SemNExT: A Framework for Semantically Integrating and Exploring Numeric Analyses

Evan W. Patton[1], Elisabeth Brown[2], Matthew Poegel[2], Hannah De Los Santos[2], Chris Fasano[3], Kristin P. Bennett[1,2], and Deborah L. McGuinness[1]

[1] Dept. of Computer Science, Rensselaer Polytechnic Institute, Troy, NY USA
[2] Dept. of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY USA
[3] Neural Stem Cell Institute,Rensselaer, NY USA

Abstract. Combining statistical techniques with semantic data representations holds the potential to enhance understandability of scientific results. It can augment scientific findings with existing data sources in a reproducible manner through provenance capture, as well as enable further analysis and deduction through computer and human understandable definitions of terms. We present a framework for semantically integrating and exploring numerical analyses. We call our work SemNExT for Semantic Numeric Exploration Technology. We apply our approach to data analysis aimed at improving understanding of human brain development that leverages the Cortecon RNA-Seq data repository. Our approach supports enrichment of Cortecon data through combinations with structured data sources available via SQL or SPARQL from the web to provide semantically enhanced analyses combined with statistical analyses. Our results are encoded as RDF graphs that may be used as input to reasoners and may drive provenance-aware visualizations. We introduce our infrastructure, describe its use on transcriptomic data analysis of a model of cerebral cortex development, and discuss some emerging suggestions for best practices and future research challenges.

Keywords: modeling, statistical processes, provenance, linked data

1 Introduction

Our work at the intersection of semantics and data analysis began after discussions with successful data analysts detailed the level of human investigation typically required. We hypothesized that some of the human interpretation and vetting of results could be automated with semantic technologies and structured web-available resources. We explore this hypothesis in the setting of brain development data analysis. Recent work by [9], among others, has explored the application of machine learning and statistical techniques to transcriptomic data from an in-vitro stem cell model of the development of the human cerebral cortex called Cortecon.[4] Roughly, Cortecon captures snapshots over 77 days of RNA presence in an experiment that models "human brain development in a dish." In

[4] `http://cortecon.neuralsci.org`

the model, embryonic stem cells differentiate and then begin forming the 6 layers of the cerebral cortex. Through Cortecon we can ethically examine human brain development and how diseases such as autism and Alzheimer's may emerge from gene mutations and exposure to toxins. We build on this work and explore techniques for annotating and publishing statistical analyses and their provenance as linked data. We explore the potential of leveraging knowledge from structured resources to enhance the statistical output with unambiguous term definitions, links to related content, and connections to query and reasoning services.

Our Semantic Numeric Exploration Technology (SemNExT) framework is designed to facilitate statistical analyses that can be easily linked to external structured resources and published as linked data. This framework serves as an exemplar for complementing the work of statisticians with the use of provenance and linked data best practices.

2 Related Work

BioGPS provides a pluggable workspace for interacting with different genomic and proteomic datasets available on the World Wide Web [14]. However, this framework does not provide a machine readable representation of the data, making it difficult to repurpose and link with additional resources unless they can be easily expressed via a URL templating scheme.

Structured scientific workflows such as Kepler [10] provide inspiration for this work. Kepler provides a provenance module to capture and query workflow execution history using a SQL-based schema.

Mathematicians have also produced modeling languages for mathematics, such as MathML [1] and the Mizar Mathematical Library.[5] MathML is a W3C recommended standard for representing mathematical formulae on the web using a markup language based on XML. The Mizar Mathematical Library provides a web accessible, machine-verifiable library of mathematical functions. For a thorough review of ontologies and representations for mathematical knowledge on the semantic web, we refer readers to [6].

STATO is an ontology for modeling statistical tests and their applications, as well as probability distributions, variables, spread, and variation metrics.[6] Our modeling efforts complement the work done by the STATO authors, as we introduce additional tests and provide an exemplar model that combines statistical processes with the W3C's Provenance Model [8].

