An Exploratory Study to Understand Knowledge-Sharing in Data-Intensive Science

Jongsoon Park and Joseph L. Gabbard

Virginia Bioinformatics Institute, Industrial & Systems Engineering
Virginia Tech
{babirong,jgabbard}@vt.edu

Abstract. This paper describes a challenges associated with data-intensive research processes, knowledge-sharing phenomena, and end-users' expectations in the field of bioinformatics. We developed a questionnaire to support deeper understanding of user experiences with knowledge sharing activities. The results reveal that there are several challenging issues biologists encounter when using bioinformatics resources. A much smaller number of biologists have engaged in passive knowledge sharing within their research fields than we had expected. However, most biologists expressed their willingness to share their own knowledge with others. This result reinforces the need for more user-centered design approaches for supporting knowledge-sharing in rapidly emerging fields of data-intensive science. At the same time, our results suggest that more work is needed to examine how to best motivate users to further engage and contribute knowledge in online scientific communities.

Keywords: User-centered Design Approach, Knowledge-Sharing, Data-Intensive Science.

1 Introduction

In the past two decades, we have seen an exponential increase in the size and breadth of available scientific data, demanding new integrated solutions to explore and elicit valuable insights more efficiently (Kelling et al., 2009). As such, there is growing interest in data-intensive scientific discovery through data integration, simulation, visualization, and validation across distributed networks of heterogeneous resources. A notable example is in biological sciences, which is currently undergoing a rapid paradigm shift to data intensive science (Bell, Hey, & Szalay, 2009). Diverse bioinformatics resources (e.g. online resources that integrate biological data and analysis tools) have been developed, potentially enabling biologists to analyze huge and complex data sets faster and more efficiently as compared to traditional methods (Bull, Ward, & Goodfellow, 2000; Katoh, 2002; Yarfitz, 2000). However, most of these resources have been developed by applied scientists (i.e. computer scientists and bioinformaticians), and are not intuitive or rich enough to address most bench biologists' goals. At the same time, bench biologists are beginning to appreciate the power and potential of bioinformatics resources, despite their poor usability.

M. Kurosu (Ed.): Human-Computer Interaction, Part III, HCII 2013, LNCS 8006, pp. 217–226, 2013.

Generally speaking, biologists and other researchers in data-intensive fields are grappling with how best to deal with “big data” while HCI researchers (collaborating with bioinformaticians) are working to understand how best represent and interact with such data. The question remains: how can we best close this gap to ensure that these resources are both powerful and intuitive to varying user classes. There have been various attempts to bridge gaps between technology-driven bioinformatics resources and the broader spectrum of biologists’ work practices, such as creating more goal-oriented solutions for data collection and storage, and conducting task analysis and usability studies (Joan Bartlett, Ishimura, & Kloda, 2011; J. Bartlett & Neugebauer, 2005; D. Bolchini, 2009; Davide Bolchini, 2009; Javahery, 2004; Mirel, 2009; Tran, Dubay, Gorman, & Hersh, 2004). In parallel, we see web-based “knowledge-sharing platforms” as a growing trend to support data-intensive discovery research by allowing scientists to exchange data, ideas, expertise, and scientific literature online to improve the effectiveness of their processes and validity of their outcomes (De Roure, Goble, & Stevens, 2009; Li, 2012; McIntosh et al., 2012; Parnell, 2011). These knowledge-sharing platforms in scientific communities may provide a timely mean to assist biologists working with large and consistently growing diverse data sets. However, to date, little attention has been paid to a comprehensive understanding of end-users’ characteristics and attitudes about knowledge sharing based on the culture of these data-intensive research domains.

To support useful and seamless knowledge-sharing and reuse in data-intensive research, we need to examine a series of higher level questions from the user’s perspective, such as: What are the major shortcomings in current online bioinformatics resources? To what extent do end-users have experience with knowledge-sharing activities (e.g., knowledge-sharing and -reuse)? What are end-users’ expectations of knowledge-sharing activities? What are users’ perceptions of challenges and opportunities in knowledge-sharing environments? Examining these questions will provide us with meaningful insights not only to identify unmet needs and opportunities, but also to support cross-disciplinary scientific research in data-intensive fields such as biology.

