Understanding Metadata Functional Requirements in Genome Curation Work

Hong Huang
School of Information
University of South Florida,
Tampa, FL, 33620
honghuang@usf.edu

Jian Qin
School of Information Studies,
Syracuse University,
Syracuse, NY 13244
jqin@syr.edu

ABSTRACT

The proliferation of genomic data and their widespread data reuse pose new challenges to effectively manage and curate genomic data. This study contributes towards better understanding of 156 genomics scientists' perception and priorities for metadata functional requirements in genome curation work. Our study was guided by previously identified twenty two metadata functional requirements (Willis, Greenberg, & White, 2012), and intended to define a context-sensitive model of groupings for metadata goals in genome curation. Analysis of the results revealed that genomics scientists recognize specific sets of metadata functional requirements in the genome-curation context. These metadata goals were reduced to six factor constructs. The rankings of these constructs in decreasing order are Portability, Reusability, Manipulability, Sufficiency, Interoperability, and Modularity. The findings indicated that genomics scientists need both domain independent and dependent metadata functional requirements that are primarily related to data comparison, integration, and reuse across platforms and databases. The constructs defined by this study advance the understanding of metadata requirements and their relationships. In addition, the resulting metadata requirement model can serve as a valuable resource to genome scientists, data curators and administrators for designing metadata schemes and developing data-curation policies.

Keywords: Genome curation, metadata goals, metadata schemes, domain

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

INTRODUCTION

The increasing use of ultra-high-throughput sequencing methods generated a massive amount of raw genomic sequencing data. Metadata describing the sequencing data were attached to them to support data discovery, access, reuse, and dissemination functions standardized curation procedures, tools, and data quality models were developed to support curation work and meet the needs of the genomics research community (Klimke et al., 2011; Huang et al., 2012; Yang et al., 2011). The complexity of the genomic data and related yet disparate information resources poses challenges in developing metadata schemes that are both easy to use and effective in facilitating genomic data documentation, exchange, archiving, and reuse (Pagani et al, 2012).

The creations of genomic metadata and related schemes are dynamic and typically done on an ad hoc basis. The genomics research community has accumulated a long list of curation tools and metadata schemes that support the constant changes of curation needs (Klimke et al., 2011). Some of these schemes are incomplete or too specific for local databases. They have to be constantly updated by adding new metadata or dropping the out-date ones. Better metadata products are imperative to support the metadata practice and curation needs.

Identification of the desired uses of metadata helps develop goal-oriented metadata schemes. Metadata functional goals (e.g., scheme flexibility, extensibility etc.) across domains were identified by reviewing the literature on discipline-specific metadata schemes and related community driven

activities (Willis, Greenberg, & White, 2012). Many of them were domain-independent, but they reflected different data processes and requirements across domains; significant relationships were also found between the domains and objectives of these metadata schemes (Willis, Greenberg, & White, 2012). Each domain might have its own configuration of metadata classification. Additionally, the metadata schemes and their data elements need to be designed to suit specific metadata needs of a community (White, 2010). For instance, the genomics community had defined a minimal set of core metadata typically for genome description (Pagani et al., 2012).

This survey study was designed to gather metadata functional goals/requirements for genomic data curation and sharing. The responses from genomic researchers and data curators not only offer valuable insights into the metadata requirements but also help understand the goals, motivations, and unique practice for organizing their data. These goals could guide the community to develop domain-feasible metadata schemes that can enhance data discovery and reuse, and maximize the capacity for data sharing across disciplines (Qin, Bell, & Greenberg, 2012).

