ISSUES IN RESEARCH SOFTWARE

A Framework for Discussing e-Research Infrastructure Sustainability

Daniel S. Katz* and David Proctor*

e-Research infrastructure is increasingly important in the conduct of science and engineering research, and in many disciplines has become an essential part of the research infrastructure. However, this e-Research infrastructure does not appear from a vacuum; it needs both intent and effort first to be created and then to be sustained over time. Research cultures and practices in many disciplines have not adapted to this new paradigm, due in part to the absence of a deep understanding of the elements of e-Research infrastructure and the characteristics that influence their sustainability. This paper outlines a set of contexts in which e-Research infrastructure can be discussed, proposes characteristics that must be considered to sustain infrastructure elements, and highlights models that may be used to create and sustain e-Research infrastructure. We invite feedback on the proposed characteristics and models presented herein.

Introduction

Reusable infrastructure, systems and components created by one or more people and intended to be used by others, have become essential for many types of research over the last century, from microscopes to telescopes, and from sequencers to colliders. Over the past few decades, most of the interfaces to research infrastructure, and in many cases, the infrastructure itself, has become digital. In this paper we discuss e-Research infrastructure, also called cyberinfrastructure, which has been defined by Craig Stewart as consisting of "… computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible." [1]

While research infrastructure as a whole is important, it is useful to consider infrastructure elements as well, as the elements comprise the overall infrastructure. Each element can be discussed in different contexts:

- Technical
  - Architecture – How does it fit into the overall infrastructure? How does it interact with other infrastructure elements?
- Social
  - Developers – Who has developed the element?
  - Users – Who uses the element?
  - Purpose – What is the intended use of the element?

- Political
  - Funders – Who funds the development and maintenance?
  - Scope – Is the element local, regional, national, or international?
  - Impact – How is the element valued by researchers and funders?

Understanding how a particular infrastructure element can be created and sustained requires answering two pairs of questions: What resources are needed to create it, and how can those resources be assembled and applied? What resources are needed to sustain it, and how can those resources be assembled and applied? In this paper, we focus on the second half of the two questions, since the amount and type of needed resources vary with the specific element being discussed.

There are also a large number of emerging challenges that the e-Research infrastructure must respond to. The infrastructure must support today's collaborative science, which involves increasingly larger teams, more disciplines, and more countries. It must support increasing amounts, rates, and complexity of data, while promoting interoperability, both at the system and policy levels. The computing systems that make up the infrastructure have more cores, are composed of more types of architectures (e.g., embedded cores, CPU cores, GPU cores), and have more levels of memory hierarchy. Bandwidths have increased faster than latencies, changing the balances throughout the systems. In terms of building systems, in the past it was common to build as large a system as one could afford to purchase, but now, the limit may be what one can afford to operate. Considerations when purchasing a system have changed

* National Science Foundation, Arlington, VA, USA
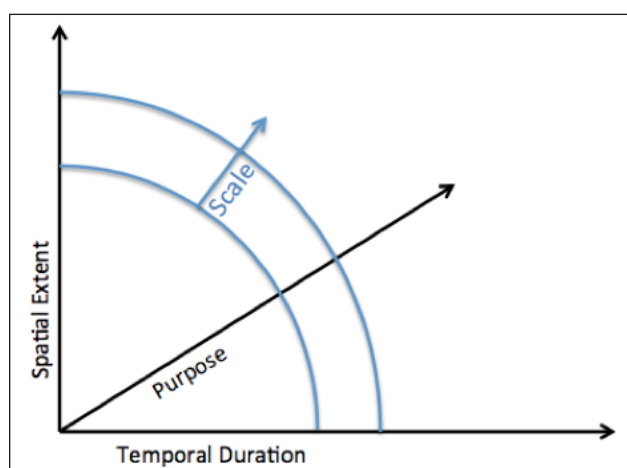dkatz@nsf.gov, dproctor@nsf.gov

Figure 1: The Space of e-Research Infrastructure Elements.

due to the increased availability of clouds. And as more systems are available in more places, networks become increasingly important, and concomitantly, security and privacy are becoming primary considerations.

