Explorations into the Provenance of High Throughput Biomedical Experiments

Jamie P. McCusker and Deborah L. McGuinness

Tetherless World Constellation
Department of Computer Science
Rensselaer Polytechnic Institute
110 8th Street Troy, NY 12180, USA
{mccusj,dlm}@cs.rpi.edu
http://tw.rpi.edu

Abstract. The field of translational biomedical informatics seeks to integrate knowledge from basic science, directed research into diseases, and clinical insights into a form that can be used to discover effective treatments of diseases. We demonstrate methods and tools to generate RDF representations of a commonly used experimental description format, MAGE-TAB, mappings of MAGE documents to two general-purpose provenance representations, OPM (Open Provenance Model) and PML (Proof Markup Language). We show through a use case simulation that the data represented in MAGE documents can be completely represented in OPM and PML through use of round trip analysis of certain examples. The success in mapping MAGE documents into general-purpose provenance models shows that promise in the implementation of the translational research provenance vision.

1 Introduction

Translational biomedical research focuses on translating findings in basic science into advances in treatment and diagnosis of diseases for patients in the clinic, and has become a major research priority in the last five years. [1,2] Translational research requires the coordination and collaboration of a number of different disciplines, including basic science, clinical research, and increasingly, biomedical informatics. [3] As the scale and complexity of biomedical experiments has increased, so has the role of biomedical informatics. It plays an active role in the design, execution, and analysis of most biomedical research. The translational research pipeline, often thought of as a cycle of knowledge from the experimental "bench" to the clinical "bedside" and back, requires the management of many different kinds of data and artifacts by specialists in their disciplines. This includes information about the collection, management, and disposition of human, animal, and xenographic biomaterials, collection and management of participants in clinical research and trials, management of patient histories and charts, data from lab results, diagnostic imaging at the radiological and histopathological scales, as well as experiments using high-throughput technologies such as

D.L. McGuinness, J.R. Michaelis, and L. Moreau (Eds.): IPAW 2010, LNCS 6378, pp. 120–128, 2010.

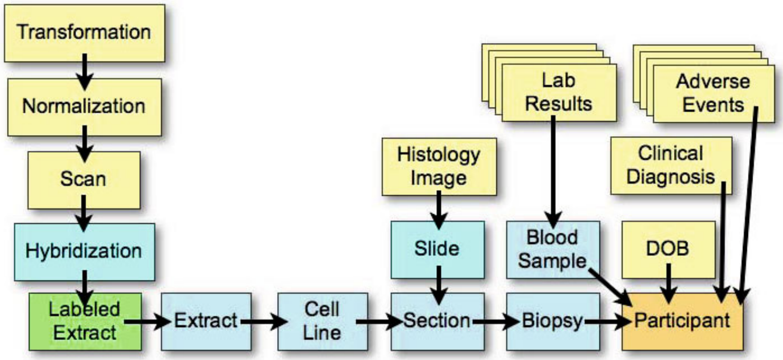

Fig. 1. Some common experimental and clinical artifacts that are created or used in the process of translational biomedical research. This example is common of translational cancer research.

microarray assays and high throughput sequencing. Common derivations and artifacts from the translational research pipeline are shown in Figure 1 on page 121.

1.1 The Translational Research Provenance Vision

The vision is relatively simple: It should be possible for a research scientist, clinician, patient or legal guardian to be able to query, assess, and collate the knowledge needed to make decisions about research and patient care. In order to be able to do this, the information that is used to make these decisions must have at hand the provenance of those materials, so as to be able to judge the relevance and veracity of the information they need. For this to happen, there must be a consistent model of provenance that can be used regardless of the origin, domain, or format of the information at hand.

By accomplishing this, we make it possible to gain a complete picture of how experiments were conducted, if they are comparable, and how well confounding variables have been controlled for. In the longer term, it also offers an opportunity to build experiments based on previous work, by understanding what kinds of methods have been used on certain kinds of problems, and to find new avenues of research. Provenance also makes electronic health records much more portable, as it becomes possible for clinicians to assess if lab work performed at other institutions is comparable with work done at their own. For patients, access to a consistent model of the provenance of their medical state means being able to take control of their own health, and to understand the reasoning behind clinical treatments and advice.

