



Science Gateways
Community Institute

PresQT



US
Research
Software
Sustainability
Institute

Research Software and Science Gateways: Addressing Sustainability, Usability and Reproducibility Challenges to Enhance Research

Sandra Gesing
sandra.gesing@nd.edu

Webinar at
NITRD Program's Software Productivity, Sustainability, and Quality
Interagency Working Group

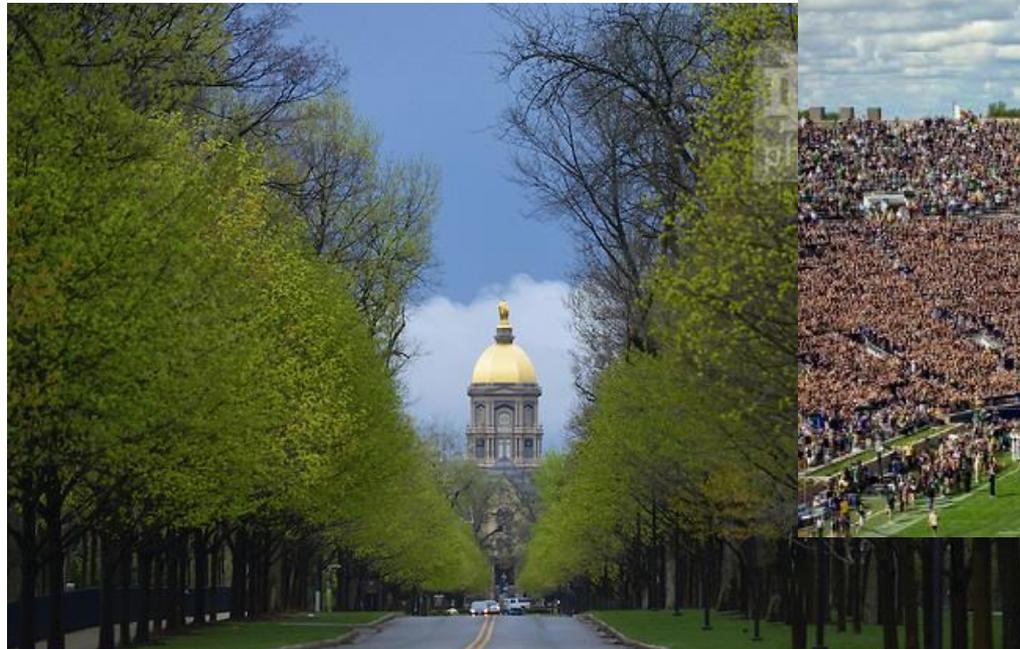


December 6, 2018



University of Notre Dame

- In the middle of nowhere of northern Indiana (1.5 h from Chicago)
- 4 undergraduate colleges
- ~35 research institutes and centers
- ~12,000 students



Center for Research Computing

- Software development and profiling
- Cyberinfrastructure/science gateway development
- Computational Scientist support
- Collaborative research/
grant development
- System administration/
prototype architectures
- Computational resources:
25,000 cores+
- Storage resources: 3 PB
- National resources (e.g., XSEDE)
- ~50 researchers,
research programmers,
HPC specialists



CRC and OIT building



CRC HPC Center (old Union Station)

<http://crc.nd.edu>

Software Sustainability

Sustainable software is software which is:

- Easy to evolve and maintain
- Fulfils its intent over time
- Survives uncertainty
- Supports relevant concerns (Political, Economic, Social, Technical, Legal, Environmental)

(Patricia Lago at WSSSPE4)



WORKING TOWARDS SUSTAINABLE
SOFTWARE FOR SCIENCE:
PRACTICE AND EXPERIENCES



The importance of sustainability

Sustainability means that the software you use today will be available - and continue to be improved and supported - in the future.

Better science through superior software

Our work is focussed around four themes we believe are fundamental to doing research correctly in the digital age. These are related to **our manifesto**.

The first of these is **Skills and Training**: creating a capable research software community by enabling access to software development training for all researchers and teaching them methods to advance their research.

Recognition and Reward promotes and contributes to systems of credit for good software development and reuse practice.

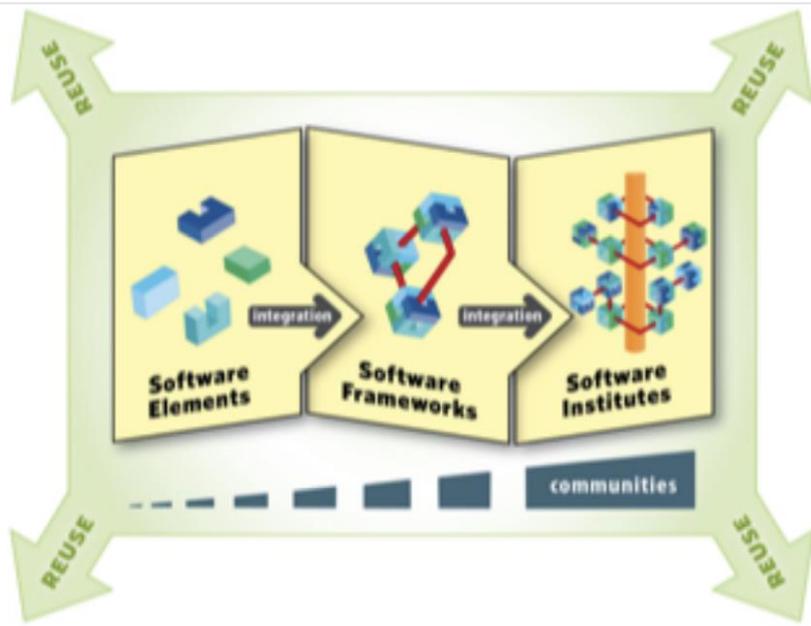
Career Paths recognises and champions the varied job roles associated with research software; with a primary focus on the academic sector but suggesting industrial practice where applicable.

Finally, **Reproducible Research** promotes the fundamental place of software in supporting confidence in the research process and its results.

Taken together, these enable the efficient and effective use of software to tackle both the grand challenges that push the boundaries of human knowledge to day-to-day research software tasks.

<https://www.software.ac.uk/about>

Sustainability for Cyberinfrastructure - NSF



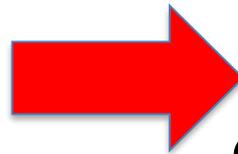
Elements: Small groups - create & deploy robust capabilities for demonstrated need to advance science & engineering.

Framework Implementations: Larger teams organized around the development and application of common infrastructure aimed at solving common research problems, resulting in a sustainable community framework serving a diverse community or communities.

Planning Grants for Community Cyberinfrastructure: Focus on long-term capabilities in cyberinfrastructure to serve a research community of substantial size and disciplinary breadth.

Community Cyberinfrastructure Implementations: Focus on long-term hubs of excellence in cyberinfrastructure and technologies, to serve a research community of substantial size and disciplinary breadth.

SI2
Software Infrastructure for
Sustained innovation



CSSI
Cyberinfrastructure for Sustained
Scientific Innovation

Sustainability for Cyberinfrastructure - NSF

Sustainability Institutes and Excellence Hubs are funded to support the CI and research community

Conceptualizations

- US Research Software Sustainability Institute (URSSI)
- Geospatial
- ...

Implementations

- Science Gateways Community Institute (SGCI)
- The Molecular Sciences Software Institute (MolSSI)
- Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)

Research Software



Use

90%

95%

Can't
continue
without

70%

63%

Research Software

> 50% neither formal nor informal training in software engineering



Use	90%	95%
Can't continue without	70%	63%

Research Software



Use

90%

95%

Can't
continue
without

70%

63%

Research Software



How to cite software?

Use	90%	95%
Can't continue without	70%	63%

<http://doi.org/10.5281/zenodo.843607>

Areas of Concern

- Functioning of the individual and team
- Functioning of the research software
- Functioning of the research field itself



Developing a pathway to
research software sustainability

Functioning of the Individual and Team

- Training & education
- Ensuring appropriate credit for software development
- Enabling publication pathways for research software
- Fostering satisfactory and rewarding career paths for people who develop and maintain software
- Increasing the participation of underrepresented groups in software engineering

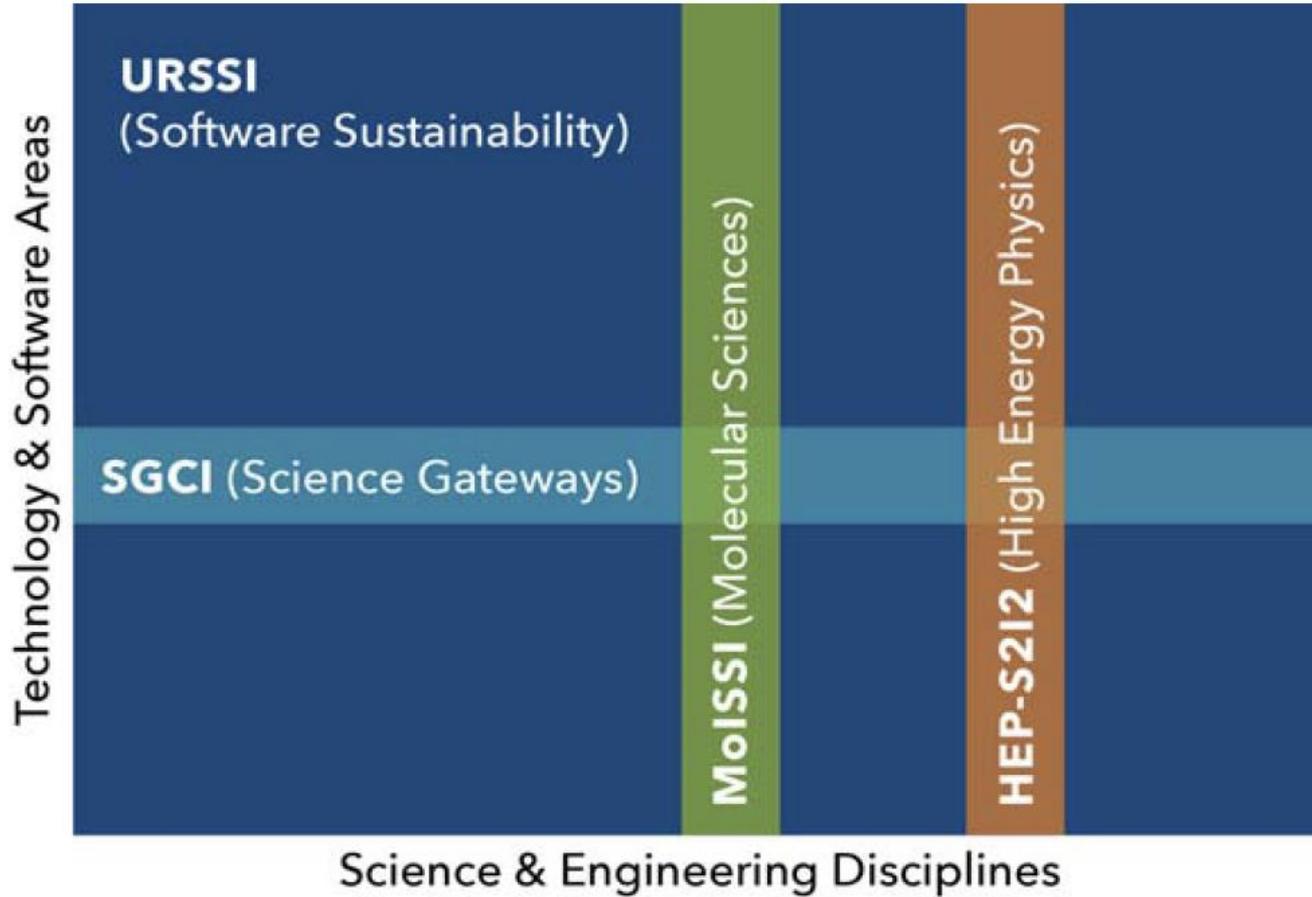
Functioning of Research Software

- Supporting sustainability of the software
- Growing community, evolving governance, and developing relationships between organizations, both academic and industrial
- Fostering both testing and reproducibility
- Supporting new models and developments (e.g., agile web frameworks, Software-as-a-Service)
- Supporting contributions of transient contributors (e.g., students)

