Towards a Common Informatics Framework for Biorepositories

Paul A. Fearn, Nicholas R. Anderson, James F. Brinkley, Peter Tarczy-Hornoch
Biomedical and Health Informatics, University of Washington, Seattle, WA

Abstract

Many biomedical research laboratories, departments and organizations struggle to manage data in biospecimen repositories that supply basic and translational research. Biorepository information systems have been developed from a variety of perspectives, and are often difficult to integrate or network within and across organizations due to lack of structural and semantic alignment with standards.

Biospecimen science is the study of the collection, processing and handling factors that affect the quality and characteristics of samples, including the their effects on the results and reproducibility of biological and biomedical investigations. To account and control for variation in samples, biorepository systems need to incorporate both workflow support and provenance information.

By leveraging existing and emerging best practices, data models, data exchange formats and vocabularies, informatics can facilitate and advance the quality and reproducibility of research. This paper reviews and synthesizes requirements and standards for biorepositories and biospecimen science, and proposes a common framework.

Introduction

Clinical and translational research, particularly research involving genomic, proteomic and metabolomic analysis, is dependent on a ready supply of high-quality human biospecimens and associated clinical data. Biorepositories that make human tissue, blood, cells, and fluids available for research are essential suppliers for academic and commercial consumers, who use these materials to develop and improve diagnostics and treatments for a wide variety of diseases. Because of the increased need to enhance these samples with associated clinical and processing data, biorepositories have become a nexus for local and national development of research systems and policies. As the scale, complexity and strategic importance of biorepositories increases, development and implementation of sustainable informatics infrastructure to support them has become a significant challenge for biomedical research laboratories, departments and organizations.

In 2009, biobanks were recognized in Time Magazine as one of the top ideas that are changing the world.[1] Large grants such as the National Cancer Institute (NCI) designated cancer centers and Specialized Programs of Research Excellence (SPORE) have funded development and support of numerous tissue and data resources.[2-3] National efforts to develop biospecimen and data repositories are emphasized in the National Institute of Health (NIH) Roadmap, and are a priority focus of the Clinical and Translational Science Awards initiative, as well as the American Recovery and the Reinvestment Act of 2009. Large biorepository development efforts have been funded and widely publicized, including the $25M Kaiser-Permanente Research Program on Genes, Environment and Health, and the $60M NCI cancer human biobank.[4-5] Because a significant amount of biomedical research relies on the supply of high quality human samples with associated clinical and protocol data, biorepositories have become a nexus for local and national development of research systems and policies. As the scale, complexity and strategic importance of biorepositories increases, development and implementation of sustainable informatics infrastructure to support them has become a significant challenge for biomedical research laboratories, departments and organizations.

The most acute problems and requirements for biorepository informatics are prospective workflow management, documentation according to standard operating procedures (SOPs), and retrospective determination of sample provenance and pre-analytic variation. Our goal is to describe the problem scope, the relevant standards, and how existing and emerging standards can be used in a common informatics framework to capture, store, retrieve and link information from the entire specimen lifecycle. We propose that the informatics framework should support 1) retrospective information retrieval for quality assurance, biological and biomedical research, and biospecimen science; 2) prospective data and work flow of biorepository operations; 3) full tracking of sample provenance; and 4) patient privacy through reflexive stewardship. The framework should help identify trends, gaps and opportunities for research and development.

Biospecimen Science. Biospecimen science is the study of the collection, processing, and handling factors that affect the characteristics and quality of samples, including the their effects on the results and reproducibility of biological and biomedical investigations, and the NCI Office of Biorepositories

and Biospecimen Research (OBBR) was created to systematically study and address these issues.[6] Until we have evidence based protocols to manage of sample factors that impact sensitive molecular studies, the vision for genomics and personalized medicine will not be translated into routine clinical practice and patient care.[7]

