# Uncertainty About the Long-Term: Digital Libraries, Astronomy Data, and Open Source Software

Peter T. Darch
School of Information Sciences
University of Illinois at Urbana-Champaign
Champaign, IL, USA

Ashley E. Sands
Department of Information Studies
University of California, Los Angeles
Los Angeles, CA, USA

*Abstract*— **Digital library developers make critical design and implementation decisions in the face of uncertainties about the future. We present a qualitative case study of the *Large Synoptic Survey Telescope* (*LSST*), a major astronomy project that will collect and make available large-scale datasets. LSST developers make decisions now, while facing uncertainties about its period of operations (2022-2032). Uncertainties we identify include topics researchers will seek to address, tools and expertise, and availability of other infrastructures to exploit LSST observations. LSST is using an open source approach to developing and releasing its data management software. We evaluate benefits and burdens of this approach as a strategy for addressing uncertainty. Benefits include: enabling software to adapt to researchers' changing needs; embedding LSST standards and tools in community practices; and promoting interoperability with other infrastructures. Burdens include: open source community management; documentation requirements; and trade-offs between software speed and accessibility.**

***Keywords— Astronomy; Big data; Big science; Data management; Data curation; Knowledge infrastructures; Long term; Open source; Scientific data***

## I. Introduction

Recent decades have seen rapid improvements in technologies for the production, processing, management, and accessibility of scientific data. Many large-scale projects exist whose primary objective is to produce large collections of digital data for researchers. These projects are examples of digital libraries in the sense defined in [1], i.e. they collect, manage, and preserve high-quality, rich digital content. Further, they provide functionality (such as the ability to access, query, aggregate, and integrate data) to meet researchers' specialized needs. Codified policies determine who is able to access and use these libraries' content, and under what conditions. Notable examples include the *Human Genome Project* (*HGP*) in bioinformatics and the *Large Synoptic Survey Telescope* (*LSST*) in astronomy, a major sky survey scheduled to begin collecting data in 2022 [2].

Digital library developers make critical decisions during conception and development of these projects, months or years before projects will collect and make available data. One set of decisions relates to managing uncertainty about the long-term. The success of digital libraries for scientific data rests on meeting requirements, such as users' research objectives and interoperability with other research infrastructures. However, the dynamic nature of research means these requirements can change significantly over time. While addressing stakeholder requirements has already been a core concern of work on digital libraries for scientific data [3], less attention has been paid to how digital library developers can strategize to meet unpredictable future stakeholder requirements.

A second set of decisions relates to policies that determine accessibility of the library's content. Research on digital libraries has largely focused on accessibility of data [4]. However, projects such as HGP and LSST also produce large amounts of code that underpin operations of their digital libraries. These projects must decide whether and how to make this code available to users. Diverging from previous sky surveys, LSST has chosen to develop and release its data management code open source. A key rationale for LSST's choice to adopt an open source approach is to help mitigate uncertainty about the future.

This paper explores how these sets of decisions intersect, and evaluates the potential of an open source approach for addressing uncertainty, by attending to the following questions:

1. What sources of uncertainty about the future concern developers of digital libraries for scientific data?
2. How does an open source approach to developing software address sources of uncertainty?

## II. Background

In common with other digital libraries [1], operating digital libraries for scientific data involves multiple components, including: identifying and defining users (human and non-human); meeting user requirements for content, quality, and functionality; making policy decisions about accessibility; and ensuring interoperability with other systems. Difficulties in addressing these components are exacerbated when digital libraries serve researchers over the long-term.

### *A. Digital Libraries for Scientific Data*

Multiple technical, scientific, and social factors determine the success or failure of digital libraries for scientific data [3]. Many factors relate to meeting user requirements by providing access to, and retrieval of, data meeting researchers' needs [5]. For example, the library must be designed to mirror the scientific understandings of users. Ontologies built into the underlying database structure need to mirror closely the domain's understandings of how scientific phenomena relate to each other, including classification schemas and underlying physical laws. Digital library users also require metadata sufficient to interpret and assess the quality of data [6].

Sponsor: Alfred P. Sloan Foundation (#20113194, #201514001)

978-1-5386-3861-3/17/$31.00 ©2017 IEEE

The success of a digital library for scientific data also relies on interoperability with tools and services used by researchers, and other components of research infrastructure [7]. For instance, data formats should be compatible with researchers' preferred data processing and analysis tools. The digital library may also be interoperable with other digital libraries to enable researchers to search across, and integrate data from, multiple collections. Further complicating digital library development is heterogeneity in the research priorities and preferred tools of users [5].

### B. Research Infrastructures and Change

Each factor outlined above is subject to change over time, complicating the development of digital libraries for scientific data. For example, researchers' scientific understandings can be transformed by discoveries of new objects or phenomena, or by studies that suggest new relationships between already-known phenomena. Meanwhile, tools and methods used by researchers change as technologies evolve.