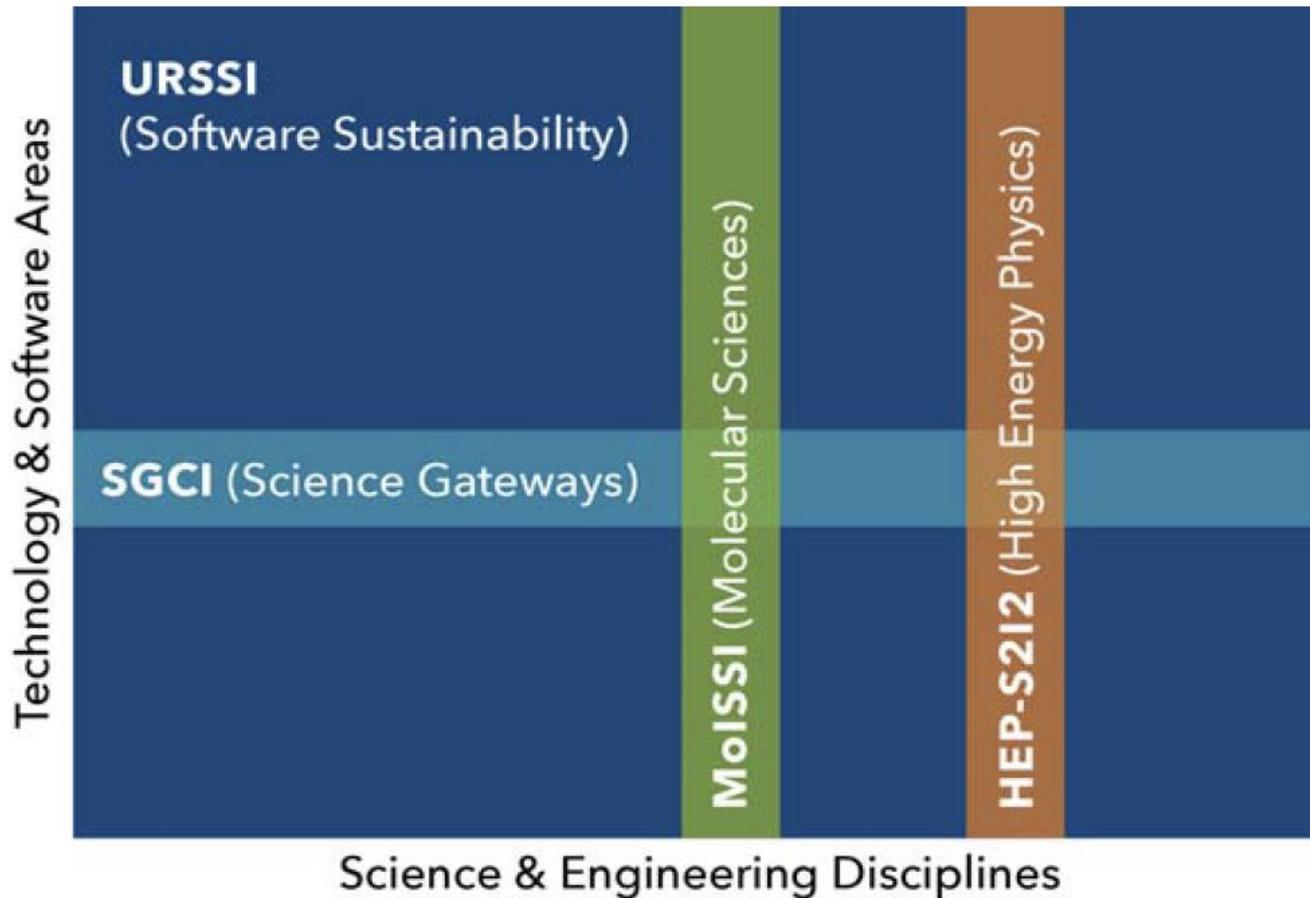
Functioning of the Research Field Itself

- Growing communities around research software and disparate user requirements
- Cataloging extant and necessary software
- Disseminating new developments
- Training researchers in the usage of software
- Understanding and improving pipelines of diverse developers and maintainers

URSSI and Other S2I2 Projects



URSSI and Other S2I2 Projects



Goal: Close collaboration and fill in gaps on each axis

Partner with URSSI

We don't want to reinvent the wheel but partner with existing initiatives!

- UK SSI
- Software and data carpentries
- ACI-REF VR
- ...



Software
Sustainability
Institute

Online sustainability evaluation

The following evaluation is a short, free, online version of the full sustainability evaluation that the Institute can perform for your project.

It takes about 15 minutes to complete the questionnaire, which gives you the opportunity to review the main issues that affect the sustainability of your software. At the end of the evaluation, a report will be generated and emailed to you with sustainability advice that is tailored to your project.

All questions are mandatory and need to be completed before you can progress through the evaluation.



Initial Straw Man

	Supporting software	Supporting science	Supporting community	Science Impact
Development support	x			x
Incubator	x			x
Training	x	x		x
Policy		x	x	x
Community	x	x	x	x

Conceptualization

- Workshops
 - First workshop took place in April in Berkeley
 - Second workshop took place in October in Chicago
 - Software credit workshop will take place in January in Santa Barbara
 - Incubator workshop will take place in February in Maryland
- Survey with about 1200 answers – in analysis
- Ethnographic studies
- Mission and vision working group

How to Engage with URSSI

- Watch the website <http://urssi.us/>
- Repos for website and workshops <https://github.com/si2-urssi>
- Blog posts <http://urssi.us/blog/>
- Join the mailing list <http://urssi.us/>
- Discuss <https://discuss.urssi.us/>
- Twitter <https://twitter.com/si2urssi>
- If you have questions, want to suggest something, want to volunteer, email us: contact@urssi.us

Technology-Enhanced Research

- Increased complexity of
 - today's research questions
 - hardware and software
 - skills required
- Greater need for openness and reproducibility
 - Science increasingly driving policy questions
- Opportunity to integrate research with teaching
 - Better workforce preparation

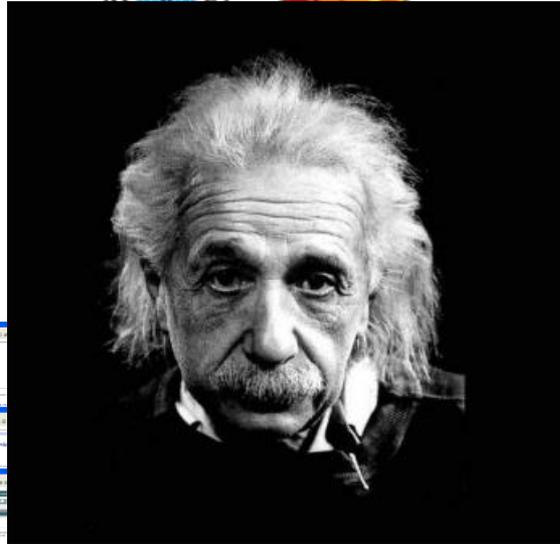
*We need end-to-end solutions that provide **broad access to advanced resources and allow all to tackle today's challenging science questions***

➔ Science Gateways

Data and compute-intensive problems



Web-based agile frameworks

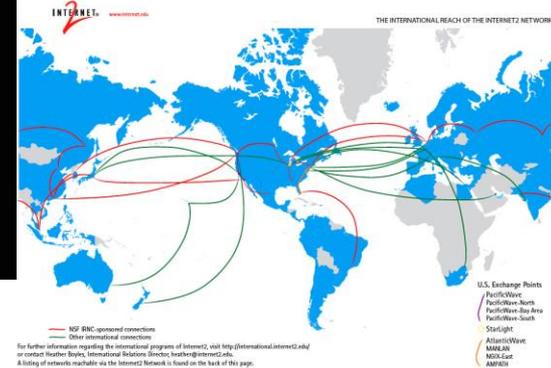


Users generally not IT specialists

Distributed data and computing infrastructures



Tools and workflow engines



High-speed networks

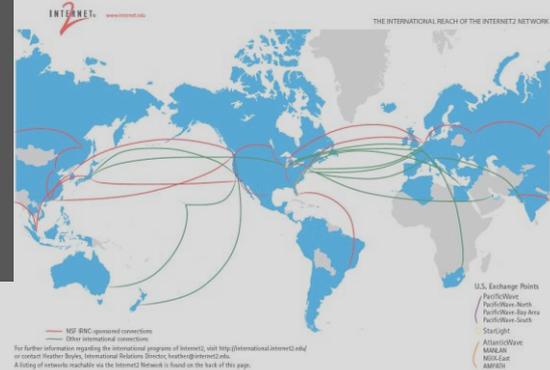
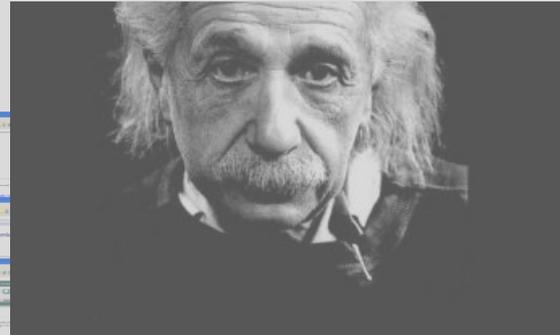
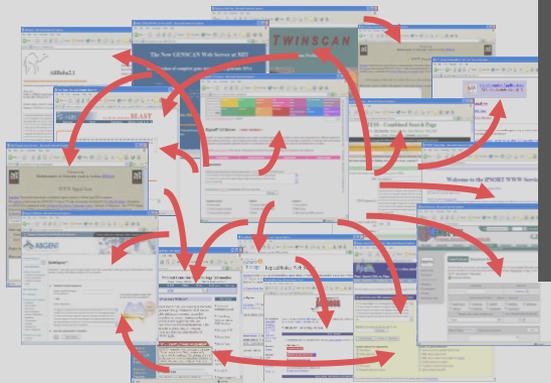
Data and compute-intensive problems

Web-based agile frameworks

Distributed data and computing infrastructures



Need for science gateways!



