Nanopublications for exposing experimental data in the life-sciences: a Huntington's Disease case study

Eleni Mina[1], Mark Thompson[1], Rajaram Kaliyaperumal[1], Jun Zhao[2], Zuotian Tatum[1], Kristina Hettne[1], Erik A. Schultes[1], and Marco Roos[1]

[1] Human Genetics Department, Leiden University Medical Center, NL
{e.mina,m.thompson,r.kaliyaperumal,z.tatum,k.m.hettne,e.a.schultes,m.roos}@lumc.nl
[2] Department of Zoology, University of Oxford, Oxford, UK jun.zhao@zoo.ox.ac.uk

Abstract. Data from high throughput experiments often produce far more results than can ever appear in the main text or tables of a single research article. In these cases, the majority of new associations is often archived either as supplemental information in an arbitrary format or in publisher-independent databases that can be difficult to find. These data are not only lost from scientific discourse, but are also elusive to automated search, retrieval and processing. Here, we use the nanopublication model to make scientific assertions that were concluded from a workflow analysis of Huntington's Disease data machine-readable, interoperable, and citable. We followed the nanopublication guidelines to semantically model our assertions as well as their provenance metadata and authorship. We demonstrate interoperability by linking nanopublication provenance to the Research Object model. These results indicate that nanopublications can provide an incentive for researchers to expose mass data that is interoperable and machine-readable.

Keywords: Huntington's disease, nanopublication, provenance, research object, workflows, interoperability, data integration

1 Introduction

The large amount of scientific literature in the field of biomedical sciences makes it impossible to manually access and extract all relevant information for a particular study. This problem is mitigated somewhat by text mining techniques on scientific literature and the availability of public online databases containing (supplemental) data. However, many problems remain with respect to the availability, persistence and interpretation of the essential knowledge and data of a study.

Text mining techniques allow scientists to mine relations from vast amounts of abstracts and extract explicitly defined information [1] or even implicit information [2], [3]. Because most of these techniques are limited to mining abstracts, it is reasonable to assume that information such as tables, figures and supplementary information are overlooked. Moreover, recent attempts to mine literature

for mutations stored in databases, showed that there was a very low coverage of mutations described in full text and supplemental information [4]. For our case study, we found that the association that we inferred between the HTT gene, which mutant form causes Huntington's Disease, and BAIAP2, a brain-specific angiogenesis inhibitor (BAI1)-binding protein, was present in a table in a paper by Kaltenbach *et al.* [5]. However, it is not explicitly in any abstract which makes it hard to retrieve from systems such as PubMed. This means that valuable findings and intermediate results become lost and are no longer available to the scientific community for further use.

This is partly remedied by making data public via online databases. However, this by itself does not guarantee that data can be readily found, understood and used in computational experiments. This is particularly problematic at a time when more, and larger, datasets are produced that will never be fully published in traditional journals. Moreover, there is no well-defined standard for scientists to get credit for the curation effort that is typically required to make a discovery and its supporting experimental data available in an online database. We argue that attribution and provenance are important to ensure trust in the findings and interpretations that scientists make public. Additionally, a sufficiently detailed level of attribution provides an incentive for scientists, curators and technicians to make experimental data available in an interoperable and re-usable way. The Nanopublication data model [6] was proposed to take all these issues into consideration. Based on Semantic-web technology, the nanopublication model is a minimal model for publishing an assertion, together with attribution and provenance metadata.

In this paper we present a case study that involves an analysis based on scientific workflows to help explain gene expression deregulation in Huntington's Disease (HD) (E. Mina *et al.*, manuscript in preparation). We show how the results of this case study can be represented as nanopublications and how this promotes data integration and interoperability.

The remainder of this paper is organized as follows. Section 2 gives the background of the Huntington Disease case study. Section 3 explains the nanopub model for our experimental data. Section 4 demonstrates the potential for data integration by means of SPARQL queries. In Section 5 we discuss how the proposed model works as a template for similar datasets and how we can further improve integration in the future. Section 6 concludes the paper.

2 The Huntington's Disease case study

Huntington's Disease is a dominantly inherited neurodegenerative disease that affects 1/10.000 individuals of European origin and thus making it the most common inherited neurodegenerative disorder [7]. The genetic cause for HD was already identified in 1993, but no cure has yet been found and the exact mechanisms that lead to the HD phenotype are still not well known. Gene expression studies revealed massive changes in HD brain that take place even before first symptoms arise [8]. Our experiment consists of a computational approach to

investigate the relation between HD deregulated genes and particular genomic regions.

