Standards Barriers to Bioinformatics Research

J. I. E. Fernando

***Abstract*— Although new and emerging information technologies (IT) can enable the analysis of rapidly expanding bioinformatics data, no standards exists. Standards validate a technology or process against a compilation of consolidated best practice specifications. Standards development represents an effective way to retrieve textual evidence, work collaboratively, and integrate bioinformatics with global e-health initiatives. Thus, standards barriers can impede otherwise productive research efforts.**

***Index Terms*—bioinformatics, bio-ontology, mashup, standards, e-health**

I. INTRODUCTION

THE capacity to share data can facilitate multi-national bioinformatics research approaches for asking new and different types of biological questions. Bioinformatics melds molecular biology with computer science, using genomic information in understanding human diseases and new molecular targets for drug discovery (1). Collecting, organizing, presenting, and disseminating biological knowledge is a challenging process (2, 3). Yet none of this activity is standardized on a local, national, or international basis. Very little of it feeds into or connects with other bioinformatics systems or e-health initiatives. Thus, this paper suggests the integration of existing systems and initiatives will reduce the standards barriers that impede otherwise productive research approaches.

II. BACKGROUND

Standards evolve over time; they are revised as change occurs. Although beginning to emerge, to date we can not locate many research attempts to devise a formal standard. Formal standards are generally ratified by national standards organizations or the International Standards Organization (ISO) (26). The "ISO/TS 18308 Health Informatics - Requirements for an electronic health record reference architecture" standard for instance, provides the documented example of a specification against which a series of benchmarks or best practices for a process or technology can be evaluated (26). A review of publications for this work indicated a paucity of research projects enquiring about standards for bioinformatics work (27, 11). Thus, researchers have apparently overlooked questions about a formal standards development process.

Interconnecting bioinformatics systems is a challenging and complex task. The plethora of ontologies and open source software options currently available to scientists has aggravated the process (3, 24). In recognition of this, several bodies, such as the OBO (Open Biomedical Ontologies) Foundry, the Bio-Ontologies IMSB (International Society for Computational Biology) special interest group and Epoch are trying to build comprehensive frameworks that feed into existing ontologies [19, 23, 24]. These efforts, along with those of the Ontology for Biomedical Investigations (OBI) or the OWL web ontology language, acknowledge the need to align their efforts with existing ontologies (6, 14). Many bioinformatics experts are apparently aware of the standards barriers to data integration.

Some bioinformatics experts have focused on devising new technological approaches to data integration in order to address the standards barrier. The approaches include ^{My}Grid, MOBY-services, Semantic-Moby, and Bio2RDF and are based on the Internet and associated technologies (31-33). These technologies shift the focus from ad hoc computational biology to emphasize standardized web-based semantic networks for ontological knowledge representation.

In order to support development of the semantic networks, the ISO have ratified a number of Topic Maps, for example the "ISO/IEC 19763-3 Information technology - Metamodel framework for interoperability (MFI) - Part 3: Metamodel for ontology registration" (30). Topic Maps standardize knowledge-sharing representations, emphasizing the "findability" of information (31). Various other languages for ontology specification, such as RDFs (resource description frameworks) and OWL ontologies, also support the construction of useful knowledge representation models. As with Topics Maps, RDFs and OWL ontologies are at the forefront of efforts to develop standardized semantic networks.

Many scientists in the semantic web services community, led by the W3C (the World Wide Wed Consortium), support the notion of OWL ontologies combined with RDFs for application to knowledge representation (32). Web version 3.0 serves as a foundation for semantic web services approaches. Version 3.0 is more robust and secure than Web versions 1.0 or 2.0. It has the capacity to convert RDF documents and OWL ontologies into integration tools, some of which are

Juanita Fernando is with the
Mobile Health Research Group (MHRG)
Faculty of Medicine, Nursing and Health Sciences
Monash University Vic 3800
Australia
Phone: +61 (0)3 9905 8537
Fax: +61 (0)3 9905 8134
Email: juanita.fernando@med.monash.edu.au

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008, Supplementary Conf. Proc., pp.261-264, 2008.

sufficiently mature to apply to actual production systems (32). Mashups, semantic web applications that combine content from more than a single source into an integrated experience, are clearly at the cutting edge of numerous efforts to eliminate standards barriers to bioinformatics research (31, 32).