Tools and workflow engines

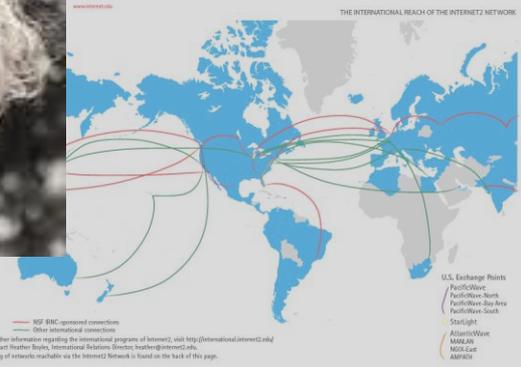
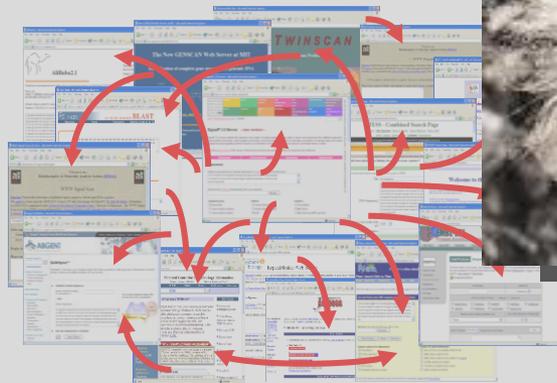
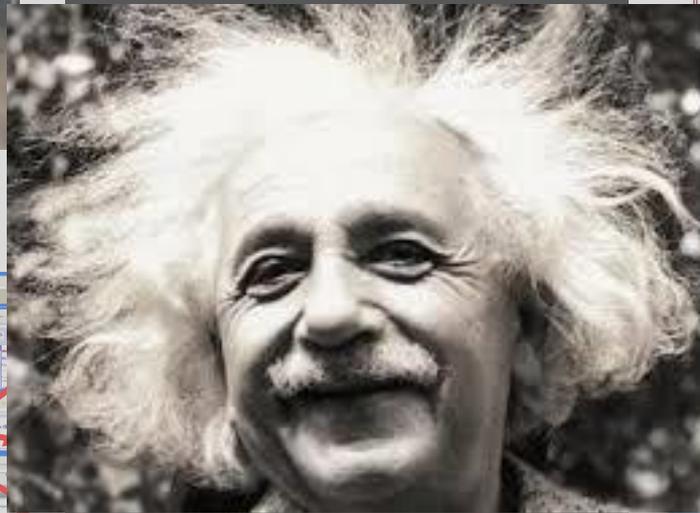
Users generally not IT specialists

High-speed networks

Data and compute-intensive problems

Web-based agile frameworks

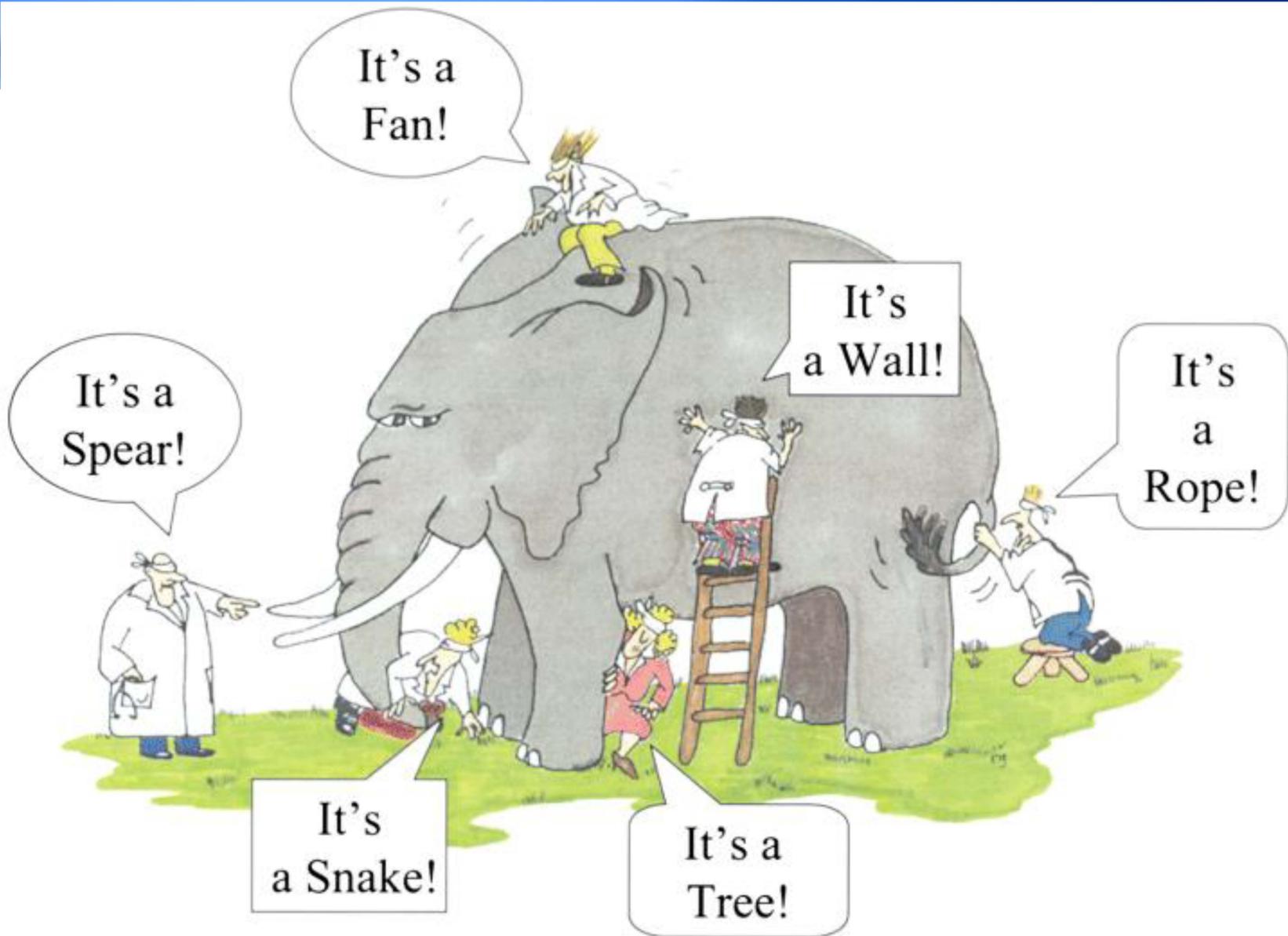
Distributed data and computing infrastructures