There is evidence for altered chromatin conformation in HD [9] and therefore we chose to work with two genomic datasets that are associated with epigenetic regulation, concerning CpG islands in the human genome [10] and chromatin marks mapped across nine cell types [11]. Identifying genes that are associated with these regions and gene deregulation in HD can give insight into chromatin-associated mechanisms that are potentially at play in this disease. We implemented our analysis as a workflow using the Taverna workflow management system [12, 13]. As input we used gene expression data from three different brain regions from normal and HD-affected individuals [14]. We tested for gene differential expression (DE) between controls and HD samples in the most highly affected brain region, caudate nucleus, and we integrated this data with the two epigenetic datasets that are publicly available via the genome browser [15].

We decided to model and expose as nanopublications two assertions from the results of our workflow: 1) deregulated genes in HD that we therefore associate with the disease and 2) genes that overlap with a particular genomic region. Note that these natural language statements would typically be used in a caption for a figure, table or supplemental information section to describe a dataset in a traditional publication. Considering the problems with automatic retrieval and interpretation of such data, we aim to expose these assertions in a way that is more useful to other scientists (for example to integrate our results with their own data). Moreover, we have to give provenance containing the origin and experimental context for the data in order to increase trust and confidence. The next section shows in detail how we do this with nanopublications.

3 The nanopublication model

The nanopublication guidelines document [6] provides details of the nanopublication schema and recommendations for constructing nanopublications from Life Science data. The schema defines that a nanopublication consists of three parts, each implemented as a RDF model in a named graph: assertion, provenance and publication information. The assertion graph contains the central statement that the author considers valuable (publishable) and for which she would like to be cited (attribution). It should be kept as small as possible in accordance with the guidelines. The provenance graph is used to provide evidence for the assertion. It is up to the author to decide how much provenance information to give, but in general, more provenance will increase the trustworthiness of the assertion, and thus the value of the nanopublication. The publication info graph provides detailed information about the nanopublication itself: creation date, licenses, authors and other contributors can be listed there. Attribution to curators and data modelers are part of the nanopublication design to incentivize data publishing. Our nanopublications are stored in the AllegroGraph triple store [16] for which the SPARQL endpoint and browsable user interface is available here:
http://agraph.biosemantics.org/catalogs/ops/repositories/HD_GDE_genomic_overlap

Logging in with username "test" and password "tester" will show the queries used in this paper under the menu "Queries → Saved".

3.1 Assertion

As authors of this nanopublication, we wish to convert the following natural language statements to RDF: "gene X is associated with HD, because it was found to be deregulated in HD" and "gene Y is associated with a promoter, and this promoter overlaps with a CpG island and/or a particular chromatin state", and we wish to refer to the experiment by which we found these associations. We decided to model our results into two nanopublications. By further subdividing those statements, we see the RDF triple relations appear naturally:

Nanopublication assertion 1:

1. There is a gene disease association that *refers_to* gene X and Huntington's Disease

Nanopublication assertion 2:

1. Gene Y is *associated_with* promoter Z
2. Promoter Z *overlaps_with* a biological region [3]

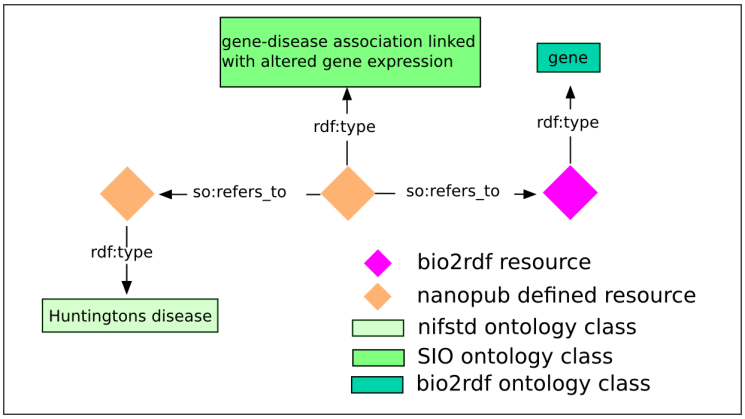

Fig. 1: Differential gene expression assertion model

The assertion templates for our models are shown in Figure 1 and Figure 2. For some of the terms in these statements we found several ontologies that defined classes for them. For example, "promoter", "gene", and "CpG island"

[3] in our case the biological region is: CpG island or one of the chromatin states, active /weak /poised promoter or heterochromatic (see Figures 1,2)