There is ongoing work at the Gene Ontology Consortium to develop a term enrichment protocol using REST principles and a JSON-based interchange format.[7] The protocol allows for metadata exchange to describe the supported inputs and outputs supported by enrichment servers. There is also a JSON-LD context to enable interoperability with linked data clients. It lacks, however,

[5] `http://mizar.org/library/`
[6] `http://stato-ontology.org/`
[7] `https://github.com/cmungall/term-enrichment-protocol`

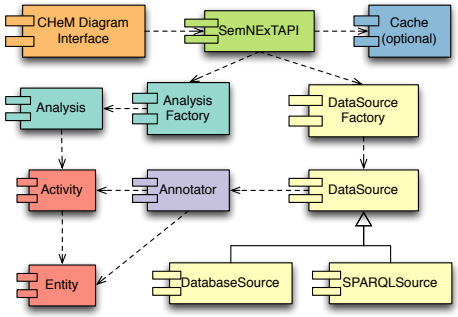

Fig. 1. A conceptual model of the SemNExT architecture. Related concepts are grouped by color.

a clear ability to expose provenance information to the client and a means of incorporating additional or alternative analyses.

Capadisli, Auer, and Riedl explore linked statistical data using RDF Data Cube [5] to enable a semantic web frontend to the R statistical language [4]. It uses distributed SPARQL queries to integrate data from multiple linked data repositories for performing analysis. PROV-O was used to model data provenance. We build on these ideas by incorporating the ability to interface with SQL-based data sources and aim to provide a more general programming mechanism for semantic integration across multiple data sources.

3 Architecture Overview

The SemNExT architecture is divided into six major components: (*1*) Data Source, (*2*) Analysis, (*3*) Annotator, (*4*) Provenance, (*5*) Web Service, and (*6*) Visualization. The interaction of these components is shown in Fig 1. Briefly, SemNExT instantiates a set of *data sources* and *analyses* using *factories*. Each data source provides one or more *annotators* that exploit the semantics of the underlying dataset to extract relevant attributes and links about *entities*. The act of annotating and analyzing entities are captured as *activities* and the results of these operations are *cached* and returned to the *CHeM diagram interface*.

Data Source. SemNExT is an extensible framework for integrating datasets sourced from across the web. We describe the datasets used for our analyses and visualizations of transcriptomic data in Section 6.

Analysis. SemNExT can be set up to perform a variety of analyses on numeric data provided by data sources via the rpy2 Python package. Section 4 describes at a high level the different analyses used on the Cortecon genetic data.

Annotator. SemNExT data sources export Annotator objects that the framework combines to extract information about entities across many datasets. Annotators provide rules to help the framework order operations appropriately at

runtime. For example, UMLS provides relations and attributes for instances of type UMLS-SN:GENE_OR_GENOME, i.a., so it contributes an Annotator that annotates an entity in that class when an entity of that type is detected. Via subclass relationships or inverse functional properties, the framework identifies appropriate entities, extracts the associated relations and attributes, and returns the enriched entity in the result. Subsection 6.5 presents an example.

Provenance. SemNExT heavily relies on the W3C recommended PROV ontology [8] for its provenance data model. Analyses and annotators are modeled as PROV:ENTITY and their applications as PROV:ACTIVITY. Section 5 details the RDF and PROV models for SemNExT.

Web Service. SemNExT provides a RESTful web services interface that is used for searching for diseases, genes, and KEGG pathways[8] and for obtaining semantically enriched networks through the integration of datasets.

Visualization. SemNExT provides a hybrid visualization called chord-heat map (CHeM) diagrams built using the JavaScript library D3. These diagrams will be discussed more in Section 7.

4 Statistical Methods

The Cortecon data set is an in-vitro stem cell model of human cerebral cortex development [9]. The data consist of RNA expression captured by RNA-Seq counts measured over 9 time points (days 0, 7, 12, 19, 26, 33, 49, 63, and 77). The data contain 14065 significant differentially expressed genes, which were filtered using criteria decided upon by the Neural Stem Cell Institute. Our goal was to gain a clearer understanding of the biological processes underlying human corticogenesis and to provide insight into the developmental pathologies of neurological disorders. By analyzing genes with known mutations linked to specific diseases, with respect to temporal gene transcription pattern in the data revealed by Singular Value Decomposition (SVD) and understanding their role in the stages of cortical development revealed by fuzzy C-means (FCM) clustering, we can help identify the stages of corticogenesis where the root causes of the diseases may be found.

