BIOMEDICAL DATA SHARING, SECURITY AND STANDARDS

Belinda Seto and James Luo*

National Institute of Biomedical Imaging and Bioengineering, 6707 Democracy Blvd, Suite 202, Bethesda, Maryland 20892-5477
E-mail: setob@mail.nih.gov*; luoja@mail.nih.gov

ABSTRACT

The National Institutes of Health (NIH) implemented a policy on data sharing in 2003. The policy reaffirmed the principle that data should be made as widely and freely available as possible while safeguarding the privacy of research participants, and protecting confidential and proprietary data. Restricted availability of unique resources upon which further studies are dependent can impede the advancement of research and the delivery of medical care. Therefore, research data supported with NIH funds should be made readily available for research purposes to qualified individuals within the scientific community.

One approach to sharing data is to establish a network of databases. However, there are a number of barriers to creating successful networks, which can include fundamental differences in informatics infrastructure and communication tools used at various research sites. Solutions will entail standards for data collection, processing, and archiving to allow interoperability among the databases and the ability to query data across databases. Open architectures for data collection as well as software to facilitate communication across different databases are needed.

An important requirement for sharing data is the protection of the privacy of individuals who participate in the research and the data. Privacy protection hinges on maintaining the confidentiality of the data and the security of the databases. There must be clear policies for data security, which may include data encryption, coding, and establishing limited access or a tiered approach to data access.

Keywords: Data sharing policy, Data confidentiality, Interoperability, Open source, Open architecture, Security

1 INTRODUCTION

Progress in scientific research depends on the free flow of information and ideas. As a matter of policy, the NIH is explicit in stating that “restricted availability of unique resources upon which further studies are dependent can impede the advancement of research and the delivery of medical care”. To ensure that future research can build on previous efforts and discoveries, the NIH has developed a data sharing policy that has been in effect since October 1, 2003. The policy expects final research data, especially unique data, from NIH-supported research efforts to be made available to other investigators. In implementing this policy, NIH is cognizant of the need to protect the privacy of individuals and thus data security becomes paramount as one considers means to share data. Successful implementation of this policy is also dependent on technology needs such as software tools and database architecture issues.

2 DATA SHARING: PRIVACY CONCERNS AND SHARING METHODS

Protecting the privacy of human participants in research studies should be a top priority for any researcher. Investigators, Institutional Review Boards (or bioethics review), and research institutions have an obligation to protect participants’ rights and welfare, including individual privacy protection and confidentiality of data. Privacy and confidentiality are particularly important for studies with very small sample size. Steps should be taken to avoid inadvertent identification of participants through deductive approaches when the sample size is small. For example, for a study involving a small community, one might be able to identify a participant based on just a few personal characteristics or attributes, without even knowing the individual’s name, address or telephone number. Similarly, there are caveats in sharing data from studies collecting sensitive data. However, even in these situations data sharing is possible without compromising confidentiality, provided that identifiers are removed from data. In addition, data sharing agreements can be used to restrict the transfer of data to others and to specify the appropriate uses for shared data.

Investigators should take into consideration possible restrictions from local, State, and Federal laws, such as the Privacy Rule, a U.S. Federal regulation under the Health Insurance Portability and Accountability Act (HIPAA). The HIPAA privacy rule mandates that an individual's written authorization is required for the use or the disclosure of protected health information unless the requirement is waived by an institution's privacy board. Furthermore, a decedent's information is protected, even though for obvious reason, his/her authorization cannot be obtained. In these situations, the next-of-kin must be contacted for authorization. For research purposes, researchers may obtain authorization to use protected health information if the information is used for a specific research study; HIPAA does not allow for future unspecified research use. It should be noted that authorization may be given to create a repository or a database. Research use of protected health information without authorization is allowed if the protected information has been completely de-identified. De-identification involves removal of all 18 identifiers as defined by HIPAA, or statistically, such that a statistician certifies that there is a very small risk that the information could be used to identify an individual. HIPAA also affords some flexibility to use a limited data set for research purposes under a data use agreement. With such an agreement, limited types of identifiers such as name of city, or state, Zip codes, and elements of dates, can be released with protected health information, but no unique identifiers can be released.