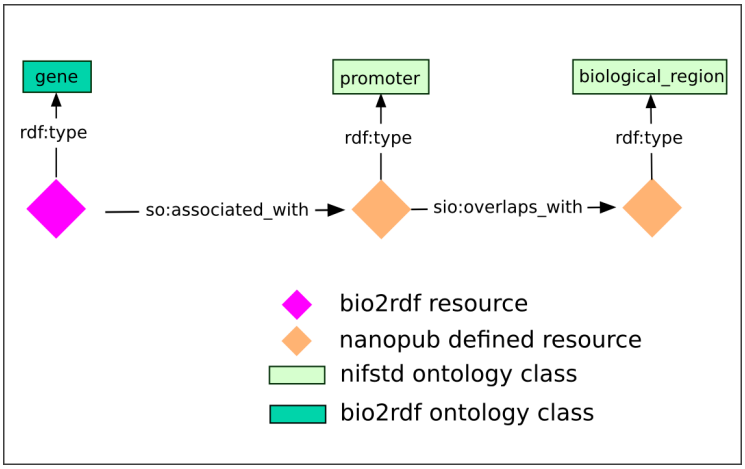

Fig. 2: Genomic overlap assertion model. Orange diamonds refer to a RDF resource that was defined by this nanopublication, whereas the gene (pink diamond) is defined by a bio2rdf resource. The Semanticscience Integrated Ontology (SIO) and Sequence Ontology (SO) were used for the predicates overlaps_with and associated_with.

appear (among others) in the following ontologies: NIF Standard ontology (NIFSTD), NCI Thesaurus (NCI) and the Gene Regulation Ontology (GRO)[4]. We chose to use NIFSTD for HD, because it covers an appropriate domain and it uses the Basic Formal Ontology (BFO), which can benefit data interoperability and OWL reasoning (e.g. for checking inconsistencies).

We chose to use bio2rdf instances for the associated genes [17] because they provide RDF with resolvable resource URIs for many different biomedical resources. To describe the gene-disease association linked with altered gene expression we used the class with that label from the SemanticScience Integrated Ontology (SIO) [18]. The SIO predicate "refers to" was used to associate each differentially expressed gene with HD. There were also terms that we did not find in an available ontology. These were the ones that described the type of the chromatin state that a promoter of a gene can be in, "active promoter state", "weak promoter state", "poised promoter state" and "heterochromatic". We decided to create our own classes to describe these terms. Being aware of interoperability issues, we defined them as subtypes of classes in the Sequence Ontology (SO). We defined the class "chromatin_region" as a subclass of "biological_region" in SO. We defined another class "chromatin_state" as a subclass of "feature_attribute". Subclasses of "chromatin_state" are the states "active_promoter", "weak_promoter", "poised_promoter" and "heterochromatic", Figure 3. The definition for the classes we created is presented in Table 1. We

[4] All ontologies mentioned in this paper are available through `http://bioportal.bioontology.org/ontologies`

also defined an object property "has_state" which has domain chromatin and range "chromatin_state".

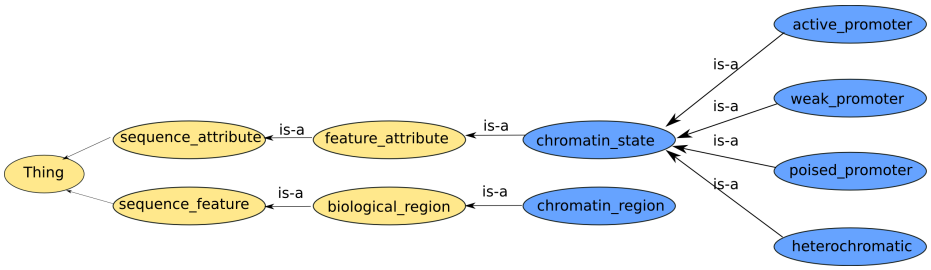

Fig. 3: Schematic representation of the extension of SO with our own defined classes. In yellow are depicted the SO ontology classes and in blue the classes we defined in our case study.