To conduct the analysis, we first normalize the data by gene into z-scores using the mean and sample standard deviation of each gene across time. We then performed SVD and FCM clustering. SVD, a form of dimensional reduction, reveals waves of gene transcription by plotting the genes in the space spanned by the first two principal components. The angle of the genes within produces a temporal ordering of the genes. FCM extract clusters of related genes based on their temporal activity resulting in six clusters, which are correspond to stages of development Pluripotency, Ectoderm, Neural Differentiation, Cortical Specification, Deep Layers, and Upper Layers. The analysis reveals a "transcriptomic

[8] `http://www.genome.jp/kegg/pathway.html`

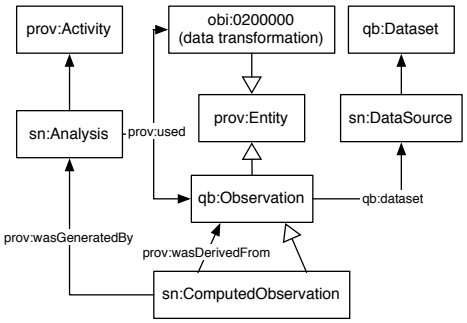

Fig. 2. An overview of the mapping from SemNExT concepts into RDF, the Data Cube vocabulary, PROV-O, and the Ontology for Biomedical Investigations (OBI).

clock" which allows one to understand the role of genes with respect to these developmental stages which are analogous to hours in a clock. All stages except Ectoderm were identified by van de Leemput et al. [9]. Characterization of this newly identified Ectoderm cluster is ongoing [2].

5 Modeling SemNExT processes in RDF

In order to link these data and statistical analyses with structured resources, we begin by converting it from its original tabular form into an RDF graph structure using the semantic techniques of Lebo et al. [7]. Observations are modeled as QB:OBSERVATION and we introduce a subclass SN[9]:COMPUTEDOBSERVATION for modeling the statistical results. Fig. 2 provides an overview of the relevant portion of the SemNExT ontology.

The code implementing the statistical analyses from Section 4 is modeled as PROV:ENTITY and applications of analyses to observations are captured using PROV:ACTIVITY. We make extensive use of STATO terminology (subclasses of OBI:0200000 in Fig. 2) and include extensions to STATO where appropriate.[10]

Consider the task of normalization. The code to compute the z-score is modeled as a method in an R file that, given set of observations, computes the mean and standard deviation of the set, and returns the z-score for each observation. When this task is applied to the dataset of interest, the SN:ANALYSIS used to model this activity is an instantiation of the transformation class STATO:0000104 that uses the specified R code and input data. The output set of z-scores is derived from the input observations and modeled as SN:COMPUTEDOBSERVATION.

The resulting RDF products are then made available as a dataset using a combination of the Vocabulary for Interlinked Datasets (VOID) and the W3C-recommended Data Catalog Vocabulary (DCAT) and Data Cube Vocabulary.

[9] `https://semnext.tw.rpi.edu/ontology/semnext#`
[10] `https://semnext.tw.rpi.edu/docs/STATO-extensions.html`

6 Linking Statistical Analyses to External Resources

6.1 Databases

The DatabaseSource object is used by SemNExT to interface with relational databases. We briefly describe each database incorporated by SemNExT.

Cortecon The Cortecon project from the Neural Stem Cell Institute provides gene read data for a stem cell culture differentiated into neurons during the first 77 days of development [9].

Ensembl. Ensembl is a product of the EMBL - EBI and Wellcome Trust Sanger Institute to automatically annotate eukaryotic genomes.[11] SemNExT uses its mappings from genes to transcripts to proteins to link different datasets.

