Curation Roles and Perceived Priorities for Data Quality dimensions and Skills in Genome Curation Work

Hong Huang
School of Information
University of South Florida
honghuang@usf.edu

Besiki Stvilia, Corinne Jörgensen
School of Library Information Studies
College of Communication and Information
Florida State University
{bstvilia, cjorgensen}@.fsu.edu

ABSTRACT
Genome curation work included users with different curation roles as end-users, curators and dual-role users. Genomics scientists with certain role might focus on different data quality aspects and skills requirements. There is a lack of understanding of scientists' perceptions and requirements which hampers the development of systematic and tailed approaches to genome data curation. This research surveyed 158 genomics scientists on their perception and priorities for data quality dimensions and quality sills. The study's findings show that depending on the roles played in scientific data sharing and curation process, genome scientists may have different priorities for and ways of assessing data quality. Curators valued higher the direct quality assessment criteria, while end-users preferred the quality criteria that could be assessed indirectly. Likewise, end-users assigned higher priorities to the data quality assessment skills and the skills needed to identify useful information, while curators valued higher the skills needed to make data useful.

Keywords
Genomics, data curation, data quality, data quality dimensions, data quality skills, bioinformatics.

INTRODUCTION

Genome curators play an important role in providing data curation and preservation support for the genome research community. With the massive accumulation of the genomic data and the number of curation tools rapidly increasing in parallel, it is very challenging for genomics scientists to identify right data curation infrastructure components and tools and define curation strategies and priorities(Howe et al., 2008). There is a lack of research on the relationships among the perception on the one hand, and data curation roles on the other hand, which makes it difficult to systematize quality assurance practices in genomics data curation. This study addresses the above-mentioned gap by examining the following research questions: 1) What are the relationships between data quality perceptions and the roles played in data curation process? 2) What are the relationships between the perceived importance of the data quality skills and the roles palyed in data curation process?

METHODS

To collect data, the study used a survey method. The survey's instrument was adapted from DQ dimensions and skills requirement questions from previous data quality surveys found in the literature (Wang & Strong, 1996; Chung, et al., 2002; Stvilia et al., 2007; Huang et al., 2012). To provide context for the questions, the survey used two scenarios conceptualizing genome curation related activities (Huang et al.,2012).The survey participants (N=158) were asked to rank the top five of the data quality dimensions and data quality assurance skills by their importance in genome curation. The participants identified themselves with three groups: end-users (87), curators (42) and those who played both roles (18) in relation to genomics data. The study used the Qualtrics software (http://www.qualtrics.com) to distribute the survey and collect data. The data was analyzed with STATA 11 software (College Station, Texas, USA) to produce descriptive and Chi-square analysis.

ASIST 2012, October 26-30, 2012, Baltimore, MD, USA.

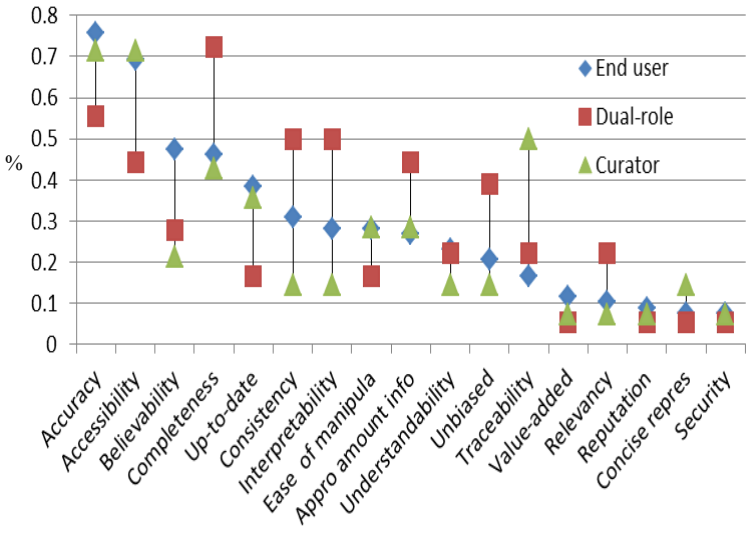

Fig 1. Percentage of the top 5 ranking for DQ dimensions.

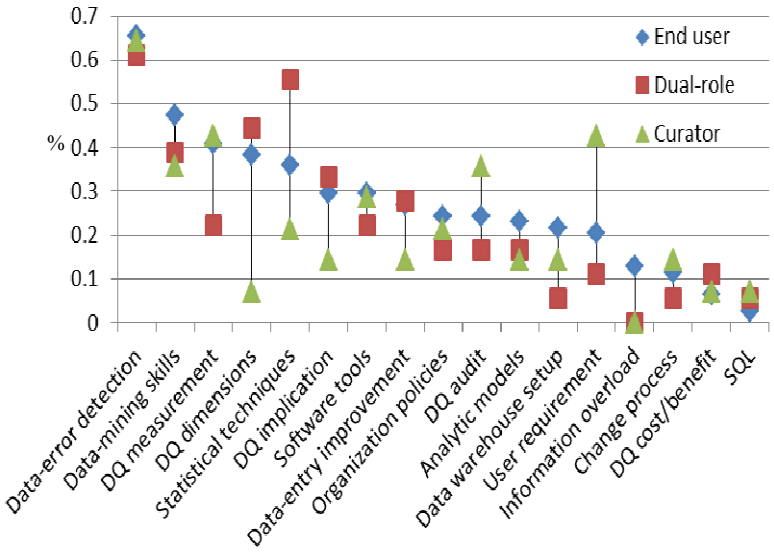

Fig 2. Percentage of the top 5 ranking for DQ skills.