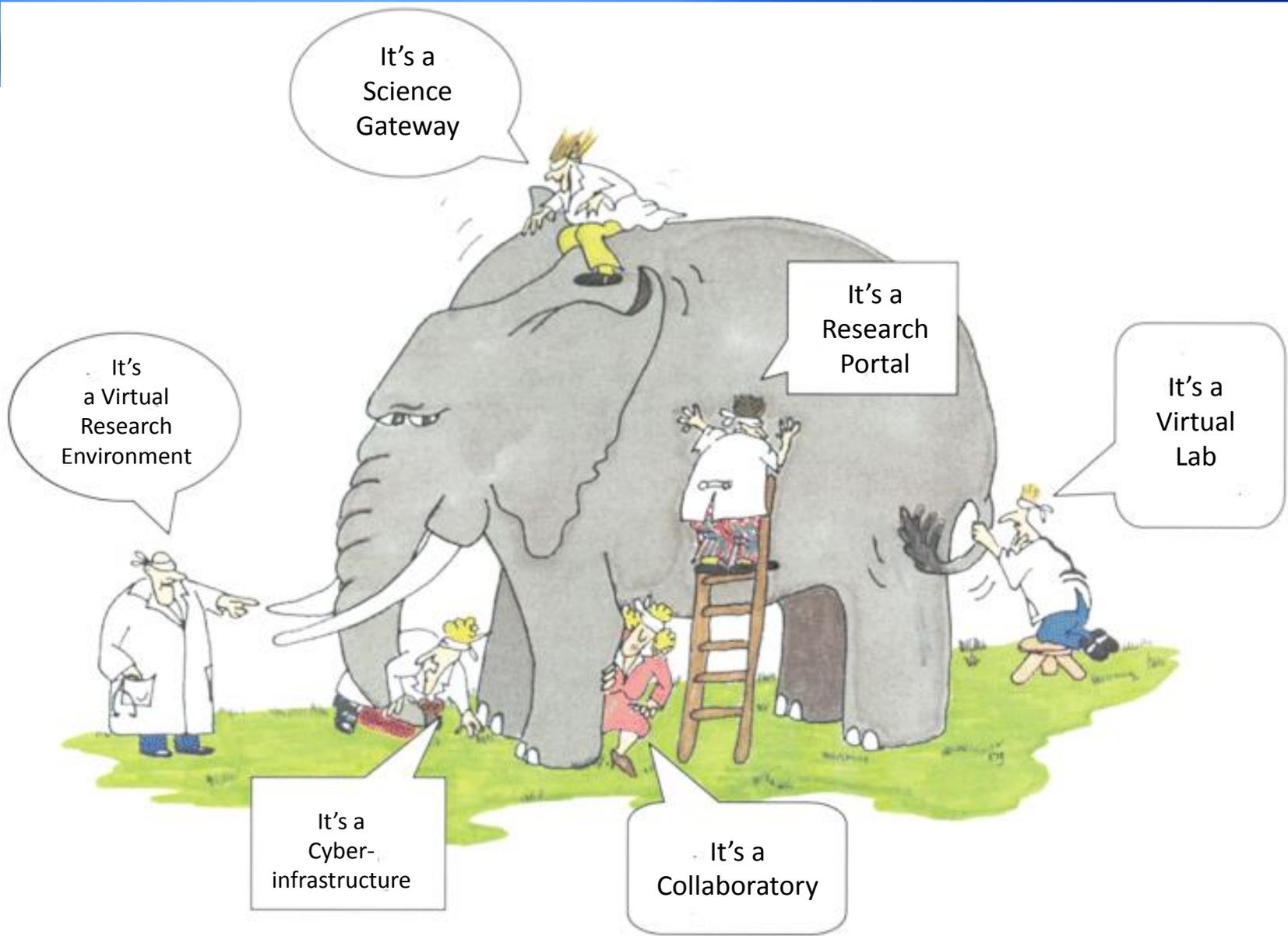


Tools and workflow engines

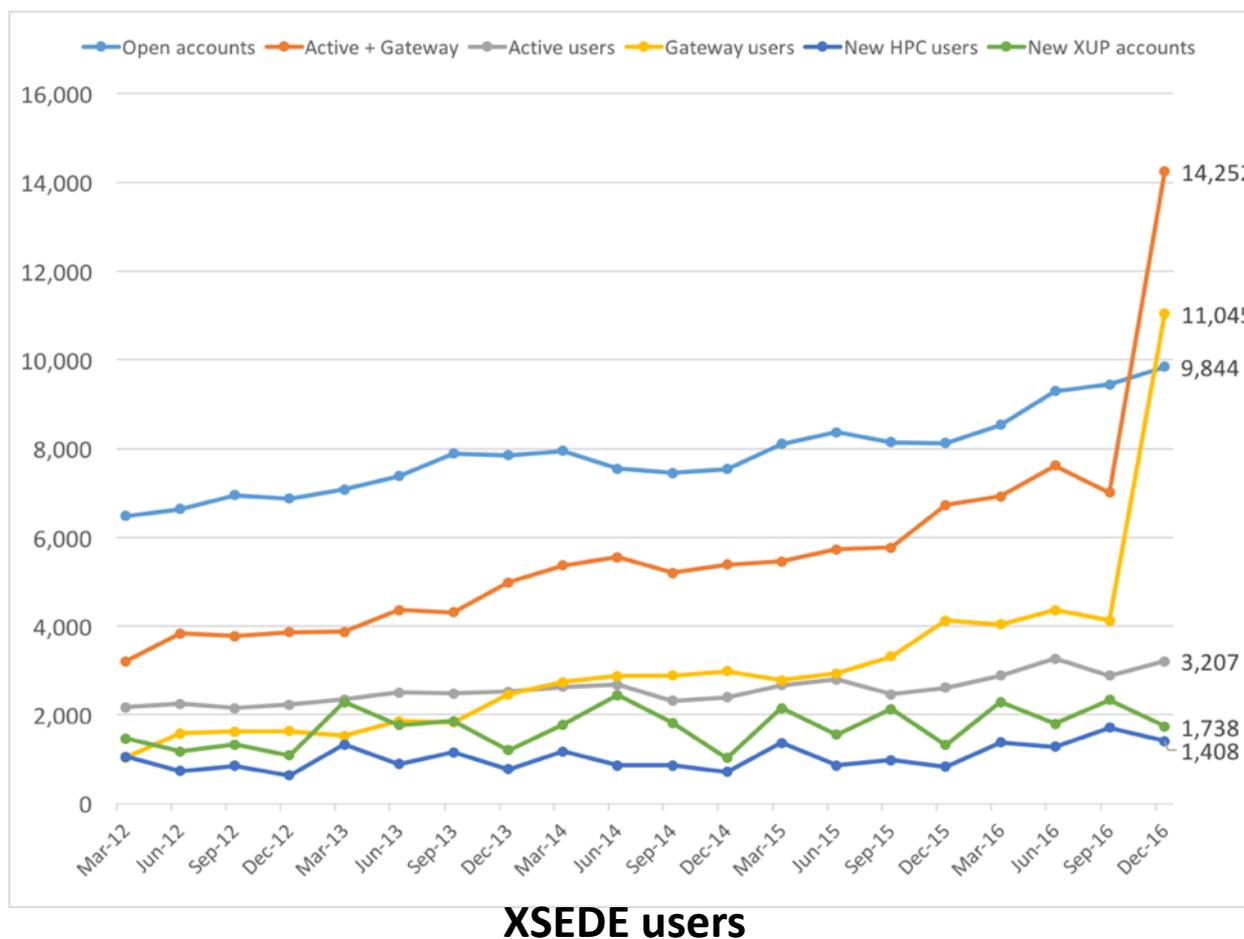
Users generally not IT specialists

High-speed networks





Gateway users are 77% of active XSEDE users in Q4 2016



All users

Gateways

Login

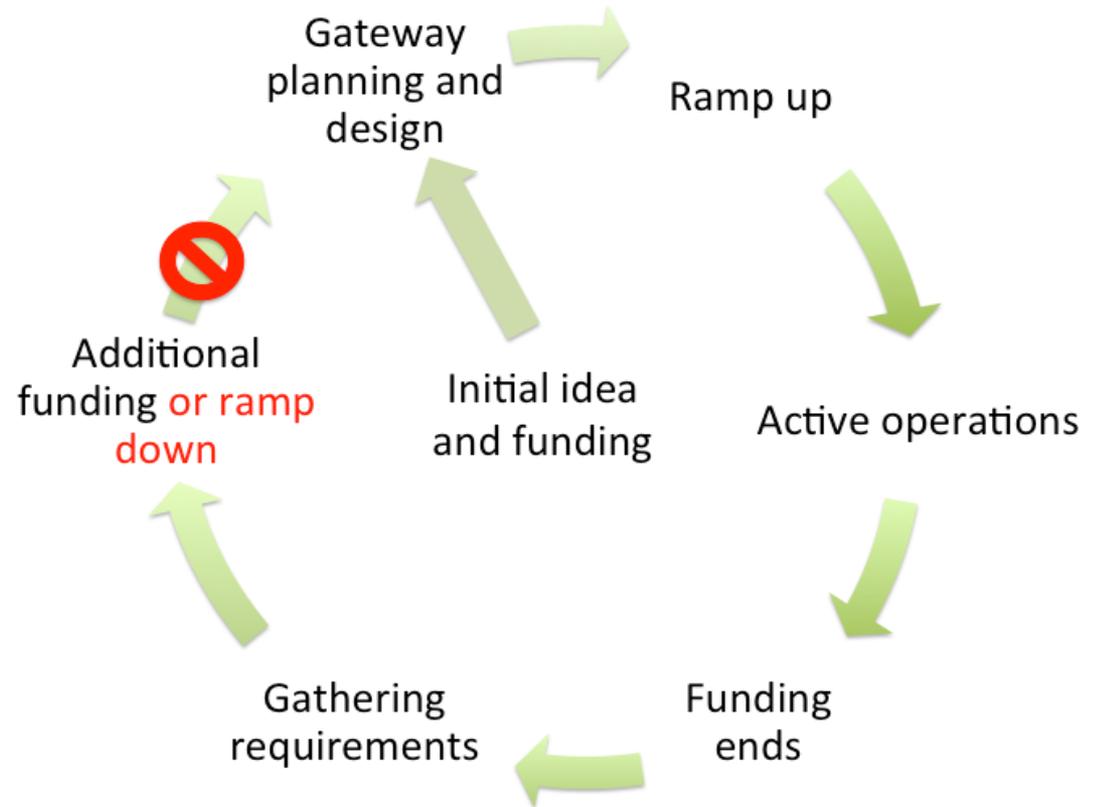
This is largely due to the CIPRES and I-TASSER gateways, but others are gaining

Life Cycle of a Science Gateway

Developers typically

- work in isolation
- must bridge to variety of resources
- need building blocks in order to focus on higher-level functionality
- struggle to secure sustainable funding

Sounds familiar?



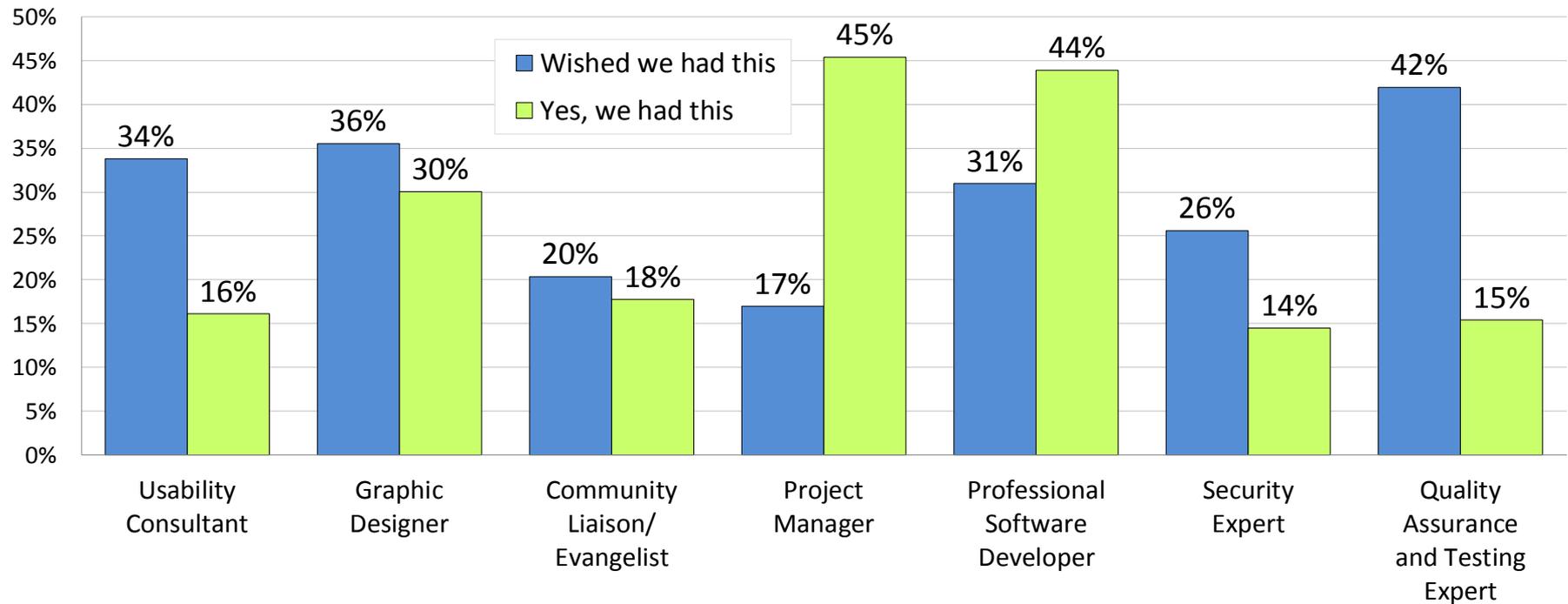
Science Gateway Survey 2014

- sent out to 29,000 persons
- 4,957 responses from across domains
- 52% from life, physical or mathematical sciences
- 32% from computer and information sciences or engineering
- 45% develop data collections
- 44% develop data analysis tools

What services would be helpful?

