Assessment of Collaboration and Interoperability in an Information Management System to Support Bioscience Research

Sahiti Myneni [a], MSE and Vimla L. Patel [a, b], PhD, DSc

[a] Center for Decision Making and Cognition, Department of Biomedical Informatics, Arizona State University
[b] Department of Basic Medical Sciences, University of Arizona College of Medicine – Phoenix, in partnership with Arizona State University

Abstract
Biomedical researchers often have to work on massive, detailed, and heterogeneous datasets that raise new challenges of information management. This study reports an investigation into the nature of the problems faced by the researchers in two bioscience test laboratories when dealing with their data management applications. Data were collected using ethnographic observations, questionnaires, and semi-structured interviews. The major problems identified in working with these systems were related to data organization, publications, and collaboration. The interoperability standards were analyzed using a C^4I framework at the level of connection, communication, consolidation, and collaboration. Such an analysis was found to be useful in judging the capabilities of data management systems at different levels of technological competency. While collaboration and system interoperability are the "must have" attributes of these biomedical scientific laboratory information management applications, usability and human interoperability are the other design concerns that must also be addressed for easy use and implementation.

Introduction

Bioscience research when coupled with the high speed processing technologies results in highly detailed datasets. These laboratories are data intensive which is evident from the publicly accessible immense databases generated by the Human Genome project [1]. The information to be processed in a genomic laboratory ranges across the DNA sequence, mutation, expression arrays, assays, antibodies, oligonucleotides etc., to name a few. The challenge to genomic medicine is to analyze and integrate these diverse and voluminous data sources to elucidate normal and abnormal physiology [2]. With recent National Institute of Health (NIH) priority for translational research, organization of the basic laboratory data has become significant. Current laboratory data management methods primarily include handwritten laboratory notebooks, paper files, homegrown small databases, and spreadsheet files [3]. The main aim of this study is to identify the nature of the problems, if any, with the existing data management practices followed in bioscience research laboratories with additional focus on the level of collaboration and interoperability supported by these systems.

Theoretical Background

Prior research emphasized on the design of information management systems with a common layer of interoperability while providing a spectrum of options that could be used to support individual researcher needs [3]. Such systems can indeed realize secondary benefits of interoperability such as collaboration. Interoperability may be defined as the ability of two or more systems to exchange information and use this exchanged information [4]. The Connection, Communication, Consolidation, Collaboration (C^4) Interoperability (I) Framework (C^4IF) was initially proposed for business information systems as a classification typology. The C^4I framework transfers linguistic concepts to information systems' design using multi-phenomena such as sounds (phonetics and phonology), word formation and word endings (morphology), word combinations (syntax), meaning (semantics) and language use (pragmatics) [5], based on which, it defines the four interoperability levels, C^4I. Connection refers to the means of data exchange (e.g. via disks or broad band), communication and consolidation refer to data format, data schema and meaning, while collaboration realizes combined action/behavior towards a shared goal. We apply this framework to analyze the interoperability standards of the existing information management practices in a typical bioscience laboratory at granular levels lying underneath the system.

Methods

To understand the influence of the existing data management applications on research in a typical laboratory (lab), we investigated two such scientific labs. Of the six candidate labs considered, two test labs were selected based on their responsiveness, motivation of the lab Principal Investigator (PI) and the richness of lab environment in terms of its ability to represent the manifold changes of use of information technology to improve scientific productivity and satisfaction in the realm of bioscience research. Ethnographic observations were carried out, during which a researcher unobtrusively observed the activities at different times in the test labs taking observational notes. Ethnography is an ideal research method for the given purpose and setting [6]. The important concepts identified during the ethnographic phase were used to design web-based questionnaires and semi-structured face-to-face interviews. Two questionnaires (Q 1, Q2) were used in this study with the lab PIs. Q 1 was administered to all six candidate lab PIs during the test lab selection process, while Q 2 was given only to the PIs of the two selected test labs. Both the questionnaires included open-ended as well as closed specific questions. Unlike the questionnaire framework, where detailed questions were formulated ahead of time, semi structured interviews began with more general unstructured questions [7]. A number of new questions were generated during the interview, allowing both the interviewer and interviewee to probe for details of any particular issue. The four interview areas of interest were data storage, data management, queries on stored data and collaboration in the test labs. The interviews with PIs were framed around themes identified in their questionnaire responses. Nine test lab members in different professional roles such as lab manager, computer support specialist, and bench molecular biology investigators were interviewed in a more open format emphasizing their work and task descriptions. The interview data were audio recorded and transcribed for analysis.

