Converting neXtProt into Linked Data and nanopublications

Chichester, Christine; Karch, Olivier; Gaudet, Pascale; Lane, Lydie; Mons, Barend; Bairoch, Amos Marc

How to cite

CHICHESTER, Christine et al. Converting neXtProt into Linked Data and nanopublications. In: Semantic web, 2015, vol. 6, n° 2, p. 147–153. doi: 10.3233/SW-140149

This publication URL: https://archive-ouverte.unige.ch//unige:75377
Publication DOI: 10.3233/SW-140149

Converting neXtProt into Linked Data and nanopublications

Editor(s): Jens Lehmann, AKSW, University of Leipzig, Germany; Oscar Corcho, Universidad Politécnica de Madrid, Spain

Solicited review(s): Prateek Jain, IBM Research, USA; Amrapali Zaveri, AKSW, University of Leipzig, Germany; one anonymous reviewer

Christine Chichester[*a1], Oliver Karch[*b], Pascale Gaudet[a], Lydie Lane[a], Barend Mons[c], and Amos Bairoch[a]
[a] *CALIPHO group, Swiss Institute of Bioinformatics, CMU 1 rue Michel Serve, 1211 Geneva 4, Switzerland*
[b] *Biomarker Technologies, Discovery Bioinformatics, Merck Serono, Frankfurter Str. 250, 64271 Darmstadt, Germany*
[c] *Netherlands Bioinformatics Centre, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands*

Abstract. neXtProt provides a comprehensive knowledgebase on human proteins complemented by an extensive cross incorporation of annotations from many databases. With the diversity of published data, provenance information becomes critical to providing reliable and trustworthy services to scientists, thus the tracking of provenance in open, decentralized systems is especially important. Since the nanopublication system addresses many of these challenges, we have developed the neXtProt Linked Data by serializing in RDF/XML annotations specific to neXtProt and started employing the nanopublication model to give appropriate attribution to all data. Specifically, a use case demonstrates the handling of post-translational modification (PTM) data modeled as nanopublications to illustrate how the different levels of provenance and data quality thresholds can be captured in this model.

Keywords: biological database, linked data, semantic web, nanopublication, post-translation modification

[1]Corresponding author. E-mail: christine.chichester@isb-sib.ch
[*]These authors contributed equally to this work.

1. Introduction

With a sustained and ever growing increase in the number of biological databases, the information required for analysis and interpretation of experimental results is frequently scattered over multiple sources. This accentuates the need for methods to integrate and query the many disparate databases. Even though the data are often well structured and the benefits of programmatic access are indisputable, the existence of a specialized API for each data set creates a landscape where significant effort is required to integrate each novel data set into an application. Therefore the shift towards Linked Data sets accelerates the move toward easily accessible open data and facilitates data sharing by establishing links between items in different data sources to create a single global data space. Data providers who offer their information as Linked Data present exciting opportunities for the next generation of Web-based applications: data from different providers can be aggregated and fragmentary information from multiple sources can be integrated to achieve a more comprehensive view of a given field.

One of the advantages of knowledge integration systems is that the system provides a complete yet concise overview of existing data without requiring the end-user to access all data sources separately. Likewise, neXtProt is a protein knowledge platform that aims to support end-user research on human proteins, similar to the role that many Model Organism Databases play for model species [12, 26]. neXtProt facilitates usability by integrating all the manually annotated information from UniProtKB/Swiss-Prot human entries [3] along with highly relevant information curated from other resources as well as a wide range of quality-filtered data from high throughput studies. The fact that neXtProt is restricted to information specific to the realm of human proteins focuses the annotations efforts to allow a greater number of sources for human data to be evaluated. There are several approaches for these information expansions: 1) manual curation from the biomedical literature by experts; 2) automated data extraction from biomedical literature with text mining methods (e.g. GeneRIFs) [32]; 3) computational inference based on sequence annotations, often derived from data in model organisms; and 4) data integration from various experimental or computational sources. Among other things, the content annotated specifically to neXtProt includes: micro-array, cDNA, and protein expression information [27, 30-32, 35], subcellular localization [33, 34], high-quality mass spectrometry-derived proteomics information and, in particular, a number of published sets of N-glycosylation and phosphorylation sites [9], Gene Ontology (GO) [2, 13] annotations [4], and the mapping of proteins to their genomic context [10]. To help users rapidly assess the most reliable data, neXtProt provides quality rankings for different levels of data. “neXtProt Gold” data are of highest quality, according to the biocurator’s judgment. When it is possible to assess the data quality through quantitative criteria, the threshold for inclusion in the Gold category are with error rates estimated at less than 1%. “neXtProt Silver” are good quality data, also according to biocurator’s judgment, and if quantitative criteria can be applied, the threshold is an estimated error rate of less than 5%. Silver data are marked as such in the annotations. All other data are assigned the “neXtProt Bronze” quality rating, and not incorporated.