Studies of infrastructures for scientific research reveal infrastructure success is affected by the ability to anticipate and adapt to change [8]: 1) during development, when decisions are made in the face of uncertainty about future user requirements; and 2) during operations, when infrastructures may have to adapt to rapid changes in user requirements.

Infrastructure developers must manage the possibility of stakeholder requirements shifting in the future. One strategy is to reduce the prospect of future change by promoting standardization of practices among users, such as tools and methods for data collection and analysis [9]. However, excessive standardization can stifle the ability of researchers to conduct innovative work.

Another strategy is to enable greater infrastructure adaptability in response to change. Research infrastructure may adapt in the following senses: existing components of the infrastructure can be modified; new features can be built from scratch and added to the infrastructure; and workarounds can be devised to overcome existing limitations [10]. Developers may adapt infrastructure as changes in user requirements occur. However, as this strategy can place an unmanageable burden on developers, an alternative approach to adaptability is to develop infrastructure in a way that empowers users to adapt and reconfigure infrastructures themselves [11].

### C. Open Source Software for Science

Developers of digital libraries for scientific data must also make critical decisions about policies relating to accessibility of their library's content. While research on these libraries has largely focused on access to data products [4], libraries also often produce large quantities of code to support operations. Digital libraries must consider whether and how to make this software available beyond the development team.

Open source approaches to scientific software promise benefits for scientific research [12], like promoting research integrity: access to software underpinning published results can facilitate reproducibility or verification. Open source software also allows researchers to use computational tools without having to pay expensive licensing fees, while empowering researchers to adapt software to their needs. Adaptations made by individuals can even be incorporated into future software releases. The potential for researchers to adapt software to their own requirements suggests adoption of open source approaches by digital libraries for scientific data offers these libraries a possible method to address uncertainty about the future. However, open source approaches also impose many burdens on developers, including generating, training, and supporting the community of software users [13].

## III. Large Synoptic Survey Telescope

The Large Synoptic Survey Telescope (LSST) was initially conceived in the mid-1990s, research and development began in 2003, and construction in 2014. The projected overall budget for LSST is around $1.1 billion, primarily from US federal sources. LSST aims to generate 15 terabytes of data per night during its operations period (2022-2032). These data will be made openly accessible to researchers and the public in the USA and Chile, with access elsewhere negotiated on a country-by-country basis. User interfaces will enable access to, and use of, the data. Data will support research in many astronomy subdomains, such as studies of dark energy, the Milky Way, the solar system, and transient phenomena [2]. The LSST Data Management team is developing software that will underpin operations. This *Data Management Software Stack* is released open source through GitHub, with new releases scheduled every six months.

## IV. Case Study Methods

This paper presents findings from an eighteen-month case study of LSST. The study followed standard qualitative methods, including interviews (n=60), observation, and document analysis [14]. The interview sample comprised: LSST Data Management (DM) leaders, team managers, scientists, and software engineers; other managers within LSST whose work interfaces with the DM team; and members of LSST leadership. Interviews lasted between 45 minutes and two hours, with most between 60 and 75 minutes. Interview transcripts totaled 1227 pages. The authors spent 14 weeks observing LSST project members at their home institutions as well as in sub-team and project-wide meetings. A corpus of LSST operating documents, totaling 1380 pages, was assembled. Data were coded and analyzed using NVivo 9, a software package supporting qualitative research.

## V. Findings

### A. Sources of Uncertainty About the Future

LSST faces multiple sources of uncertainty about the circumstances in which it will operate from 2022-2032.

#### 1) Uncertainty about users' research priorities

LSST team members face difficulties anticipating what questions researchers will use LSST data to address. The team expects major shifts in researcher priorities both between now and the start of operations, as well as during operations. These shifts may even relate to astronomical phenomena not yet discovered. Recent decades have seen discoveries of novel phenomena, such as dark energy and exoplanets, that have provoked deep shifts in understandings of the universe; LSST team members expect this trend to continue. Researchers will also use LSST data to address questions about already-known

phenomena, but in ways that cannot yet be predicted. Theories that characterize the behavior of these phenomena often change over time. Classification schemas can also change: one example is the reclassification of Pluto as a "dwarf planet." The topics of interest to researchers will also shift as funding agency priorities evolve. Researchers will require that LSST data infrastructure adapts to all of these sources of change.

*2) Uncertainty about tools and expertise*

A second source of uncertainty relates to the computational methods researchers will use to access and analyze LSST data. Between now and when operations conclude, the LSST team expects many advances in the computational tools available to researchers. LSST data and software must remain interoperable with changing researcher workstation operating systems.