The World Wide Web Consortium (W3C) has chartered a incubator group for provenance representations[1] and has developed a number of biomedical use

[1] http://www.w3.org/2005/Incubator/prov

cases. Many of these use cases describe specific steps along the path to realizing this translational research provenance vision. Additionally, existing automated computational workflow systems, such as Wings/Pegasus [4], Taverna/myGrid [5], and VisTrails [6], have started to converge towards a common interchange language for provenance.

1.2 High Throughput Experiments and Provenance

Scientists look to provenance information, such as experimental workflow, to learn about experiments and results in their field. Critical to their understanding is: (1) how the experiment was performed, and (2) what needs to be known to be able to repeat it. As such, it is vital that systems that support the sciences provide a framework for incorporating provenance information at every step of the research chain. This is especially true for high throughput assays such as microarrays. Each microarray can measure hundreds of thousands to more than a million nuclear material hybridizations. Because of the scale of measurement, an experimental design must overcome a potentially high False Discovery Rate (FDR) [7] through use of many biological replicates for each experimental condition. Additionally, the context that is provided by richly encoded provenance can be used to automate certain aspects of scientific research.

In bioinformatics and computational biology, the problem of provenance has been an issue for some time. Goble [8] identifies a number of provenance-related issues in re-use and propagation of database information. The rapid growth and evolution of experimental techniques has makes it ever more difficult for scientists to evaluate the soundness and validity of the data at hand. This growth has resulted in the establishment of a standard for describing microarray-based experiments. The MIAME (Minimal Information About a Microarray Experiment) and MAGE (MicroArray and Gene Expression) standards [9] established metadata requirements for microarray experiments in informal (MIAME) and formal (MAGE) terms. MAGE currently has a number of representations, including MAGE-ML (MAGE Markup Language) and MAGE-TAB (MAGE TABle). These standards, combined with data sharing requirements from most funding institutions such as the National Institutes of Health in public databases such as the National Center for Biotechnology Information's (NCBI) GEO (Gene Expression Omnibus) [10] and the European Bioinformatics Institute's (EBI) ArrayExpress [11], along with those databases' adoption of the MAGE and MIAME standards, have resulted in thousands of microarray experiments stored in a consistent standardized format.

However, this format is designed specifically for microarray experiments. New assay types, such as tissue microarrays [12], high-throughput sequencing, and other low or medium throughput experiments require a more generalized data model. Additionally, information about findings is absent from MAGE and MIAME, as is the detailed information about the biospecimens that were used in the experiment gathered and managed by the biospecimen bank. This is all valuable information that can benefit from a common data model, if one were available. Integration with other data sources in the translational pipeline, such as biospecimen management tools, Laboratory Information Management Systems (LIMS), and computational workflow automation tools through the use of a common model of provenance can provide a complete picture of the provenance of experiments. A first step in this process is to convert experimental data into a common provenance representation. We accomplish this by implementing a simulation the following use case:

MAGE Data Sharing Use Case: Two databases, A and B, are repositories for microarray experiments that conform to the MAGE standard. B would like to load some experiments from A, which publishes a web service that describes its experiments using a general purpose provenance model. B should be able to re-create the information about the experiments it retrieves from A without loss.

Implementing this use case using a general purpose provenance model would demonstrate that it is possible to transform MAGE-compliant experimental metadata into that provenance model without losing any information. We create implementation simulations of this use case using the Open Provenance Model (OPM) [13] and Proof Markup Language (PML) [14]. Through these implementations, we show that it is possible to represent microarray experimental metadata fully and without loss in two common general purpose provenance models, and with the continuing adoption of general purpose provenance models by computational workflow systems, establishes the first link in the chain of provenance for the translational research provenance vision.

2 Related Work

There is a significant amount of related work to this topic. We highlight three areas, which the following subsections each discuss. The first area is work related to the MAGE object model. MAGE is a standard used to describe microarray experiments in a consistent manner. The second area is work related to analysis of data format compatibility using round trip analysis. We use this type of analysis to determine the suitability of representing MAGE-based experiment metadata in general purpose provenance models. The third area is work related to general purpose provenance models, specifically OPM. OPM has been used as a common representation for provenance interchange at two provenance challenges [15,16].

