Next Generation Cancer Data Discovery, Access, and Integration Using Prizms and Nanopublications

James P. McCusker [1,3], Timothy Lebo [2], Michael Krauthammer [3], and Deborah L. McGuinness [1,2]

[1]Department of Computer Science, [2]Department of Cognitive Science, Rensselaer Polytechnic Institute, 110 8th Street Troy, NY 12180, USA http://tw.rpi.edu
[3]Department of Pathology, Yale School of Medicine, 300 George St., New Haven, CT, 06510, USA http://krauthammerlab.med.yale.edu
{mccusj,lebot}@rpi.edu,dlm@cs.rpi.edu,michael.krauthammer@yale.edu

Abstract. To encourage data sharing in the life sciences, supporting tools need to minimize effort and maximize incentives. We have created infrastructure that makes it easy to create portals that supports dataset sharing and simplified publishing of the datasets as high quality linked data. We report here on our infrastructure and its use in the creation of a melanoma dataset portal. This portal is based on the Comprehensive Knowledge Archive Network (CKAN) and Prizms, an infrastructure to acquire, integrate, and publish data using Linked Data principles. In addition, we introduce an extension to CKAN that makes it easy for others to cite datasets from within both publications and subsequently-derived datasets using the emerging nanopublication and World Wide Web Consortium provenance standards.

1 Introduction

Peer-reviewed publications remain the principal means for exchanging cancer research information, despite the critical need for other researchers to access supporting data so that they may progress their own (or others') investigations. Critical ancillary data, such as gene expression data, are usually shared at time of publication, but there is a paucity of data sharing outside the realm of publications and it is usually limited to large consortia (ENCODE, TCGA), or government-mandated data sharing (data.gov). The National Institutes of Health and National Science Foundation both pass data-sharing mandates on to their awardees, but leave the implementation of those mandates to the awardees. An easy solution to data sharing would help federal grantees comply with award requirements and also help create more open, shareable data resources. Additionally, from our own experience there is a wealth of data that is rarely shared, such as ancillary data that does not make it into publications, negative findings, and findings from investigations that were not fully completed due to resource issues. Many institutions lack the expertise to transform local data into accepted

data standards. There are also data that are ready to be shared (such as lists of specimens, and annotations), but few institutions have the technical means to host it for others using a grid-enabled system.

Efforts to facilitate data sharing are common, but few are truly successful. We believe that most data sharing initiatives do not adequately address two key ingredients for a working data sharing environment: few constraints on how to share data, and a recognized avenue for receiving academic recognition (such as recognized citations). Most data sharing initiatives are built on data standards, which promise seamless data exchange at the expense of flexibility. Such initiatives (such as caBIG [1]) can also be overly technical without offering avenues for straightforward data sharing. Finally, few initiatives specify how academic credit is established for shared content. One reason that the scientific community is not sharing data fully is that there are no commonly accepted standards to publish and cite researchers' data-level contributions. We propose a new mode of data-sharing that we believe will be successful for the following two major reasons: First, the use of natural language provides a low barrier to entry for authors to express their research findings; and second, authors value publications as they offer the standard accepted proof of their academic work.

Towards this end, we are building a data sharing infrastructure with the following key features: first, a flexible data sharing setup, which allows for the sharing of plain text, excel, and other similar documents, with the ability to gracefully add metadata when needed; and second, the use of nanopublications, tiny and highly standardized statements that are useful for establishing provenance and academic credit, and for expressing high-level insights into the shared data. Our architecture is built upon Semantic Web technology, and is thus compatible with existing linked data sharing efforts.