METHODS

The study used a survey questionnaire to collect data, which was constructed based on the twenty two metadata functional requirements identified in Willis, Greenberg, & White (2012). To provide the context for the questions, the survey used a representative scenario to conceptualize the activities related to metadata scheme development in genome data curation. The population for this study consisted of scientists that conduct genomic research and are familiar with metadata issues and practices. A total of 4,012 authors (with their email addresses) of the genomic research articles related to genome annotation and metadata practice from 2010/12/01 to 2012/12/01 were extracted from the PubMed database, 800 of which were randomly selected for inclusion in this study. The survey finally collected 156 responses with a response rate as 20%. The Qualtrics software (http://www.qualtrics.com) was used to distribute the survey and collect data. Respondents were asked to rate the importance of metadata functional requirements on a seven-point Likert scale. The descriptive statistics and factor analysis reports were generated by the statistical software SPSS.

FINDINGS

The survey received 156 responses. 84% (n=131) of the participants indicated that they had experience with applying metadata or standard vocabularies in their research work. They self-identified their curation roles as follows: end users (n=104, 67%) and curators (n=52, 33%). Over half of the participants (n=88, 58%) had a biology background, worked in higher education in the U.S. or Canada (n=89, 57%), and held a doctorate (n=117, 75%).

Table 1. Survey participants' ranking of metadata requirements by mean importance.

Metadata requirements	*# of responses*	*Mean*
Data comparability	150	6.14
Data portability	153	5.92
Data retrieval	148	5.76
Scheme simplicity	150	5.75
Data interchange	147	5.73
Data publication	147	5.46
Data validation	145	5.39
Data documentation	146	5.38
Data archiving	146	5.34
Scheme extensibility	144	5.30
Sufficiency (minimal set)	145	5.25
Technical stability	148	5.23
Provenance	145	5.21
Inter-scheme modularity	146	5.16
Conceptual stability	149	5.07
Data lifecycle	149	5.01
Scheme harmonization	144	4.98
Element refinement	143	4.82
Comprehensiveness	143	4.78
Intra-scheme modularity	144	4.76
Core set	148	4.75
Scheme flexibility	145	4.73
Abstraction	142	4.34

Note. Responses of "not applicable" were not included in the statistical analysis.

The descriptive statistics of the metadata requirement rankings are given in Table 1. Mean was calculated for each metadata functional requirement. On average, the participants ranked Data comparability as being of the highest importance and Abstraction the lowest. Other functional goals among the top-five list are Data portability, Data retrieval, Scheme simplicity, and Data interchange. This reflected that the curated genome data was expected to be unique and highly heterogeneous and in a large scale.

Multiple metadata schemes exist in disparate genomic databases and repositories (Klimke et al., 2011). Scientists need to develop metadata as simple and convenient ways to access, compare and collect, and integrate disparate pieces of scientific data sets across the databases and repositories (Qin, Bell, & Greenberg, 2012).

To identify the grouping structure for metadata functional goals reflected by the 156 survey respondents' rankings, the study conducted exploratory factor analysis using principal-components analysis as the extraction method and varimax with Kaiser normalization as the rotation method. The factor loading revealed that these metadata functional requirements could be reduced into six factor constructs. Table 2 showed these six factor constructs and their arithmetic average of the mean ratings.

Table 2. The six factor constructs generated from the metadata requirements, and ranked by the arithmetic averages of the mean ratings of the individual requirements loaded on the constructs.

Metadata constructs	*Avg*	*Metadata requirements*
Portability	5.49	Conceptual stability, data portability, Scheme simplicity, Technical stability
Reusability	5.45	Data lifecyle, Data archiving, Data publication, Data interchange, Data retrieval, Data documentation
Manipulability	5.33	Data comparability, Element refinement, Scheme harmonization, Data validation
Sufficiency	5.00	Core set, Sufficiency (minimal set)
Interoperability	4.87	Comprehensiveness, Provenance, Scheme extensibility, Scheme flexibility
Modularity	4.86	Inter-scheme modularity, Abstraction, Intra-scheme modularity

Table 2 showed that Portability construct was ranked the highest, followed by the Reusability, Manipulability, Sufficiency, Interoperability, and Modularity constructs. This indicates that the goal for genomic community is to capture related metadata easily, and the scheme should be technically stable and independent across software tools and operation systems. If the metadata schemes are complicated and have a deep-layered structure, this makes the automatic metadata generation difficult (Qin, Bell, & Greenberg, 2012).