Differences in sample SOPs have been implicated in the variation in results of biological and biomedical investigations, but it is difficult to find data to explain the variation because of lack of accessible, consistent and complete documentation. With inadequate information systems, investigators must query clinical or repository databases, search though paper files, and seek information from other staff to obtain provenance and pre-analytic variation information about individual samples. Projects that require data sharing with other laboratories, departments or extramural groups experience information problems more acutely, impeding collaborative research. Requests for data sharing may provoke reactive cultural and organizational issues, as well as intellectual property, competition, human subjects, privacy, security, and project sponsor issues.[8] While NIH and other federally funded agencies urge data sharing, individual laboratories, departments, sponsors and cooperative groups traditionally maintain independent datasets with strict access policies. At many sites, determining sample provenance or comparing sample processing to support biospecimen science or quality control (QC) is a time consuming manual task involving document review, discussion, and educated guesswork.

Since the publication of the National Biospecimen Network Blueprint in 2003, there has been a surge in literature about the importance of biospecimen information for translational research.[9] The Blueprint highlighted the importance of biorepository informatics, and recent work has affirmed the importance and the critical contribution of information systems to support biospecimen research. Moreover, growth of participation in data sharing networks is a strong trend in biorepository informatics. Larger local, regional and national research networks such as caBIG, CTSA and SPORE collaborations are likely to drive biorepository informatics requirements. Widespread participation in biorepository data and materials sharing networks will require further development and adoption of formal standards for that support interoperability. Although there has been some work in integration of biorepositories through the NCI caBIG project, there are still gaps in the implementation of best practices, common data models, data exchange formats and ontologies that describe sample provenance, processing and pre-analytic variation.

Biorepository Operations / Workflows. The confluence of workflows, diversity and complexity of information, staffing and budget constraints in many biorepositories make this a challenging environment for system and process redesign, and the difficulties are compounded in higher-volume, faster-paced operations. Although people and processes may be stressed or disrupted by innovations in biorepository informatics, the current information problems in biorepositories are in a state where intervention and improvement is necessary.

A number of common workflow questions arise in advancing informatics support of biospecimen science. What is the history of the source or donor prior to and following collection of a biospecimen? Which samples shared a common feature of provenance or processing (i.e. storage container, centrifuge, batch of additive, date of processing, technician)? Which experimental results originate from a particular source, process or batch of samples? Ultimately the entire biorepository research supply chain from biospecimen collection to experimental methods and results must be tracked to establish an explicit and formal informatics framework.

Existing Standards. Over the past decade, communities of scientists who investigate tissue, blood, cells and sample derivatives have labored to develop data and reporting standards.[10] Vocabularies, data exchange formats, and object models for genomics and proteomics experiments have evolved and matured, and informatics researchers have developed data integration methods, ontologies, and tools to facilitate integration of biological data. Sponsored and grassroots efforts have developed popular standards that are applicable to this domain. Given the strength and coverage of existing standards and community efforts, there does not appear to be a need to develop new standards for biorepository and biospecimen information management. However, many people are still unfamiliar with existing standards and tools for this domain and how they might be integrated. This paper fills a gap in the literature by synthesizing requirements, existing standards and reference systems for this important but relatively unstudied domain of biomedical informatics.

Approach

To develop an informatics framework for biorepositories and biospecimen science, it was necessary to systematically explore the boundaries and intersecting aspects of this domain. The overall approach for framework development was to review literature using a tentative framework that spans from

source of specimen to biorepositories through to publication of the results of studies using samples.

Conceptually, we should be able track detailed provenance of samples from the original source, population and environment through collection, processing and handling, to the consumption of those samples in biomedical investigations. Moreover, we should be able to track data, information and knowledge produced from investigations into publically accessible databases and biomedical literature. Ideally, we should be able to track consumption of samples in technology and product development. Figure 1 is a conceptual model depicting five components organized along boundaries between biorepository data and systems.