Table 1: Definition of new classes

chromatin_region	A region of chromatin, likely to be involved in a biological process
chromatin_state	Annotation of chromatin states, defined by combinations of chromatin modification patterns (described in publication by Ernst et al. Nature, 2011)
active_promoter	Open chromatin region, associated with promoters, transcriptionally active, defined by the most highly observed chromatin marks : H3K4me2,H3K4me3, H3K27ac, H3K9ac
weak_promoter	Open chromatin region, associated with promoters, weak transcription activity, defined by the most highly observed chromatin marks : H3K4me1, H3K4me2,H3K4me3, H3K9ac
poised_promoter	Open chromatin region, associated with promoters, described as a bivalent domain that has strong signals of both active and inactive histone marks. Most highly observed histone marks: H3K27me3, H3K4me2, H3K4me3
heterochromatic	Closed chromatin formation, transcriptionally inactive. It is associated with none histone marks

For the predicates we considered the use of the Relation Ontology (RO), because its use of BFO. However, we found that the OWL domain and range specifications did not match our statements. Instead of extending RO with the appropriate predicates, which could be better for interoperability and reasoning in the long term, we decided to use predicates from the also popular Sequence Ontology (SO) and Semanticscience Integrated Ontology (SIO) [18] that also seemed appropriate for our assertions. This is a typical trade-off between quality and effort that we expect nanopublishers will have to make frequently. We can

justify this for two reasons: 1) releasing experimental data as linked open data using any standard ontology is already a huge step forward from current practice and 2) interoperability issues at the ontology level is a shared responsibility with ontology developers and curators who provide mappings between ontologies and with higher level ontologies.

3.2 Provenance

In the provenance section of the nanopublication we would like to capture as accurately as possible where the assertion came from and what the conditions of our experiment were. In our case the experiment is in-silico: a workflow process that combines existing data sources to expose new associations. Details and references to the original datasets, the workflow process itself and the final workflow output are interesting provenance as they increase trust in the assertion and make it possible to trace back the results of the experiment.

An extra benefit of using workflows is that provenance information can be automatically generated by the workflow system and additional tools can be used to associate a workflow with additional metadata and resources. We used Taverna to build and execute our workflows [12]. Taverna provides an option to export the provenance of a workflow execution in prov-o [19]. On top of this, models, tools, and guidelines are being developed for bundling workflows with additional resources in the form of workflow-centric Research Objects (ROs) [20]. Additional resources may include documents, input and output data, annotations, provenance traces of past executions of the workflow, and so on. ROs enable in silico experiments to be preserved, such that peers can evaluate the method that led to certain results, and a method can be more easily reproduced and reused. Similar to nanopublications, the RO model is grounded in Semantic Web technologies [21]. It is comprised by a core ontology and extension ontologies. The core ontology reuses the Annotation Ontology (AO) and the Object Reuse and Exchange (ORE) model to provide annotation and aggregation of the resources. The extension ontologies keep track of the results and methods of an experiment (wfprov), provide the descriptions of scientific workflows (wfdesc) and capture the RO evolution process (roevo) [22]. ROs extend the already existing functionality of myExperiment packs. We created ROs using the RO repository sandbox, which offers a user friendly interface for creating ROs either by importing an already existing pack from myExperiment, or uploading a .zip archive or creating a research object manually [23].

An overview of the connection between the Nanopublication model and the RO is given in Figure 4. In the nanopublication provenance graph we include a simple provenance model that describes the context of the workflow process: in particular the relation of the nanopublication assertions as the origin of the experiment outputs. Note that the workflow activity links to the RO and each of the input/output entities link to the corresponding entity in the RO. This way, the nanopublication provenance serves as a proxy for the RO, such that larger nanopublication collections can be queried without downloading all ROs. Moreover, we increase interoperability by using the standard Prov-o ontology in