Proposed Service	% Interest
Evaluation, impact analysis, website analytics	72%
Adapting technologies	67%
Web/visual/graphic design	67%
Choosing technologies	66%
Usability Services	66%
Visualization	65%
Developing open-source software	64%
Support for education	64%
Community engagement mechanisms	62%
Keeping your project running	62%
Legal perspectives	61%
Managing data	60%
Computational resources	59%
Mobile technology	59%
Database structure, optimization, and query expertise	59%
Data mining and analysis	58%
Cybersecurity consultation	57%
Website construction	57%
Software engineering process consultation	53%
Source code review and/or audit	51%
High-bandwidth networks	45%
Scientific instruments or data streams	44%
Management aspects of a project	38%

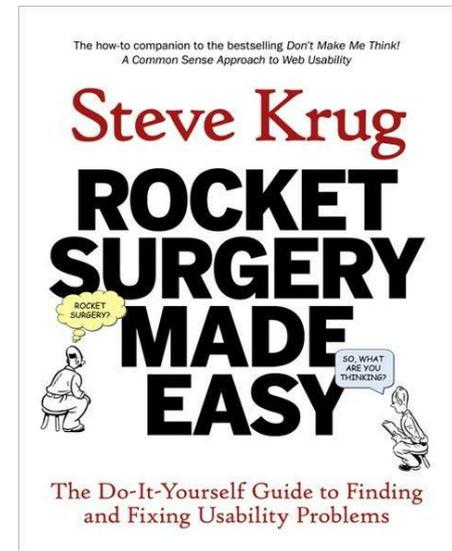
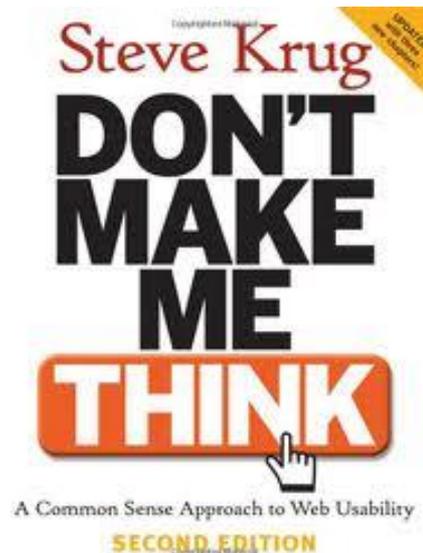
Well-designed gateways require a variety of expertise



Usability

“After all, usability really just means that making sure that something works well: that a person ... can use the thing - whether it's a Web site, a fighter jet, or a revolving door - for its intended purpose without getting hopelessly frustrated.”

(Steve Krug in “Don't make me think!: A Common Sense Approach to Web Usability”, 2005)



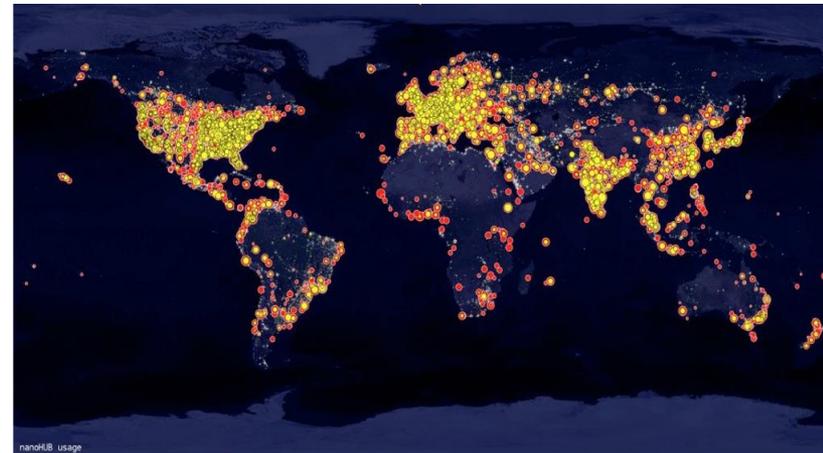
Technologies

- Widely used complete frameworks (Galaxy, HubZero, Globus Online etc.)
- RESTful APIs and support of multiple programming languages in widely used frameworks (Apache Airavata, the Agave platform, etc.)
- Reused interface implementations such as the one of CIPRES with its RESTful API (CIPRES has served more than 20,000 users to date)
- Science gateways as a service with provision of hardware in the background such as SciGap (Science Gateway Platform as a Service)

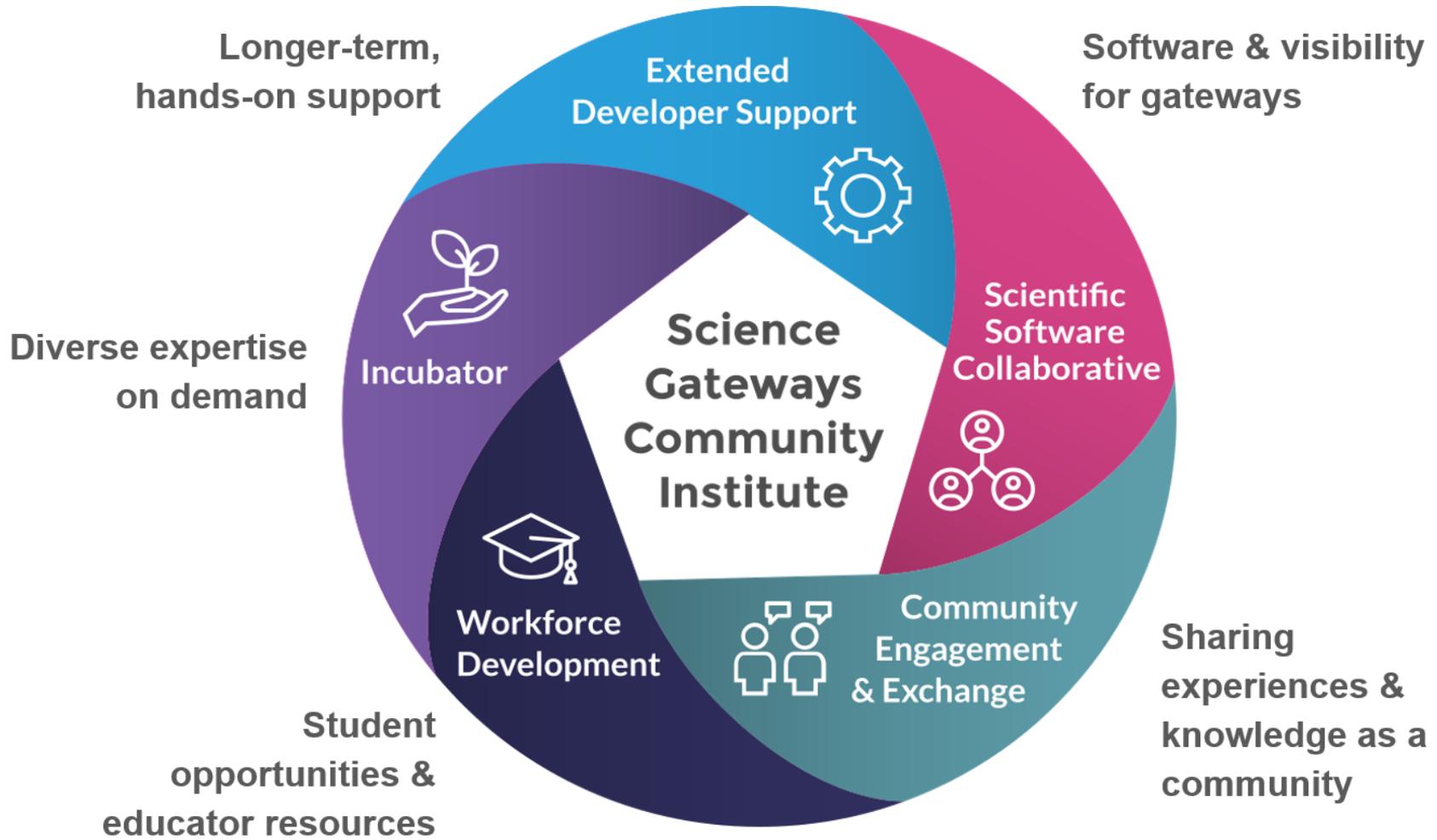
Lessons learned: approaches should be technology agnostic, using APIs and standard web technologies OR deliver a complete solution

Community Engagement is key

Hubzero instances world wide



Science Gateways Community Institute



Sustainability via On-Campus Teams

On-campus teams

It is a **centralized** team at your

institution –

irrespective whether you are part of

a university, a national lab,

an organization, a consortium or

a company...

Local teams vs. distributed and

remote teams:

For local teams it is **still easier** to

build more **trust**, to be more **efficient**

and to create a **strong culture**.

Addressing Software Sustainability on Your Campus



Is your campus seeing an increasing number of research projects that include web-based applications? Does each group have to hire developers independently? This can be time consuming and inefficient.

You are not alone.

THERE IS A SOLUTION

Creating a central pool of expertise on your campus offers many benefits including:

- Great visibility for the institution's research activities
- Synergy between projects
- Shared resources, costs and expertise across departments
- Expertise that is otherwise difficult for individual projects to obtain
- Lower learning curves
- Ability to retain top-quality research computing support by providing interesting projects

NOW IS THE RIGHT TIME! WE CAN HELP YOU!

- We can provide supplemental expertise where you don't have it.
- We can provide support for your journey to creating a campus-based group.
- We can provide ongoing advice based on campuses who have successfully created their own groups.

HOW TO START?

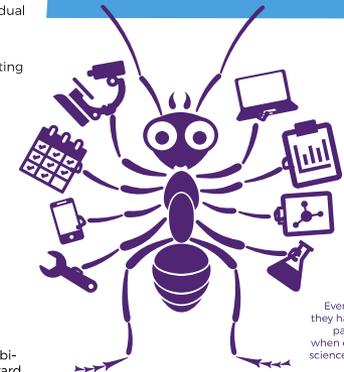
Contact us to request a free consultation, webinar, or on-campus visit to start your path toward sustainable gateway development.

INTERESTED? CONTACT US!

<http://sciencegateways.org/campusgroups>
help@sciencegateways.org

Science gateways are online, end-to-end solutions that provide broad access to advanced resources. They provide a community space for science and engineering research and education, allowing all to tackle today's challenging science questions.

Gateways are an increasingly common component of funded activities by many agencies. Individual PIs find it challenging to recruit and sustain teams that offer the diversity of expertise necessary for developing gateways.