Contribution to the scientific record via publication of articles is one avenue in which researchers can contribute to the scientific body of knowledge. A second possibility for contribution is the submission to, and curation of, biological databases. Previously, database annotations have not received the appropriate attribution to incentivize their contribution. The nanopublication schema [14] as a novel publication model may serve to alter this trend. The nature of the nanopublication is such that it does not need to be related to a full scientific article (although this is also an option) but it can also refer to a database from which it is derived. Attribution or provenance information present in a nanopublication about who created and published the data, and how it was produced, will also provide several means for quality assessment, an important concern with Linked Data sets. Therefore to improve the value proposition in neXtProt Linked Data, we have provided the established neXtProt quality thresholds,

silver and gold, and evidence codes, which explicitly document the information generation process, modeled as nanopublication provenance [29]. Additionally, using the neXtProt quality assessments, a judgment on the reliability of the information can be presented to the end user, eliminating the requirement that all users have expertise in all fields.

Depositing the neXtProt annotations as nanopublications in a stable and accessible format in open repositories will allow mining for citations for database annotations, as well as scientific assertions contributed by an individual researchers or research groups, and allow for filtering on quality assessments. Such efforts incentivize researchers to submit data to a central repository knowing that they will receive the appropriate attribution, not only due to the nanopublications but also by the citation of the publication describing the data sets.

Using the standard triple model of the Resource Description Framework (RDF) of the Semantic Web, this paper provides a case study description for initial exploration of the neXtProt Linked Data with the following two primary objectives:

1) To show an initial conversion of the neXtProt data to RDF to provide interaction with the potential Semantic Web user community (starting with the Open PHACTS project), and

2) To provide the nanopublication approach for connecting semantics with the sequence positions of post translational modifications (PTMs) present in protein isoforms with the corresponding provenance information.

2. Modeling neXtProt RDF

The Linked Data principles are a set of standards and practices for publishing structured data on the Web [5]. They provide guidelines for uniform data access, uniform syntax, and a uniform data model, but do not address the issues of heterogeneous schemas, duplicate records, and uncertain information quality. We attempted to minimize the duplication of data by focusing on annotations specific to neXtProt and excluding the integrated UniProtKB information, to reduce redundancy with the UniProtKB Linked Data.

Data sources may publish with a mixture terms from widely used vocabularies as well as proprietary terms. It is best practice to use terms from widely deployed vocabularies to represent data wherever possible. The first step in the creation of the neXtProt Linked Data set was to specify exactly what information should be encoded, and represent this ontologically using a variety of vocabularies (Table 1). In this initial conversion of the neXtProt data set, many of the ontology choices followed the UniProtKB Linked Data model [25]. The most significant deviations from the UniProtKB

