Published in final edited form as:
Sci Transl Med. 2012 August 29; 4(149): 149fs32. doi:10.1126/scitranslmed.3004032.

Designing a Public Square for Research Computing

Daniel R. Masys[1,*], **Paul A. Harris**[2], **Paul A. Fearn**[1], and **Isaac S. Kohane**[3]

[1]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98195, USA

[2]Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

[3]Center for Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA

Basic, translational, and clinical research is becoming increasingly data intensive. To accelerate productivity, the U.S. National Institutes of Health (NIH) have for decades provided public support for computerized tools that permit the acquisition, management, communication, and analysis of research data (1–5). But the adoption of NIH supported computing tools by biomedical research institutions and investigators has varied widely. Here, we compare three contemporary efforts that have experienced divergent outcomes with respect to uptake by their intended users over the past five years. Our analyses stem from the perspective of our roles in each effort: as members of the Informatics Working Group (IWG) of the National Cancer Advisory Board (D.R.M. and P.A.F.), project directors for the Research Electronic Data Capture (REDCap) (P.A.H.), and the Informatics for Integrating Biology & the Beside (i2b2) data warehouse and workbench applications (I.S.K.).

BEHEST AND BIASES

The genesis of the analysis described here came in early 2012 with a charge from National Cancer Institute (NCI) Director Harold Varmus to the IWG chair (D.R.M.) to identify a recipe for success for future research software efforts. We appreciate that there are many equally valid opinions regarding the success or failure of large public research software programs, in particular those that provide payments to many individuals involved in software development and adoption. We also are aware that any author who is also a software developer is at risk of a self-promotional bias in attempting to opine on why one computer application achieves better adoption than another. With those caveats, we offer the following as a personal but informed perspective on publically funded tools for research computing.

We focused on three examples of efforts to create what is commonly referred to as enterprise-level software rather than the now ubiquitous concept of apps—small computer programs designed for personal use. Enterprise software has the characteristic of affecting the workflow of many different individuals within and among organizations and requires some level of professional information technology support by personnel dedicated to its installation, configuration, and use. Thus, enterprise software is usually associated with an institutional commitment to support multiple individuals and groups within an organization.

[*]Corresponding author. dmasys@u.washington.Edu.

Competing interests:
D.R.M. and P.A.F. are members of the National Cancer Advisory Board Informatics Working Group; D.R.M. is chair of the i2b2 Extramural Advisory Board. P.A.H. is project director for REDCap, and I.S.K. is project director for i2b2.

For purposes of this paper, we define success as adoption and implementation of enterprise software by at least 50% of the intended target group of institutions. For REDCap and i2b2, we used as the target group the 60 academic organizations that have received NIH Clinical and Translational Science Awards (CTSAs) (6); for caBIG, we used the 66 NCI-supported Cancer Centers.

CASE 1: REDCap

As a software application designed to collect and manage data for translational research studies, REDCap (7) uses a study-specific data dictionary to eliminate programming requirements for the creation of electronic case-report forms and patient-focused surveys. To use REDCap, researchers create a set of database variables to be captured along with the characteristics of those variables (called metadata) using a Web interface or spreadsheet software. The REDCap application then automatically creates the underlying database tables, Web-enabled secure data entry forms, and utilities for conducting data quality control and to update and download in formats amendable to statistical analysis. REDCap was initially developed in 2004 at Vanderbilt University with funding support from NIH's National Center for Research Resources (NCRR) General Clinical Research Center program. Since 2004, additional funding has included support from Vanderbilt's CTSA program award and an NIH/NCRR program development grant. In 2005, the university began sharing the software using a consortial collaboration-and dissemination model. Software and support are provided at no cost to academic institutions and nonprofit partners (www.projectredcap.org). As of this writing, the REDCap consortium contains 386 institutional partners across six continents with a user base of more than 50,000 researcher end-users, including installations at all of the current institutions (60) that have received CTSAs.

CASE 2: i2b2

As one of the NIH-supported National Centers for Biomedical Computing that were started in 2004, i2b2 uses the informational byproducts of health care and data from biological materials accumulated through the delivery of health care to provide computational tools for researchers (8). i2b2's major product is a Web-enabled research data warehouse server and associated search and analysis software. Implementation of the software is more complex than that of REDCap because of the need to import data into the research data server from existing local applications, such as electronic medical records systems. i2b2 software is free and open source (that is, it can be modified by users) and has been adopted in more than 60 academic health centers in the United States, including more than half of the CTSA institutions and 12 international medical centers. Although the software is free, each academic health center has to invest substantial local resources to install, support, and maintain the software and data extracted from electronic health records. i2b2 currently supports an Academic Users' Group of more than 300 members from more than 75 institutions that meet biannually for code workshops, discussion of application issues, preview of coming software, and networking (www.i2b2.org/work/aug.html).