Results

The PIs of all the six candidate labs were asked to summarize, with respect to their own labs, productivity, satisfaction, and organization on a scale of one to four (Poor-Excellent). Of the 18 responses, only one lab was identified as excellent in terms of PI satisfaction. Organization of lab data was rated as the most problematic compared to productivity and satisfaction by five of the six PIs, indicating that there was room for improvement of data management in the test labs. This derivation was further bolstered by the findings from the analysis of the semi-structured interviews conducted with the test lab PIs and lab members. Analysis of interview data helped us identify a number of problems emerging from the current information management methods used to organize data in the two test labs. These problems are elaborated below and sample quotes from the semi-structured interviews are included in Table 1, illustrating the issues with the information management practices in the test labs.

Problems of data maintenance by an individual

In the test labs, researchers often kept scientific data in spreadsheets, handwritten notebooks, and logs. Although the content was fully intelligible to the creator of the notes and organized for a great deal of personal efficiency, the structure was not transparent to other researchers. This can be understood from Section A, Table 1. Such idiosyncratic ways of data organization followed by the members in the two test labs lacked an established convention, thus rendering the research data cryptic to the co-researchers.

Limitations to publication success

The primary means of knowledge dissemination across the scientific field is through publishing. The inability of the test labs to organize and record their research activities in a structured fashion sometimes led to loss of information. This can be inferred from the sample quotes presented in Section B of Table 1. Because of such data loss, it was apparent that publishing the related findings would not be an easy task. Besides, loss of such data may lead to an unsubstantiated finding, which is one of the main challenges for conducting translational research [8].

Problems of collaboration within the laboratory

There were portions of data that were created and maintained by some members, (e.g. lab managers) that were shared with the other members of the test lab. The sharing of scientific data and experimental results was done in a weekly lab meeting in the form of formal verbal presentations or informally in personal discussions among colleagues. Certain procedures were followed for diagnosis assessments in the test labs as mentioned in the first quote of Section C in Table 1. Such inconsistent methods had every possibility of something going wrong. Indeed, there were concerns about database access permissions, security, and protection of individual contributions with the existing approaches.

Problems of collaboration outside the laboratory and with experts in diverse domains

Generally in the two test labs, databases were kept at each of the collaborating sites and copies were transformed (e.g., format conversion) and forwarded in the form of spreadsheets or delimited files for interpreting and integrating with destination databases. But there were always problems in representing and communicating context, which is crucial for working with collaborators in the same domain as well as with experts in other domains (e.g. biostatisticians). Data sharing with the experts from other domains was much affected because of lack of options to communicate context in the information management practices followed in the test lab. Also, there were problems when data was shared with the collaborators in the same domain because of inadequate common terminology. These collaborative issues with the current management practices can be construed from Section C, Table 1. The representational heterogeneity across the databases resulting from the decentralized scientific community frustrates efforts to integrate them [9]. With data sharing and integration across multiple sites being so ineffective, the collaborative research may not be rewarding.