StringDB. StringDB is a protein-protein interaction (PPI) database containing both known and predicted interactions based on literature textmining.[12] SemNExT uses StringDB as one of two background PPI knowledge bases.

Unified Medical Language System (UMLS). UMLS provides a metathesaurus integrating healthcare and bioinformatics databases.[13] SemNExT uses UMLS as a link set to bridge databases.

6.2 Linked Data

SemNExT integrates directly with a number of linked data resources via the built-in SPARQLSource data source. The framework makes heavy use of *owl:sameAs* links as well as inverse functional properties on identifiers to infer sameness between entities in different datasets and drive the annotation process.

Bio2RDF. We make use of a number of linked life science resources through the Bio2RDF project [3]. Of particular interest are the curated Uniprot Gene Ontology Annotations and descriptions of KEGG pathways.

ReDrugS. The Repurposing Drugs with Semantics (ReDrugS) project[14] is a rich linked data resource that links Online Mendelian Inheritance in Man (OMIM), IRefIndex, and DrugBank from Bio2RDF to identify likely paths for off-label drug uses [11]. It provides a rich scoring mechanism based on how evidence is obtained, e.g. via randomized controlled trials.

Uniprot. We use the SPARQL endpoint provided by the Uniprot Consortium[15] to obtain curated Gene Ontology annotations and expression data [13].

[11] `http://ensembl.org/`
[12] `http://string-db.org/`
[13] `http://www.nlm.nih.gov/research/umls/`
[14] `http://redrugs.tw.rpi.edu`
[15] `http://beta.sparql.uniprot.org/`

6.3 Linking strategies

Reuse of existing identification schemes for linking is a primary goal of SemNExT. Our URI schemes make extensive use of database identifiers for genes, e.g. Entrez IDs and Ensembl IDs. It is not always the case that a clear identification scheme is available for a particular class, however. SemNExT therefore uses three different strategies to perform text-based entity linking (e.g., by matching *rdfs:label*s), the most trivial being exact text and substring matching.

The matching algorithm extensively uses type hierarchies across data sources to limit the search space and return relevant matches. The SemNExT ontology is a hand curated ontology mapping that integrates relevant concepts from our input data sources to enable this work.

When multiple concepts in a dataset match an entity, the framework will exploit broader relationships in an attempt to find a root, or archetype, node encompassing all of the matched nodes. This allows us to answer, for example, a single disease when many genetic variations may exist. If the user provides a specific variation name, the more general concept will not be used to override it.

6.4 Enrichment analysis using linked data resources

After the data sources have been linked, we can use the structured resource to perform a gene set enrichment analysis [12] against the Gene Ontology.[16] This is accomplished by generating a log-odds matrix and computing the p-value for each Gene Ontology term using a hypergeometric distribution. Such a GO enrichment analysis helped reveal the role of the newly described Ectoderm stage. Examining the enrichment of stages using genes with mutations causing neurological diseases suggests hypotheses for how the disease pathologies arise during corticogenesis.

We also attempted to perform a similar analysis on the Tissue Specificity Annotations provided by Uniprot. Unfortunately, the annotations provided are textual rather than URI based, which leads to very poor results without additional processing in the form of natural language processing.

6.5 Linking Example: Septin-9

Consider the gene Septin-9 (SEPT9), which is an important gene in cell division and its mutations are implicated in some cancers. To build a network with this gene, SemNExT begins by searching all of its datasets for the identifier SEPT9 (e.g., *dc:identifier*). It then queries for relations and attributes of SEPT9. For example, the system finds SEPT9's Entrez ID (10801) in Cortecon and adds that information to the gene's shared in-memory model. A UMLS database annotator then executes because it can map the Entrez ID to the UMLS concept identifier for SEPT9. Extracted information includes the fact that SEPT9 is part of the process "Cytokinesis" and that it has an Ensembl Gene ID (ENSG00000184640).