2 Objective and Research Question

This study has three objectives: 1) to understand end-users’ perspectives on shortcomings of current online bioinformatics resources, 2) to identify to what extent end-users have experience with knowledge-sharing activities to support their research processes, and, 3) to elicit specific end-users’ concerns and expectations of knowledge-sharing.

3 Methods

We developed a questionnaire to elicit users’ experience with, perceptions of, and attitudes towards knowledge-sharing activities. The questionnaire is based on previous studies (Bock, Zmud, Kim, & Lee, 2005; Preece, Nonnecke, & Andrews,

2004) and feedback from domain experts who are familiar with biological "wet-lab" experiments and have worked in fields of biology for five years or more.

The first set of questions elicits participant demographics (e.g., age, gender, current work/academic role) and background (e.g., usage frequency of bioinformatics resources). The second set of questions elicits information on various aspects of users' experiences, challenges, and expectations with current bioinformatics resources and knowledge-sharing related to their research processes.

After data were collected, we employed statistical analysis to describe characteristics and behavior of users' current knowledge-sharing activities. The findings of this survey are intended to help identify and prioritize distinguishing web resource features needed to support online knowledge-sharing in data-rich scientific processes.

3.1 Participants

We collected responses from participants of workshops offered by the Virginia Bioinformatics Institute and from PhD students in the Virginia Tech Genomics, Bioinformatics, Computational Biology graduate program. However, the calculation of a response rate was difficult because we do not know how many total PhD students in the program were invited to participate. Eighty-one of eighty-four total responses were usable (three were incomplete or incomprehensible).

More than half of the participants (55.6%) are male, and 72.9% are between 20 and 39 years of age. Almost all characterized their main research role as biologist (63), with others self-reporting roles of bioinformatician (6), chemist (4), computer scientist (4), mathematician (3), and other (11) such as microbiologists, biochemists, and clinician (note: participants were asked to select all items that apply). Slightly over 60% of participants (n = 49) have over five years of research experience in biology, while about 40% (n = 32) have been conducting research for no more than five years.

The frequency of use of bioinformatics resources ranges from every day to less than once a month. A majority of participants (70.3%) reported using bioinformatics resources more than once a week, 13.6% reported use as more than once a month, and 13.6% reported using bioinformatics resources around once a month.

3.2 Results

We performed data analysis using SPSS (version 18.0), defining statistical significance at $\rho < 0.05$. In general, we observed similar response patterns among participants with no more than 5 years of research experience. Those with over 5 years research experience also showed similar response tendencies. As such, in the following discussion, we consider two broad classes of participants; those with no more than 5 years of research experience and those with over 5 years of research experience.

3.3 Limitations in Online Bioinformatics Resources

Our results show that most participants are currently challenged by the lack of integration and inconsistent results across online bioinformatics resources (e.g. different gene naming conventions, different annotations for the same gene). In the same vein, they repeatedly highlighted limitations due to the poor quality of genomic sequences and metadata. Some participants noted strengths of bioinformatics resources such as multiple views on the same data and multiple comparisons across different genomes. Inconsistency in user interfaces and general lack of usability were cited as major difficulties for a number of participants, implying a steep learning curve (i.e., long learning times) as a key usability issue. In addition, some participants had trouble accessing data due to complex navigation structures typical of bioinformatics resources. Lastly, data security was noted as an important issue, since many researchers are leveraging these resources to support hypotheses generation, publications or grants.

3.4 Important Factors of Bioinformatics Resources

To examine users' expectations of bioinformatics resources, we asked participants which resource features are the most important or valuable. Multiple responses were categorized and tallied using the multiple-dichotomy frequency analysis. We constructed a cross tabulation table to analyze the most dominant participants' response. Table 1 presents the cross-tabulation frequencies by years of research experience.