Prefix:	Namespace URI (http://):	Ontology describes:
biblo	purl.uniprot.org/core/	Bibliographic information
bp	www.biopax.org/	Biological pathways
dcterms	purl.org/dc/terms/	Metadata for resources
eco	purl.obolibrary.org/obo/eco.owl#	Evidence codes
foaf	xmlns.com/foaf/0.1/	People
nextprot	www.nextprot.org/db/entry/	neXtProt data set
np	nanopub.org/nschema#	Nanopublications
owl	www.w3.org/2002/07/owl#	The Web Ontology Language
pav	purl.org/pav/	Provenance, authoring, versioning
prov	www.w3.org/TR/prov-o/#	Provenance
psimod	bioportal.bioontology.org/ontologies/1041?p=terms&conceptid=#	Protein residue modifications
ptm	www.nextprot.org/db/term/#	PTMs
range	sadiframework.org/ontologies/GMOD/RangedSequencePosition.owl#	Sequence features
rdf	www.w3.org/1999/02/22-rdf-syntax-ns#	XML syntax for RDF
rdfs	www.w3.org/2000/01/rdf-schema#	RDF vocabulary description language
sio	semanticscience.org/resource/	Types and relations of objects
skos	www.w3.org/2004/02/skos/core#	Taxonomies and other vocabularies
uni	purl.uniprot.org/uniprot/	UniProt core vocabulary
wi	purl.org/ontology/wi/core#	Preferences (interests) within contexts
xsd	www.w3.org/2001/XMLSchema#	XML

Table 1. Prefixes, URIs, and descriptions of the most common namespaces in the dataset

vocabularies were made in the representation of the citations for publications and the PTM annotations. With the continued subsequent improvements of the neXtProt Linked Data set, it is envisioned to move towards more standards-based ontologies and well-known vocabularies.

The design decisions for the model of neXtProt Linked Data followed the neXtProt schema. Unlike UniProtKB entries, which are based on a master sequence, the data model for neXtProt was developed to present the data corresponding to the specific protein isoform. The data model design represents a tight integration of semantically similar data because all information, for example, the position-specific annotations for functional relationships on the amino acid level and sequence polymorphism information are presented in the context of the specific protein isoform.

The neXtProt data set can be applied in a myriad of different scenarios in the biological domain, particularly in respect to resolving queries concerning protein isoforms, granularity that is missing from other Linked Data resources. A sampling of the data that can be retrieved include lists of proteins or protein isoforms with specific characteristics such as: secreted proteins with length greater than 100 amino acids but not containing cysteines; proteins highly expressed in brain but not in testis; proteins located in mitochondrion and lacking a transit peptide; proteins expressed in liver and involved in transport; proteins having a protein kinase domain but no kinase activity; proteins acetylated and methylated and located in the nucleus; proteins interacting with more than 50 other proteins and not involved in a disease; etc.

3. neXtProt Linked Data production

The neXtProt XML [11] was used as the basis for the generation of the Linked Data. For publishing the data set, different serialization formats have been used. NeXtProt isoform entries leverage UniProtKB's RDF/XML format whereas TriG [6] format is used to represent the named graphs of nanopublications.

The current multi-stage conversion procedure (Figure 1) is using a template-driven approach, which allows the data model to gradually evolve in a modular and flexible fashion. This will accommodate the gradual replacement of UniProtKB vocabularies and inclusion of additional nanopublications. The implementation of the template assembly tool is realized as a light wrapper utilizing Apache Velocity template engine [22] at its core and Velocifero [23] for object-relational mapping. By separating access to the data-model from templates driving the transformation process the tool also supports conversion of datasets other than neXtProt. It is available under an open-source license from [19].

The entire procedure is controled by a continuous integration system (Jenkins/Hudson [18]) which automatically downloads the neXtProt XML data from the ftp-site [11]. As new neXtProt releases become available, they are converted and deposited on the download site. In the first stage, the XML data is transformed into a relational data model utilizing a parallelized processing pipeline. The relational data model facilitates efficient querying and traversal over the large neXtProt dataset.

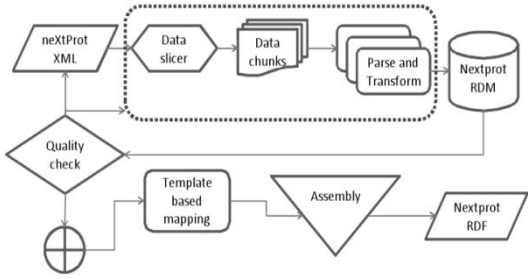

Fig. 1: Flow chart of the conversion procedure initially producing a neXtProt relational data model (RDM) which is subsequently being queryied to generate the RDF.