World-wide governments are also aware of the standards barrier. For instance, a recent report to the Australian Health Information Council (AHIC), the basis of advice to ministers, suggests that although non-standardized developments may allow scientists to meet their own needs, it is clear from both national and international analyses that bioinformatics benefits can only be realized when systems are integrated and data can be shared and aggregated (5).

The standards shortcoming is exacerbated by conflict between computer scientists and biomedical experts which sometimes occurs on the basis of perceived notions of best practice (6). A wider understanding of the computer science perspective amongst biomedical scientists may partly address the conflict. Nonetheless, the application of computers to health knowledge representation is a fact of modern life so research practice needs to accommodate this (5).

At the same time, computer experts must understand the needs of biomedical scientists from a wide range of complex and comprehensive disciplines. Bioinformatics applications need to be usable, useful and contextual. The usability discipline concerns how easy, controllable, intuitive and satisfactory it is for users to work with systems and devices in a specific environment [7, 8, 9]. Usability errors concern machine interfaces that do not reflect the context of biomedical work, are not based on any contextual metrics, have navigation problems and feature confusing displays (10, 11). Unusable informatics applications are not useful or specific to scientific circumstances and molecular biology experts are rightfully wary of them. Thus, a consolidated set of best practices needs to specify useable as well as interoperable bioinformatics standards.

Standards can enable bioinformatics interfaces with a range of end-user applications, platforms, and database products, including proprietary systems and systems that have been built for differing purposes. The standards creation and ratification processes usually incorporate a working group that includes many recognized experts in the field. Thus, the standards process addresses concerns that either computer scientists or biomedical experts could dominate standards development.

III. The Interconnectivity Benefits

Interoperable standards developments in two key areas would benefit the bioinformatics discipline. The first area of benefit concerns the effective retrieval of textual evidence. Secondly, standards development could integrate bioinformatics with global e-health initiatives and so link the disciplines. After outlining the core ideas herein, each of the areas is discussed in turn.

A. A draft standards model

"Figure 1: Meta view of a draft bioinformatics and e-health interoperability model" illustrates the objectives of formal standards development. The figure shows various biomedical ontologies and technological platforms linking to a common language library mashup. The mashup builds links between the data, but does not warehouse any of it [32]. The language library facilitates systematic reviews and data mining algorithms, resulting in the complete capture of relevant information. Middleware, which can also be supported by data mashups, fosters interoperability within biomedical research disciplines as well as between them and between biomedical research and e-health frameworks. Albeit simple, Figure 1 crudely models the central standards development ideas expressed in this paper.

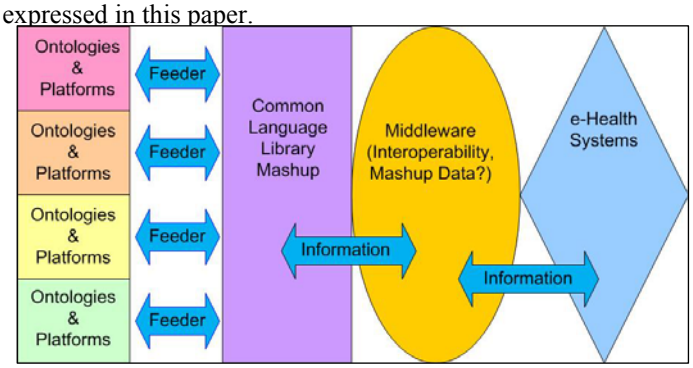

Figure 1: Meta view of draft bioinformatics and e-health interoperability model

B. The effective retrieval of textual evidence.

The effective retrieval of textual evidence is sub-divided into two core areas. They are systematic reviews and data mining.

Systematic reviews

Systematic reviews concern a structured literary search for objective information on a topic potentially answered by an analysis of evidence that has been systematically identified, critically appraised and summarized according to predetermined criteria. Reviews synthesize data from evidence to formulate study findings (12). Systematic reviews and meta-analysis of the data have become important approaches for evaluating associations across different study platforms and populations (13). The caliber of evidence collected by systematic reviews underpins the quality of the research findings. Therefore, a key factor in assessing the quality of such reviews is the complete capture of relevant studies (13).