Table 1. Important factors of bioinformatics resources

	Research Experience (yrs)		**Total**
	No more than 5	**over 5**	
Speed and responsiveness of resource	23 (71.9%)	36 (76.6%)	59 (74.7%)
Wealth of available data	22 (68.8%)	30 (63.8%)	52 (65.8%)
Breath of resource tools and functions	16 (50.0%)	29 (61.7%)	45 (57.0%)
Degree of data integration	17 (53.1%)	26 (55.3%)	43 (54.5%)
Ease of use	16 (50.0%)	26 (55.3%)	42 (53.2%)
Ability to upload my own data	13 (40.6%)	25 (53.2%)	38 (48.1%)
Ability to ask questions related to my research	12 (37.5%)	23 (48.9%)	35 (44.3%)
Ability to create publication quality images	10 (31.3%)	22 (46.8%)	32 (40.5%)
Advanced visualizations	13 (40.6%)	19 (40.4%)	32 (40.5%)
Ability to collect knowledge from others researchers	9 (28.1%)	21 (44.7%)	30 (38.0%)
Ability to share knowledge with other researchers	6 (18.8%)	9 (19.1%)	15 (18.9%)
Total # of participants	32	47	79

As expected, performance-related factors common to most web-based systems ranked relatively high in "important factors of bioinformatics web resources". Namely, participants valued "high speed and responsiveness of resource" (74.3%), followed by "wealth of available data" (65.8%), "breadth of resource tools and functions" (57.0%), "degree of data integration" (54.5%), "ease of use" (53.2%), and "ability to upload my own data" (48.1%).

Interestingly, nearly half of the participants with over 5 years of research experience selected "ability to ask questions related to my research" (48.9%) and "ability to collect knowledge from others" (46.8%) as an important resource features. In contrast, a much smaller proportion of these participants (19.1%) appear interested in sharing their knowledge with others. It can be inferred from these results that experienced participants are more interested in enhancing the overall quality (and performance) of their research by making use of others' shared knowledge than sharing their accumulated knowledge and skills. In comparison with the above findings, participants with no more than 5 years of research experience showed little interest in sharing and collecting knowledge as compared to other features.

3.5 Knowledge Sharing Experience

Our knowledge-sharing results suggest significant, but limited, online knowledge-sharing activity among our sampled user population (See Fig. 1).

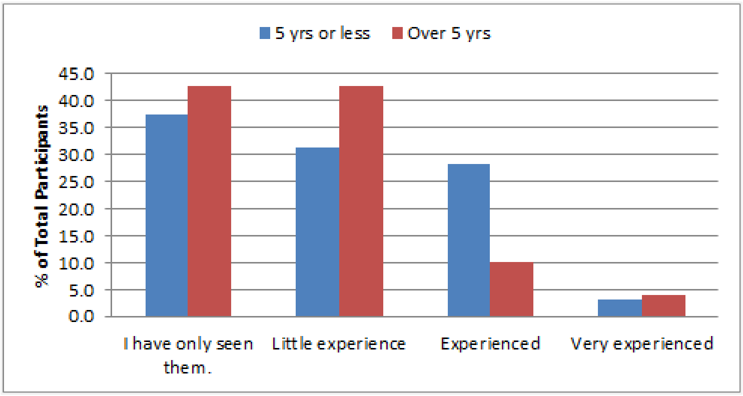

Fig. 1. Knowledge sharing experience

Specially, all participants reportedly engage in knowledge-sharing activities to some degree (e.g., knowledge-sharing or -reuse), but nearly 80 percent of respondents reported participating in knowledge-sharing activities in a passive manner. That is, most users rarely share their knowledge, but instead tend to seek and use knowledge shared by others. This is a typical example of a lurker; an individual who consumes information and requests specific questions from others, but does not explicitly contribute to the shared knowledge base (Preece et al., 2004).

To examine potential correlations among age, years of research experience, and experience level in knowledge-sharing activities, we performed a Kendall's tau correlation analysis. As might be expected, we found a positive, statistically significant correlation between age and research experience (Kendall's tau b = 0. 608, ρ = 0.000). Contrary to our expectation, however, we found no significant correlation between years of research experience and the experience level in knowledge-sharing activities (Kendall's tau b = -0.154, ρ = 0.123). Moreover, there was no correlation between age and knowledge-sharing experience (Kendall's tau b = -0.019; ρ = 0.847). Even though one might expect younger researchers to be more attuned to online knowledge-sharing opportunities.

We next examined the types of knowledge consumed in knowledge-sharing activities; including specific examples of both implicit and explicit knowledge (Bock et al., 2005; Choo, 2000). These results provide valuable insight into what types of knowledge participants seek from others in support of their research (See Table 2).