In the next stage, the templates driving the assembly process query the relational data-model to produce RDF. The data model mapper (*$db* reference in Fig. 2) utilizes named queries to decouple access to the model from content creation enabling seamless exchange of the underlying query and storage engine. The templates draw on a library of macros (Fig 2), which dynamically load and render sub-templates at

runtime, which simplifies the production of RDF modeled with different attributes.

The current RDF Linked Data set contains approximately 97 million triples attributed to over 38,000 protein isoform entries. The RDF dump is openly and freely available [16].

To enlighten the data consumer to the details of the neXtProt data set, we have published a Vocabulary of Interlinked Datasets (VoID) [1] file along with the full data set. The VoID provides basic metadata about neXtProt, including the licensing, using the Dublin Core Metadata (dcterms) [17] and Provenance, Authoring and Versioning (pav) [20] vocabularies. The metadata with details about the location of the data dumps and the structural metadata concerning the data modeling are described using VoID and Data Catalog Vocabulary (dcat) [24].

As well as offering an open freely accessible data set for download under the Creative Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license, the complete neXtProt data set is envisaged to be integrated into the Open PHACTS Discovery platform [36], which maintains a sustainable infrastructure with open APIs to facilitate querying various biomedical datasets relevant for drug discovery.

```
#macro( nanopubs $npid $isoid )
  #set( $nanoCont = "" )
  #set( $annos = $db.getWithParams( 'nanopub_ptm',
    { 'npid' : $npid, 'isoid' : $isoid } ) )
  #set( $uniacc = $npid.substring(3) )
  #set( $isoacc = $isoid )
[…]
  #foreach( $anno in $annos )
    #set( $resLink =
      "${anno.isoid}_${anno.accession}_${anno.pos_start}")
  <uni:annotation rdf:resource="http://[...]/$resLink"/>
    #set( $gRead = $render.eval( $nano_ptm_trig_vm ) )
    #set( $nanoCont = "$nanoCont$gRead" )
  #end
#end
```

```
   […]
:${resLink} np:hasAssertion :${resLink}_assertion ;
    np:hasProvenance :${resLink}_provenance ;
    np:hasPublicationInfo :${resLink}_publicationInfo .
:${resLink}_assertion a np:Assertion ;
    rdfs:comment "$anno.anno_desc modification
        of amino acid $anno.pos_start of
        $anno.isoid"^^xsd:string .
:${resLink}_assertion { :${resLink}_PTM-Modif
    bp:participant nx:$anno.isoid, ptm:$anno.accession ;
      sio:SIO_000008 [ a range:RangedSequencePosition ;
      a so:SO_0001089 ;
      sio:SIO_000053 [ a range:StartPosition;
        sio:SIO_000300 $anno.pos_start ] ;
      sio:SIO_000053 [ a range:EndPosition; sio:SIO_000300
        $anno.pos_end ] ; ] ;
    a <http://www.variationont[...]?term=VariO:0024> ;
    a <http://id.loc.gov/[...]/sh88005292>.
    ptm:$anno.accession
      owl:sameAs <http://identifiers[...]/MOD:00048> . }
[…]
```

Fig. 2 Upper part shows an excerpt from a velocity macro producing the PTM nanopublications for a given protein isoform. Velocity **directives** are in bold, *variables and functional objects* (e.g. data model mapper *$db*) are italic. The lower part presents a passage from a dynamically loaded template to render an individual nanopublication in RDF TriG format.

4. Modeling of neXtProt PTM nanopublications

A nanopublication is the smallest publishable unit that adequately expresses a single assertion and its provenance. The nanopublication schema (http://nanopub.org/nschema) is composed of 3 named graphs: assertion, provenance, and publication information (Figure 3). Named graphs [6] are models that allow entire groups of triples to be assigned a URI. The assertion graph is composed of one or several triples that state the scientific assertion. The provenance graph contains statements supporting the assertion, such as the methods used to create the assertion, as well as attribution for the assertion. The publication information (publicationInfo) graph contains provenance statements referring to the creation of the nanopublication itself. The complete neXtProt Linked Data set includes over 46 million nanopublication triples.