2.1 MAGE

The MAGE object model and related representations has been in wide use for a considerable period [9] in bioinformatics. Currently, the most commonly used MAGE format is MAGE-TAB [17], a delimited text-based format encompassing a number of file formats, of which we use information from the Investigation Design Format (IDF), which contains global information about an experiment, including submitters, publications, protocols, experimental factors, etc.; Sample and Data Relationship Format (SDRF), which describes the experimental workflow and how samples and other data and physical artifacts relate to each other.

2.2 Round Trip Analysis

Round-Trip analysis of data representations, especially meta-models, have been a gold standard for validating the expressivity of those models. Farquhar *et al.* [20] uses a similar method validate conversion from various ontology languages into a common format. Antkiewicz *et al.* [21] discusses round-trip engineering, or using round-trip analysis to show that Framework-Specific Modeling Languages (FSMLs) can be shown to reliably represent Domain-Specific Modeling Languages (DSMLs) in the Java Eclipse platform.

2.3 General Purpose Provenance Models

A commonly used provenance interlinguas is the Open Provenance Model. OPM has its roots in the workflow world, where it has evolved as a proposed common interchange language for computational workflow management and execution tools, and was used as a standard interchange for the second and third Provenance Challenges. A number of scientific workflow applications have participated in these challenges, which involved the generation of OPM graphs by each team for query by the other team members. Because of this, Taverna [22] and Pegasus/Wings [4] now support the export of provenance information in OPM. This support makes it very attractive as a first link between disciplines within the translational pipeline, as bioinformaticians often use computational workflow automation tools for research.

3 Methods

We simulate the MAGE data sharing use case in a semantic web environment using the MGED Ontology [23] as a foundation for the representation of MAGE documents in RDF. The overall process flow is seen in Figure 2 on page 125. We start by converting MAGE-TAB documents into RDF using MAGETAB2MAGERDF[2] and feed the resulting RDF into an OPM processor that infers the relevant OPM structure from the original RDF. A separate engine extracts the statements relating to OPM so that the resulting document is a pure instance of provenance data in OPM. A second processor attempts to reverse the initial step, taking the provenance data and generating the original MAGE RDF document. Finally, a comparison processor takes the difference between the original MAGE RDF and the regenerated MAGE RDF and outputs the statements that are missing in the regenerated document. The process of converting MAGE-TAB to RDF and the mapping from MAGE-RDF to OPM is discussed in depth in McCusker and McGuinness [24].

3.1 Evaluation

To evaluate our mappings, we perform an extraction of OPM-specific information from the resulting output RDF graphs and reverse the mapping process discussed

[2] http://magetab2rdf.googlecode.com

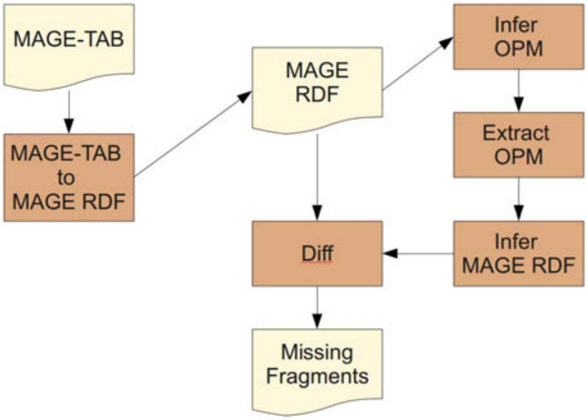

Fig. 2. The round trip analysis process. A MAGE-TAB file is converted to RDF and is fed into a processor that infers OPM from the original MAGE RDF. The OPM is extracted, ensuring only information represented in the provenance model remains. A processor then re-creates the original RDF to the best of its abilities from OPM. The resulting RDF is compared to the original to find missing statements.

above. We then compare the graphs to determine what statements are missing from the reconstructed graph as compared to the original.