Our infrastructure, called Prizms, is built entirely on open source software, leveraging existing data exchange software such as CKAN.[1] We have deployed instances of CKAN and Prizms at melagrid.org to serve the SPORE in skin cancer institutes to sharing melanoma related data.[2] The SPOREs have an active data sharing culture, and have recognized the need for exchanging research information. We are using the Prizms infrastructure (lod.melagrid.org) to extend the existing MelaGrid data portal (data.melagrid.org), used for sharing SPORE-related data. To encourage the use of data.melagrid.org by the melanoma community, we have populated it with melanoma-related datasets from ArrayExpress using a CKAN harvester we developed.[3] We currently have over 331 datasets in our repository.

The Prizms architecture leverages the Linked Data philosophy: use identifiers for things (URLs) that are addresses where consumers can get more information. When a human visits that address, they get a human-readable web page, with useful information, visualizations, and links to other resources. When a machine visits the page, it gets an RDF representation of the thing identified by the

[1] http://ckan.org
[2] http://trp.cancer.gov/spores/skin.htm
[3] https://github.com/jimmccusker/ckanext-arrayexpress

URL. The RDF should re-use existing resources that also follow the Linked Data philosophy, thereby providing aggregate benefits to both resources [2]. We will show how we provide a simple means of dataset discovery and citation for scientists and present a framework we use, composed of proven semantic technologies, to provide on-demand enhancement of that data into high-quality Linked Data.

2 Requirements: Levels of Data Sharing

Our experience suggests that only a few basic levels of data description are needed to promote successful data sharing. We want to make the value received from data description to be at least linearly related to the effort put into that description, and we want the value to pay off even at very simplistic levels of description. We therefore propose 5 levels of data sharing that will take data providers from very little effort (Level 1) to fully integrated and semantically enriched data that is easy to discover, integrate, and use (Level 5). Each of these levels serves as a broad use case for data sharing based on increasing levels of sophistication.

Level 1: Basic data sharing Basic data sharing consists of users 1) posting data somewhere, 2) telling the world about it (such as where it is, when it was modified, who controls it, or a simple description to make it more searchable). This information, often called provenance [3], consists of the basic information about data, such as who controls it, what is it about, when was it created, where can one get it, why was it created, and how was it created and used?

Level 2: Automated Conversion Using no domain knowledge, tools can create "naive", or non-knowledge driven, conversions of tabular data into structured formats such as RDF to provide basic search, browsing, and data integration.

Level 3: Semantic enhancement Semantic enhancement is performed using tools that allow users to specify improved data representations beyond what a computer can provide without additional knowledge. This can be by the data originator or other parties.

Level 4: Semantic eScience Further annotation and enhancement can be performed by describing the metadata for the dataset using vocabularies with well understood semantics. This provides a foundational component of Semantic E-Science, and corresponds to caBIG-style data sharing.

Level 5: Community-Based Standards By providing a framework for communication and discovery of consensus ontology use, a system can assist communities to converge on standard representations of data that result in interoperability across organizations. Further, by giving credit to contributors, the system can make it easier to find a community member that is able to assist in data representation challenges, which enables content-oriented collaborations among geographically or organizationally disparate community members.

3 Nanopublications for Datasets: Datapubs

MelaGrid reuses the existing open-source cataloging system CKAN to list and describe publishers' datasets. CKAN accounts for a majority of the basic Level 1 data sharing information that we identify in the previous section. However, it is incomplete, only providing information about dataset publication dates, data locations and hosting, but does not provide a means to describe how the data was produced, nor does it provide a sophisticated mechanism for identification of data owners. We have extended the CKAN RDF publication template to make better use of the available metadata in CKAN using DCAT, DC Terms, and PROV-O. This generates a novel form of nanopublication [4] we call a data-publication, or datapub. We have also included an interface (see Figure 1) that makes it easy to cite published datasets using plain text for non-technical users such as biologists and clinical researchers, BibTeX, PROV, or direct use of a nanopublication [4]. This functionality is available as an Open Source CKAN extension in GitHub called ckanext-datapub.[4] We have manually uploaded a dataset from a recent publication [5] and have cited it here using BibTeX. All citation modalities, including plain text, provide a Linked Data URL that provides human and machine-readable representations of the dataset using content negotiation.