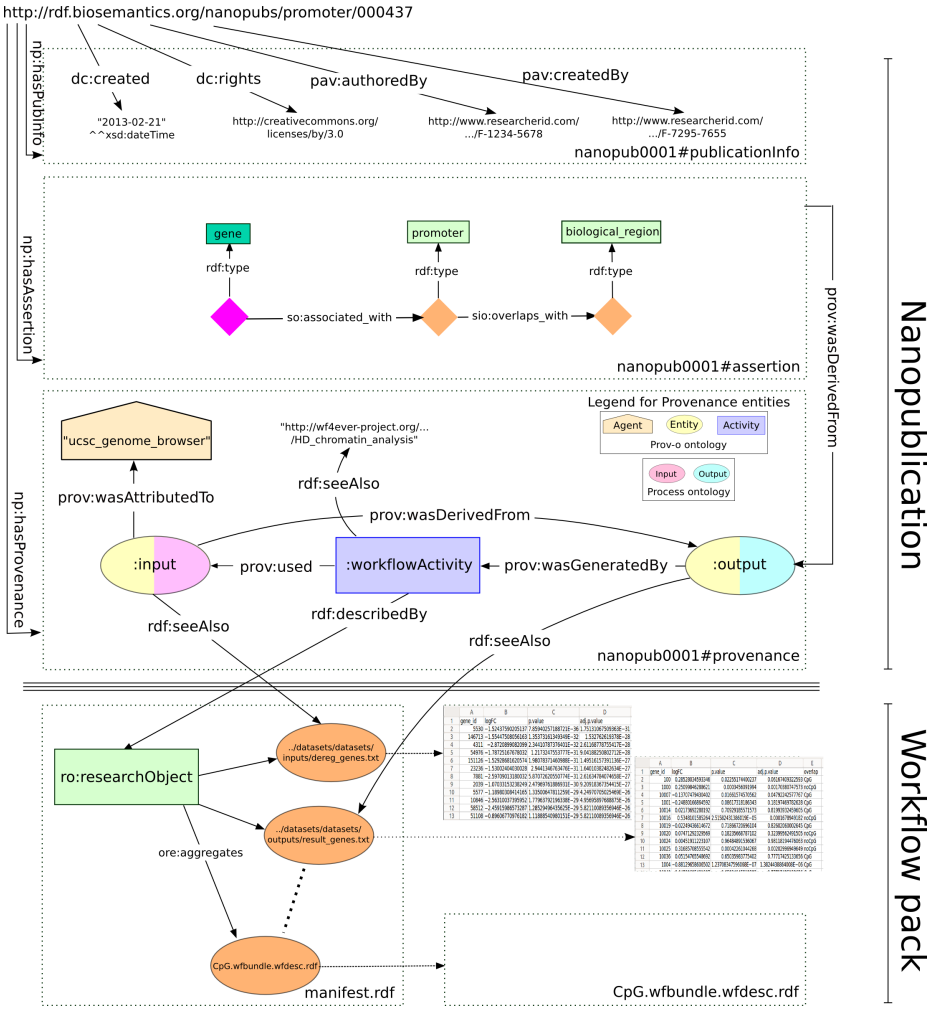

Fig. 4: Connecting nanopublication (top) and ResearchObject (bottom)

the nanopublication provenance, to which the RO ontology is aligned. Futhermore we increase the semantics of the input/output entities by using the domain specific Process ontology [24]. In summary, the strength of linking the entities in a nanopublication provenance to a RO, is to augment the experimental context information which is key evidence for the statement made in the assertion.

3.3 Publication information

In this section we capture details that is required for citation and usage of the nanopublication itself. The authors of the nanopublication and possible contributors are described here, and represented by a unique research identifier to

account for author ambiguity. The timestamp of the nanopublications creation is also recorded in this part, as well as versioning details. Finally, information about usage rights and licence holders is included.

4 Data integration

Because the nature of biological data is by default complex, data interoperability is a challenge. The choice of Semantic Web standards for the implementation of Nanopublications facilitates interoperability. In HD research, diverse working groups recruit a variety of disciplines that produce data encompassing brain images, gene expression profiles in brain and blood, genetic variation, epigenome data, etcetera, with the common goal to identify biomarkers to develop effective treatment to slow down disease progression. Nanopublications provide an incentive to expose this data such that we can more easily combine them with each other. Following standardized templates to model information ensures data interoperability that can facilitate complex queries for discovering new information. In addition, the attached provenance information will give neccessary information related to the experiment, to ensure trust but also to be able to reuse the scientific protocol and replicate the results. We applied simple sparql queries to our set of nanopublications to demonstrate how data integration with nanopublications can occur in practice. These canned queries are stored in our nanopublication store and the user can browse and execute them using the login account mentioned previously in this paper.

Storing data in RDF format provides an easy way for data integration because of the use of same URIs, something that is not guaranteed in relational databases. This saves a lot of effort from taking the time to understand and map external data sources in order to join information and retrieve the results that are related to our query. To check and reassure the interoperability of our nanopublications we did a simple integration query with the bio2rdf resource (see Figure 5 and Figure 6). Our goal was to query for all drug targets that are associated with a set of genes that is differentially expressed in HD and overlaps with CpG islands. Using the gene_ID of our set of genes in the bio2rdf endpoint, we can list all predicates associated with those gene ids and retrieve the gene nomenclature for each of them. This can be then used to query drugbank and retrieve drugtargets and drugnames.