Shared bio-ontologies, as defined by a formal standard, can support the capture of all relevant studies and so improve the quality of evidence for a systematic review. Yet much relevant information is currently not available to scientists for meta-analyses of study findings when they need it. For instance, standard queries of databases such as PubMed, the largest publicly available biomedical literature database, are labour-intensive and are not as productive as they might be due to bio-ontology standards shortcomings (13). The lack of standards means gathering information about genetic association studies is necessarily time-consuming and labour intensive (11).

Data mining

Data mining concerns the way scientists store, access, model, and ultimately describe and understand data sets (15). As with systematic reviews, data mining algorithms and their applications concern very large data sets and depend upon the complete capture of relevant information. A range of extensive, interrelated and complex foci are included in the biomedical discipline, from plant life to animal and human biological data-types [23, 24]. Competing or overlapping data mining tools magnify the demands of an already challenging discipline.

A review of publications indicates the proliferation of independently developed data mining tools, such as the BioWeka project, the Open Bioinformatics Foundation, MOBY iterations or Bio2RDF (16,17,32) The tools focus on semantic web formats, different classification systems, independent families of databases, the National Library of Medicine's Unified Medical Language System (UMLS) or annotation systems, and other computerized pattern matching tools such as OWL - DL (web ontology language - descriptive language) (3,18-19). Standards-based ontologies can reduce the duplication of work and the knowledge loss that results from this variety of independently developed data mining tools (11-18). Hence, the development of formal interoperable bioinformatics standards will improve the time-consuming, labor-intensive cost of scientific work while facilitating the complete capture of relevant data.

C. Bioinformatics and global e-health initiatives

Global e-health initiatives and biomedical research need to be linked. Technical hindrances to global sharing of information, such as low bandwidth, have essentially been resolved so impediments to true interoperability no longer exist (20). Patient histories taken by clinicians focus on diagnosing what was done or what happened, while bioinformatics research studies and constructs the 'why' part of patient care (2). Together, biomedical scientists and clinicians alike contribute to the development of comprehensive medical records (2). Devising an interoperable standard that could link e-health initiatives with bioinformatics studies would be a significant medical achievement. To this end, a number of international proposals for integrating patient histories, supported by standards-based e-health frameworks, have been underway for some time.

The international proposals include the OMG (Object Management Group)/CORBAmed (Common Object Request Architecture) service interface definitions, the UCL (University College London) SynEx telematics project, the CEN (Comité Européen de Normalisation**)** program, HL7 (Health Layer 7) proposals and OpenEHR (Electronic Health Record) (20-21). The proposals have contributed valuable material to standards development achievements.

A merger of the openEHR team with SynEx along with some convergence involving CEN's 13606 EHR standard, HL7 version 3, CDA (clinical document architecture), and other harmonization bridges into the openEHR group are advancing the creation of an interoperable open source implementation. The implementation will define the generic structural components of e-health systems and is made up of archetypes. 'Archetypes' are reusable elements that facilitate interoperability and evolve as knowledge and practice do (2, 20-21). Archetypes, supported by the model depicted in Figure 1, can advance the links between health care and biomedical research.

Archetypes may also present a way forward in terms of alleviating the best practice tensions between computer scientists and biomedical experts. OpenEHR's archetype efforts are open to software developers, students, researchers, scientists, governments, and Informaticians alike as they work to devise a knowledge-orientated computer framework that incorporates useable and interoperable standards for professionals across health care, biomedicine and related disciplines (22).

The construction of security standards to protect intellectual property and patient privacy presents another challenge that might be addressed by the applications of standards. Data security centres on three main pillars. The first pillar, data confidentiality, consists of all measures that protect information privacy. The second pillar is data availability, which concerns the ability to obtain information when and where it is required. The third and final pillar centers on data integrity or the accuracy and completeness of stored information [25, 27]. International proposals to achieve e-health frameworks incorporate work to achieve the data security standards. Thus, a standards-based link between e-health frameworks with biomedical research may address intellectual property concerns.

The effective retrieval of textual evidence and integration of bioinformatics with e-health initiatives are important contributions to bioinformatics research. Standards barriers impede the potential for productive and collaborative biomedical and clinical research efforts. Collective input into formal standards development from biomedical scientists, clinicians and computer experts will pioneer effective and interoperable research tools.