[16] `http://geneontology.org/`

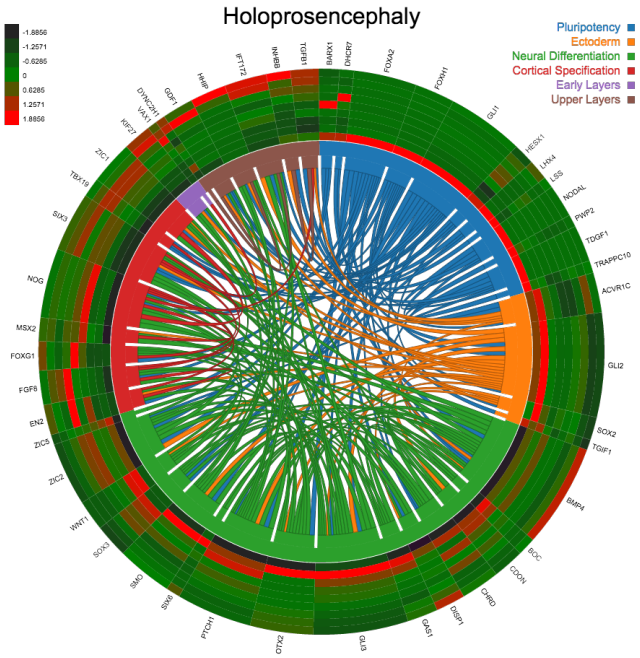

Fig. 3. Sample CHeM visualization suggesting how gene mutations in holoprosencephaly alter corticogenesis by primarily altering the neural differentiation stage. The outer ring shows the expression of each gene across time with genes ordered using SVD analysis and colored by developmental stages identified by ECM. The links show PPI.

Ensembl's annotator triggers on the addition of the Ensembl Gene ID to the shared model, which allows the system to pick up the Gene Ontology annotations provided by Ensembl, for example "cytoskeleton." This procedure continues until no new information can be obtained from the background knowledge sources.

7 Semantic Statistical Visualizations using JSON-LD

SemNExT provides a RESTful interface that produces JSON-LD descriptions of a genetic network that drives visualizations such as our Chord and Heat Map (CHeM) diagrams for understanding how mutations associated with disease may produce changes in cerebral cortex development implemented using D3.js.[17] Figure 3 shows a CHeM diagram for Holoprosencephaly, a disease in which the brain fails to form two hemispheres. Genes are organized around the outside of the diagram according to the transcriptomic clock revealed by SVD. The stages of development found by FCM are shown by different colors in the inner band.

[17] A demonstration is available at `https://semnext.tw.rpi.edu/chords.html`

Table 1. Summary of dataset mapping coverage. 2-Way coverage indicates number of concepts overlapping in two datasets. N-way coverage indicates number of concepts overlapping in all datasets.

Dataset	Diseases	Genes	Proteins	Protein Interactions Dir.	Protein Interactions Undir.
Cortecon	4,566	16,980	–	–	–
Ensembl	–	23,044	101,436	–	–
KEGG (via Bio2RDF)	1,360	30,680	–	–	–
ReDrugS	3,886	152	76,883	383,752	10,806
StringDB	–	–	22,523	328,514	2,425,314
UMLS	92,984	34,928	–	–	–
Uniprot	4,233	–	145,892	–	82,608
Total	97,937	34,928	145,892	712,266	2,518,728
2-way Coverage	5,128	34,928	101,436	N/A	N/A
N-way Coverage	1,042	152	9,508	N/A	N/A

Chords between genes are PPI extracted from structured sources of protein-coding genes. The width of each gene varies with the number of connections. The diagram's outer ring shows the normalized RNA-Seq counts of each gene as a heatmap for each day of measurement, with day 0 closest to the center and day 77 in the outermost ring.

We chose to use JSON-LD as our primary serialization to exploit the D3 visualization toolkit while maintaining interoperability with semantic web capable agents. The framework publishes JSON-LD contexts for its network representation that makes use of best practice ontologies for linked statistics, including STAT-O, RDF Data Cube, PROV-O, and the Gene Ontology.

