JESLIB 2012; 1(1): 51-56
doi:10.7191/jeslib.2012.1000

Journal of eScience Librarianship
putting the pieces together: theory and practice

Description and Annotation of Biomedical Data Sets

Jen Ferguson*

Harvard School of Public Health, Boston, MA, USA

Abstract

Deposition of biomedical data sets is on the rise as more scientists submit experimental data to accompany their publications. Scientists are also increasingly reusing these publicly available data sets in their own work. Despite these developments, lack of both context and metadata can create barriers to understanding and repurposing these data sets. Researchers from the Bioinformatics Core Group in the Harvard School of Public Health attempted to address this issue by assembling a team of data curators who used the open source software suite ISA tools to annotate and contextualize microarray data sets.

This paper describes the workflow and software used in curating these data sets, discusses similarities and differences in the approaches of team members to the work, and suggests possible roles for librarians in similar data curation projects.

*Jen Ferguson is presently the Data Services Librarian at Northeastern University, Boston, MA, USA.

Biomedical data deposition is on the rise as more scientists make their experimental data openly available (Piwowar and Chapman 2010). This phenomenon can likely be attributed in part to increasing pressure from publishers and funding agencies to encourage and even mandate data deposition to accompany publication. In a recent survey, more than 40% of peer reviewers for the journal *Science* indicated that they routinely access or use the data sets that accompany publications (Science 2011). Researchers use these data sets in a variety of ways, including validation and testing of statistical models, and critical evaluation of data discussed in publications. Some works rely heavily upon this body of publicly available data sets, employing data mining for much of their investigative basis. In perhaps the best known example, Mootha and colleagues (2003) successfully identified the human genetic defect that gives rise to Leigh syndrome by first mining publicly available data.

Despite these developments, lack of context and metadata can still create obstacles to understanding and reuse of data sets. Certain types of biomedical data, such as sequence data, can be interpreted fairly simply; little additional context aside from the sequence itself is necessary to make use of the data. Gene expression microarray data, on the other hand, require thorough understanding of the experimental context and conditions that produced it. As a result, comprehension and reuse of microarray data

Correspondence to Jen Ferguson: j.ferguson@neu.edu
Keywords: curation, microarray, eScience, ISA, gene expression

sets, in particular, can suffer from lack of consistency and detail in associated metadata (Ochsner et al. 2008, Ventura 2005).

Researchers from the Harvard School of Public Health attempted to address these issues by assembling a team of curators to annotate and contextualize NCBI Gene Expression Omnibus (GEO) microarray data sets deposited in conjunction with published articles (NCBI 2007). Staff from the Bioinformatics Core Group in the Harvard School of Public Health (HSPH/HBC) initiated contact with Boston-area graduate students in late 2010, requesting assistance with a data curation project. I learned of their recruiting efforts through a life sciences graduate student listserv at Brandeis University, where I worked as a science librarian. HSPH staff agreed to add me to the curation team, which consisted of about six life sciences graduate students and postdoctoral fellows from several local universities.

The curation team met at the School of Public Health in January 2011 for an initial training session with members of the research staff. This session introduced team members to the problems being addressed by the project, and included an overview of the ISA tools software (ISATeam, n.d.; Rocca-Serra et al. 2010) to be used in curation. From this point on, most work was done remotely. Team members used the project management tool Basecamp (37 Signals, n.d.) extensively as a way to interact with research staff and with each other, discuss problems, and share sample curated records, screenshots, and assignments. A member of the ISA tools development team also fielded software questions and suggestions through the Basecamp site.

Curators were assigned a previously published paper available in PubMed with affiliated GEO microarray data sets. Curators read the paper closely to understand the experimental approach and research protocols in detail. Particular care was taken in examining the Materials and Methods section, as this yielded much of the metadata used in curation and annotation. Curators retraced the experimental steps taken by the authors, correlating their descriptions in the journal article with the data sets they had deposited as GEO files in PubMed.