We extract data relevant to a particular ontology in order to ensure that only data relating to that ontology or its imports remains. Only statements that use properties from the ontology import closure and individuals with a rdf:type of a class in the ontology import closure are extracted, everything else is filtered out. We perform the data extraction using three Jena models: (1) a base comparison model with only the ontology import closure, (2) an input model with the ontology import closure and the input graph to extract from, and (3) an output model with starts with the ontology import closure. The algorithm then iterates over all the classes and properties in the base comparison model and copies all statements relating to those classes and properties.

For the conversions back to MAGE from OPM, we reverse the rules discussed in McCusker and McGuinness [24]. These rules are described in opm2mage.rules[3]. The same generic inferencing processor, using the Jena API and rules engine, is used to run the conversion of OPM back into MAGE.

We use the Jena API to compare the reconstructed MAGE RDF graph with the original by creating a Jena models for the original and reconstructed graphs, and take difference of the original graph against the reconstructed graph. This results in a model that contains all statements that are in the original graph that are not in the reconstructed graph.

[3] https://scm.escience.rpi.edu/svn/public/mageprovenance/rules/opm2mage.rules

4 Results

For the mapping of MAGE to OPM, we performed a round-trip analysis on the ArrayExpress experiment E-MEXP-986[4], a small-scale but well-annotated exemplar experiment on *Arabidopsis thaliana.* We report that on conversion of the experiment to OPM and back to MAGE, there are no missing statements from the reconstructed RDF graph.

5 Discussion

A successful mapping of one of the most widely-used experimental description formats to a general purpose provenance model suggests two things: (1) descriptions of high-throughput experiments can be successfully represented using a general purpose model, and (2) OPM is sufficiently mature as a model of provenance to support real-world descriptions of experimental workflows in the biological sciences. Given this successful mapping, it is now possible to support a wide range of biomedical experiments within existing provenance models without a need for domain-specific extensions. It also means that the vision of consistent provenance representations across the translational research pipeline is possible, and points to interesting future work in representing biospecimen history and clinical information using general-purpose provenance models.

5.1 Future Work

We are currently working on scaling the declarative mapping for OPM and developing a procedural mapping for Proof Markup Language, another provenance representation [14]. The MAGE object model represents a small part of the derivational history of biospecimens that are used in these experiments. Future work of providing biospecimen history and analysis, as well as patient clinical history in a provenance model, we look to realize the translational research provenance vision laid out in this paper. More generally, each part of the translational research pipeline represents future work that is needed to realize the translational research provenance vision. Finally, research is needed in visualization and search of large graphs of provenance before generalized provenance models can be used effectively. Biomedical informaticians already use graph-based tools such as Cytoscape to visualize large molecular interaction graphs, but clinicians and patients will probably require a different perspective.

6 Conclusion

We proposed a vision of provenance for translational biomedical research that supports the integration of clinical and research artifacts across the translational research pipeline, provided an overview into the translational research pipeline,

[4] http://www.ebi.ac.uk/microarray-as/ae/files/E-MEXP-986

and showed how high throughput experiments provide a critical role in current biomedical research. We also demonstrated mappings of MAGE descriptions of experiments onto a general purpose models of provenance, OPM, and showed that the mappings are faithful and complete using an analysis of the round trip mapping of exemplar data. We also provided a framework for analyzing conversions of data from one RDF model to another. We discussed the advantages and pitfalls of declarative and procedural mappings. Finally, we gave a window into future work in implementing the translational research provenance vision.