Table 2. Types of shared knowledge employed

		Research Experience (yrs)		**Total**
		no more than 5	**over 5**	
	None	0 (0%)	6 (18.8%)	6 (11.1%)
Implicit Knowledge	**General ideas**	14 (63.6%)	15 (46.9%)	29 (53.7%)
	Accumulative research experiences	9 (40.9%)	15 (46.9%)	24 (44.4%)
	Unique opinions	6 (27.3%)	7 (21.9%)	13 (24.1%)
Explicit Knowledge	**Articles published in books, websites, and documents**	13 (59.1%)	16 (50.0%)	29 (53.7%)
	Products, patents, databases, tools, and prototypes	8 (36.4%)	7 (21.9%)	15 (27.8%)
	Rules, routines, or operating procedures	10 (45.5)	5 (15.6%)	15 (27.8%)
	Total # of participants	22	32	54

The most frequently reported knowledge used is “general ideas” (53.7%), “articles published in books, websites, and documents” (53.7%), followed by “accumulative research experiences” (44.4%). The overall pattern of responses indicates that our participants employ implicit and explicit knowledge in an evenly balanced way.

We also found that participants with no more than 5 years research experience tended to rely on "general ideas" (63.6%) more than participants with over 5 years research experience (46.9%), with open-ended responses suggesting that the relatively inexperienced participants use shared knowledge extensively to generate research questions and confirm hypotheses. In addition, these relatively inexperienced researchers depend on others to confirm rules or procedures (45.5%) as compared with more experienced participants (15.6%).

3.6 Knowledge Sharing Intention

Next, we assessed participants’ intention to share knowledge using questions adapted from Bock et al. (2005). Participants indicated their agreement or disagreement with statements using a seven-point Likert-type scale (where scores of 7 suggest a strong willingness to share). Surprisingly, almost all participants (95%) reported a willingness to share knowledge with others (Fig. 2). This result suggests there are ample opportunities to promote and grow knowledge-sharing in data-intensive sciences such as biology.

Fig. 2. We used a one-way MANOVA to determine whether there are any differences between our two research experience groups on more than one kind of knowledge. We found no significant group effects for the types of knowledge on their intention to share knowledge, *F* (14, 117) = 0.738, ρ =0.743; *Wilk's* λ = 0.728, *partial* $\varepsilon 2$ = 0.54.

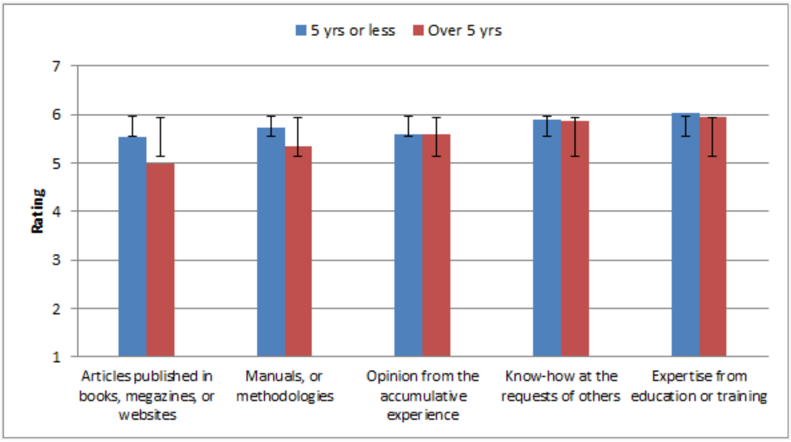

Fig. 2. Intention to share specific knowledge (7=strongly agree, 1=strongly disagree)

Next, we used an independent t-test to investigate differences in knowledge-sharing intentions between the two research experience groups, and found no significant differences between them across any of the five variables: articles published in books, magazines, or website, $t(75)$ = 1.539, ρ = 0.128; manuals, or methodologies, $t(73)$ = 1.316, ρ = 0.192; own opinion from the accumulative experience, $t(75)$ = -0.023, ρ = 0.982; know-how at the requests of others, $t(75)$ = 0.124, ρ = 0.901; and expertise from education or training, $t(75)$ = 0.314, ρ = 0.755.