Even ants wish they had an extra pair of hands when developing science gateways!

The **Science Gateways Community Institute (SGCI)** is an online and physical resource that supports science gateways with free services, including community building, consulting, and opportunities for sharing expertise, technologies, and practices.

Connect with SGCI

Incubator Sustainability Bootcamp

- <https://sciencegateways.org/engage/bootcamp>

I have an idea! 

Articulate the value of your gateway and how it's distinctively different from what already exists.

Who benefits? 

Identify audience and stakeholder groups and consider how they impact your success.

Where does it fit in? 

Establish where your gateway solution fits within the existing market landscape of partners and competitors.

How do I make it happen? 

Define measurable goals for success and sustainability. Consider multiple needs such as technology, security, project management, usability, and funding.

How do I sell it? 

Spread the word!
Plan how to tell the unique story of your gateway.

- 5 full days
- Teams on projects
- Interactivity
- Community formation
- Putting away the normal daily routine
- Homework

- twice per year
- additional ones can be booked (travel expenses for presenters)
- adapted to feedback

Connect with SGCI

Incubator Sustainability Bootcamp

- <https://sciencegateways.org/engage/bootcamp>

Work with us

- <https://sciencegateways.org/consulting/work-with-us>

Yearly Conference

- <https://sciencegateways.org/engage/annual-conference>

Become involved as a partner or affiliate

- <https://sciencegateways.org/about/partners>

Software/Gateway Catalog

- <https://catalog.sciencegateways.org/>

Train students in internships

- <https://sciencegateways.org/engage/student-focused>

Webinars, blogs, newsletter, Twitter, LinkedIn etc.

<https://sciencegateways.org>

Interagency Workshop 2019



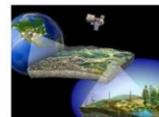
Contact:
Nancy Wilkins-
Diehr
wilkinsn@sdsc.edu



Funding agencies

OAC supports Research Cyberinfrastructure to uniquely enable collaboration and discovery frontiers at all scales

Shared resources, capabilities & services across the scientific workflow



CI-Enabled Instrumentation



Computing Resources



Gateways, Hubs, and Services



Cloud Resources & Services



Data Networks, Cybersecurity



Coordination & User support



Software, Applications, Workflow Systems



Bridging the Gap to Data Sharing



Image Credit Peter Alfred Hess (CC BY 2.0)

Researchers

“the local academic community struggles to effectively manage its assets which manifested itself in a number of challenges, and as for researchers, they lacked storage capacity and data curation processes, and the institution lacked standard metadata and indexing technologies, as well as tools that would support the whole research workflow” - Digital Asset Strategy Committee, DigitalND, 2011

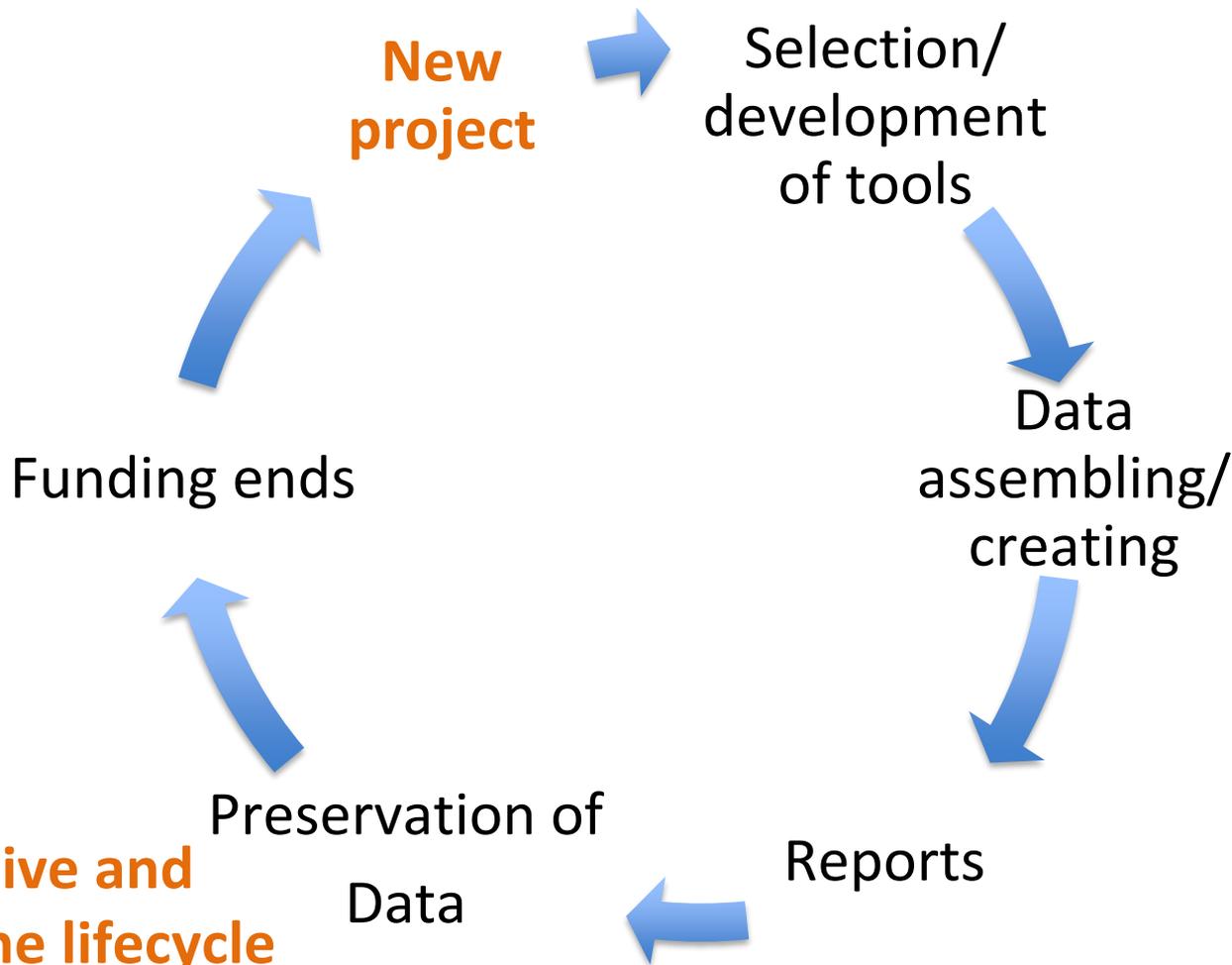
Libraries

Typically, data curation happens retroactively, and as a result data is either not captured at all or available metadata is incomplete.

Pressures from the Outside

“...digitally formatted scientific data resulting from unclassified research supported wholly or in part should be stored and publicly accessible to search, retrieve, and analyze.” - White House OSTP Public Access Memo, Feb. 2013

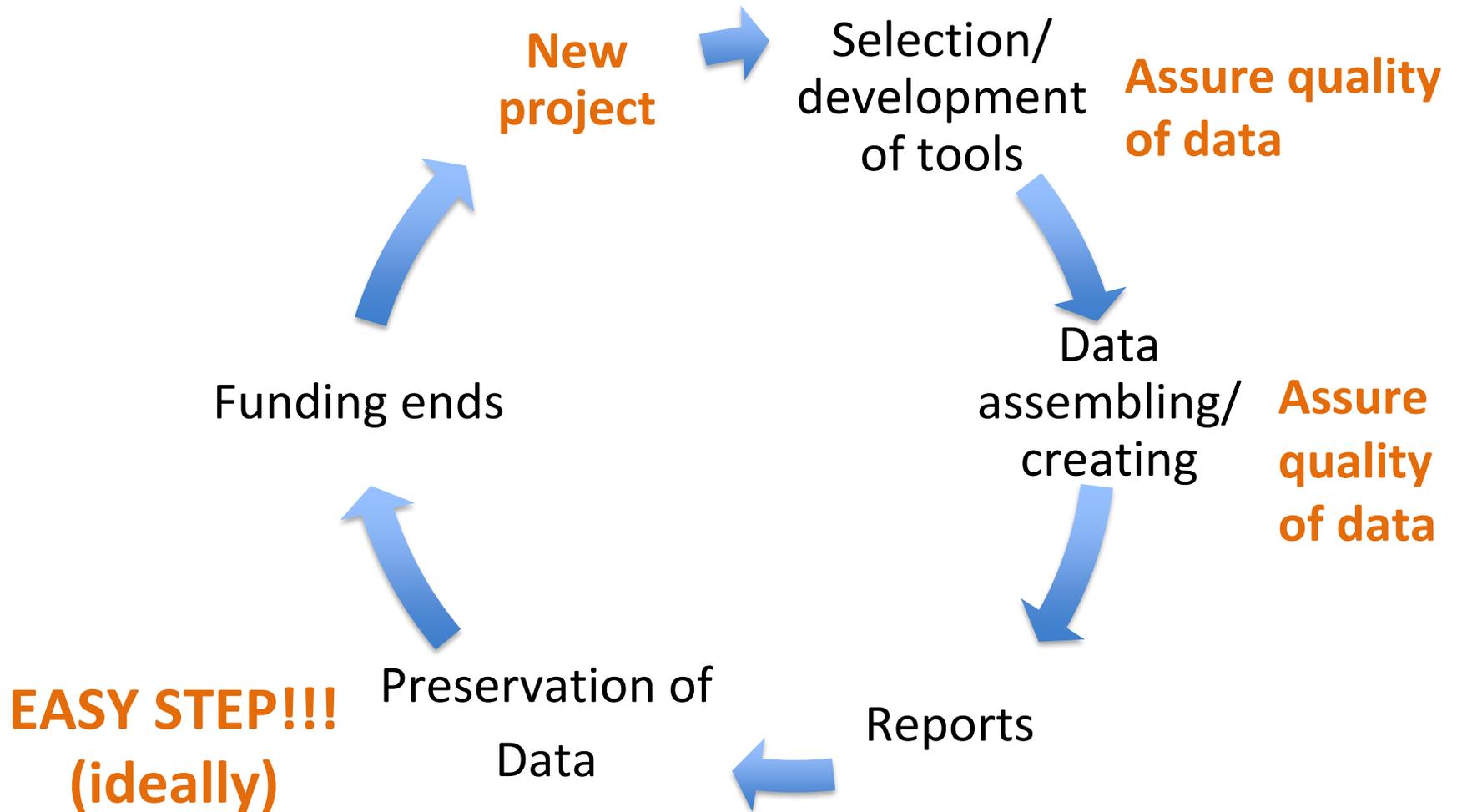
Current Lifecycle of Research Projects



Work-intensive and too late in the lifecycle



Target Lifecycle of Research Projects



PresQT

A collaborative design effort to enhance reproducibility and more open sharing of research data through open source development (July 2018-June 2020) of **Tools and RESTful Services to Improve Preservation and Re-use of Research Data & Software.**



<https://www.imls.gov/grants/awarded/lg-72-16-0122-16>

<https://www.imls.gov/grants/awarded/lg-70-18-0082-18>



UNIVERSITY OF
NOTRE DAME

Hesburgh Libraries



Secure | <https://presqt.crc.nd.edu>

PresQT Preservation Quality Tool

Search PresQT

About | People | Workshops | Resources

PresQT engages stakeholders in a collaborative planning effort to enhance reproducibility and more open sharing of research data through open source development of a **Research Data & Software Preservation Quality Tool**.