5 Discussion

In this paper we presented an example case study for exposing Life Science data into machine readable associations, along with their provenance metadata and publication information. The process of nanopublication modeling is a one-time effort, which can be greatly simplified by using examples or *templates*. The nanopublication models presented in this paper can serve as templates to expose similar assertions. For example, the investigation of gene differential expression under specific conditions is a very common analysis and those results

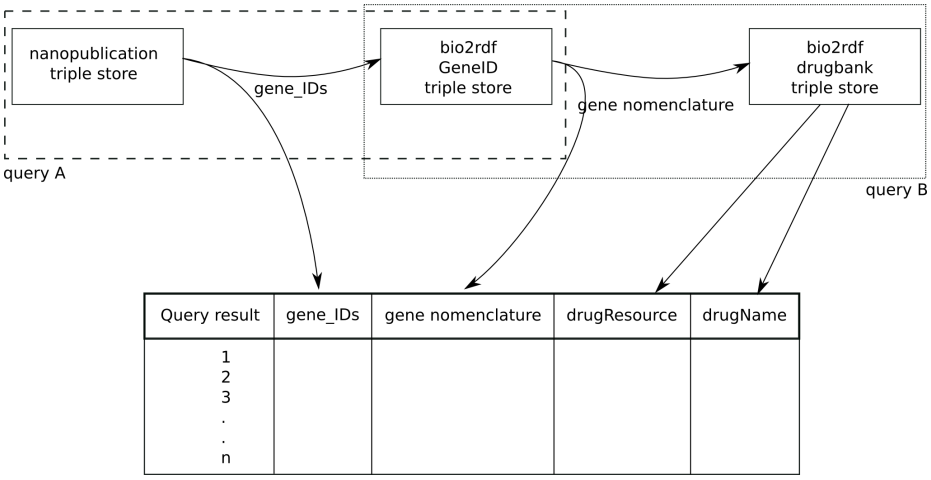

Fig. 5: Retrieving drug targets and drug names from drugbank that corresponds to our set of differentially expressed genes in Huntington's Disease that overlaps with CpG islands. **Query A**: retrieve all gene_IDs of deregulated genes of which the promoter overlaps with a CpG island. **Query B**: given a set of gene nomenclature names, query drugbank to retrieve drug targets and given those drug targets obtain drugResource and from drugResource get the drugName

could also be modeled based on our template for gene differentiall expression. We demonstrated reusability of our own template by exposing 5 different types of nanopublications, concerning genomic overlaps using the corresponding template. Vice versa, the reuse of templates improves interoperability of scientific results beyond the interoperability that RDF already provides. Therefore templates facilitate and encourage scientists to make their own discoveries public and therefore help to make large amounts of experimental results accessible while giving evergrowing data integration opportunities.

In Section 4 we presented examples of nanopublication integration on the assertion level in order to combine and extract information that is stored in our nanopublication store, but also in other triple stores (bio2rdf geneID, bio2rdf drugbank). In addition to that, nanopublications can facilitate even more sophisticated queries to integrate data based on their provenance information. For example, we could query for genes that are differentially expressed in Huntington's Disease, in both blood and brain tissue to identify potential blood biomarkers that could be used in brain. Querying provenance information could also relate to the methodology that was used as for example to retrieve all other nanopublications that have been using our workflow implementation. Another option could be to use the information stored in the publication info graph and retrieve information related to that. This way, we could determine the most frequently cited nanopublication creators and authors, for example in order to calculate some kind of impact factor.

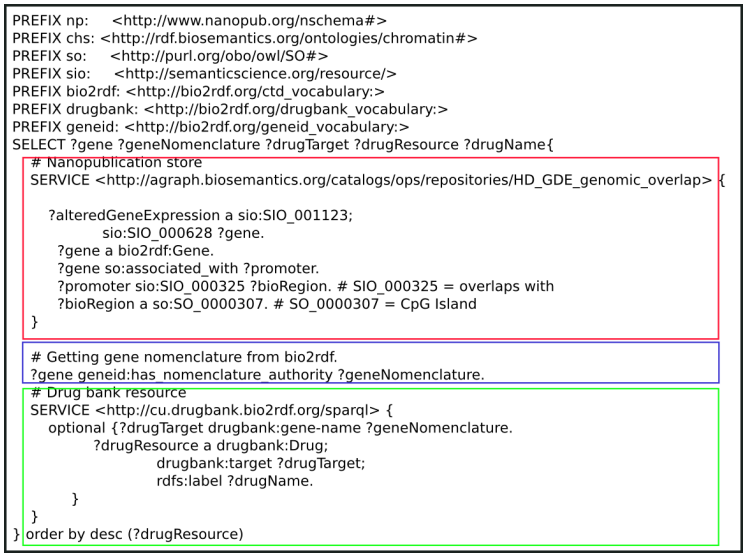

Fig. 6: The example SPARQL query that retrieves drug targets and drug names from Drugbank that are associated with the genes that we identified as differentially expressed in Huntington's Disease and overlapping with CpG islands. Red: query A from Figure 5; blue and green: query B of the same figure.