IV. CONCLUSION

The findings outlined herein are not new. However, this paper represents the first time the evidence has been assembled into a single paper. The development of formal bioinformatics standards can address the challenge of integrating medical research with clinical e-health efforts, foster the construction of a shared bio-ontology for complete data capture and lessen the barriers to collaborative research.

In order for formal standards development to benefit research efforts across the disciplines, descriptions based on end-user tasks and service models are required (33). It is not clear whether it is possible (or even desirable) for all services in the bioinformatics, health care and associated domains to be

standardized. To this end, we are establishing an informatics community across the Faculty of Medicine, Nursing and Health Sciences at Monash University.

The community of scientists aims to advance collaborative bioinformatics and health informatics research. We also hope to survey community members as to research tasks and contextual service models. The data will be collated with research looking at other data integration efforts, as are described above. Over the long term, the institutional, national and international creation of communities like these will address standards barriers to bioinformatics research.

ACKNOWLEDGMENT

The author thanks Mick Foy, Technology Services Group, Faculty Medicine, Nursing & Health Sciences, Monash University, for his review of ideas for this paper.

REFERENCES

[1] Fox J. What is bioinformatics? *The Science Creative Quarterly*. 2008 Sept 07- April 08(3).

[2] Rector A, Rogers J, Taweel A, Ingram D, Kalra D, Milan J, et al. *CLEF – Joining up healthcare with clinical and post-genomic research*. 2003 [cited 6 June 2008]; Available from: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/061.pdf

[3] Harris MA, Parkinson H. Standards and ontologies for functional genomics: Towards unified ontologies for biology and biomedicine. *Comparative and Functional Genomics*. 2003; 4(1):116-20.

[4] Spasić I, Schober D, Sansone S-A, Rebholz-Schuhmann D, Kell DB, Paton NW. *Facilitating the development of controlled vocabularies for metabolomics technologies with text mining*. Proceedings of the 10th Bio-Ontologies Special Interest Group Workshop 2007 Ten years past and looking to the future; 2007 20 July; Viena, Austria; 2007. p. S1.

[5] Australian Health Information Council. *Future Directions Briefing Paper for AHMAC meeting 4 October 2007*. [Government e-document] 4 October 2007 [cited 8 April 2008]; Available from: http://www.health.gov.au/internet/main/publishing.nsf/Content/B58193B329DDD77ACA25740F001F32D0/$File/AHIC%20eHealth%20Fut.Dir%20brief.paper-final.pdf

[6] Aranguren ME, Bechhofer S, Lord P, Sattler U, Stevens R. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC Bioinformatics* 8:57. 2007 February 20; 8(57).

[7] Schneiderman B. *Designing the user interface*. Third ed. Boston; MA: Third Ed: Addison Wesley 1998.

[8] Nielsen J. Medical usability: How to kill patients through bad design. *Alertbox* April 11 2005 [cited 13 April 2005]; eNewsletter]. Available from: http://www.useit.com/alertbox/20050411.html

[9] Johnston J, Eloff JHP, Labuschagne L. Security and human computer interfaces. *Computers & Security*. 2003; 22(8):675-84.

[10] Koppel R, Metlay JP, Cohen A, Abaluck B, Localio AR, Kimmel SE, Strom BL. Role of computerized physician order entry systems in facilitating medication errors. *JAMA*. 2005 March 9, 2005; 293(10):1197-203.

[11] Fernando J. Factors that have contributed to a lack of integration in health information system security. *The Journal on Information Technology in Healthcare*. 2004; 2(5):313-28.

[12] National Institutes of Health. National Information Center on Health Services Research and Health Care Technology. 2008. *HTA 101 Glossary* [cited 7 June 2008]; Available from: http://www.nlm.nih.gov/nichsr/hta101/ta101014.html

[13] Wei Yu W, Clyne M, Dolan SM, Yesupriya A, Wulf A, Liu T. GAPscreener: An automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*. 2008 22 April; 9(205).

[14] Berkeley Bioinformatics Source Project. *The open biomedical ontologies*. 2008.[cited 5 June 2008]; Available from: http://obofoundry.org/

[15] Hand DJ, Mannila H, Smyth P. (Eds) *Principles of data mining* MIT Press Cambridge MA. 2001.