FINDINGS

The survey results for DQ dimensions showed that all three user groups ranked the Accuracy, Accessibility, and Completeness as most important DQ dimensions in genome curation work. At the same time, the Believability and Up-to-date were ranked higher by end-users than by curators, suggesting that end-users tended to assess data quality indirectly based on the source's reputation. Curators, on the other hand, tended to value Traceability, Accessibility higher pointing to the importance of data provenance metadata and open data standards in genomic data curation (Fig. 1).

The Chi-square analysis confirmed (Table 1) that the differences in data quality perception between curators and end-users were statistically significant. Compared to end-users, curators ranked Traceability higher, and Believability Consistency, Understandability lower. Dual role users ranked Interpretability and Completeness higher, and Accessibility lower than end-users and curators. These differences in the data quality priorities could be linked with the ways these groups may evaluate data. Curators may have the necessary knowledge and access to data and assess its quality directly, while end-users may often have to resort on indirect evaluations.

Table 1. Analysis of DQ dimensions ranking for users with different curation roles.

Attribute	End user(A) Ranked by	Both(B) Ranked by	Curator (C) Ranked by	χ^2 A vs B	χ^2 A vs C	χ^2 B vs C
Accessibility	59 (67.8%)	8 (44.4%)	30 (78.9%)	3.528	0.173	***3.95***
Believability	41 (47.1%)	5 (27.8%)	9 (23.7%)	2.268	***7.88***	0.284
Complete-ness	39 (44.8%)	13 (72.2%)	18 (47.4%)	***4.48***	0.045	***4.35***
Consistency	27 (31%)	9 (50%)	6 (15.8%)	2.381	***4.17***	***8.57***
Interpreta-bility	22 (25.3%)	9 (50%)	6 (15.8%)	***4.38***	2.017	***8.57***
Traceability	13 (14.9%)	4 (22.2%)	21 (55.3%)	0.582	***17.9***	***4***
Understanda-bility	19 (21.8%)	4 (22.2%)	6 (15.8%)	0.001	***3.74***	2.414

Note. Bold/Italics: Chi-Square scores were statistically significant (p <0.01).

As for the DQ skills, all user groups perceived Data error detection skills and Data mining skills are the most important (see Fig. 2). Furthermore, end-users had higher priorities for data quality literacy skills (DQ dimensions, DQ measurement), and Statistical techniques. Interestingly, curators ranked User requirement and Data quality audits as more important than did end-users. Compared to curators and end-users, users with dual-roles valued stronger data quality literacy skills such as DQ dimensions and DQ implication (Fig 2).

Chi-square analysis (Table 2) revealed that the differences in the quality skills priorities between curators and end-users were statistically significant. Curators had higher priorities for User requirement, and Structural Query Language (SQL) when compared to those of end-users. End-users valued higher the skills needed to deal with Information overload, and DQ literacy skills such as DQ dimensions. Users with dual-roles have higher priorities for Statistical techniques than did other groups.

Table 2. Analysis of DQ skills ranking for users with different curation roles.

Attribute	End user(A) Ranked by	Both(B) Ranked by	Curator (C) Ranked by	χ^2 A vs B	χ^2 A vs C	χ^2 B vs C
Data-quality dimensions	33 (37.9%)	8 (44.4%)	3 (7.9%)	0.266	***13.34***	***11.71***
Information overload	10 (11.5%)	0 (0%)	0 (0%)	2.287	***5.233***	N/R
Statistical techniques	31 (35.6%)	10 (55.6%)	9 (23.7%)	2.487	2.671	***6.782***
SQL	0 (0%)	1 (5.6%)	3 (7.9%)	4.88	***6.362***	0.051
User requirement	17 (19.5%)	2 (11.1%)	18 (47.4%)	0.715	***7.789***	***5.714***

Note. Bold/Italics: Chi-Square scores were statistically significant (p <0.01).

CONCLUSION AND FUTURE WORK

This research examined genomics scientists' perception and priorities for data quality and quality skills. The study's findings showed that depending on the roles played in scientific data sharing and curation process, genomics scientists may have different priorities for and ways of assessing data quality. Curators valued higher the direct quality assessment criteria, while end-users prioritized the quality criteria that could be assessed indirectly. Likewise, end-users assigned higher priorities to the data quality assessment skills and the skills needed to identify useful information (i.e., "finding a needle in a haystack"), while curators valued higher the skills needed to make data usable.

To enable effective genomics data evaluation and use, it is essential that data repositories would take into account these differences in data quality perception and support both direct and indirect (e.g., heuristics based) ways of quality assessment. Furthermore, the repositories should not only organize and curate data, but also provide the tools needed to discover the data that meets the scientist's needs and use context.

The future research will include the operationalization of the data quality model examined in the current study. In particular, the quality metrics for genomics data will be developed and tested. Data quality models will be created to tailor the curation needs from different user groups.

REFERENCES

Chung, W., Fisher, C., & Wang R. (2002, November). What skills matter in data quality? Paper presented at the 7th International Conference on Information Quality (ICIQ-02), Boston, MA.

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.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L.,…,Rhee, S.Y. (2008). Big data: the future of biocuration. Nature, 455, 47–50.

Lee, Y., Pipino, L., Funk, J., & Wang, R. (2006). Journey to data quality. Cambridge, MA: MIT Press.

Stvilia, B., Gasser, L., Twidale, M., & Smith, L. (2007). A framework for information quality assessment. Journal of American Society of Information Science and Technology, 58(12), 1720–1733.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5–3.