Curators then used the open source software suite ISA tools to record and annotate the experimental descriptions and data sets affiliated with the paper. The ISA tools suite consists of several Java-based desktop components that can be used independently or in tandem. For this project, curators used the ISAcreator (Figure 1) and ISAvalidator components. Curators first used ISAcreator to curate investigations, producing a tab-delimited ISA-Tab record. This record supplies metadata for the investigation as a whole. Within the ISA-Tab record, curators also annotated and described most subsets of the experimental work, breaking down published accounts with increasing granularity into investigations, studies, and assays. This structure cleverly mimics the format of the experimental work as it is carried out in the laboratory, while providing enriched context and clarification of the precise relationship of the data sets to the published paper. Annotated data associated with an investigation typically included both raw (e.g. DNA microarray data) and derived (e.g. gene lists) data types within an ISA-Tab record.

Completed ISA-Tab records (Figure 2) were then analyzed using another software tool called ISAvalidator. ISAvalidator examined the new record for inconsistencies or errors in metadata added by curators, and flagged records for further follow up by members of the research staff when necessary. Upon successful validation, the completed ISA-Tab record was sent to an internal data management system. As of the time of this writing, HSPH/HBC has collected over 50 annotated studies comprising more than 900 assays. Ultimately, the project aims to create a collection of records that clearly tie curated, metadata-enriched data sets to published works. The ISA-Tab records that contain

Figure 1: Curation in progress: example of a record in process in ISAcreator. Curators analyzed PubMed papers and associated GEO datasets, then created ISA-Tab records annotating and contextualizing experimental data. At the pictured stage in the process, curators supplied metadata for the investigation as a whole. Later stages in the workflow involved annotation at the more granular assay and protocol levels.