[16] Open Bioinformatics Foundation. *Latest news and information about O|B|F affiliated projects*. 2008. [cited 5 June 2008] Available from: http://news.open-bio.org/

[17] Gewehr JE, Szugat M, Zimmer R. BioWeka extending the Weka framework for bioinformatics. *Bioinformatics*. 2007 March 1, 2007;23(5):651-3.

[18] Soldatova L, Batchelor C, Liakata M, Fielding H, Lewis S, King R. *ART: An ontology based tool for the translation of papers into Semantic Web format*. Proceedings of the SIG/ISMB07 Ontology Workshop. [cited 5 June 2008]; Available from: http://bio-ontologies.org.uk/index.php

[19] National Center for Biomedical Ontology. *Wiki: Workshop on clinical trial ontology*. 2008. [cited 5 June 2008]. Available from: http://esw.w3.org/topic/HCLS/ClinicalObservationsInteroperability/

[20] Schloeffel PD, Beale T, Heard S, Rowed DD. *Background and overview of the Good Electronic Health Record* (GEHR). 2001 [cited 30 May 2003]; Available from: http://www.gehr.org/Documents/BackgroundOverview_of_GEHR.htm

[21] Ocean Informatics. *CEN, CDA and openEHR*. 2006 17 March 2006 [cited 6 October 2006]; Available from: http://oceaninformatics.biz/CMS/index.php?option=com_content&task=view&id=31&Itemid=34

[22] OpenEHR. *Welcome to OpenEHR*. 2008.[cited 16 March 2005] Available from: http://www.openehr.org/about_openehr/t_home_aims.htm

[23] Schober D. *OBO Foundry naming conventions survey*. 2007.[cited 5 June 2008}: Available from: www.plantonlogy.org

[24] Goble, C, Stevens, R, & Karp, P. *Satellite meetings and SIGs*. 2008 [cited 5 June 2008] http://www.iscb.org/ismb2008/sigs.php

[25] Cheong I. Privacy and security of personal health information. *Journal of Informatics in Primary Care*, March, pp.15-18.

[26] International Standards Organization (ISO) *Why standards matter.* 2006. [cited 6 October 2006] Available from http://www.iso.org/iso/en/aboutiso/introduction/index.html#one

[27] Williams, P.A.H. *The role of standards in medical information security: An opportunity for improvement.* (2006*)* Proceedings of the 2006 International Conference on Security and Management (SAM '06), Las Vegas, Nevada; June 26-29. pp. 410-415

[28] National E-Health Transition Authority (NEHTA). *Delivering e-health's foundations.* (2006) [cited 13 February 2006] Available from: http://www.nehta.gov.au/component/option,com_docman/task,doc_download/gid,61/Itemid,139/

[29] Coiera, E., Westbrook, J. Should clinical software be regulated? In *eMJA,* 184(12.) (2006) [cited 19 June 2006] Available from: http://www.mja.com.au/cgi-bin/s.cgi?q=Coiera+2006&o=1

[30] Joint Technical Committee ISO/JTC 1. *ISO/IEC 19763-3:2007 Information technology - Metamodel framework for interoperability (MFI) - Part 3: Metamodel for ontology registration* (2006) [cited 625 June 2007] Available from http://www.saiglobal.com.ezproxy.lib.monash.edu.au/online/autologin.asp

[31] Librelotto, G. R., Machado, H. T., Martins, M., Ferreira, P. G. D., Ramalho, J. C., & Henriques, P. R.., *Topic maps applied to PubMed*. (2007). Proceedings of the Extreme Markup Languages 2007® conference, Montréal, Québec, August 7-10

[32] Belleau, F., Nolin, M-A., Tourigny, N, Rigault, P. & Morissette, J. *Bio2RDF: Towards a mashup to build bioinformatics knowledge system*. (2007) Proceedings of the 16th International World Wide Web Conference (WWW 2007), Banff National Park, Canada. May 8-12 2007

[33] Lord, P., Bechhofer, S., Wilkinson, M. D., Schiltz, G, Gessler, D., Hull, D., Goble, C., Stein, L. *Applying Semantic Web Services to bioinformatics: Experiences gained, lessons learnt*. (2007) [cited 25 June 2008] Available from: http://www.cs.man.ac.uk/~hulld/papers/applying_semantic_web_services_to_bioinformatics.pdf