We would like to comment on nanopublications regarding the issue of reproducibility (and lack-thereof) in traditional journal publications. For example Ioannidis *et al.*, pointed out that they could not reproduce the majority of the 18 articles they investigated describing results from microarray experiments, including selected tables and figures [25]. Nanopublication does not guarantee full reproducibility, but at least – as a model for combining data with attribution and provenance in a digital format – it makes it possible to trace the origin of scientific results. The provenance section of a nanopublication ties the results (the nanopub assertion) to a description of an experiment and the associated materials, conditions and methods. In our case, we elaborated the proveanance with the Research Object model, showing that Nanopublication enables reuse of provenance information that may be already available. However, deciding the amount and relevance of provenance information to be included in the nanopublication remains to be decided by the nanopublication author.

Finally, we provided an endpoint that can give access to our nanopublication store. However, this implies that the user is already aware of the online location of this endpoint and is familiar with SPARQL. The Semantic Web implementation of the Nanopublication concept by itself does not provide a complete solution to their discoverability. Therefore, we argue that future work on tools for nanopublication should include a registry that permits easy discovery and use of nanopublications. Furthermore, we are working with the nanopublication community on the definition of a nanopublication API to further facilitate the

development of tools such as for creating, browsing and querying nanopublications.

6 Conclusion

To date there is an enormous amount of valuable information that has been produced by expensive experiments, but remains lost in databases and other repositories that are not easily accessed or processed automatically. This results not only in replicating experiments that have already been performed, but also in preventing all those associations from being tested or reused for building new hypotheses. This paper presents a method that enables life scientists to (i) expose the results from an analysis as scientific assertions, (ii) claim these as their contribution and (iii) provide provenance of the analysis as reference for the claimed assertions. We demonstrated an example from research in Huntington's Disease. In addition, we presented simple examples of nanopublication integration in the context of HD, and examples of how nanopublications can facilitate more sophisticated queries, integrating datasets from different research domains. The models for these nanopublications can be used as templates to create similar nanopublications, while the extension to the RO model can also be used to aggregate resources from other experiments that do not involve scientific workflows. Nanopublication provides an incentive for scientists to expose the results from individual experiments. This ultimately facilitates research across datasets that we anticipate will provide new insights about disease mechanisms. Research can become more efficient and go beyond monolithic journal publication [26].

7 Acknowledgement

We gratefully acknowledge Stian Soiland-Reyes and Graham Klyne for help with Research Objects and Paul Groth for his help on provenance. The research reported in this paper is supported by grants received from the Netherlands Bioinformatics Centre (NBIC) under the BioAssist program, the EU Wf4Ever project (270129) funded under EU FP7 (ICT-2009.4.1), and the IMI-JU project Open PHACTS, grant agreement n 115191.

References

1. D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. v. Mering, "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, pp. D561–D568, Nov. 2010.
2. R. Jelier, M. J. Schuemie, P.-J. Roes, E. M. van Mulligen, and J. A. Kors, "Literature-based concept profiles for gene annotation: The issue of weighting," *International Journal of Medical Informatics*, vol. 77, pp. 354–362, May 2008.