LSST team members are also uncertain about the skill and comfort some researchers will have with computationally- and data-intensive methods during its operations period. Although a number of astronomy subdomains have seen a surge in recent years in use of these methods, distribution of relevant expertise is uneven across researchers and subdomains.

*3) Uncertainty about other instruments*

Identifying and characterizing *transients* is one of LSST's main science drivers. Transients are astronomical phenomena that have a limited visual lifespan, even as short as seconds or minutes. LSST leadership anticipates that many of the project's major outcomes will relate to transients. During operations, LSST will broadcast an alert to researchers and observing instruments external to the project within sixty seconds of observing a transient. LSST aims to establish a global network of these instruments to provide rapid follow-up observations of transients in response to these alerts. Some follow-up instruments already exist; others are under construction or being considered for funding. Forming this network is critical for LSST's success. However, the LSST team faces uncertainty about future availability of instruments to join the network. The funding environment for astronomy is unpredictable: extant instruments and those under construction are vulnerable to loss of funds, while those being considered for funding may not be built. Other uncertainties relate to whether instruments that are operational will be able to follow-up LSST alerts. In some cases, these instruments will be in use for other purposes. In other cases, instruments' computer systems may not be interoperable with LSST software and data.

### *B. Open Source Software and Uncertainty*

LSST leaders decided early on to develop and release the project's Data Management Software Stack open source. The LSST team hopes that many researchers and large-scale instruments and facilities will begin using these tools for their own work in advance of LSST operations. This strategy has many potential benefits, but also imposes multiple burdens on LSST team members and resources.

*1) Benefits of an open source strategy*

Making the software stack open source promises to help address sources of uncertainty about the future in multiple ways. One way is addressing uncertainty about the future scientific priorities of researchers. By enabling users to employ the software stack in their own research, and request or build new features as necessary, the LSST software stack can co-evolve with changes in the scientific priorities of researchers.

Open source software can also help address uncertainty about researchers' future tools and expertise. By making software available to researchers during construction, LSST hopes many researchers will develop skills and comfort with the computational methods needed to fully exploit LSST data. Open source development also means software can be adapted by users to changes in their workstation operating systems.

Finally, open source software can reduce uncertainty about instruments available for follow-up observations. Making LSST software available to these instruments can promote interoperability between LSST and these instruments' systems.

*2) Burdens of an open source strategy*

While an open source approach to the LSST software stack appears a promising strategy for addressing uncertainty, fully realizing this promise places significant burdens on LSST team members, resources, and infrastructure. One burden is the work involved in recruiting and managing a community of researchers who can use and build upon the LSST software stack. This work involves devising strategies to promote the software stack to potential users; implementing systems for users to seek technical support, report bugs, and make requests for new features; and reviewing code produced by users before deciding whether to incorporate it into future releases.

A second burden is the level of documentation required for LSST code. This documentation must be accessible to a range of users, who may not have in-depth software engineering knowledge and familiarity with LSST development practices. Accordingly, LSST has devised, and is enforcing, stringent standards for documentation for its software engineers. A number of these engineers worked previously on the *Sloan Digital Sky Survey* (*SDSS*), a sky survey whose data management code was not released fully open source. These engineers spoke of how SDSS requirements for code – which needed to be understood only by other SDSS software engineers familiar with SDSS practices – required less work to conform to than LSST requirements.

A third burden is the prospect of reduction in the speed of the software stack. LSST has faced a trade-off between processing speed and ensuring the stack remains accessible to as many users as possible. While the core of the stack is written in the C++ programming language, the control layer (the part of the stack with which the end-user is more likely to interact) is written in Python. This decision was made because Python is generally easier to learn and use than C++, making it accessible to more users. However, Python runs many times slower than C++.

## VI. Discussion and Conclusions

The task of building digital libraries for scientific data is fraught with many difficulties. Not only must developers meet a range of stakeholder requirements [3], but these requirements are typically subject to significant scientific, technological, and infrastructural change during both the development and operations phases of the digital library. The challenges of anticipating future change are particularly acute for big data

projects, given their development and operations periods often unfold on the order of decades, rather than years.

The case study presented here demonstrates three particularly significant sources of uncertainty facing the Large Synoptic Survey Telescope's (LSST) team: the scientific topics researchers will use LSST data to address; the tools and expertise researchers will have at their disposal; and the available instruments for follow-up on transient alerts. The first two sources are already known as posing significant challenges to developers of research infrastructure [8]. However, LSST also faces uncertainty about the future availability of instruments external to the project. Digital libraries do not exist in isolation; their value proposition often relies on the extent to which they are interoperable with other research infrastructure [7]. No digital library can assume the future existence and interoperability of this external infrastructure.