8 Evaluation

We evaluated our integration by determining the overlap of concepts across the data sources that SemNExT interfaces with. Table 1 presents a breakdown of the number of concepts identified in each dataset. Queries to match genes and proteins were limited to the species *Homo Sapiens*. Coverage was computed by pairwise selecting datasets containing a shared concept (disease, gene, protein, directed and undirected protein interaction) by using the techniques identified in Section 6 to infer *owl:sameAs* links. Intersections of all datasets sharing a concept were computed in a similar fashion, except that the entity being linked was required to exist in every dataset containing its concept to be counted.

We found that differentiation of classes in different resources was a challenge. For example, Cortecon uses high level concepts such as Disease that in UMLS are decomposed further into concepts such as "Disease or Syndrome," "Congenital Abnormality," and "Pathologic Function." Where appropriate, we attempted to limit linking to relevant classes to improve search and integration responsiveness.

Lastly, edge extraction for the chord diagrams uses any available data source that asserts an edge between two proteins coded by genes. However, systems such as ReDrugS that we incorporate have more robust scoring mechanisms validated by biostatisticians that could further refine appropriate edge choices.

9 Discussion

The combination of statistics and semantics can provide robust error checking and synchronicity between members of a project. While migrating the CHeM visualizations from a Matlab version to the JSON-LD based API written in Python, a project developer noted discrepancies in the values of z-scores between the two implementations.

Listing 1.1. An excerpt of a provenance trace from a SemNExT analysis where z-score values were mismatched between Matlab and Python analyses.

```
@base <https://semnext.tw.rpi.edu/analysis/> .
@prefix sn: <https://semnext.tw.rpi.edu/ontology/semnext#> .
<z-score.py#zscore> a prov:Entity, sn:Analysis;
  prov:wasDerivedFrom <std.py#std>.
<std.py#std> rdfs:comment ```ddof : int, optional
    Means Delta Degrees of Freedom. The divisor used in
        ↪ calculations is N - ddof, where N represents the
        ↪ number of elements. By default ddof is zero.'''@en .
<z-score.m#zscore> a prov:Entity sn:Analysis;
  prov:wasDerivedFrom <std.m#std>.
<std.m#std> rdfs:comment ``By default, the standard deviation
    ↪ is normalized by N-1, where N is the number of
    ↪ observations''@en .
```

Comparing the provenance of the two methods (Listing 1.1) made it clear that z-scores from the Matlab implementation were based on the corrected sample standard deviation[18] whereas the Python implementation z-scores were based on an uncorrected sample standard deviation.[19] This difference in behavior was the result of Matlab's implementation defaulting to a normalization of $N - 1$ whereas NumPy's normalized by a default of N.

9.1 Limitations

Due to the use of manual ontology mappings between data sources, our approach is limited in its direct applicability to other domains without a time investment by knowledge representation experts. However, the mapping we generated contained a limited number ($\sim$100) of axioms so its construction was relatively short (on the order of hours). Larger, more complex domains may face scaling issues unless automated ontology mapping approaches are employed.

[18] `https://www.mathworks.com/help/matlab/ref/std.html`
[19] `http://docs.scipy.org/doc/numpy/reference/generated/numpy.std.html`

Using relational databases is another limitation. Annotators accessing databases must be hand-coded whereas SPARQL-based accesses make use of class hierarchies, class mappings, *owl:sameAs* assertions, and reuse of URIs. We leave exploration of tools to replace hand-coded SQL accesses (e.g., D2R) to future work and note that more resources are becoming available via SPARQL.

10 Conclusions & Future Work

We modeled statistical analyses of genomic data using best-in-class ontologies. These statistical outputs were linked with additional structured resources using linked open data techniques. The resulting RDF graphs were then visualized using a combination of chord diagrams and heat maps. We evaluated our mapping techniques by comparing individual mappings inferred by our API to the total number of individuals across all datasets for a relevant selection of classes. Lastly, we identified some benefits to being able to represent statistical and linked data transformations and their provenance in RDF.