Software is also becoming more complex, with applications and frameworks that handle more types of physics, or more types of data. This leads to challenges in how the software is programmed, with goals of novel programming models that support new abstractions for science, data, and systems. As the infrastructure elements and systems of elements become more complex, issues such as validation, verification, and reproducibility are essential, and are often handled in software. Also, the increasing complexity and scale of the infrastructure means that resiliency to the faults that may occur and be exposed as errors must be handled, again often in software. Finally, people are elements of the infrastructure, and they need education and training to become productive, career paths to stay motivated, and credit and attribution to be recognized and rewards and to move along their career paths.

The space of e-Research infrastructure

We believe elements of e-Research infrastructure can be placed in a three-dimensional space, as shown in **Figure 1**, and that doing this will lead to increased understanding of issues related to creating and sustaining these elements.

The first axis is the temporal duration of the element. This ranges from 5 years for computer systems, to about 10 years for networks and instruments, 20 years for production software, 40 years for people, and infinity for data, including publications, which can be viewed as a subset of data. Note that these values are approximate and can be debated; they do not completely define the duration of the elements. They are points of reference, and any given infrastructure element might have a shorter or longer duration than the point of reference given above. (In particular, the idea of a temporal duration for an instrument is unclear, but one can certainly consider an instrument as having a lifetime during which it is useful to a research community as shared infrastructure.) However, the key point is that decisions that are made about creating and sustaining infrastructure elements need to include awareness of the expected lifetime

of the element. The second axis is the spatial extent of the element. In an academic setting, this might range from a particular laboratory to a department, college or school, university, university system or regional alliance, nation, and beyond. This could also be thought of for general research institutions, which might have alternative administrative units in place of departments, colleges or schools, such as divisions or directorates.

The third axis is the purpose of the infrastructure element. This ranges from the element being used for one particular problem—though in this case it's unlikely to be infrastructure—to it being used for a variety of problems in one discipline (e.g., climate data from Arctic ice cores), to it being used for a variety of problems across a set of disciplines (e.g., molecular dynamics software), to it being used generally across all disciplines (e.g., a network, an HPC system). There are linkages between the temporal duration of an element and its purpose, e.g., although the lifetime of a given software element may be 20 years when just considering its technical context, if the element ceases to be useful to its user community then the lifetime will be shortened.

Note that the number of users of a given element should be larger the farther the element is from the origin in any direction, as should the cost. These two elements (number of users and cost) can be generically called 'scale' in this context. Scale is thus a metric of the space, though it is not orthogonal to any of the three axes.

Creating and sustaining e-Research infrastructure elements

Each infrastructure element first needs to be created, then needs to be sustained. However, before we can consider models for the assembly and application of resources to create and sustain infrastructure elements, we must define sustainability; what is meant by the creation of e-Research infrastructure elements is clear, what is meant by sustainability is not. (Note that we, along with other organizers of the WSSSPE1 workshop [2], have begun a survey to understand how the community defines software sustainability. It is expected that this survey will gather one or more consensus definitions, and lead to a short paper discussing them, as well as the level of consensus.) We use sustainability to mean that the element will continue to be supported as changes occur to other infrastructure elements, the user communities and their abilities and needs, and the underlying principles upon which the element was built. Specifically, we ask the following questions, each of which hints at a environmental change to which an element must respond, in order to make our definition of the term sustainability more concrete:

- [Dependent Infrastructure] Will the infrastructure element continue to provide the same functionality in the future, even when the other parts of the infrastructure on which the element relies change?
- [Collaborative Infrastructure] Can the element be combined with other elements to meet user needs, as both the collaborative elements and the individual elements change?