Digital library developers must also determine who can access its underlying code, and under what conditions. In addition to the usual rationales advanced for open source scientific software, such as promoting reproducibility and making tools accessible to researchers [12], this case study considers another rationale: the management of uncertainty about the future. LSST has chosen to develop and release its software open source. This approach combines strategies of promoting standardization and enabling adaptability [9], [11]. By releasing software open source, LSST seeks to promote its software practices as standards within the astronomy community. If LSST practices become community standards, the project will be better able to predict future methods employed by end-user researchers, and to ensure future interoperability with other astronomy infrastructures. LSST also intends to enable adaptability of its own infrastructure by allowing users to request or develop new features of LSST software to meet their changing needs.

Despite its promise, developing and releasing software open source is not an easy solution to address uncertainty. The LSST software development team faces major burdens such as increased documentation requirements and management of users. These are burdens commonly found in open source development efforts in general, and not just limited to cases involving digital libraries for scientific data [13]. The LSST case study also revealed another significant burden, namely a trade-off between the software's technical performance and its accessibility. This burden is particularly applicable to digital libraries for scientific data. To promote standardization across the scientific domain it serves, and to enable adaptability to a wide range of changing requirements, the software must be accessible to as wide a range of researchers as possible. Future work will involve more precisely quantifying these benefits and burdens and exploring the extent to which they arise in digital libraries across other scientific domains.

Digital libraries for scientific data face uncertainty about the future. Releasing and developing software open source has potential benefits for addressing this uncertainty, as well as helping to advance open science more generally. However, this approach also imposes significant resource burdens and trade-offs on digital libraries and their developers.

Digital library stakeholders should consider early in a project whether they have the capacity to develop and release collections through an open source environment. While open source development can mitigate many sources of uncertainty about the future, library budgets and staff may not be able in the short- and medium-term to devote necessary resources. When available resources are sufficient, this study has shown that open source approaches to software can reinforce the success of digital library investments into the future.

## ACKNOWLEDGMENT

We thank LSST team members for participating in our study, and other members of the UCLA Center for Knowledge Infrastructures for discussion on topics in this paper.

## REFERENCES

[1] LSST Science Collaboration *et al.*, "LSST Science Book, Version 2.0," arXiv e-print, Nov. 2009.

[2] C. Lagoze and K. Patzke, "A research agenda for data curation cyberinfrastructure," in *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, 2011, pp. 373–382.

[3] C. L. Borgman, P. T. Darch, A. E. Sands, J. C. Wallis, and S. Traweek, "The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management," in *2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*, London, 2014, pp. 257–266.

[4] L. Candela *et al.*, "The DELOS digital library reference model. Foundations for digital libraries," DELOS - A Network of Excellence on Digital Libraries, 2007.

[5] J. C. Wallis, M. S. Mayernik, C. L. Borgman, and A. Pepe, "Digital libraries for scientific data discovery and reuse: from vision to practical reality," in *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, Gold Coast, Queensland, Australia, 2010, pp. 333–340.

[6] I. M. Faniel and T. E. Jacobsen, "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data," *J. Comput. Support. Coop. Work*, vol. 19, no. 3–4, pp. 355–375, Sep. 2010.

[7] D. Ribes, K. S. Baker, F. Millerand, and G. C. Bowker, "Comparative Interoperability Project: Configurations of Community, Technology, Organization," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 2005. JCDL '05*, New York, 2005, pp. 65–66.

[8] D. Ribes and J. B. Polk, "Flexibility Relative to What? Change to Research Infrastructure," *J. Assoc. Inf. Syst.*, vol. 15, no. 5, pp. 287–305, 2014.

[9] S. B. Steinhardt and S. J. Jackson, "Anticipation Work: Cultivating Vision in Collective Practice," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, New York, NY, USA, 2015, pp. 443–453.

[10] M. J. Bietz and C. P. Lee, "Adapting cyberinfrastructure to new science: tensions and strategies," in *Proceedings of the 2012 iConference*, 2012, pp. 183–190.

[11] L. K. Johannessen, D. Gammon, and G. Ellingsen, "Users as designers of information infrastructures and the role of generativity," *AIS Trans. Hum.-Comput. Interact.*, vol. 4, no. 2, pp. 72–91, 2012.

[12] D. S. Katz *et al.*, "Report on the Second Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE2)," *J. Open Res. Softw.*, vol. 4, no. 1, Feb. 2016.

[13] E. H. Trainer, C. Chaihirunkarn, A. Kalyanasundaram, and J. D. Herbsleb, "From Personal Tool to Community Resource: What's the Extra Work and Who Will Do It?," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015, pp. 417–430.

[14] M. Hammersley and P. Atkinson, *Ethnography: Principles in Practice*. London, UK: Routledge, 2007.