This tool will provide for reuse of preserved software applications, improve technical infrastructure, and build on existing data preservation services. It aims to fill an essential niche in the technical stewardship portfolio, and its collaborative open source development will improve and support the national digital platform.

PresQT addresses several timely data reuse issues and will have a lasting impact on the field by affording researchers and data curators methods to:

- Better represent digital workflow methodologies
- Improve data and software provenance
- Automatically enhance metadata
- Perform schema validation
- Improve file format recognition, interoperability and data integrity
- Facilitate scientific reproducibility

The project will design a **Research Data & Software Preservation Quality Tool**, which supports interoperability with existing platforms and solutions and improves the quality of preserved scientific digital content making it more reusable and reproducible, aligning well with the *Institute of Museum and Library Services* (IMLS) goal to promote the use of technology to facilitate discovery of knowledge.



STAKEHOLDERS → ENGAGEMENT → DELIVERABLES → IMPLEMENTATION → USER COMMUNITY

GRANT PERIOD (Left side): Stakeholders (Domain Researchers, Data Curators, Repository Managers, Librarians, Software Developers, Workflow Tool Developers, Linked Data Community, Journals) engage through **SURVEYS** and **WORKSHOPS** to produce **TOOL DESIGN** (Reports & Paper).

FUTURE DIRECTION (Right side): Implementation of an **OPEN SOURCE TOOL** (community-building and ongoing collaborative development) leading to a **USER COMMUNITY** (Domain Researchers, Data Curators, Repository Managers, Librarians, Software Developers, Workflow Tool Developers, Linked Data Community, Journals).

Open Science Framework
Everything produced for the PresQT project is shared on the Open Science Framework (OSF).
[Explore Project Resources on OSF](#)
DOI: 10.17605/OSF.IO/D3JX7

Two Workshops & the Needs Assessment answered by 1740 stakeholders have been completed.

All Resources avail online

<http://presqt.crc.nd.edu/>

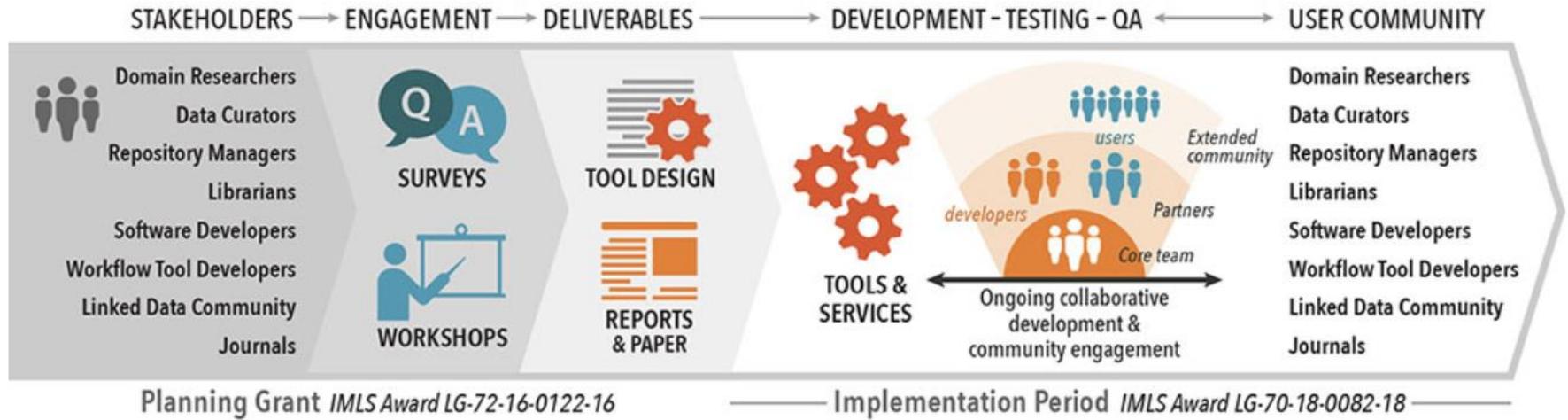


UNIVERSITY OF NOTRE DAME

Hesburgh Libraries



Collaborative Effort



Where we are now

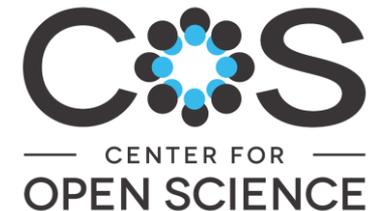


PresQT OSF Project

The screenshot shows the OSF project page for "PresQT Data and Software Preservation Quality Tool Planning Project". The browser address bar shows the URL <https://osf.io/d3jx7/>. The page header includes the OSFHOME logo and navigation links for Search, Support, Donate, Sign Up, and Sign In. The project title is "PresQT Data and Software Preservation Quality Tool Planning Project". Below the title, it lists contributors: John Wang, Sandra Gesing, Rick Johnson, Natalie Meyers, and Jeffrey R. Spies. It also lists affiliated institutions: University of Notre Dame and Center For Open Science. The date created is 2016-05-30 08:09 PM and the last updated is 2017-07-11 10:37 AM. Identifiers include DOI 10.17605/OSF.IO/D3JX7 and ARK c7605/osf.io/d3jx7. The category is "Project". The description states: "The goal is to collaboratively design interoperable and repository agnostic data and software preservation quality tools." The page is divided into several sections: "Wiki" with a diagram titled "Research Data & Software Preservation Quality Tool Planning Effort" showing a process flow from Stakeholders to User Community; "Citation" with the URL osf.io/d3jx7; "Components" listing three items: "PresQT Sept 18, 2017 Workshop", "PresQT May 1-2, 2017 Workshop", and "Outreach Presentations"; and "Files" with a search filter and a list of files including "PresQT Data and Software Preservation Qual..." and "Google Drive: Agendas and Minutes for P...".

An open project with all stakeholder input, workshop materials, and meeting info shared on Open Science Framework.