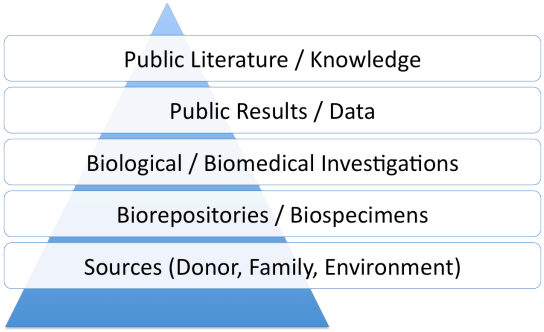

Figure 1: Five informatics framework *components*

Component of Proposed Informatics Framework

Public Literature / Knowledge. Ideally, a supply of high quality specimens should end up in experiments and investigations that yield public data, knowledge, technology and validated products. To measure the expression of consumed biospecimens in the public literature, we need to systematically mine the literature for articles resulting from sample analysis. The volume and attributes of articles by time and by search terms from specimen types, analytes and analysis platforms can be used to generate a dashboard of for funding agencies and individual repositories to measure their impact.

To ensure that biospecimen science could be supported and reflected by this informatics framework component, the NCI OBBR Biospecimen Research Database (BRD) was used to identify sample types, analytes and technologies for biological and biomedical research using specimens.[6] From BRD terms and discussions with biologists and biomedical researchers, a search of PubMed was performed to measure the overall volume of literature generated by different types of specimens per year from 2000 to 2008. Other output generated from biospecimens (i.e. new biotechnology and validated pharmaceutical products) were not assessed in this project. The frequency of articles by specimen type, technology and analyte is a proof of concept and proxy measure of relative impact of specimens in public knowledge. Figure 2 shows that studies of protein and peptides produced orders of magnitude more public literature than RNA and DNA over the last decade. In terms of specimen types, tissue produced twice the amount of literature as blood, and for technology, polymerase chain reaction (PCR), immunohistochemistry, and spectrophotometry were the top three platforms, each associated with more than twice the articles generated from DNA microarray analysis. Further work is needed to refine search strategy, assess the performance of these metrics compared with documented utilization of specimens from individual biorepositories, and to assess the ability to track detailed provenance of samples from literature back to public databases, repository systems, and original sources / donors.

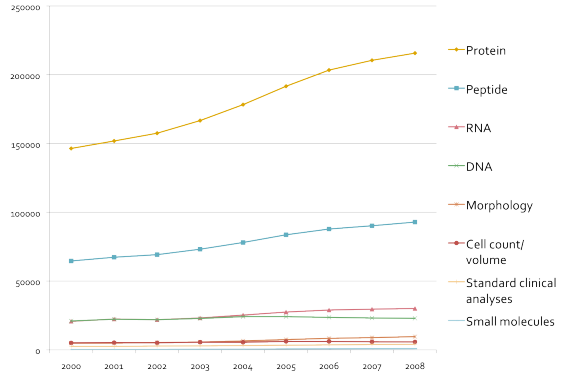

Figure 2: Analytes by year in PubMed, 2000-2008

Public Results / Data. Some (hopefully large) portion of data produced from biological and biomedical investigations and experiments using biospecimens ends up in public databases (i.e. GenBank, RefSeq, Peptidome, GEO, Protein Data Bank, UniProt).[11] As with the literature metrics, the volume and types of data in these data sources can be mined to estimate trends in output of biorepositories and utilization of specimens. The BRD categories for specimen type, analyte and technology platform can be organized and employed to index and search public datasets. Can we also track provenance and biospecimen collection, processing and handling associated with these results data? Developing metrics and dashboards to evaluate and monitor impact in public databases for biological and biomedical research, and assessing the provenance of these data back to repositories and sources / donors is the second component of an informatics framework to support biorepositories and biospecimen science.

Biological / Biomedical Investigations. Over the past decade, the biological and biomedical research

communities have engaged in a wave of standards development. The bellwether was the MIAME standard for DNA microarray experiments, which describes guidelines for reporting experimental details and data.[12] The Minimum Information for Biological and Biomedical Investigations (MIBBI) project is an umbrella for data and experiment reporting guidelines and checklists, and they roughly correspond to BRD technology platforms.[10] The informatics implementations of MIBBI guidelines and checklists have recently converged around a subset of object models, data exchange formats and ontologies. While coverage of potential usage of biospecimens is incomplete in these guidelines and standards, they are an ongoing community based model and reference for biospecimen science. Most of these standards describe sample collection, processing and handling parameters, but the annotation and treatment of sample characteristics under the MIBBI umbrella varies. There is a need and opportunity to analyze, harmonize and fill gaps to provide complete coverage and mapping between biospecimens, the BRD and the standards for reporting biological and biomedical investigations.