Section	Issue	Sample Quotes from the semi-structured interviews
A	**Problems of data maintenance by individual**	**" I mean she's very smart and she keeps good notes, but first she will go to the computer here and then she'll go to her written notes, I mean without her, it would be very hard to back trace."** **"Yeah, I'd have to train somebody, and that's a big concern for me. I have started writing down protocols for different actions taken by the database, but I haven't certainly completed it."**
B	**Limitations to publication success**	**"Not really recording exactly how they did do it, but they'll get as close as they can in the publication because they don't have good records…usually the level of detail…there are many things that labs cannot even attempt… because of their lack of organization."** **"But we've re-made a lot of things just because either we don't know where something is, or even if we find it, it's about papers, but a little more trivial detail, we don't know exactly what sequence is in there, we don't know exactly what restricted enzymes. So that is frustrating and it's a big waste of time. ….that's ridiculous that shouldn't happen, but it does happen"**
C	**Problems of collaboration within as well as outside the laboratory**	**"Well, we have then XX and I meet once a week and we review what we've done. … And at that point, we assign a final diagnosis to the patient and she…you know, I say out loud what the final diagnosis is and she confirms it and we put it in the database."** **"I can give data that I think are appropriate to answer a question to a biostatistician, but when they look at it, they see it from a different point of view…. and that spread sheet does not really encapsulate where it came from very well, how was it generated, was it random, how was this data collected."** **"Their person in Europe wants the identified phenotype information... he's not going to know what our variables mean. So, what do I do? I send him an email and I say, "These are our forms so you can see how it's attached to the tables, but what exactly do you want? Basically, I have to keep kind of playing with it until I give them what they need"** **"The only common context which we have is just basic language, that is, in terms of disease terminologies, which of course are slippery.... There's no common framework. There are still many gene names that are being changed."**

Table 1: Sample quotes from interviews illustrating the problems identified with the existing data management practices in the test laboratories.

In summary, inefficient data organization can potentially lead to substantial data loss, misinterpretation, and degraded security and privacy levels resulting from limited data sharing options, thus compromising the productivity of a scientific research laboratory. The nature of the problems identified with the existing laboratory practices in the two test labs revealed that the collaboration (in the same domain or in other fields) promoted by these methods was minimal. Improved collaboration can be facilitated by adopting appropriate use of technology enabled communication strategies [10]. Additional level of interoperability analysis of these systems was performed using C^4I framework.

Interoperability Issues

Analysis of the information management practices using C^4I framework gave us an insight into the interoperability standards of the data management applications used in the test lab. As discussed earlier, interoperability can be addressed at different levels such as connection, communication, consolidation, and collaboration using this framework. Advancement in each of these areas can influence but cannot determine the advancement of the other areas. A biomedical information management system can have a high degree of interoperability at communication and consolidation level, while being low at the other two levels. For example, using advanced technologies such as wireless broadband network to exchange data instead of using manual techniques (e.g. compact disks) can be deemed essentially as advancement in the connection area and this cannot automatically assure semantically rich terminology at the communication/consolidation level and vice versa. An analysis of the degree of interoperability in the information management practices followed by the test lab at different levels of C^4I framework is included in Table 2.

Collaboration	**Support for action towards a shared goal**	**Low**
Consolidation ***Communication***	**Commonly accepted entity relationships, data schema, data format, data meaning**	**Low**
Connection	**Data exchange via compact disks (CDs) or broadband**	**Medium**

Table 2: Analysis of system interoperability supported by the information management practices in the test lab using C^4I framework

Data sharing options were limited, however, the test lab members were using some means (such as email) to transfer their research data to their collaborators at the distributed research sites. As reflected on table 2, this showed that the connection level of this system had a medium (M) level of interoperability. But data ontology and format were not considerably standardized, thus giving a low (L) interoperability in the consolidation and communication domain. Similarly, collaboration was minimally supported by the system, where this was achieved usually through formal meetings, electronic mail and shared documents. For an information management system to succeed in collaborative interoperability, common ideas on work flow patterns and functions needs to be established [5], which were not realized in this case. Hence, it was given a low (L) rating for interoperability at the collaboration level. With such a comparative analysis of interoperability within different layers of the system, we were able to identify and prioritize the issues that needed to be addressed. For instance, consider the mapping of analysis from semi-structured interviews to the C^4I layered framework as outlined in Table 2. This approach allowed us to identify the high priority tasks such as developing standard ontology and common terminology, creating distributed workflows and functional procedures. Such initiatives can bolster the interoperability standards at the collaboration, communication and consolidation areas. Consequently, interoperability at these levels can be elevated to medium (M) to be on par with the interoperability standards of connection domain.