3.7 Concerns about Knowledge Sharing

One of the significant results to emerge from our responses is that most participants worry about source reliability (i.e. poster's expertise, data integrity, experimental verification, quality of information). Similarly, results indicate concerns about encountering "incorrect information or annotations". Other concerns include the need for an easy way to contribute their own knowledge, suggesting that participants have encountered usability issues when they have previously tried to share information. Additionally, participants noted response time, frequency of knowledge update, copyright, and privacy as potential barriers to engaging in knowledge-sharing activities.

3.8 Expectations

We found the following three themes of expectations concerning knowledge-sharing from the free text comments. First, there were many responses that referred to source creditability. For example, participants expect to be able to authenticate contributors' expertise as well as shared knowledge. Second, about 50% of participants said they often need technical support to make use of bioinformatics tools (e.g. troubleshooting advice, application tips, how-to guides). Participants noted the lack of well-organized information repositories of shared knowledge by domain, technique, methodology, etc. Third, there was some evidence that participants are interested in reusing knowledge from others to improve their research processes and outcomes. Frequently mentioned expectations include implicit knowledge related to protocols used for conducting biological experiments such as, small "bench work tricks", protocols of best practices for cutting edge -omics research, and information about negative data.

4 Discussion

The results of this research indicate not only existing barriers to use of bioinformatics resources but also opportunities to address users' unmet needs in data-intensive scientific communities. We found no significant differences across age groups and years of research experience; suggesting that designs to support knowledge sharing should consider other user class characteristics.

Our results indicate that biologists struggle to utilize bioinformatics resources regardless of years of research experience mainly due to inconsistent results and poor user interfaces. Moreover, users indicate that the greatest shortcomings of current resources are often associated with the most important features. A possible explanation may be that many online bioinformatics resources employ a system-oriented development approach rather than user-centered design approach that aims to better understand users' unmet needs.

We found approximately 80% of participants do not actively engage in knowledge-sharing than we expected, regardless of age or years of research experience. Only a small portion of participants have actively engaged in research-related knowledge-sharing. A possible explanation might be that current links exist between knowledge-sharing platforms and biologists are limited or ad hoc despite the prevalence of online knowledge-sharing resources in scientific communities. In other words, we can suppose that many knowledge-sharing platforms are built with a focus on current technological trends rather than user experience factors, which affect users' motivation to engage in knowledge-sharing activities. Another possible explanation is that characteristics of the scientific culture may influence an individual's propensity to engage in knowledge sharing and reuse. Thus, more study is needed to understand how best to foster knowledge sharing and reuse in scientific communities.

Results from this study also show that participants with no more than 5 years research experience rely more on implicit knowledge shared by other practitioners than explicit knowledge. It seems possible that self-efficacy caused by accumulated expertise may contribute to the differences found in usage patterns of implicit knowledge.

The most significant finding is that nearly 95% of participants are willing to share their knowledge, contradicting the very low levels of current involvement in knowledge-sharing activities. The lack of quality user experiences to support seamless and easy knowledge contribution suggests a need to develop knowledge-sharing platforms that embrace user-centered design approaches.

Lastly, this work identifies several issues that must be ensured to facilitate knowledge-sharing and reuse in data intensive settings aiming to support scientific discovery. A majority of participants were concerned about the quality of knowledge and the degree to which they can trust shared knowledge. These results are consistent with those of previous studies (Golbeck, 2008; Levin, Cross, Abrams, & Lesser, 2002) and suggests that source credibility has a considerable impact on attitudes towards knowledge-sharing. More study is therefore needed to better understand how to cultivate trust in, and increase motivation to use, knowledge-sharing activities.

5 Conclusion

This research is one of the first studies to investigate knowledge-sharing in emerging data-intensive sciences such as biology. Our results imply significant opportunities to support knowledge-sharing in these communities, but that careful attention needs to be taken to users' perceived and actual needs. This initial study focuses on eliciting basic user experiences with, and perceptions of, bioinformatics resources and online knowledge-sharing activities. Results presented herein may inform future studies to explore user experiences and knowledge-sharing activities in data-rich environments.

This study was mainly conducted among biologists. The outcomes of the study could be strongly influenced by the culture of experimental science. Hence, to determine whether these findings can be applied to a wide range of knowledge-sharing platforms for rapidly emerging fields of data-intensive science, there needs to be further study with additional participants from different background (e.g. applied scientists) across other data-intensive fields (e.g., visual analytics, meteorology).