CASE 3: ca BIG

The most ambitious case in terms of scope and complexity is that of the caBIG program, which began in 2004 as a high-profile research infrastructure initiative of NCI (5). caBIG sought to meet the "Tower of Babel" challenges of the growing volume and heterogeneity of research data by developing shared vocabularies, data models for research-data exchange, and a set of computer applications that embodied common data elements, standards, and methods for computers to communicate data via a grid architecture. Computer grids form a network that is capable of functioning as one large computer spread over many different

machines at geographically dispersed sites connected by the Internet. An early caBIG survey effort to identify the kinds of computerized tools needed yielded a list of dozens of potential applications spanning basic, translational, and clinical research. caBIG has been managed as a set of contracts, with strong central guidance to contractors on the technical factors needed to make computer programs compatible with the overall caBIG architecture. NCI also provided financial incentives to NCI-supported cancer centers to adopt caBIG tools as they became available and promoted participation in caBIG as a criterion for renewal of cancer center core grants.

By 2010, more than $300 million of public funding had been committed to the caBIG program, and Harold Varmus in his capacity as the newly appointed NCI director convened a program review by a group of technical experts who interviewed representatives from more than 50 NCI cancer centers. The review group's report, issued in March 2011 (9), found that caBIG had catalyzed progress in the development of community-driven standards for research data exchange and interoperability and had provided valuable support for investigator-initiated computer applications relevant to cancer research. A total of 32 computer applications had been supported, with program investments ranging from $100,000 to more than $9 million. However, the committee also found limited adoption by the intended community of users and a somewhat counterintuitive inverse relationship that tools with lower development costs had higher adoption by users. A comparison of the costs of individual applications and their subsequent usage can be found in (9); in no case was an enterprise-level software tool adopted by more than half of NCI-supported cancer centers. Data sharing and interoperability goals were far from being achieved, and the overall impact of the program was not commensurate with the level of public investment.

The report called out the main problems with the caBIG approach: A "cart-before-the-horse" grand vision; a technology-centric approach to data sharing; unfocused expansion; a one-size-fits-all technical approach; an unsustainable business model for both NCI and users; and lack of independent scientific oversight. The report recommended creation of an external oversight group (http://deainfo.nci.nih.gov/advisory/ncab/workgroup/caBIG/mission-Statement.pdf), which was formed in July 2011 and continues its work of assessing the value of ongoing and proposed NCI informatics projects.

SECRETS OF SUCCESS

The divergent natural histories of these three NIH-sponsored research-computing efforts offer insights into common success factors for similar kinds of projects. We group these insights into the following categories: standards, scope and complexity, and path to uptake.

A nonstandard lesson about standards

Anyone who has purchased an electrical device with confidence that it can be plugged into an outlet or has obtained cash from a remote automated bank machine can appreciate the utility of standards. In research, data standards lower the barrier to data sharing and offer the hope that insights might be gained by linking dissimilar data types that are related to the same biological system. But standards are a double-edged sword when applied to partially understood complex systems because they not only embody the notion of an approved set of names for biological objects and concepts, but also can convey an explicit set of relationships between those objects and concepts that can be inhibiting to scientific creativity or, worse, found to be wrong as science progresses. REDCap and the i2b2 workbench take a permissive stance on the use of data standards for both variable names and the allowable values that those variables can take. This makes the use of standards an optional, value-added activity that can be exercised when the benefits outweigh the additional costs of finding, selecting, and entering data in accordance with a relevant naming

standard. caBIG's laudable vision of interoperability came with a requirement for use of a variety of standards at multiple levels. These included data standards that have been well accepted but also complex technical standards for communications between computers that many developers and potential users found difficult to implement and out of balance with the perceived benefit. As well, the process for extending existing standards and creating new standards within caBIG was perceived by many as onerous and excessively time consuming.

In the perhaps inevitable tension between the perceived functionality provided by the computer software and its requirements for the use of standard methods of naming and manipulating data, the examples described here suggest that sponsors and developers should not place a higher importance on the use of standards than on providing functionality that is directed to researchers' immediate needs. The best software plays to both themes by making adoption of standards easier rather than harder as compared with non-standardized alternatives, but this is not always possible. The best standards are those vetted by the marketplace as being genuinely useful rather than those mandated by a sponsor. The one-size-fits-all criticism of the caBIG approach also suggests that it may be best to allow intra-organizational and inter-rganizational needs and technologies to converge or diverge as needed so as to maximize scientific productivity. Standards are particularly important when sharing data beyond the borders of the lab or institution that created them; thus, it appears prudent to focus standards efforts designed to facilitate data sharing on those scientific problems in which no single lab or institution can solve the problem without the collaboration and data provided by other labs and organizations—recognizing that in an era of global science, this scenario may apply in the majority of cases.