3. H. H. H. B. M. van Haagen, P. A. C. 't Hoen, A. Botelho Bovo, A. de Morre, E. M. van Mulligen, C. Chichester, J. A. Kors, J. T. den Dunnen, G.-J. B. van Ommen, S. M. van der Maarel, V. M. Kern, B. Mons, and M. J. Schuemie, "Novel protein-protein interactions inferred from literature context," *PLoS ONE*, vol. 4, p. e7894, Nov. 2009.
4. J. Y. Antonio and K. Verspoor, "Towards automatic large-scale curation of genomic variation: improving coverage based on supplementary material,"
5. L. S. Kaltenbach, E. Romero, R. R. Becklin, R. Chettier, R. Bell, A. Phansalkar, A. Strand, C. Torcassi, J. Savage, A. Hurlburt, G.-H. Cha, L. Ukani, C. L. Chepanoske, Y. Zhen, S. Sahasrabudhe, J. Olson, C. Kurschner, L. M. Ellerby, J. M. Peltier, J. Botas, and R. E. Hughes, "Huntingtin interacting proteins are genetic modifiers of neurodegeneration," *PLoS Genetics*, vol. 3, no. 5, p. e82, 2007.
6. "Guidelines for nanopublication." http://nanopub.org/guidelines/working_draft/.
7. C. Landles and G. P. Bates, "Huntingtin and the molecular pathogenesis of huntington's disease," *EMBO reports*, vol. 5, pp. 958–963, Oct. 2004.
8. J.-H. J. Cha, "Transcriptional dysregulation in huntingtons disease," *Trends in Neurosciences*, vol. 23, pp. 387–392, Sept. 2000.
9. E. A. Thomas, G. Coppola, P. A. Desplats, B. Tang, E. Soragni, R. Burnett, F. Gao, K. M. Fitzgerald, J. F. Borok, D. Herman, *et al.*, "The HDAC inhibitor 4b ameliorates the disease phenotype and transcriptional abnormalities in huntington's disease transgenic mice," *Proceedings of the National Academy of Sciences*, vol. 105, no. 40, p. 15564, 2008.
10. M. Gardiner-Garden and M. Frommer, "CpG islands in vertebrate genomes," *Journal of molecular biology*, vol. 196, pp. 261–282, July 1987. PMID: 3656447.
11. J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature*, vol. 473, pp. 43–49, Mar. 2011.
12. D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn, "Taverna: a tool for building and running workflows of services," *Nucleic acids research*, vol. 34, pp. W729–732, July 2006. PMID: 16845108.
13. K. Wolstencroft, R. Haines, D. Fellows, A. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. Nieva de la Hidalga, M. P. Balcazar Vargas, S. Sufi, and C. Goble, "The taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud," *Nucleic acids research*, May 2013. PMID: 23640334.
14. A. Hodges, "Regional and cellular gene expression changes in human huntington's disease brain," *Human Molecular Genetics*, vol. 15, pp. 965–977, Jan. 2006.
15. "UCSC genome browser home." http://genome.ucsc.edu/.
16. "AllegroGraph RDFStore web 3.0's database."
17. F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: towards a mashup to build bioinformatics knowledge systems," *Journal of biomedical informatics*, vol. 41, pp. 706–716, Oct. 2008. PMID: 18472304.
18. "SIO - semanticscience - the semanticscience integrated ontology (SIO) - scientific knowledge discovery - google project hosting."
19. "Taverna workflow provenance." http://www.w3.org/2011/prov/wiki/TavernaProvenance.
20. "Workflow4ever." http://wf4ever-project.org.
21. K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garca Cuesta, J. M. Gomez-Perez, G. Klyne, K. Page, M. Roos,

J. E. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C. Goble, "Workflow-centric research objects: First class citizens in scholarly discourse," in *Proc. Workshop on the Semantic Publishing (SePublica)*, Proc. Workshop on the Semantic Publishing (SePublica), (Crete, Greece), 2012.

22. K. Belhajjame, J. Zhao, D. Carijo, K. M. Hettne, R. Palma, O. Corcho, J. M. Gmez-Prez, S. Bechhofer, G. Klyne, and C. Goble, "The research object suite of ontologies: Sharing and exchanging research data and methods on the open web," 2013. Manuscript submitted for publication.
23. J. Zhao, G. Klyne, P. Holubowicz, R. Palma, S. Soiland-Reyes, K. Hettne, J. Ruiz, M. Roos, K. Page, J. Gómez-Pérez, *et al.*, "Ro-manager: A tool for creating and manipulating research objects to support reproducibility and reuse in sciences," *Proceedings of the 2nd International Workshop on Linked Science*, 2012.
24. "OWL-S: semantic markup for web services."
25. J. P. A. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort, "Repeatability of published microarray gene expression analyses," *Nature Genetics*, vol. 41, pp. 149–155, Jan. 2008.
26. F. Harmelen, G. Kampis, K. Brner, P. Besselaar, E. Schultes, C. Goble, P. Groth, B. Mons, S. Anderson, S. Decker, C. Hayes, T. Buecheler, and D. Helbing, "Theoretical and technological building blocks for an innovation accelerator," *The European Physical Journal Special Topics*, vol. 214, pp. 183–214, Dec. 2012.