Acknowledgments. This work was supported in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900040C.

References

1. Bartlett, J., Ishimura, Y., Kloda, L.: Why choose this one?: Factors in scientists' selection of bioinformatics tools. Information Research 16(1), 15 (2011)
2. Bartlett, J., Neugebauer, T.: Supporting information tasks with user-centred system design: The development of an interface supporting bioinformatics analysis. Canadian Journal of Information and Library Science 29(4), 486–487 (2005)

3. Bell, G., Hey, T., Szalay, A.: Beyond the data deluge. Science 323(5919), 1297–1298 (2009)
4. Bock, G.W., Zmud, R.W., Kim, Y.G., Lee, J.N.: Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. MIS Quarterly, 87–111 (2005)
5. Bolchini, D.: Better bioinformatics through usability analysis. Bioinformatics (Oxford, England) 25(3), 406–412 (2009), doi:10.1093/bioinformatics/btn633
6. Bolchini, D., Finkestein, A., Paolini, P.: Designing Usable Bio-information Architectures. In: Jacko, J.A. (ed.) HCII 2009, Part IV. LNCS, vol. 5613, pp. 653–662. Springer, Heidelberg (2009)
7. Bull, A.T., Ward, A.C., Goodfellow, M.: Search and discovery strategies for biotechnology: the paradigm shift. Microbiology and Molecular Biology Reviews 64(3), 573–606 (2000)
8. Choo, C.W.: Working with knowledge: how information professionals help organisations manage what they know. Library Management 21(8), 395–403 (2000)
9. De Roure, D., Goble, C., Stevens, R.: The design and realisation of the Virtual Research Environment for social sharing of workflows. Future Generation Computer Systems 25(5), 561–567 (2009), doi:10.1016/j.future.2008.06.010
10. Golbeck, J.: Weaving a web of trust. Science 321(5896), 1640–1641 (2008)
11. Javahery, H.: Beyond power making bioinformatics tools user-centered. Communications of the ACM 47(11), 58 (2004), doi:10.1145/1029496.1029527
12. Katoh, M.: Paradigm shift in gene-finding method: From bench-top approach to desk-top approach (review). Int. J. Mol. Med. 10(6), 677–682 (2002)
13. Kelling, S., Hochachka, W.M., Fink, D., Riedewald, M., Caruana, R., Ballard, G., Hooker, G.: Data-intensive science: a new paradigm for biodiversity studies. t BioScience 59(7), 613–620 (2009)
14. Levin, D.Z., Cross, R., Abrams, L.C., Lesser, E.L.: Trust and knowledge sharing: A critical combination. IBM Institute for Knowledge-Based Organizations, 1–9 (2002)
15. Li, J.W.: SEQanswers: an open access community for collaboratively decoding genomes. Bioinformatics (Oxford, England) 28(9), 1272–1273 (2012)
16. McIntosh, B.K., Renfro, D.P., Knapp, G.S., Lairikyengbam, C.R., Liles, N.M., Niu, L., Hu, J.C.: EcoliWiki: a wiki-based community resource for Escherichia coli. Nucleic Acids Research 40(D1), D1270–D1277 (2012), doi:10.1093/nar/gkr880
17. Mirel, B.: Supporting cognition in systems biology analysis: findings on users' processes and design implications. BioMed Central Ltd. (2009)
18. Parnell, L.D.: BioStar: An Online Question & Answer Resource for the Bioinformatics Community. PLoS Computational Biology 7(10), e1002216 (2011)
19. Preece, J., Nonnecke, B., Andrews, D.: The top five reasons for lurking: improving community experiences for everyone. Computers in Human Behavior 20(2), 201–223 (2004), http://dx.doi.org/10.1016/j.chb.2003.10.015
20. Tran, D., Dubay, C., Gorman, P., Hersh, W.: Applying task analysis to describe and facilitate bioinformatics tasks. Research Support, N I H, Extramural (2004)
21. Research Support, U S Gov't, P H S. Stud Health Technol. Inform 107(Pt 2), 818–822
22. Yarfitz, S.: A library-based bioinformatics services program. Bulletin of the Medical Library Association 88(1), 36 (2000)