Strategies for Sustainable Open Source Projects for Clinical and Translational Research: Lessons from the Trenches

Elizabeth K. Nelson, PhD¹ (Moderator), Leon Rozenblit, PhD, JD², Michael Mendis, BS³, Ben Bauman, BA⁴, Mark Igra, BS¹
¹LabKey Software, Seattle, WA; ²Prometheus Research, New Haven, CT; ³Harvard Partners, Boston, MA; ⁴OpenClinica, Waltham, MA;

Abstract

Theoretically, taking an open source approach can broaden the public benefits of grant-funded software projects; increase the leverage of informatics investments; draw upon a wider pool of contributors and expertise; and improve transparency, reproducibility, and extensibility. However, as Dr. Isaac Kohane has warned, open source software is ‘‘...free like a pony. You still have to feed it and clean up after it’’ (TEDMed, 2013). Furthermore, simply making software open source does not ensure that it will become immediately useful to others. This panel will cover practical strategies for generalizing, sustaining, and evolving open source software developed for clinical and translational research. Panel members will address sustainable business models, feasibility of grant support, implications of different open source licenses, modes of dissemination (including community norms for attracting open-source evangelists), community-building approaches, practical trade-offs, and unexpected challenges. Panelists represent open source platforms that have proven useful across multiple organizations and shown sustainability over time. Platforms include LabKey Server (http://labkey.org), RexDB (http://rexdb.org), i2b2 (https://i2b2.org/), and OpenClinica (https://openclinica.com/).

Background

‘‘i2b2 is free and open source. Free like a pony. You still have to feed it and clean up after it.’’

- Isaac Kohane, TEDMed Conference, Washington DC, April 19, 2013

‘‘Open source has the half life of a graduate student.’’

- Don Listwin, Xconomy Forum, Seattle, Washington, May 12, 2010

Despite the idealistic promise of open source software, actually delivering broad, lasting public benefit requires solid business strategy, long-term vision for software architecture, evangelism verging on mania, and gallons of elbow grease. Otherwise, as Don Listwin quips, an open source tool will have the ‘‘half life of a graduate student.’’

The increasing scope of clinical and translational research keeps raising the bar for informatics tools, both in the features delivered and the confidence investigators need in software longevity before considering adoption. For example, the longer the time frame researchers expect a study to continue, the longer software must be sustained, supported and extended to meet evolving needs.

For open source tools for clinical and translational research to become more widely accepted and adopted, the informatics community needs to better understand how to build, sustain and share open source projects in a way that supports field-specific needs and expectations.

To fulfill the promise of multiplying returns from informatics funding, open source tools need to be designed and developed in a way that meets the needs of a broader community, not just initial inventors. Furthermore, reusing proven open source, not reinventing the wheel, needs to become a norm for the field.

For biomedical open source projects to remain in good health (and cheerfully supported) after initial grants expire, the research community needs to explore and share business/funding models that have proven practical.

Panelists will address the value of open source approaches in clinical and translational research, the potholes that impede success, and strategies that have worked for their teams in establishing sustainable, widely used open source platforms.

Panelists and Presentations

Panelists represent open source platforms that have proven sustainable over time and useful across multiple clinical and translational research organizations. As of 2013, the software systems represent 29 collective years of experience as public open source projects. The platforms and their dates of first release are: LabKey Server (2005), RexDB (2005), i2b2 (2007), and OpenClinica (2005).
Each panelist will provide an overview of the open source project he/she represents, describe the project’s “special sauce,” and explore lessons learned. For ease of comparison, panelists will also address a predefined list of core points. The different platforms’ strategies will be summarized incrementally on slides whose content are built across the presentations. Issues and metrics covered for each platform will include: (1) target users, (2) core value for users, (3) business/funding model, (4) challenges of funding model, (5) open source license, (6) number of active installations, (7) team size, (8) origins (academic or otherwise), (9) dissemination approach, and (10) community-building strategies. This will allow clear, consistent comparison of the platforms and strategies used to grow, maintain and evolve them, plus the tradeoffs made by each platform according to its focus.

**Platform #1: LabKey Server**

LabKey Server is an open source platform for large-scale, translational research. The system helps teams of researchers collaborate smoothly and make sense of the flood of complex data produced by modern biomedical research, from novel assay and ‘omics results to clinical reports to specimen information. LabKey Server supports web-based integration, analysis and secure sharing of diverse data types within distributed research teams. Upon research publication, the system can serve as a portal for public, interactive exploration of published analyses and de-identified data, opening doors to validation and extension of results.

Installations of the LabKey Server platform serve leading scientific organizations all over the world, including the Immune Tolerance Network (ITN TrialShare: http://itntrialsshare.org), the Statistical Center for HIV & AIDS Research at the Fred Hutchinson Cancer Research Center (FHCRC) (Atlas: http://atlas.scharp.org) and NWBioTrust (http://www.nwbiotrust.org). There are currently over 100 active installations of the platform.

LabKey Server originated within the FHCRC, with its first public release in 2005. In the same year, the project’s developers founded LabKey Software to support, extend, and sustain the platform beyond what was possible within a research institute. Today, research organizations use LabKey Server’s rich API to develop new features independently or in partnership with LabKey Software. Code contributed by users (such as an electronic health record system for primate centers) undergoes code review by the LabKey Software team before addition to the source depot.