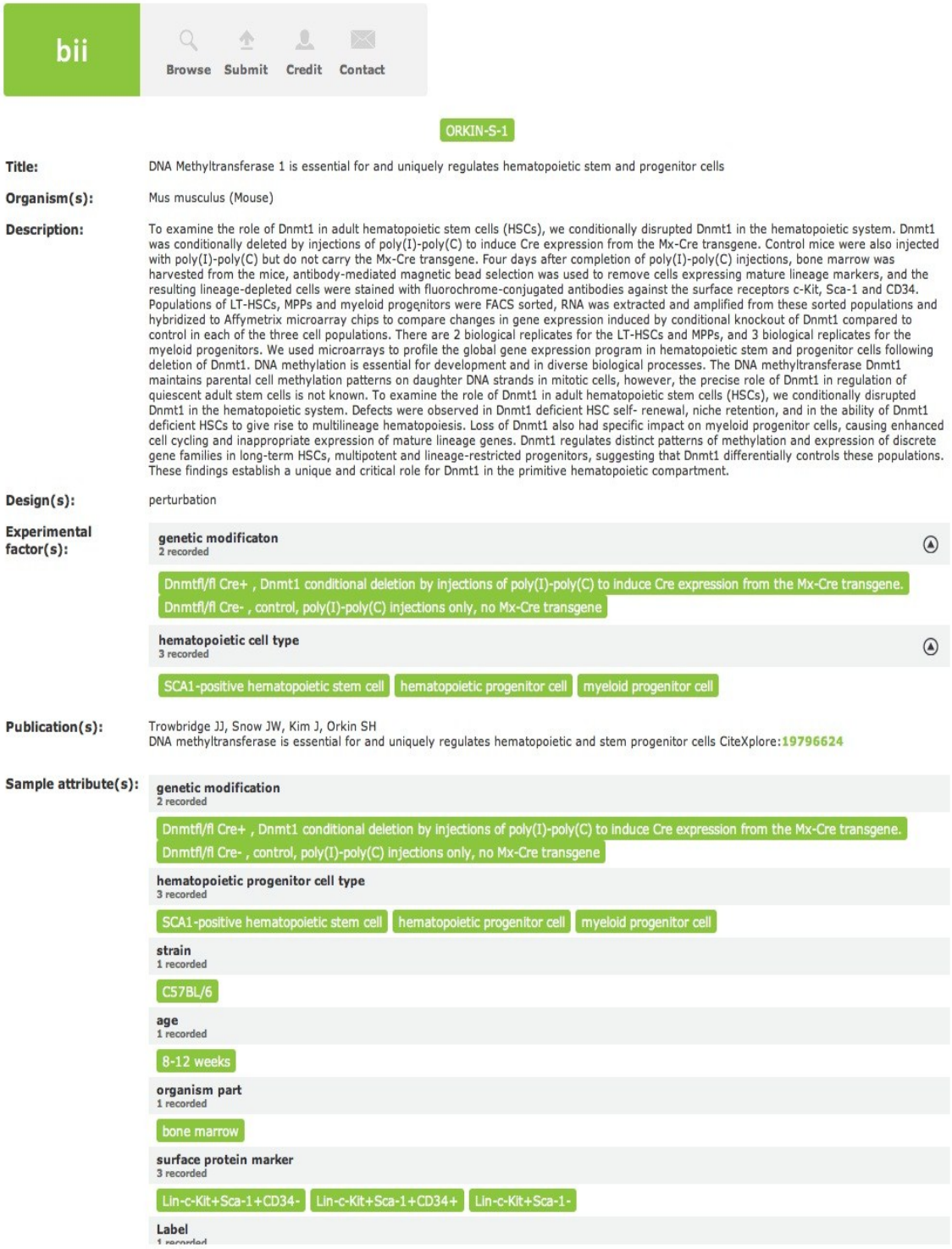
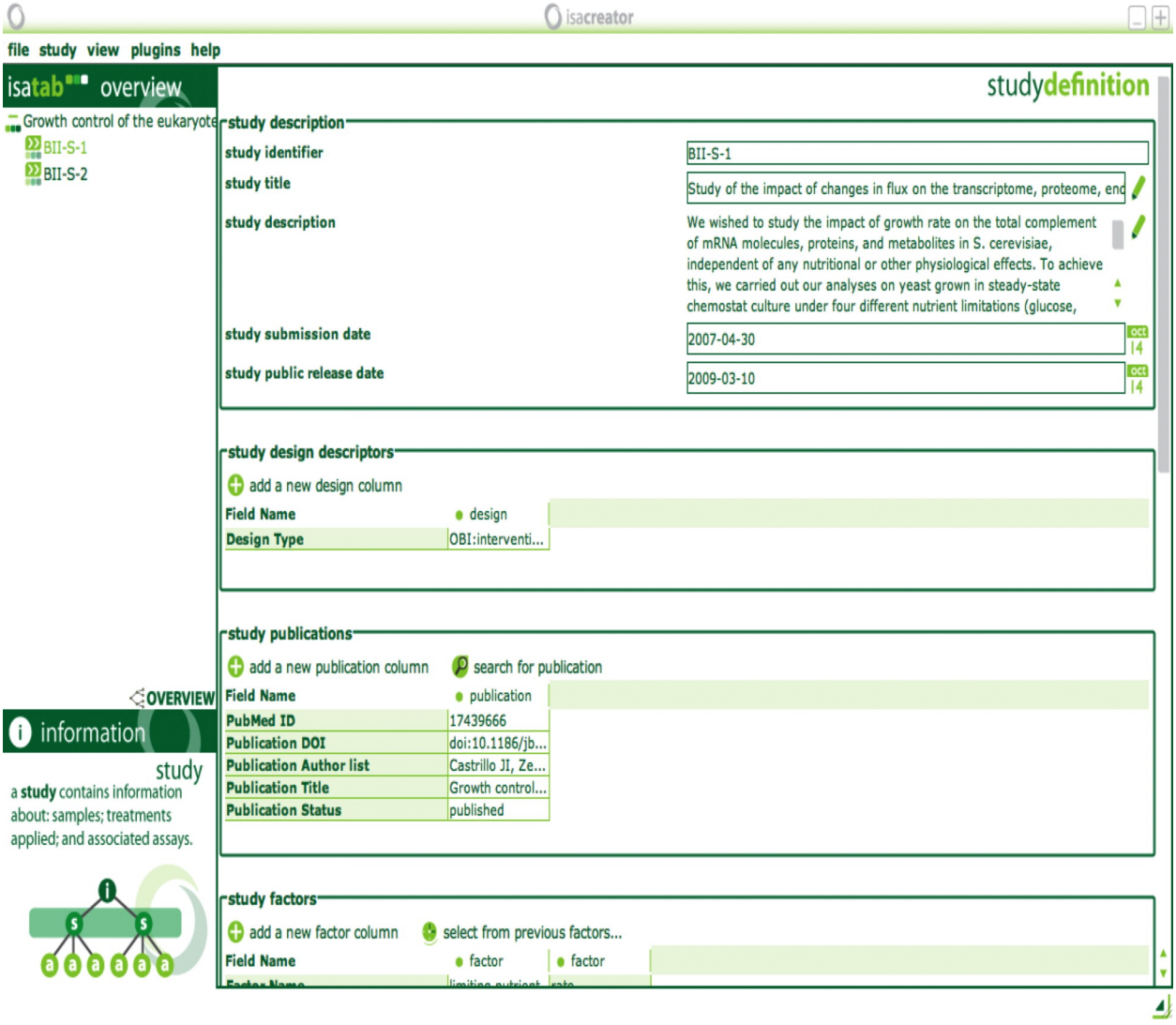