- [New Users] Is the functionality and usability of the infrastructure element clearly explained to new users? Do users have a mechanism to ask questions and to learn about the element?
- [Existing Users] Does the infrastructure element provide the functionality that current users want? Is it modular and adaptable so that it can meet the future needs of the users?
- [Science] Does it incorporate and implement new science and theory as they develop?

Using this definition of sustainability, the following models are commonly used for the assembly and application of resources to *create* and then *sustain* an infrastructure element:

- Open source: a leader (or a set of leaders) promotes a goal of creating an infrastructure element in a public manner and a community voluntarily forms to work together on this goal.
- Closed partnership: a set of partners works together to create an infrastructure element, but the partnership is not open to external contributions.
- For profit: a group creates an infrastructure element using its own resources with the goal of later selling, leasing, or licensing the element or its design to recover the expended resource and make a profit.
- Dual licensing: a group creates an infrastructure element using its own resources with the goal of allowing academic free use (and depending on the license, perhaps gaining further free contributions from that academic community), while also selling, leasing, or licensing the element or its design to industry in order to recover the expended resource and perhaps make a profit, or at least, break even. This model also often has an implicit goal of not allowing other companies to financially profit directly from the element.
- Open source and paid support: a group supports an open source element in exchange for resources from the users of that support. The support can include helping the users with the existing element, and adding features to the element for the supported users, though these added features become available to all users, not just those who have paid for support.
- Foundation or government: one or more groups convince an organization that promotes public advancement that creating an infrastructure element will be a public good that should be supported.

Note that while open source has been thought of as applying to software, it can also apply to other types of infrastructure elements, including data (e.g., citizen science) and computing and other hardware systems (e.g., Arduino and the maker community [3])

Particularly in the case where open source is involved, but to a lesser extent in the other models, governance is an important factor. Governance tells the community how the project makes decisions and how they can be involved. The community can consist of the users, developers, advocates, or any combination thereof. Two examples of open source governance models are a benevolent dictatorship, as is used in the Linux kernel, and a meritocracy, as is used in Apache Foundation projects. These can be considered top-down and bottom-up governance, respectively. Note that these are orthogonal to top-down and bottom-up development, the cathedral and the bazaar models respectively. [4]

Which models work in what part(s) of the space of e-Research infrastructure?

We believe that research is needed to correlate the success and failure of various models with the different portions of this space. Some questions we would like to answer are:

- Do the cathedral or bazaar governance model correlate with successful projects along any or all axes? For example, perhaps one works better at small scale, and the other works better at large scale.
- Do particular resource assembly and application models cluster along any or all axes? For instance, government funding may be found at large values of temporal duration, while a mix of models are found at middle values, and closed partnerships are found at small values

Conclusions and next steps

We hope this paper encourages thought about how e-Research infrastructure elements and e-Research infrastructure itself should be considered, in terms of how different elements may have commonalities and differences across types of element, user communities, etc. The purpose of this paper is to begin a discussion about these issues. We are eager to receive feedback, and suggest the following discussion questions:

- Are the axes we've suggested meaningful?
- Are the resource assembly and application models we've suggested complete?
- Can we find correlations or clusters between them?

References

1. **Stewart, C A** 2010 What is Cyberinfrastructure? In: ACM SIGUCCS 2010 Annual Meeting. Available at: http://hdl.handle.net/2022/13987 [Last accessed 01 April 2014].
2. **Katz, D S, Allen, G, Chue Hong, D, Parashar, M** and **Proctor, D** 2013 First Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE): Submission and Peer-Review Process, and Results. Available at: http://arxiv.org/abs/1311.3523 [Last accessed 02 May 2014].
3. **Wilcher, D** 2014 *Make: Basic Arduino Projects: 26 Experiments with Microcontrollers and Electronics.* Maker Media, Inc.
4. **Raymond, E S** 1999 *The Cathedral & the Bazaar.* O'Reilly. Available at: http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/index.html [Last accessed 01 April 2014].