **was** addressed in subsequent events. And I hope those reports will be published in a more timely fashion!

Is the topic of the opinion article discussed accurately in the context of the current literature?

Partly

Are all factual statements correct and adequately supported by citations?

No

Are arguments sufficiently supported by evidence from the published literature?

Partly

Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary genetics, genome evolution, bioinformatics and computational biology.

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research