Groups that purchase LabKey Software’s services apply a portion of this funding to maintenance of core platform infrastructure, including documentation. Successful dissemination and community-building strategies have included publishing papers, hosting a yearly user conference, hosting/supporting community message boards, and in-person outreach.

Challenges include funding core platform maintenance and installers; funding innovation that goes beyond current customer projects; growing the team quickly enough to meet user needs; simplifying the new user experience for a tool that serves a broad variety of researchers; and managing expectations for the cost of support for free software.

LabKey Server source code, compiled binaries, documentation, and tutorials are professionally maintained and freely available under the Apache 2.0 license at [http://www.labkey.org](http://www.labkey.org).

**Platform #2: RexDB**

The Research Exchange Database (RexDB®) is an extensible, web-native software platform that helps researchers securely collect, integrate, manage and share data. RexDB embraces the flexibility required to support the needs of dynamic scientific collaborations. RexDB systems can be regularly updated to meet the changing needs of ongoing research projects, allowing many modifications to be performed by nontechnical software users. RexDB is built on top of sustainable open-source components, ensuring that research data will never be marooned in a proprietary format or a legacy system. A unique, layered architecture allows novice and advanced users, data managers, statisticians, and local IT staff to effectively interact with the system, while a highly granular privileging model ensures easy compliance with local and federal regulations.

RexDB is presently the data collection, data warehousing, and data sharing software platform for numerous longitudinal multidisciplinary research studies, including the Simons Foundation’s Simons Simplex Collection and Variation in Individuals Project, the Yale Child Study Center’s Yale Autism Research Database, and the centralized data solution at Emory University’s Marcus Autism Research Institute. Dr. Rozenblit will address the following points.

- Users: Research organizations that generate data with high potential reuse value
- Sales pitch: Reduce the cost of data integration and repurposing across all your studies, data types and sites; transform your data into a value-generating asset.
• License strategy: Affero General Public License version 3 (AGPLv3), contributor covenants; special “permissive” licensing on high-value components

• Licensing challenges: managing tension between a for-profit’s desire to limit competition and commitment to open source

• Funding model: NSF and NIH grants; professional open source, coupled with data management services, with service delivery methods accelerated by OS technology

• Funding challenges: Professional services model limits profit margins, which limits funding for product enhancement

• Open-source community building: Conferences (technical/scientific), Meetups

• Special sauce: Promoting and releasing high-value individual components as developer tools (e.g., HTSQL, cogs)

• Lessons learned: Tightly scoped developer tools are easier to promote than integrated systems.

RexDB is available at: http://www.rexdb.org/.

Platform #3: i2b2

i2b2 (Informatics for Integrating Biology and the Bedside) is a scalable informatics framework that enables clinical researchers to use existing clinical data for discovery research.² i2b2 is funded as a cooperative agreement with the National Institutes of Health (NIH).

Both open source communities and proprietary companies have benefited from the ability to add and extend the functionality of the i2b2 software by way of modular components called “cells.” An example of an i2b2 cell would be a de-identification cell that strips names and dates from reports. Cells are exposed through web services, so proximity is not assumed. Remote cells can be hosted, so code does not have to be shared. For the open source community, i2b2 provides an infrastructure that consists of a wiki site, source code repository, and bug tracker. Two models are available for the development of other cells: a “Sponsored Projects” model, which provides the complete open source stack, and “Related Projects” model that includes a wiki site. As with all open source projects, i2b2 continues to push to engage the community and incorporate input and enhancements into the i2b2 platform. These efforts aim to broaden the community and improve the overall software. Within the last year, i2b2 has worked on running on a complete open source stack, with the last components being the use of an open source database.

i2b2 source code, virtual machines images, documentation and tutorials are available at http://www.i2b2.org under the i2b2 license.

Platform #4: OpenClinica

OpenClinica is an open source clinical trials data management platform. With thousands of implementations and a community comprising over 18,000 people, the software is used in over 100 countries across a diverse spectrum of both academic and industry research. OpenClinica helps increase the flow of clinical trial data, provides a more adaptable research IT infrastructure, and reduces the barriers to obtaining enterprise-quality electronic data capture and clinical data management systems.

Mr. Baumann will describe elements crucial to OpenClinica’s success since the project’s inception, recount certain things that have not worked as planned, and provide an overview of the OpenClinica community and commercial open source model that has allows it evolve into a sustainable initiative.

OpenClinica is distributed under the GNU Lesser General Public License (GNU LGPL) at https://community.openclinica.com.

Plan for Audience/Panel Interaction

The moderator will prepare questions for the panelists, but the focus of the interactive portion of this session will be on audience questions. In particular, this panel aims to help nascent open source projects strategize their own modes of sustainability and dissemination. The Q&A session will allow the audience to dig further into the varied strategies used by the open source projects represented on the panel.

Acknowledgements

We gratefully acknowledge grants from the NIH, including support for LabKey Server (UM1AI068618 and U01AI068635) and i2b2 (U54LM00874).

All coauthors have agreed to take part in the panel.

References