We recommend a number of practices based on our experiences: (*1*) Statistical resources should be modeled with the RDF Data Cube and annotated with appropriate provenance and frameworks should take advantage of these structured resources when available; (*2*) JSON-LD provides a balance between the verbosity of existing semantic web serializations and the succinctness of formats such as CSV to incorporate semantics into web-based visualizations; and (*3*) Capturing of provenance of statistical operators, especially when such operations may have differing default behavior between implementations, is important as it enables replicability and can aid in debugging analyses.

10.1 Future Work

We are looking to expand the number of supported analyses and to apply the SemNExT framework to datasets outside of bioinformatics. SemNExT currently does not automatically reconcile concepts with adjective variations in labels that tend to violate traditional stemming rules, for example "Spinal cancer" in Cortecon compared with "Spine cancer" in UMLS. We intend to investigate resources such as Wordnet as an initial means of resolving these conflicts and will explore other natural language processing techniques as necessary to improve the framework's ability to link resources without identifiers.

We are also interested in modeling the inputs and outputs of statistical analyses using languages such as OWL-S[20]. This will enable linking of relevant analyses at runtime rather than at code design time.

[20] `http://www.w3.org/Submission/OWL-S/`

Acknowledgements

Mr. Patton was supported by an NSF Graduate Research Fellowship and RPI internal funds. Ms. Brown, Ms. H. De Los Santos and Dr. Bennett were supported by NSF Grant 1331023 and RPI internal funds.

References

1. Ausbrooks, R., Buswell, S., Carlisle, D., Dalmas, S., Devitt, S., Diaz, A., Froumentin, M., Hunter, R., Ion, P., Kohlhase, M., et al.: Mathematical markup language (mathml) version 2.0. Tech. rep., World Wide Web Consortium (2003)
2. Bennett, K.P., Brown, E., los Santos, H.D., Boles, N.C., Kiehl, T.R., Patton, E.W., McGuinness, D.L., Temple, S., Fasano, C.A.: Temporal analysis of differentiating pluripotent stem cells using singular value decomposition. In Preparation (nd)
3. Callahan, A., Cruz-Toledo, J., Ansell, P., Dumontier, M.: Bio2RDF release 2: Improved coverage, interoperability and provenance of life science linked data. In: The Semantic Web: Semantics and Big Data, pp. 200–212. Springer (2013)
4. Capadisli, S., Auer, S., Riedl, R.: Linked statistical data analysis. In: Proc. SemStats 2013 (2013)
5. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF data cube vocabulary. Tech. rep., W3C (2014), `http://www.w3.org/TR/vocab-data-cube/`
6. Lange, C.: Ontologies and languages for representing mathematical knowledge on the semantic web. Semantic Web 4(2), 119–158 (2013)
7. Lebo, T., Erickson, J.S., Ding, L., Graves, A., Williams, G.T., DiFranzo, D., Li, X., Michaelis, J., Zheng, J.G., Flores, J., Shangguan, Z., McGuinness, D.L., Hendler, J.: Producing and using linked open government data in the twc logd portal. In: Wood, D. (ed.) Linking Government Data. Springer, New York, NY (2011)
8. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV ontology. Tech. rep., W3C (2013)
9. van de Leemput, J., Boles, N.C., Kiehl, T.R., Corneo, B., Lederman, P., Menon, V., Lee, C., Martinez, R.A., Levi, B.P., Thompson, C.L., et al.: CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. Neuron 83(1), 51–68 (2014)
10. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y.: Scientific workflow management and the kepler system. Concurrency and Computation: Practice and Experience 18(10), 1039–1065 (2006)
11. McCusker, J.P., Yan, R., Solanki, K., Erickson, J.S., Chang, C., Dumontier, M., Dordick, J., McGuinness, D.L.: A nanopublication framework for biological networks using cytoscape.js. In: Proc. 5th ICBO (2014)
12. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad Sci USA 102(43), 15545–15550 (2005)
13. UniProt Consortium, et al.: The universal protein resource (UniProt). Nucleic acids research 36(suppl 1), D190–D195 (2008)
14. Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., Hodge, C.L., Haase, J., Janes, J., Huss, J.W., et al.: BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 10(11), R130 (2009)