Scope and complexity

The ambitious agenda and complex technical architecture of caBIG suffered from the basic truth that in computing, increased functionality that is built at the expense of increased system complexity is at elevated risk of delays in development, an inability of intended users and support staff to understand the systems created, and being overtaken by other technical approaches that are similar but easier to implement. The clear message is that one framework should not attempt to encompass all clinical and translational research information technology problems. As well, the burgeoning environment of available internet-aware applications lobbies for investing in relatively simple and minimalist interfaces between applications rather than in complex architectures. If there is a program-management message from the successes of REDCap and i2b2, it is one of using scientific domain experts to lead small, nimble software development teams in solving one challenge at a time. These challenges, in turn, should represent problems that intended users really care about and which, from the users' perspective, can be successfully implemented in a few months to perhaps a year. When the needed functionality is identified, the development time to a useful first software product needs to be very short (months, not years) because of the rapid pace of change in both the biomedical sciences and information technologies (10).

Path to widespread uptake and use

All three of the examples described here had a strategy of demonstrating success first with a small group of the most advanced users or institutions, then having others follow these pathfinders. They differed in incentives for adoption, with caBIG providing financial incentives to NCI-supported cancer centers to install and use caBIG-developed applications, whereas REDCap and i2b2 were offered under similar no-cost licensing arrangements but did not include financial incentives for adoption. Getting free software of this complexity is akin to getting a free pony: The acquisition costs are dwarfed by local implementation and ongoing maintenance expenses. The caBIG experience makes it apparent that organizations that cannot afford ongoing staffing and help desk functions for software support should not

be expected to adopt software even if it is free. Furthermore, enterprise-level software that affects the workflow of many units of an organization is likely to be rejected by users or units or to become financially unsustainable, especially if key drivers for its implementation are coming from outside the institution. From the perspective of a software sponsor, the use of financial incentives to encourage adoption of software products is an unsustainable business model because the costs, particularly for software maintenance and inevitable enhancements, grow ever larger. Thus, even for publicly sponsored computing tools the most successful approach to enhancing uptake and use is to collaborate with organizations that demonstrate their willingness to invest their own assets in implementation of the software.

GUARANTEED SUCCESS?

An important limitation of these comparisons is that they are based on a convenience sample of only three contemporary NIH supported programs—albeit nationally prominent ones—and are not a systematic review of the entire field. We harbor no delusions that there is a simple formula for successful adoption of a new computational tool by clinical and translational researchers, who are subject to a complex environment of scientific opportunities, incentives, regulatory and economic constraints, and heterogeneous organizational capabilities. We offer these observations as suggestions to guide sponsors and developers in the current fast-paced milieu of biomedical research. Immediate and obvious functionality that empowers researchers to accomplish their goals is still the most dominant characteristic of any successful research computing application or infrastructure.

Acknowledgments

Funding:

i2b2 is supported by NIH award 5 U54 LM008748, and REDCap is supported by NIH award 8 U54R00012302.

References and Notes

1. Baruch JJ, Barnett GO. Real-time shared on-line digital computer operations. J. Chronic Dis. 1966; 19:377–386. [PubMed: 6004868]
2. Raub WF. The PROPHET system and resource sharing. Fed. Proc. 1974; 33:2390–2392. [PubMed: 4435244]
3. Musen MA. Stanford Medical Informatics: Uncommon research, common goals. MD Comput. 1999; 16:47–48. 50. [PubMed: 10202424]
4. Bethesda, MD: NIH; The Biomedical Information Science and Technology Initiative (BISTI). www.nih.gov/about/director/060399.htm.
5. Buetow KH. An infrastructure for interconnecting research institutions. Drug Discov. Today. 2009; 14:605–610. [PubMed: 19508923]
6. NIH CTSAs. www.ctsacentral.org.
7. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. J. Biomed. Inform. 2009; 42:377–381. [PubMed: 18929686]
8. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. J. Am. Med. Inform. Assoc. 2012; 19:181–185. [PubMed: 22081225]
9. caBIG. http://deainfo.nci.nih.gov/advisory/bsa/bsa0311/caBIGfinalReport.pdf.
10. Matta N, Krieger S. From IT solutions to business results. Bus. Horiz. 2001; 44:45–50.