Conclusions

Our study describes the problems incurred by the researchers as they manage the voluminous bioscience research data. The four major problems identified with current scientific data management methods in the two test laboratories are related to data maintenance, publications, collaboration within as well as outside the laboratory, and interoperability. Low interoperability is found with respect to data format, data schema, commonly accepted terminologies, and ability to work together on distributed workflows. The solution strategies aimed at achieving common terminologies and well-understood workflow patterns can automatically solve some problems of collaborative research in scientific laboratories. However, incorporation of additional advanced functionalities to facilitate visualization of heterogeneous research data among researchers from diverse backgrounds such as statistics, computer science, bioscience, and medicine is vital to the success of bioscience information management applications. Analysis of the interoperability standards of bioscience research data management systems using the described framework (C^4IF) can be useful in the granular scrutiny of different layers within any system. This framework aids in proper selection of information management systems suitable to a laboratory profile based on the laboratory size, amount of research data handled, and research focus. This sort of personalized selection is made possible through the clear demarcated layers of the C^4I framework model. The study begs the next question of intervention to incorporate features in the information management systems for better interoperability and collaboration between distributed research sites as well as within the laboratory.

Design Recommendations

System usability, cognitive interoperability, and human interoperability may be the other worthy aspects, which the designers should pay attention to, in course of developing an efficient information management tool. Usability is a well- understood concept in the scientific literature and can be viewed as the capacity of a system to allow users to carry out their tasks safely, effectively, efficiently, and enjoyably [11]. It can indeed act as an important factor affecting system acceptability [12]. We define human interoperability as the ability of a system's design to allow humans to use other similar systems with minimal training, showing some generalizable skills. Human interoperability can be interpreted as a dependent on cognitive interoperability. Cognitive interoperability is related to human actor's way of thinking when using a system [13], and may have significant influence on humans' learning curve for new technologies and systems. Achieving congruence in thoughts and perceptions of end users of different systems used for solving similar tasks may be the key to improve knowledge transferability, new system implementation, and acceptance.

Acknowledgements

We graciously acknowledge the support from Drs. Steven G. Bova, Donald Zack, Jian Wang, and other members of our collaborative study. We thank Jan Horsky for his help with the early stages of data collection and analysis. The study was supported in part by a grant (1R41CA105217-01A1-STTR) from National Institute of Health/National Cancer Institute (NIH/NCI).

References

[1] Consortium IHGS. The human genome. Nature. 2001: 860–921.

[2] Louie B, Mork P, Sanchez FM, Halevy A, Hornoch PT. Data integration and genomic medicine. Journal of Biomedical Informatics. 2007; 40: 5–16.

[3] Anderson NR, Lee ES, Brockenbrough JS,Minie ME, Fuller S, Brinkley J, et al. Issues in biomedical research data management and analysis: Needs and barriers. Journal of American Medical Informatics Association. 2007; 14: 478-88.

[4] IEEE. (1990). IEEE (Institute of Electrical and Electronics Engineers): Standard Computer Dictionary- A Compilation of IEEE Standard Computer Glossaries.

[5] Peristeras V, Tarabanis K. The connection, communication, consolidation, collaboration interoperability framework (C^4IF) for information systems interoperability. Interoperability in Business Information Systems.2006; 1: 61-72.

[6] Van Maanen, J. (1996). Ethnography. In A. Kuper & J. Kuper (Eds.), The Social Science Encyclopedia (2nd ed., pp. 263-265). London: Routledge.

[7] Crabtree BF, Miller WL. Doing qualitative research. Newbury Park, CA, Sage, 1992.

[8] Unger EF. All is not well in the world of translational research. Journal of American college of Cardiology. 2007; 50: 738-740.

[9] Sujansky W. Heterogeneous database integration in biomedicine. Journal of Biomedical Informatics. 2001; 34: 285-98.

[10] Patel VL, Kaufman DR, Allen VG, Shortliffe EH, Cimino JJ, Greenes RA. Toward a framework for computer-mediated collaborative design in medical informatics. Methods of Information in Medicine.1999; 38: 158-176.

[11] Preece J, Rogers Y, Sharp H. Interaction design: beyond human–computer interaction. New York: Wiley; 2002.

[12] Nielsen J. Usability Engineering. New York: Academic Press; 1993.

[13] Goldkuhl G. The challenges of Interoperability in E-government: Towards a conceptual refinement. Pre-ICIS SIG eGovernment Workshop, Paris, 2008.