Data reusability construct was ranked the second important in the genome curation metadata requirement model including the goals of Data lifecyle, Data archiving, Data publication, Data interchange, Data retrieval, and Data documentation. All of these are fundamental for metadata schemes to facilitate data sharing, exchange, use, reuse and preservation (Willis, Greenberg, & White, 2012). Historically, the genomics community may not create genome curation related metadata for global sharing. Many of these metadata were stored locally and disparately, and were designed by following the community's data standards (Huang, Stvilia, Jörgensen & Bass, 2012). This ranking result suggests that the metadata is expected to support the function of data reuse.

Manipulability construct, ranked as the third, included the requirements (e.g., Data comparability, Validation) which refer to different data practice across disciplines. In the genomic data curation process, scientists have to pull out data from various databases, compared metadata from different schemes, and validate related metadata elements. These manipulability function will be required to develop disciplinarily feasible metadata schemes to facilitate data processing and analysis.

Other constructs such as Sufficiency, Interoperability, and Modularity were related to the metadata goals that are applicable among metadata schemes across domains (Willis, Greenberg, &White, 2012). The genomics community had defined a minimal set of core metadata elements (Pagani, et al., 2012) that are sufficient to meet the minimal requirements for data manipulation in order to produce analysis-ready datasets (Qin, Bell, & Greenberg, 2012). The constructs (Interoperability and Modularity) include the goals of scheme extensibility, flexibility, and modularity ensuring the scheme sustainability and promoting the adoption as time goes on (Willis, Greenberg, &White, 2012). Interoperability construct showed the metadata schemes and its related datasets should thoroughly cover the current knowledge, adapt for future needs, and be traceable for the origin of data sources.

CONCLUSION AND FUTURE WORK

This study provided empirical data analysis to identify scientists' metadata functional goals in genome curation work. The results indicate that the metadata requirements depend on how scientists process the massive data sets, and make comparisons across disparate data resources, databases, and repositories. Genomics scientists would like the metadata to support organizing data as well as research needs for working tasks related to data curation. The result will help develop goal-oriented metadata schemes for genomics research community.

Future research will collect additional empirical data regarding the metadata requirements in the genomics community through observations and interviews which can give us further insight into the genome curation and metadata relationships. In particular, the future research will also develop goal-oriented metadata artifacts (e.g., metadata schemes, curation policies), and test metadata assessment metrics supporting data quality control, data management and reuse in genomics research community.

REFERENCE

Huang, H., Stvilia, B., Jörgensen, C., & Bass, H. (2012). Prioritization of data quality dimensions and skills requirements in genome annotation work. Journal of the American Society for Information Science and Technology, 63(1): 195-207.

Klimke, W., O'Donovan, C., White, O., Brister, J. R., Clark, K., Fedorov, B., ... & Tatusova, T. (2011). Solving the Problem: Genome Annotation Standards before the data deluge. Standards in genomic sciences, 5(1), 168.

Pagani, I., Liolios, K., Jansson, J., Chen, I. M. A., Smirnova, T., Nosrat, B., ... & Kyrpides, N. C. (2012). The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. Nucleic acids research, 40(D1), D571-D579.

Qin, J., Ball, A., & Greenberg, J. (2012). Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data. 2012-10-22]. http://dcevents. dublincore. org/index. php/IntConf/dc-2012/paper/view/107/61.

White, H. C. (2010). Considering personal organization: Metadata practices of scientists. Journal of Library Metadata, 10(2-3), 156-172.

Willis, C., Greenberg, J., & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. Journal of the American Society for Information Science and Technology, 63(8), 1505-1520.

Yang, X., Ye, Y., Wang, G., Huang, H., Yu, D., & Liang, S. (2011). VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery. Physiological genomics, 43(8), 457-460.