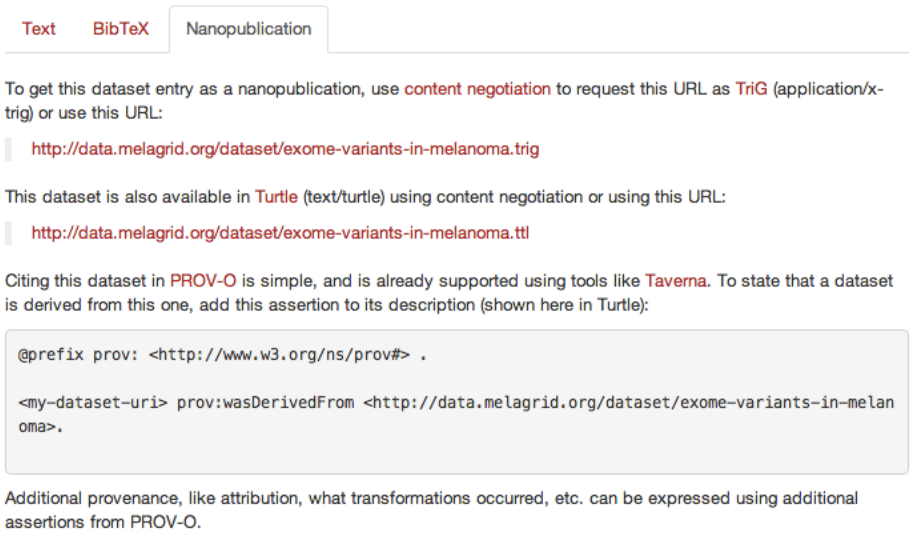

Fig. 1. Citing a datapub dataset using plain text, BibTeX, or PROV

[4] https://github.com/jimmccusker/ckanext-datapub

4 The Prizms Architecture

The Prizms architecture provides the technical foundation to support the remaining four levels of data sharing that we outline above. Prizms combines tools that the Tetherless World Constellation has developed during the past several years for use both internally and externally in many semantic web applications of scientific domains, such as a population science project that integrated health data, tobacco policy, and demographic data [6] and a system for the HHS Developer Challenge developed to integrate a wide variety of health data. The overall workflow of how MelaGrid uses the Prizms architecture and the Datapub extension is shown in Figure 2.

