

Basis for remarks at Conference on***Designing Cyberinfrastructure for Collaboration and Innovation*****Washington****January 29, 2007*****“Is ‘Designing’ Cyberinfrastructure - or, Even, Defining It - Possible?”*****Peter A. Freeman****National Science Foundation**

Note: The views expressed herein are the personal views of the author and are not necessarily those of the National Science Foundation or the U.S. Government.

Abstract: *In the technical world, 'design' often implies having a set of specifications to which the resulting design replies. This, in turn, often implies that you know what it is that you are designing - i.e. that you have a definition of it. The current state of cyberinfrastructure calls into question both of these assertions. This brief paper will suggest some important considerations.*

Context

It is important in any rational discussion of a subject on which there may be subjective opinions – which certainly includes the topics to be discussed here – to understand not only the linguistic utterances of the author or speaker, but also to know their perspective (context) and viewpoint (where they stand in looking at the subject at hand). Only then is it fair to critique their message.

The author of this brief note, a computer scientist, comes from the community of toolmakers that create cyberinfrastructure. After his initial activity as a scientific programmer starting forty-five years ago, he has not been a user of scientific cyberinfrastructure other than in his research on design process and software engineering, and in educational activities. As an educator, he taught courses on the social impacts of computing for over thirty years, which provides an important part of his context. As a government official at NSF in the late 1980's, he was a member of the small, interagency writing team that developed the first public strategy for high-performance computing and communications (HPCC) – one of the primary progenitors of current cyberinfrastructure activity. When he returned to NSF in 2002, he became the official recipient

of the report of the Blue Ribbon Panel on cyberinfrastructure (the Atkins Report)¹, and from then until 2005 was responsible for the funding of NSF's current cyberinfrastructure activities and planning for future cyberinfrastructure development – in short, initial implementation of the Atkins report. Since 2005, he has been one of the group of senior NSF officials responsible for guiding NSF's cyberinfrastructure activities.

His viewpoint then is that of one: knowledgeable in the technological foundations of cyberinfrastructure; well aware of the types of impacts of cyberinfrastructure on society, organizations, and individuals; experienced in governmental infrastructure efforts; and desirous of seeing cyberinfrastructure realize its potential for revolutionizing many areas of human activity.

In the context of this perspective and viewpoint, the objective for this paper (and the accompanying presentation at the Conference), is to comment on the process of “designing” cyberinfrastructure in the hope that this will aid other discussions of the design of cyberinfrastructure. A necessary precursor to any design process is to know what it is that one is designing – i.e. what is the definition of the desired artifact.

Is It Possible to Define Cyberinfrastructure?

One often hears: “What is cyberinfrastructure?” “Do we already have a cyberinfrastructure?” “Will my cyberinfrastructure work in Europe? Or, in California?” “How does it differ from supercomputing?” “Does it include the PC on my desk or in my lab?” Almost everyone dealing with cyberinfrastructure has struggled with his or her own version of such questions, which is good – read on to see why.

These questions suggest that the definition of cyberinfrastructure may be broad, even at the most instrumental level. For example, the cyberinfrastructure that is needed by an educational researcher working with kindergartners will likely be very different from that needed by particle physicists interested in analyzing the exabytes of data that will be generated by the projects at CERN. Thus, as a practical matter (uppermost in the minds of those attempting to build new cyberinfrastructures) when defining (specifying) a particular cyberinfrastructure, it is imperative to grapple with the intended uses of the cyberinfrastructure.

It is useful to consider the origin of the term. To do that properly, we need to go back in history. Infrastructure in the sense of roads, water supplies, and so forth is a very old and well-understood term. In relatively recent times, for our purposes here beginning with the post-WW 2 period, the concept of infrastructure for the conduct of science (labs, equipment, support personnel, etc. provided for general use going beyond a specific project) began to be explicitly considered. A good example is the NSF-supported facilities in Antarctica, dating back to the 1950's, that were built to provide a platform for a wide variety of polar studies. That effort continues today, providing one of the most important pieces of infrastructure for scientific research.