The Functional Genomics Experiment (FuGE) UML object model has emerged as the dominant data structure under the MIBBI umbrella.[13] This model explicitly deals with biological materials (biospecimens) and the protocols by which they are treated and manipulated in an experiment. The FuGE model could potentially be used to codify biorepository SOPs that describe the processes for collecting, processing and handling biospecimens.

The dominant data exchange format emerging from the research community is ISA-TAB.[14] ISA-TAB is a spreadsheet based (TAB) format for exchanging information from a variety of experiments. XML based data exchange formats are also common, but many investigators and laboratories do not have sufficient informatics support to implement and use XML based systems. TAB based formats that can be easily implemented and are increasingly popular.

Controlled vocabularies or ontologies have been developed to support a variety of investigations and technology platforms. There is an ongoing effort to consolidate biological and biomedical investigation terms in the Ontology for Biomedical Investigations (OBI), which is a participant in larger community driven efforts to refine and organize biomedical ontology efforts such as OBO Foundry.[15]

The MIBBI checklists, FuGE, ISA-TAB and OBI describe good practices and standards. Efforts to integrate these standards are ongoing, and would make an excellent foundation for biospecimen science. However, these leading standards from biological and biomedical investigation have not yet propagated into biorepository management systems.

Biorepositories / Biospecimens. Informatics for biorepositories comes from distinct lineages: clinical systems for anatomic pathology and laboratory medicine, and research based laboratory information management systems (LIMS) or biorepository management systems. Conflicting requirements and design patterns present a challenge for the development of a common framework.

The International Society of Biological and Environmental Repositories (ISBER) and the OBBR have published best practices for biorepositories, which include recommendations for informatics.[6,16] These recommendations include tracking of unique identifiers / barcodes, timestamps, collection, handling and processing information, QA/QC data, images, containers and locations, IRB and consent documents, stewardship information about patient / donor preferences and other organizational or regulatory constraints. Tracking of costs related to specimens is crucial for developing and managing sustainable financial models for repositories.

There are a number of good commercial repository LIMS available that can meet the ISBER and OBBR recommendations. To support efficient workflows in high volume biorepositories, it is generally necessary to interface or integrate with anatomic and clinical pathology information systems, which tend to converge on College of American Pathologist (CAP) protocols, as well as standard clinical terminologies (i.e. SNOMED, LOINC, ICD) which are indexed under the UMLS Metathesaurus umbrella.[17]

Through caBIG program funding, caTissue Suite has been developed as an open source reference system to manage biorepositories and support biospecimen science.[18] The caLIMS2 system is under development to link caTissue with biological and biomedical investigations data (i.e. caArray). Several vendors have collaboratively developed a minimal Common Biorepository Model (CBM) that would allow existing commercial and open source LIMS and biorepository systems to share data for research. The CBM standard defines parameters for *specimen locators* (e.g. NCI Specimen Locator) that allow searching for specimens within and across repositories.[19-20] Along with FuGE, the UML object models from caTissue, caLIMS2, CBM, and caArray are a good starting point and can be extended and harmonize as part of an overall framework.

In terms of vocabularies or ontologies, existing standards under the UMLS umbrella can be leveraged for biorepositories and biospecimen science (e.g. MeSH, NCIt, LOINC, SNOMED). As such, we believe that there is no need to develop new

vocabulary standards or ontologies.