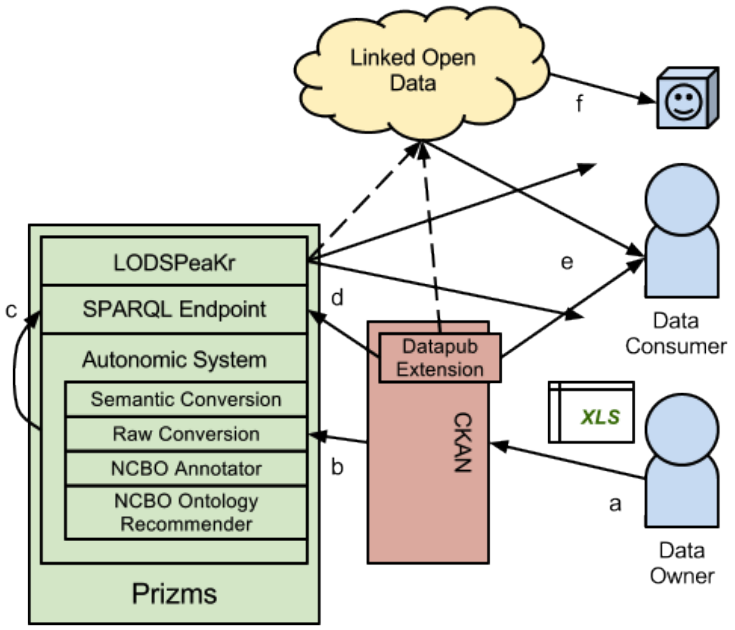

Fig. 2. Data flow through Prizms. A data owner (a) submits a dataset to a CKAN instance. This data can be in any format, including Excel (shown), CSV, XML, JSON, or other formats. The Prizms Autonomic System (b) recognizes the addition or change of a dataset and triggers tools that are "interested" in particular new datasets. It is then hosted by a standards-compliant SPARQL endpoint (c). The datapub CKAN extension then (d) generates RDF to describe the dataset as a Datapub. Human data consumers (e) can then browse the full dataset listing and access the data itself using either the traditional CKAN web interface or a Linked Data portal created using LODSPeaKr. Computational data consumers (f) can then access the data in conjunction with the Linked Open Data ecosystem.

While MelaGrid uses CKAN with the Datapub extension to address Level 1 "Basic" data sharing requirements, Prizms exposes the essential data access information as Linked Data using the W3C's Dataset CATalog vocabulary (DCAT),[5] the Dublin Core Terms (DC Terms) vocabulary,[6] and the W3C's PROV-O [7] provenance ontology. Prizms addresses Level 2 data-sharing requirements (automated RDF conversion) by using the access metadata to retrieve, organize, and automatically translate data posted to CKAN (such as Excel files) into RDF data files and hosting portions of each in a publicly-accessible SPARQL endpoint. All processing steps record a wealth of provenance described in best practice vocabularies such as Dublin Core, VoID,[7] and PROV-O, which enables transparency of any of Prizms' data products. For example, any RDF triple or RDF file can be traced back to the original data file(s) and the original publisher(s) [8]. This is important to maintain the reputability of Prizms, which serves as a third party integrator of others' data.

Prizms addresses Level 3 data-sharing (semantic enhancement) by transforming the original data to user-defined RDF. In the case of tabular data, such as Excel or CSV, transformations are specified using a domain-independent declarative description which itself is encoded in RDF. For example, one can specify that the third column in the data is mapped to a user-specified RDF class for concepts like gender or diagnosis. These concise transformation descriptions can be shared, updated, repurposed, and reapplied to new versions of the same dataset or within other instances of Prizms; they can also be maintained on code hosting sites like GitHub.com or Google Code. The transformation descriptions also serve as additional metadata that can be included as part of queries for the data (e.g., finding all datasets that were enhanced to use the class "specimen").

Reusing existing entities and vocabularies is the heart of Level 4 data-sharing (Semantic eScience), and using community-agreed ontologies and vocabularies are essential to Level 5 data sharing. We use new parameters of the same semantic conversion tools that are described in Level 2 for this purpose. In addition, datasets can be automatically augmented to produce inferences based on well-structured data that appears in Prizms' data store. For example, Prizms will augment any address encoded using the vCard RDF vocabulary[8] with the corresponding latitude and longitude (which it computes using the Google Maps API). When clients request Prizms' data elements, Prizms includes links to other available datasets based on a variety of curated and heuristic connections. These link suggestions can motivate community effort to mature the data towards more matures levels of data sharing.

At all levels of data-sharing, Prizms uses the LODSPeaKr web framework to create Linked Data applications and publish RDF data quickly and with minimal effort. LODSPeaKr provides a set of functionalities that not only improves the accessibility of the data for humans but also for machines by providing content

[5] http://www.w3.org/TR/vocab-dcat/
[6] http://purl.org/dc/terms/
[7] http://www.w3.org/TR/void/
[8] http://www.w3.org/Submission/vcard-rdf/

negotiation (i.e, the ability to return different formats depending on the client's request for the data element URL). This increases accessibility of the data while minimizing the workload for the development team. Additionally, the system allows the creation of new web pages to display particular subsets of the data that users may considered important. Data consumers can also perform query operations against the backing SPARQL endpoint.