Figure 2: Example of a finished ISA-Tab record. Some record fields were taken directly from the published paper, while curators supplied additional terms and values such as the table listing sample attributes and experimental factors. Note that this is simply a record overview; links are provided to download additional study details such as metadata records and assay data files.



this enriched information will be openly available in public repositories for examination and download.

Work on this curation project highlighted both similarities and differences in team member approaches. Some of these variations could be attributable to differences in background and expertise.

For example, controlled vocabularies are built into ISAcreator in the form of ontology lookups. These include a number of highly specific controlled vocabularies created to describe organisms, techniques, and biomedical processes, as well as some broader vocabularies (such as MeSH) that are likely familiar to many librarians. Use of these ontologies helped provide consistency in the terms curators assigned to studies. Howev-

er, ontology lookups were available only for certain record fields in ISA-Tab, but even for those fields, curators often opted to supply free text terms rather than choose controlled vocabulary terms from the ontologies. This may reflect confusion over which ontology to use, as the lookup tool presented curators with a large list of ontologies to choose from, and little guidance as to which one to use. Supplying free text terms rather than using controlled vocabularies could also reflect varying degrees of curator confidence in the capabilities of full text search. My concern, based on my library experience, is that this method, given its variations in terms and occasional data entry errors, will not be optimal for record search and discovery.

ISA developers are attentive to issues such as metadata conversion and integration with existing repositories. As a case in point, another component of the ISA software suite, ISAconverter, has recently been developed. ISAconverter can convert ISA-Tab files into other formats such as MAGE-Tab (a metadata standard for describing DNA microarray data), SRA XML (for high-throughput sequencing data), and Pride-ML (for mass spectrometry data), thus enabling submission of records to several public repositories. Still, this project in its current form seems focused on tackling data reuse problems within a fairly narrow discipline. Here I think e-science librarians, by approaching data curation from a broad perspective, can offer valuable knowledge to our scientist colleagues. E-science librarians are aware of similar efforts to curate and annotate data in a variety of other disciplines. Given our experience with issues such as file formats and interoperability, we're also thinking proactively of both the challenges and possibilities in the realm of cross-disciplinary reuse.