Data sharing can be accomplished through a number of methods. The most common method is publishing articles in scientific publications. Researchers also share data through an informal channel, by responding directly to data requests. However, as the need for data confidentiality increases, the methods of sharing become more stringent. These may take the form of a data enclave involving controlled, secure environments in which eligible researchers can perform analyses using data resources. Alternatively, data archives can be used where machine-readable data are acquired, manipulated, documented and distributed. A combination of these methods can be used to share data, each providing a different level of access.

An example of sharing sensitive research data comes from an NIH-supported survey, questionnaire study, the National Longitudinal Study of Adolescent Health. The study involved more than 20,700 adolescents in grades 7-12, who were followed from 1994 to 2002, as well as their parents and the school administrators. Independent data were collected on the neighborhoods and communities where these schools were located. Measures of health, health-related behaviors such as sexual activity and drug use, as well as determinants of health at the individual, family school, peer group and community level were included in the questionnaire. On the one hand, the challenges to data sharing from this study were the sensitive nature of the responses, the need to protect individual privacy and the danger of deductive disclosure. On the other hand, the benefits and rationale for sharing research data from this study were overwhelming because the scope of the large study made it unlikely to be replicated because of the substantial costs. The potential of learning much more beyond the primary outcome assessment of adolescent behavior and strategies for interventions was significant.

The solution to the risks of data sharing was to develop a multi-tiered system for data sharing: public-use data, contractual data sets and a "cold room" for on-site data use. Public-use data were made available through a data archive managed by a contractor. The public data set contained a small fraction of the entire data, approximately 6,500 cases where identifying information was redacted. Data from small populations that were over-sampled, such as ethnic groups, were not made publicly available. At the next tier of data security, i.e, under a data-use contractual agreement, the full data set was made available to researchers if the IRB approved the data security plan. The data-use contract must remain active with a signed agreement and the requesters must agree to cover the costs of providing the data. At the highest level of data security, a cold room was established by the NIH at the site of the grantee institution.

3 INTEROPERABILITY

Sharing data between databases requires interoperability of data formats, standards, and data types. When these elements are standardized, researchers can access and use heterogeneous information, such as molecular biology, DNA and protein sequence, genomic, proteomic, micro-array, clinical and biomedical imaging, just to name a few.

A major incentive of bioinformatics is to establish agreed-upon standards for file formats and data exchange protocols.

Vocabulary and ontology provide a common set of words and a common context and specific meanings for the words to integrate data across databases and write general purpose software for data processing. Ontology, as a computable, machine-interpretable language to represent biology, can facilitate the semantic interoperability of biological data in different domains such as genomic or clinical studies.

The NIH-funded Gene Ontology Consortium has stimulated the development and adaptation of tools for accessing databases and mapping gene products to ontology terms. These tools enable us to query the databases in a consistent manner, based on a common understanding of the definition of querying terms. Another example of ontology development and application at the NIH is the National Cancer Institute's Cancer Bioinformatics Grid (caBIG). The caBIG vocabularies are based on the NCI thesaurus. Repositories of common data elements provide data standards, including the development, promotion and support of vocabularies, and ontology to ensure that the entire caBIG community is speaking the same "language". The caBIG infrastructure achieves data sharing and interoperability through a federated model, providing a platform for researchers to access a rich collection of biomedical data using informatics tools to integrate diverse data types. Thus caBIG serves as a model for sharing and accessing data in a federated database paradigm. Programming and messaging interfaces (APIs) based on the standard vocabularies, ontology and common data elements permit sharing and exchange of data using a variety of software tools.