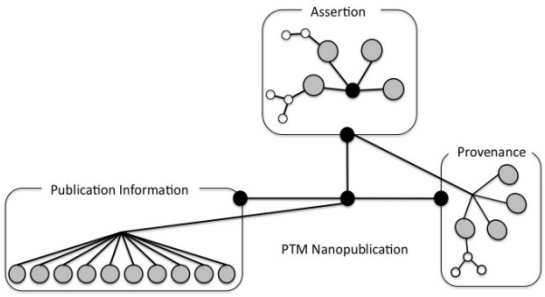

Fig. 3: Schematic of the PTM Nanopublication. The nanopublication consists of 3 named graphs: assertion, provenance, and publication information (publicationInfo). The filled black circles represent URIs internal to the nanopublication, which are uniquely minted in the scope of each graph of the nanopublication and nanopublication assertion. The grey and white circles represent the other entities within the PTM nanopublication model.

```
:NX_P35916-2_PTM-0255_1063 a np:Nanopublication.
:NX_P35916-2_PTM-0255_1063 np:hasAssertion :NX_P35916-2_PTM-0255_1063_assertion  ;
          np:hasProvenance :NX_P35916-2_PTM-0255_1063_provenance ;
          np:hasPublicationInfo :NX_P35916-2_PTM-0255_1063_publicationInfo .

:NX_P35916-2_PTM-0255_1063_assertion a np:Assertion .
:NX_P35916-2_PTM-0255_1063_provenance a np:Provenance .
:NX_P35916-2_PTM-0255_1063_publicationInfo a np:PublicationInfo .

:NX_P35916-2_PTM-0255_1063_assertion {:NX_P35916-2_PTM-0255_1063_PTM-Modif bp:participant nextprot:NX_P35916-2, ptm:PTM-0255 ;
    sio:SIO_000008 [ a range:RangedSequencePosition ;
          sio:SIO_000053 [ a range:StartPosition; sio:SIO_000300 1063 ] ;
          sio:SIO_000053 [ a range:EndPosition; sio:SIO_000300 1063 ] ;
          ] ;
    rdfs:comment "Phosphotyrosine modification of amino acid 1063 of Vascular endothelial growth factor receptor 3 isoform 2
(Homo sapiens)."^^xsd:string ;
    a bp:MolecularInteraction .
    ptm:PTM-0255 owl:sameAs psimod:MOD:00048 .
}
:NX_P35916-2_PTM-0255_1063_provenance {:NX_P35916-2_PTM-0255_1063_assertion prov:wasDerivedFrom uniprot:P35916 ;
          prov:wasGeneratedBy eco:ECO:0000218 ;
          prov:PrimarySource <http://www.ncbi.nlm.nih.gov/pubmed/18083107> ;
          wi:evidence <http://www.conceptwiki.org/concept/UUID-neXtProt Gold> .
          <http://www.conceptwiki.org/concept/neXtProt Gold> rdf:type eco:ECO:000205 ;
          rdfs:label"neXtProt Gold"^^xsd:string  .
}
:NX_P35916-2_PTM-0255_1063_publicationInfo {:NX_P35916-2_PTM-0255_1063 prov:wasGeneratedBy eco:ECO:0000248 ;

....}
```

Fig 4. RDF Turtle notation of an example nanopublication. The URI namespace prefixes are given in Table 1. The publicationInfo graph notation has been truncated and the bp:dataSource, pav:versionNumber, pav:authoredBy, pav:createdBy, dcterms:created, dcterms:rights, and dcterms:rightsHolder triples are not shown.