5 Discussion

The MelaGrid initiative provides usable, integrated informatics systems that enable collaboration, data sharing, and enhanced analysis to research groups studying skin cancer. Specimen and associated Omics data sharing is a high priority for the MelaGrid initiative. Clinical annotations and phenotyping of specimens, along with Single Nucleotide Polymorphism (SNP), transcription, methylation, and copy number are just a few of the types of data that have become important in cancer research. All of these data have representation in the ArrayExpress subset of data.melagrid.org, and we will be extending its use with additional information from tools like caTissue.

The consortium's first priority is to increase the number of shared data entities, and Prizm's flexible architecture is assisting in this goal. Melagrid has the support of all four national skin SPOREs for use of this infrastructure. Currently, all shared data is at Level 1 (raw data with associated datapubs), and Level 2 (automatic RDF conversion). We will be using the Prizms architecture for converting institution-specific data descriptions into an accepted SPORE OWL/RDF Ontology (currently CDEs, as defined on melagrid.org) as appropriate. This is Level 5 data sharing in Prizms, as it involves a community-agreed standard (Level 3 is using a locally developed ontology, and Level 4 is re-use of ontologies, but not necessarily in a community-agreed manner).

6 Future Work

Currently, Prizms can be applied to dataset collections with other content domains, and it offers the same benefits that MelaGrid provides for melanoma data. We look forward to developing Prizms as we apply it to other applications, and we expect that others will find value by doing the same. For example, we are starting a portal for clinical depression treatment based on the Prizms infrastructure. Because using CKAN and the Datapub extension with Prizms has been so useful, we expect to extend Prizms to include both of them in future versions, so that we can facilitate others' adoption of all three components. We also look forward to developing additional out-of-the box capabilities for any datasets that Prizms is used to integrate, such as better connected exploration, better overviews, and better recommendations or guidance on how the data could be better modeled using best practice modeling techniques.

7 Conclusion

We have described an infrastructure for creating and using next generation science data portals. We have used the infrastructure to create two data portals - one reported on here in melanoma data and one in response to the human health services data challenge.[9] We have described how our infrastructure supports assimilating, publishing, and enhancing science data into best practices formats. The CKAN infrastructure makes it easy to aggregate data from multiple sources through its harvester framework and we have developed and used a CKAN harvestor to obtain and populate data.melagrid.org with 330 melanoma datasets that are now published as linked data. Further, we have provided a citation method for people to cite datasets from within both publications and subsequently-derived datasets using the emerging nanopublication (via our use of datapubs) and World Wide Web Consortium provenance standards.

References

1. von Eschenbach, A.C., Buetow, K.: Cancer informatics vision: caBIG. Cancer Informatics **2** (2006) 22–24
2. Berners-Lee, T.: Linked Data - Design Issues (2006)
3. Buneman, P., Khanna, S., Tan, W.: Why and where: A characterization of data provenance. ICDT **1973**(D) (2001) 316–330
4. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services and Use **30**(1) (2010) 51–56
5. Krauthammer, M.: Variant data from "exome sequencing identifies recurrent somatic rac1 mutations in melanoma". `http://data.melagrid.org/dataset/exome-variants-in-melanoma` Last Accessed: 2013-02-16.
6. McCusker, J.P., McGuinness, D.L., Lee, J., Thomas, C., Courtney, P., Tatalovich, Z., Contractor, N., Morgan, G., Shaikh, A.: Towards Next Generation Health Data Exploration : A Data Cube-based Investigation into Population Statistics for Tobacco. In: Proceedings of the Hawaii International Conference for System Science. (2013)
7. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology. `http://www.w3.org/TR/prov-o/`
8. Lebo, T., Wang, P., Graves, A., McGuinness, D.: Towards unified provenance granularities. In Groth, P., Frew, J., eds.: Provenance and Annotation of Data and Processes. Volume 7525 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2012) 39–51

[9] http://healthdata.tw.rpi.edu