Project Partner

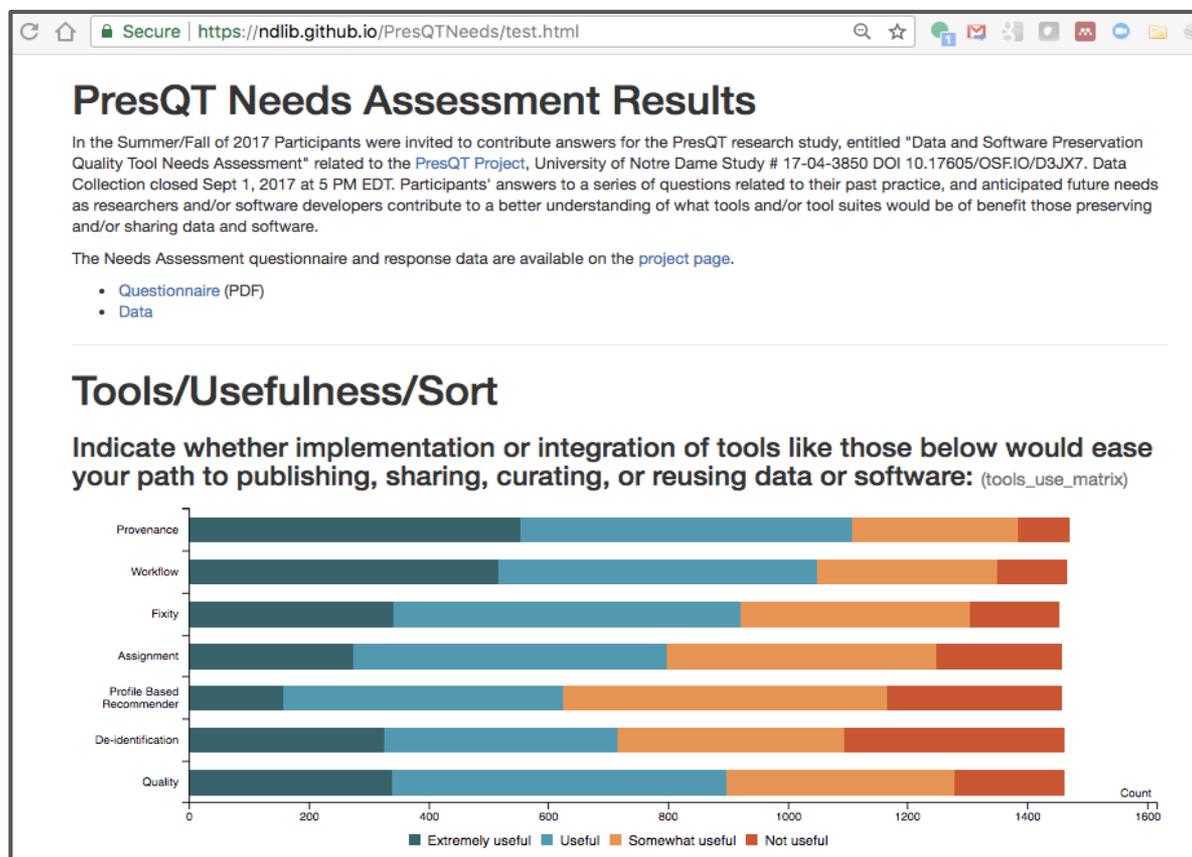


<https://osf.io/d3jx7/>

<https://cos.io/>

Needs Assessment Results - over 1700 answers

<https://ndlib.github.io/PresQTNeeds/>

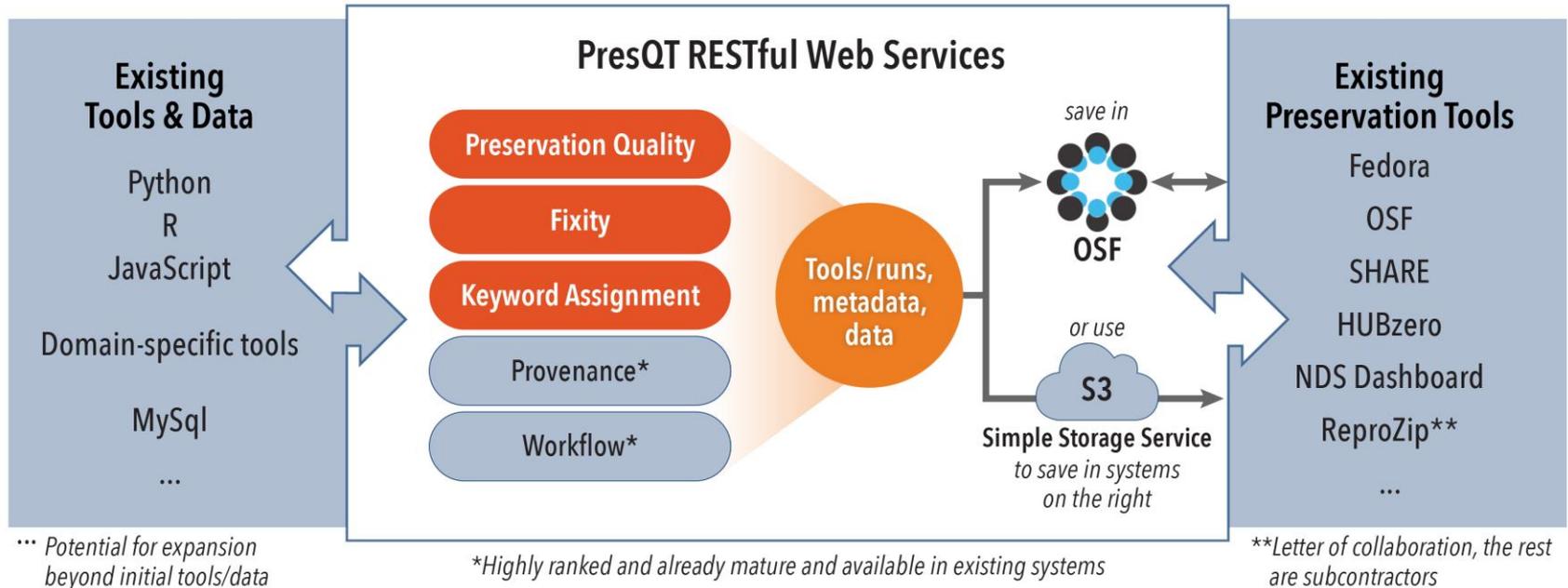


ToolsUsefulnessSort

Indicate whether implementation or integration of tools like those below would ease your path to publishing, sharing, curating, or reusing data or software:

	Extremely useful	Useful	Somewhat useful	Not useful
Provenance: Tools that show who did what when, or what changed when	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Workflow : Tools that let you preserve your own or reuse others' workflows	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fixity: Tools that help users or data curators identify whether a digital file is fixed, or unchanged.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Keyword Assignment: Tools that automate or nudge for better or easier tagging	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Profile Based Recommender: Tool that helps users identify digital resources of interest based on their profile	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
De-identification: Tools that make it easier to de-identify or anonymise data so you can share it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Quality:Tools that provide an assessment of a digital object's metadata completeness or preservation quality	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Repository and Tool Agnostic Solutions



- Open design of tools and services using standards
- Integrate with workflows, tools, and virtual environments
- Priority Focus Areas
 - Available for anyone to adopt what they need and build upon it!

Open Design Document

The screenshot shows a web browser window with the URL <https://osf.io/6eky7/>. The OSFHOME logo is in the top left, and the user's name, Sandra Gesing, is in the top right. The main navigation bar includes links for My Quick Files, My Projects, Search, Support, Donate, and a user profile. A secondary navigation bar has links for Technical Project Plan Resources, Files, Wiki, Analytics, Registrations, Contributors, Add-ons, and Settings. On the left, a sidebar shows a file explorer with folders for Google Drive, OSF Storage, and a file named 'PresQT Technical Design.docx.gd...'. The main content area displays the title 'PresQT Technical Design Document - Implementation' and the first section, '1 Community-Driven Gaps Analysis in the Preservation Landscape'. The text discusses the challenges of data and software preservation in research, mentioning the PresQT project funded by IMLS.

OSFHOME

My Quick Files My Projects Search Support Donate Sandra Gesing

Technical Project Plan Resources Files Wiki Analytics Registrations Contributors Add-ons Settings

Technical Project Plan Resources

- Google Drive: Technical Project Plan
- PresQT Technical Design.docx.gd...
- PresQT Technical Design Imple...
- OSF Storage (United States)

PresQT Technical Design Document - Implementation

1 Community-Driven Gaps Analysis in the Preservation Landscape

Preservation of data and software is a challenge that many disciplines face in research. One reason is that a variety of scientists are interested in assuring reproducibility of their results and long-term archival of their data and software. Another reason lays in demands by funding bodies to report results and assure that data and software is preserved in a way that it is accessible and reusable also after a project ends. Typically, scientists reach out to digital librarians for support for the preservation process at the end of the lifecycle of projects. The point of time creates not only a tight schedule but also risks the loss of important intermediate data. Additionally, preservation tasks are more labor intensive if they are not considered at different stages of the project life cycle but only at the end. The project PresQT (Preservation Quality Tool) funded by IMLS (Institute of Museum and Library Services) has been tackling these challenges via a collaborative planning effort and an implementation phase that started in July 2018. In the course of the planning phase two workshops and a widely distributed needs assessment answered by over

Partners and Committed Collaborations

- Sheridan Libraries, John Hopkins University
- NDS
- UC San Diego Library
- HUBzero team, Purdue University
- Yale University Library

- Libraries at Amherst College, Fontbonne University, Tuskegee University, Confederation of Open Access Repositories (COAR)
- ReproZip, Jupyter, CERN, RDA groups

- Midwest Big Data Hub, Science Gateways Community Institute, URSSI, Center for Open Science, Data Curation Network, Software Preservation Network

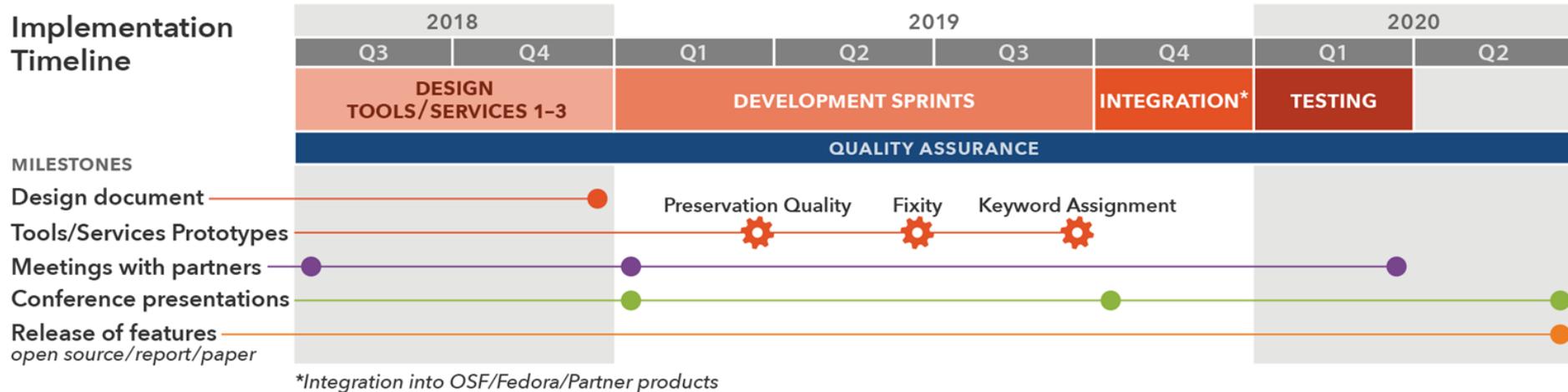
Partners and Committed Collaborations

- Sheridan Libraries, John Hopkins University
- NDS
- UC San Diego Library
- HUBzero team, Purdue University
- Yale University Library

JOIN US!

- Libraries at Amherst College, Fontbonne University, Tuskegee University, Confederation of Open Access Repositories (COAR)
- ReproZip, Jupyter, CERN, RDA groups
- Midwest Big Data Hub, Science Gateways Community Institute, URSSI, Center for Open Science, Data Curation Network, Software Preservation Network

Implementation Timeline



Contact us: presqt-contact-list@nd.edu

PresQT on the web: <https://presqt.crc.nd.edu/>

Subscribe to our newsletter!



UNIVERSITY OF
NOTRE DAME

Hesburgh Libraries

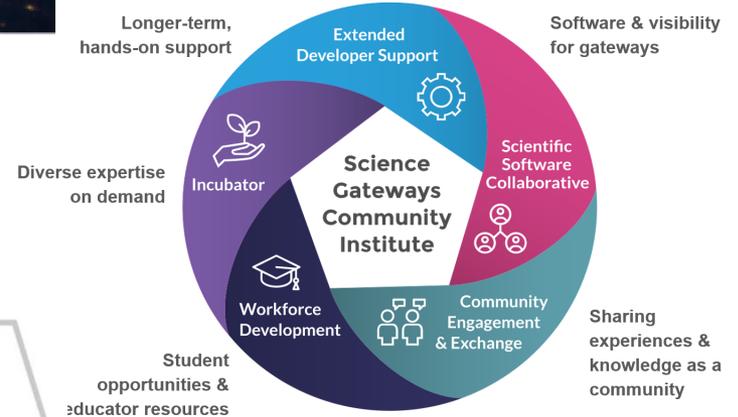
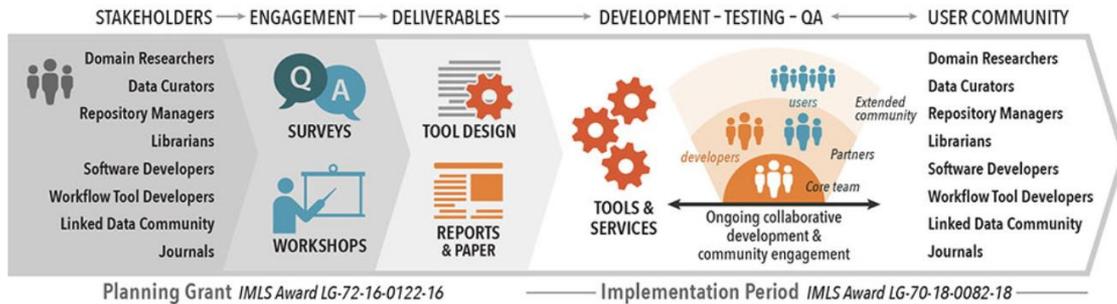


Thanks!



<http://urssi.us>

<https://presqt.crc.nd.edu>



<https://sciencegateways.org/>

sandra.gesing@nd.edu



Science Gateways
Community Institute