4 OPEN SOURCE AND OPEN ARCHITECTURE IN BIOINFORMATICS

The National Institutes of Health strongly urge their funded investigators to comply with an open-source software philosophy. The essence of an open source development model is the creation of solutions within an open, collaborative environment, with participation of the developers and the end-users. In this working model, the community (both the software developers and end users) work collaboratively to develop requirements and solutions. Each community can read, redistribute, adapt, fix bugs, modify and improve the source code for a piece of software as the software evolves. In contrast to the traditional closed development model where only a few programmers can see the source code, the open-source approach uses collective knowledge and experience of a wide community. Such collaboration promotes a higher standard of quality and validation. Open-source development facilitates the long-term viability of both data and applications that address the communities' needs. Commonly cited reasons for the growing interest, acceptance and even preference for open-source products include low cost, high value, quality and reliability, security, increased freedom and flexibility (for both hardware and software) and adherence to open standards.

Open source, as defined in the Open Source Initiative (http://www.opensource.org/docs/definition.php), is not limited to open access to the source code. The definition states that "the distribution license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources". The source code as well as the compiled program must be publicly available. One must be able to modify the open-source software and the license must allow derivatives to be distributed under the same terms as the license of the original software.

Stajich and Lapp [1] identified freely available source code for 275 projects that are intended to address a specific problem or problem domain of biological sciences using open source. Bioinformatics.org (http://bioinformatics.org) is an organization that provides and promotes open access to biological data, as well as free and open-source software. There are 314 groups or projects and more than 21,000 members have participated in Bioinformatics.org activities. Sourceforge (http://www.sourceforge.net) is the world's largest open-source software development web site, hosting more than 100,000 projects and over 1,000,000 registered users, with a centralized resource for managing projects, issues, communications and codes. It hosts approximately 750 projects under the category of "Bioinformatics", which includes projects such as the Generic Model Organism Database (GMOD), the Microarray Gene Expression Data (MGED) society and Life Sciences Identifiers (LSID). There are many projects hosted by the developers' home institutions or by other open source umbrella organizations, such as the Open Bioinformatics Foundation (http://www.open-bio.org).

The Biomedical Informatics Research Network (BIRN), supported by the National Center for Research Resources at the NIH, is a geographically distributed virtual community of shared informatics resources. It is committed to an open-source policy. BIRN provides the research community with a cyber-infrastructure, as well as software and data products. The cyber-infrastructure consists of a cohesive implementation of key information technologies and applications specifically designed to support biomedical scientists. These resources are freely available to the biomedical community via BIRN's Tools & Data page (http://www.nbirn.net/downloads/index.shtm).

Open-source software libraries also provide fundamental building blocks for algorithms and applications. An example is the Visualization ToolKit (VTK), which is an open source, freely available software system for 3D computer graphics, image processing and visualization. Thousands of researchers and developers around the world have utilized this tool for various purposes.

The open-source philosophy has become mainstream in bioinformatics. The examples described above share a common program feature of being well managed and executed, and result in robust and high-quality bioinformatics products. A flexible open architecture and well-tested, harmonized software codes are some of the critical factors that contribute to the success such open-source projects.

5 REFERENCES

1. Stajich JE and Lapp H. 2006. Open source tools and toolkits for bioinformatics: significance, and where are we? *Brief Bioinform. 7(3):*287-96.

2. GenBank web site: http://www.ncbi.nlm.nih.gov/Genbank/. Retrieved from the WWW May 8, 2007.

3. Bioinformatics.org web site: http://bioinformatics.org. Retrieved from the WWW May 8, 2007.

4. Open Source Initiative (OSI) web site: http://www.opensource.org/ . Retrieved from the WWW May 8, 2007.

5. Source forge web site: http://www.sourceforge.net. Retrieved from the WWW May 8, 2007.

6. The NCI Center for Bioinformatics (NCICB) web site: http://ncicb.nci.nih.gov/ . Retrieved from the WWW May 8, 2007

7. The Biomedical Informatics Research Network (BIRN) Tools & Data page: http://www.nbirn.net/downloads/index.shtm . Retrieved from the WWW May 8, 2007.