Regardless of background and expertise, a recurring issue for all curators was the question of how much metadata and annotation was sufficient for discovery. Many experimental protocols in this area of biological research are fairly standard and well defined (e.g. sample preparation, RNA extraction and labeling). However, most labs follow their own variations of these protocols. Is it acceptable to ignore these standard protocols when curating records, let alone the 'tweaks' made by each group of investigators? We generally elected to ignore basic protocols in generating curated records, as otherwise the time spent curating each investigation would increase significantly.

Curating a single investigation could take up to 10 hours, including time spent reading the journal article, creating the curated record, and submitting the completed ISA-Tab record for validation. This figure decreased as curators became more facile with both the subject matter and the software tools, but a significant time commitment was still required to generate each curated investigation. Outsourcing this task to the curation team did shift this burden from the researcher – and thereby helped ensure that the work was completed – but it greatly increased the time needed to become familiar with the experimental work and accurately curate the investigation, and raises questions as to the sustainability of this approach. Significant time and subject matter expertise was necessary just to relate the published work with its associated data sets. As an advocate for digital curation and preservation, it was quite educational to experience barriers to data reuse firsthand. Accordingly, cross-disciplinary data reuse at times seemed a distant possibility.

From my involvement with the HSPH/HBC project, I remain convinced that there are valuable roles for librarians to play in data curation. Some of the most worthwhile contributions that librarians can offer may occur prior to the actual curation process. Consultation with software and tool developers regarding core librarian competencies such as metadata interoperability, authority control, and consistent use of controlled vocabularies will help ensure that data is discoverable. We can encourage scientists who collect

and organize research data to consider that the visibility and usability of the work they generate – beyond just the papers they write, and even beyond their own discipline – is worth the time spent to clearly document and describe data. Librarians can also play a key role in connecting researchers across disciplines that are working on similar problems.

This is a time of opportunity for eScience librarians, as scientists are clearly also aware of the need for action to make deposited data more findable and usable. The challenge may lie in getting scientists and software developers to think of librarians as having the sort of expertise that makes us good partners for this endeavor. Librarians with subject matter background, an enterprising spirit, and the ability to cultivate strong liaison relationships can go a long way towards gaining that acceptance.

References

37signals. n.d. “Project management software, online collaboration: Basecamp”, accessed July 22, 2011, http://basecamphq.com/.

"Introduction: Challenges and Opportunities." Science 331, no. 6018 (2011): 692-93.

ISA Team. n.d. "Isa – about," accessed July 22, 2011, http://isatab.sourceforge.net/index.html.

Mootha, V. K., P. Lepage, K. Miller, J. Bunkenborg, M. Reich, M. Hjerrild, T. Delmonte, A. Villeneuve, R. Sladek, F. H. Xu, G. A. Mitchell, C. Morin, M. Mann, T. J. Hudson, B. Robinson, J. D. Rioux, and E. S. Lander. 2003. "Identification of a Gene Causing Human Cytochrome C Oxidase Deficiency by Integrative Genomics." *Proceedings of the National Academy of Sciences of the United States of America* 100: 605-10.

Ochsner, S. A., D. L. Steffen, C. J. Stoeckert, and N. J. McKenna. 2008. "Much Room for Improvement in Deposition Rates of Expression Microarray Datasets." *Nature* Methods 5: 991-91.

NCBI. 2007. “Microarrays Factsheet”, accessed November 23, 2011, http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html

Piwowar, Heather, and W. Chapman. 2010. "Recall and Bias of Retrieving Gene Expression Microarray Datasets through Pubmed Identifiers." *Journal of Biomedical Discovery and Collaboration* 5: 7-20.

Rocca-Serra, P., M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. D. Tong, and S. A. Sansone. 2010. "Isa Software Suite: Supporting Standards-Compliant Experimental Annotation and Enabling Curation at the Community Level." *Bioinformatics* 26: 2354-56.

Ventura, B. 2005. "Mandatory Submission of Microarray Data to Public Repositories: How Is It Working?" *Physiological Genomics* 20: 153-56.

Disclosure: The author reports no conflicts of interest.

ISSN 2161-3974