Sources (Donor / Family / Environment). Guidelines and standards for capturing and storing clinical data from human donors and patients have been published by CDISC, and much work has been done to develop clinical data repositories, enterprise data warehouses, and data marts that can provide adequate clinical annotation for samples.[21] However, there is a evolving body of work on the impact on the patient and community as a result of inclusion – consented or not, into biorepositories with significant potential secondary or tertiary use.[8] Though these issues are difficult to encode within a structured descriptive standard, it is likely that the visibility of resulting standards may include patient reported information – which may in turn effect the routes to making samples available in large-scale data sharing networks.

Conclusions

There is no need to develop new data models, data exchange formats, vocabularies, ontologies or guidelines for systems to support biorepositories and biospecimen science. Rather, existing standards from public databases, biological and biomedical investigations (MIBBI, FuGE, ISA-TAB, OBI), caBIG (caTissue, caLIMS2, caArray, NCIt), BRD, UMLS and CDISC can be leveraged, extended, integrated and in some cases refactored. To facilitate biospecimen science within existing pathology systems, LIMS and biorepository management system we need to align existing informatics tools and standards to track complete provenance and impact, from the source to public databases and literature. To achieve widespread adoption of this informatics framework, we will need to keep costs and barriers to use low, and balance detailed data collection for research with the reality of time and resource pressures / constraints in clinical and research operations. With a driving need to account for and control variables that impact downstream research, and to support acceleration of reproducible, inter-institutional collaborative research, we need a common informatics framework for biorepository systems and biospecimen science.

Acknowledgements

This research is supported in part by the NLM training (NIH NLM #T15 LM07442) and ITHS (NIH NCRR 1 UL1 RR 025014) grants.

References

1. Parks A. 10 Ideas Changing the World Right Now. Time Magazine, Mar 12, 2009.
2. NCI Cancer Centers Program. [cited 2009 Nov 17]; http://cancercenters.cancer.gov/.
3. Specialized Programs of Research Excellence. [cited 2009 Nov 17]; http://spores.nci.nih.gov/.
4. Kaiser Permanente "Biobank" Receives $2 Million Grant from National Institutes of Health. [cited 2009 Oct 22]; Available from: http://www.rwjf.org/grants/product.jsp?id=49613.
5. National Cancer Institute caHUB ARRA funding. [cited 2009 Oct 21]; Available from: http://www.cancer.gov/.recoveryfunding/page10.
6. NCI Office of Biorepositories and Biospecimen Research. [cited 2009 Oct 21]; Available from: http://biospecimens.cancer.gov/.
7. Compton C. Getting to personalized cancer medicine: taking out the garbage. Cancer. 2007;110(8):1641-3.
8. Fullerton SM, Anderon NR, Guzauskas G, et al. Meeting the Governance Challenges of Next-Generation Biorepository Research. Science Translational Medicine. 2010;2(15):1-4.
9. National Biospecimen Network Blueprint. Constella Group. Sep 2003.
10. Taylor C, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol.2008:26(8)889-96.
11. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2010;38(Database):D5-D16.
12. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001;29(4):365-71.
13. Jones AR, Lister AL, Hermida L, et al. Modeling and managing experimental data using FuGE. OMICS. 2009;13(3):239-5.
14. Release Candidate 1, ISA-TAB v1 (Nov 2008). [cited 2010 Mar 15]; Available from: http://isatab.sourceforge.net/specifications.html.
15. Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007;25(11):1251-5.
16. ISBER. [cited 2010 Mar 15]; Available from: http://www.isber.org/.
17. Unified Medical Language System (UMLS). [cited 2010 Mar 14]; Available from: http://www.nlm.nih.gov/research/umls/.
18. Cancer Biomedical Informatics Grid (caBIG). [cited 2010 Mar 14]; https://cabig.nci.nih.gov/.
19. NCI Specimen Resource Locator. [cited 2010 Feb 28]; http://biospecimens.cancer.gov/locator.
20. NIH Human Biospecimen Database. [cited 2010 Feb 28]; http://biospecimens.ordr.info.nih.gov/.
21. Clinical Data Interchange Standards Consortium (CDISC). [cited 2010 Mar 14]; http://www.cdisc.org/.