References

1. Zerhouni, E.A.: Translational and clinical science–time for a new vision. New England Journal of Medicine 353(15), 1621 (2005)
2. Zerhouni, E.A.: US biomedical research: basic, translational, and clinical sciences. Jama 294(11), 1352 (2005)
3. Payne, P.R.O., Johnson, S.B., Starren, J.B., Tilson, H.H., Dowdy, D.: Breaking the translational barriers: the value of integrating biomedical informatics and translational research. Journal of Investigative Medicine 53(4), 192 (2005)
4. Kim, J., Deelman, E., Gil, Y., Mehta, G., Ratnakar, V.: Provenance trails in the Wings/Pegasus system. Concurrency and Computation: Practice and Experience 20(5), 587–597 (2008)
5. Zhao, J., Goble, C., Stevens, R., Turi, D.: Mining taverna's semantic web of provenance. Concurrency and Computation: Practice and Experience 20(5), 463–472 (2008)
6. Scheidegger, C., Koop, D., Santos, E., Vo, H., Callahan, S., Freire, J., Silva, C.: Tackling the provenance challenge one layer at a time. Concurrency And Computation 20(5), 473 (2008)
7. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289–300 (1995)
8. Goble, C.: Position statement: Musings on provenance, workflow and (Semantic web) annotations for bioinformatics. In: Workshop on Data Derivation and Provenance, Chicago (2002)
9. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C., Causton, H., et al.: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nature genetics 29(4), 365–372 (2001)
10. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N., Edgar, R.: NCBI GEO: archive for high-throughput functional genomic data. Nucl. Acids Res. 37(suppl. 1), D885–D890 (2009)
11. Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., et al.: ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. In: Nucleic Acids Research (2008)
12. Berman, J., Edgerton, M., Friedman, B.: The tissue microarray data exchange specification: A community-based, open source tool for sharing tissue microarray data. BMC Medical Informatics and Decision Making 3(1), 5 (2003)

13. Moreau, L., Miles, S., Missier, P., Simmhan, Y., Futrelle, J., Myers, J., Stephan, E., Kwasnikowska, N., den Bussche, J.V., Freire, J., et al.: The open provenance model (v1. 1) (2009)
14. McGuinness, D., Ding, L., Pinheiro da Silva, P., Chang, C.: Pml 2: A modular explanation interlingua. In: Proceedings of AAAI, vol. 7 (2007)
15. Davidson, S.B., Freire, J.: Provenance and scientific workflows: challenges and opportunities. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1345–1350 (2008)
16. Moreau, L., Kwasnikowska, N., Van den Bussche, J.: The Foundations of the Open Provenance Model (2009)
17. Rayner, T., Rocca-Serra, P., Spellman, P., Causton, H., Farne, A., Holloway, E., Irizarry, R., Liu, J., Maier, D., Miller, M., Petersen, K., Quackenbush, J., Sherlock, G., Stoeckert, C., White, J., Whetzel, P., Wymore, F., Parkinson, H., Sarkans, U., Ball, C., Brazma, A.: A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics 7(1), 489 (2006)
18. Bian, X., Klemm, J., Basu, A., Hadfield, J., Srinivasa, R., Parnell, T., Miller, S., Mason, W., Kokotov, D., Duncan, M., et al.: Data submission and curation for caArray, a standard based microarray data repository system (2009)
19. Stokes, T., Torrance, J., Li, H., Wang, M.: ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. BMC bioinformatics 9(Suppl 6), S18 (2008)
20. Farquhar, A., Fikes, R., Rice, J.: The ontolingua server: A tool for collaborative ontology construction. International Journal of Human-Computers Studies 46(6), 707–727 (1997)
21. Antkiewicz, M., Czarnecki, K.: Framework-specific modeling languages with round-trip engineering. In: Wang, J., Whittle, J., Harel, D., Reggio, G. (eds.) MoDELS 2006. LNCS, vol. 4199, p. 692. Springer, Heidelberg (2006)
22. Missier, P., Belhajjame, K., Zhao, J., Goble, C.: Data lineage model for Taverna workflows with lightweight annotation requirements. In: Freire, J., Koop, D., Moreau, L. (eds.) IPAW 2008. LNCS, vol. 5272, pp. 17–30. Springer, Heidelberg (2008)
23. Whetzel, P., Parkinson, H., Causton, H., Fan, L., Fostel, J., Fragoso, G., Game, L., Heiskanen, M., Morrison, N., Rocca-Serra, P., et al.: The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 22(7), 866 (2006)
24. McCusker, J.P., McGuinness, D.L.: Representing high throughput biomedical experiments using the open provenance model. Technical report, Technical Report TW-2010-14, Tetherless World Constellation, Rensselaer Polytechnic Institute, USA (2010)
25. Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13(11), 2498 (2003)