For the neXtProt nanopublication, the assertion is the minimal unit of information to describe the PTM for a specific protein isoform. It states that there is a modification on one specific amino acid in a specific protein isoform with a specific type of post-translational modification (Figure 4). The BioPAX ontology [8] was used, which provides an information model for representing those data with formally defined semantics, including the possibility to explicitly model the relationships between the different entities in the interaction, namely the protein isoform and the PTM. Semantic science Integrated Ontology (SIO), a simple, integrated ontology (types, relations) for diverse knowledge representation across physical, procedural, and informational entities [15], was suited for bridging the gap between sequence positional claims and their physical or biological interpretation. The provenance includes attribution information pertaining to the assertion, such as the journal article as the primary source and UniProtKB as the first stage of annotation. The provenance graph includes the curated quality level associated with the assertion via the Weighted Interests Vocabulary [21] which provides a link between the assertion and the quality assessment: neXtProt gold, or silver. URIs for neXtProt gold and silver concepts were generated by adding these concepts to the ConceptWiki [7]. The ConceptWiki is a system that allows the minting of URLs that can be used as resources for concepts that are not included in any formal ontology. The publication information for the nanopublication includes the contributing authors, when it was made by a date/time stamp, and how the nanopublication can be (re)used (copyright information). Given that the nanopublication attribution could be linked to each researcher through their unique digital identifier, (e.g. ORCID identifier, researcher ID, etc) could provide the basis for the necessary metrics for measuring each scientific nanopublication contribution.

5. Conclusions

Nanopublications encoded in RDF can be more easily mined, queried, and retrieved through web services and be subject to computer reasoning, something that is not feasible for regular articles and many annotations from databases. Importantly, nanopublications may encourage potential data contributors to place their data in the public domain for other users to freely access and optimally exploit,

since nanopublications can be attributed and cited in the same way as articles published traditionally [28]. Moreover, particular attention can be given to indicate the level of data confidence in a clear way, for example, with the gold and silver ratings attributed in neXtProt since the provenance elements of nanopublication enable these elements. The eventual crowdsourcing of bronze level annotations with a simple "endorse or deny" choice could be a simple first process to involve more life scientists in contributing knowledge to databases and receiving credit for their nanopublications. Finally, the PTM assertions present in the neXtProt nanopublications can allow the generation of graphic output in the form of maps or images as the result of queries based on semantics.

In a distributed and collaborative environment like the current World Wide Web, there can be a lot of redundancy across data sources. While redundancy increases noisy or unnecessary information, it can also be an advantage, in the sense that two descriptions of the same thing can mutually complete and complement each other. To avoid redundancies and to support user applications, we have purposely attempted to incorporate in the Linked Data set only data that is specific to neXtProt database while reducing the overlap with the UniProtKB Linked Data. In cases where neXtProt shares common concepts with other databases the ontology choices followed the UniProtKB Linked Data model.

As we proceed with the future developments of the neXtProt Linked Data, which includes the conversion to more widely used attributes, the possibility of using nanopublications will be considered for all relevant neXtProt specific scientific assertions.

Acknowledgments

We greatly appreciate the comments by Paul Groth and Antonis Loizou on the Linked Data and thenanopublications. This work was supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement n° 115191 for Open PHACTS, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

References

[1] K. Alexander, R. Cyganiak, M. Hausenblas, J, Zhao. Describing Linked Datasets - on the design and usage of void, the 'vocabulary of interlinked datasets. In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (2009).

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology, The Gene Ontology Consortium. Nat. Genet., 25 (2000), 25–29.

[3] A. Bairoch, L. Bougueleret, S. Altairac, V. Amendolia, A. Auchincloss, G. Argoud-Puy, K. Axelsen, D. Baratin, M. Blatter, B, Boeckmann, et al., The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res. 37 (2009), D169-74.

[4] D. Barrell, E. Dimmer, R.P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, The GOA database in 2009–an integrated Gene Ontology Annotation resource, Nucleic Acids Res., 37 (2009), D396–D403.

[5] T. Berners-Lee, Linked Data. World wide web design issues. http://www.w3.org/DesignIssues/LinkedData.html (2006).

[6] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, Named graphs, provenance and trust. In: Proceedings of the 14th international conference on World Wide Web, WWW '05, ACM, New York, NY, USA (2005), 613–622.

[7] C. Chichester and B. Mons, Chapter 26: Collaboration and the Semantic Web, in: Collaborative Computational Technologies for Biomedical Research, S.Ekins, M. Hupcey, A. Williams, eds., John Wiley and Sons, (2011), 453-466.

[8] E. Demir, M. P. Cary, S. Paley, K. Fukuda, C. Lemer et al., The BioPAX community standard for pathway data sharing. Nat. Biotechnol., 28, 935–942 (2010)

[9] T. Farrah, E.W. Deutsch, G.S. Omenn, D.S. Campbell, Z. Sun, J.A. Bletz, P. Mallick, J.E. Katz, J. Malmstrom, R. Ossola, et al., A high-confidence human plasma proteome reference set with estimated concentrations in PeptideAtlas. Mol. Cell. Proteomics, 10 (2011), M110.006353.

[10] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, et al., Ensembl 2011. Nucleic Acids Res., 39 (2011), D800–D806.

[11] ftp://ftp.nextprot.org/pub/current_release/xml/

[12] P. Gaudet, G. Argoud-Puy, I. Cusin, P.Duek, O. Evalet, A. Gateau, A. Gleizes, M. Pereira, M. Zahn-Zabal, C. Zwahlen, et al., neXtProt: organizing protein knowledge in the context of human proteome projects. J Proteome Res., 12 (2013), 293-8.

[13] Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res., 38 (2010), D331–D335

[14] P. Groth, A, Gibson, and J. Velterop. The anatomy of a nanopublication, Information Services & Use 30 (2010) 51-56.

[15] http://code.google.com/p/semanticscience/wiki/SIO

[16] http://downloads.nbiceng.net/nextprot/currentrelease/data/

[17] http://dublincore.org/documents/2010/10/11/dcmi-terms/

[18] http://jenkins-ci.org/
[19] http://nextprot2rdf.sf.net
[20] http://purl.org/pav/2.1
[21] http://smiy.sourceforge.net/wi/spec/weightedinterests.html
[22] http://velocity.apache.org
[23] http://velosurf.sourceforge.net
[24] http://www.w3.org/TR/2013/WD-vocab-dcat-20130801/
[25] E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, E.B. Suzek, M.J.Martin, P. McGarvey, E. Gasteiger. Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinform, 10, 136 (2009).
[26] L. Lane, G. Argoud-Puy, A. Britan, I. Cusin, P. Duek, O. Evalet, A. Gateau, P. Gaudet, A. Gleizes, A. Masselot, et al., neXtProt: a knowledge platform for human proteins Nucleic Acids Res. 40 (2012), D76-83.
[27] U. Liebel, V. Starkuviene, H. Erfle, J.C. Simpson, A. Poustka, S. Wiemann, and R. Pepperkok, A microscope-based screening platform for large-scale functional protein analysis in intact cells. FEBS Lett., 554 (2003), 394–398.
[28] B. Mons, H. Haagen, C. Chichester, P.B.t'Hoen, J. Dunnen, G. Ommen, E. Mulligen, B. Singh, R. Hooft, M. Roos, et al., The value of data. Nat. Genetics, 43 (2011), 281-283.
[29] B. Mons and J. Velterop, Nano-publication in the e-science era. Eds: T. Clark, J. S. Luciano, M. S. Marshall, E. Prud'hommeaux, S. Stephens In: Workshop on Semantic Web Applications in Scientific Discourse, Washington DC, USA (2009), http://ceur-ws.org/Vol-523/.
[30] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, et al., ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res., 39 (2011), D1002–D1004.
[31] J.U. Pontius, L. Wagner, and G.C. Schuler, Ch. 21. In: McEntyre,J. and Ostell,J. (eds), The NCBI Handbook. National Center for Biotechnology Information, Bethesda, MD, 2003.
[32] E.W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, S. Federhen, et al., Database resources of the National Center for Biotechnology Information. Nucleic Acids Res., 39 (2011), D38–D51.
[33] A. Sigal, T. Danon, A. Cohen, R. Milo, N. Geva-Zatorsky, G. Lustig, Y. Liron, U. Alon, and N. Perzov, Generation of a fluorescently labeled endogenous protein library in living human cells. Nat. Protocols, 2 (2007), 1515–1527.
[34] J.C. Simpson, R. Wellenreuther, A. Poustka, R. Pepperkok, and S. Wiemann, Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. EMBO Rep., 1 (2000), 287–292.
[35] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, et al., Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol., 28 (2010), 1248–1250.
[36] A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, and B. Mons. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today, 17:1188–1198, 2012.