¹ *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure* (Atkins Report), January 2003, NSF (<http://www.nsf.gov/od/oci/reports/toc.jsp>)

When the modern computer and its usage in science in the 1950's came about, it was natural that it quickly became a part of the developing scientific infrastructure in advanced societies. There were, of course, many individual uses of computers in research labs, but a good place to mark as the start of what we now are calling cyberinfrastructure was the 1960s. It was then that the NSF undertook a program of supporting Academic Computing Centers on a number of U.S. campuses. Those facilities, open to a broad scientific community and not dedicated to specific projects, provided the initial introduction for thousands of faculty and students to the use of computers in scientific research and education.

NSF's support for scientific computing infrastructure took a number of different forms over the years that won't be detailed here. One example with which the author is familiar was the Coordinated Experimental Research (CER) Program started in the late 1980s in response to serious concern over the terrible state of computing infrastructure for the conduct of computer science and engineering research in universities. That program, which continued more or less in its original form into the 1990s and continues today in other forms, had a huge impact in energizing and enabling computer science and engineering research and education in this country. Thousands of students and faculty in CS&E departments were able to explore experimentally a variety of topics that helped lay the foundation for much of the computing revolution of the 1990s and beyond. For example, many of the basic concepts for distributed computing were developed and explored on the equipment provided in the CER Program. The impact was felt far beyond traditional CS&E since those resources were often used to explore new topics that the lack of equipment would have prevented. One particularly impactful result was the development of some of the core algorithms later used in the human genome project.²

NSF again led in the early 1980s in providing supercomputing resources for the open scientific community. The creation of the first two NSF Supercomputer Centers at the University of Illinois and the University of California, San Diego, came about in response to several studies and reports documenting the potential for science in deployment of production versions of this emerging class of computers. This activity led to the creation of additional centers, more studies, and an evolution of this activity through several forms to what NSF today spends several hundred millions of dollars annually on in the aggregate. Support for lower capability machines continued through a variety of forms directly from the research programs in each discipline.

The networking part of the story also goes back thirty-five years or more. For example, a very early local ring distributed computing project was undertaken at the University of California, Irvine, in the early 1970s by Prof. David Farber with NSF support.³ Wide-area networking support by NSF began with Theory Net, which morphed into CSNet, and eventually NSFNET. These developments and much technical work in a variety of supporting fields, along with early networking deployment by the NSF Supercomputer centers, eventually merged with the ARPAnet and under NSF guidance led in the early 1990s to what we now call the Internet. Further research and advanced

² Gene Myers, talk at NSF, September 9, 2002.

³ DCS Project, UC Irvine, 1970-1977. See http://en.wikipedia.org/wiki/Grid_computing#Origins for a short history and references to early reports.

development, including the Gigabit Testbed projects, Mosaic, and many of the search and digital signal processing fundamentals that under gird today's Internet and World Wide Web have continued up to today under NSF support.

The GENI Project⁴, mentioned again below, continues this history of pushing forward the boundaries of computation and communication capabilities. This project, currently in the planning and initial research stage, is aimed at encouraging fundamental research on networking and distributed system fundamentals by providing a major experimental instrument that will support such investigations. In addition, eventually it will also support a wider set of investigations into topics that are essential to the future of cyberinfrastructure for all areas of activity. The impetus for this project is the consensus among those that have built the current Internet, which forms the fundamental fabric on which cyberinfrastructure is built, that we are nearing the limits of those structures. The Internet, however, has become so essential and ubiquitous that it is impossible to experiment on it directly with new fundamental structures. The GENI Project aims to support the research necessary to this "reinvention of the Internet."

This brief historical excursion is meant to provide a richer context in which to consider the definition of cyberinfrastructure. This history is, as you might expect, much richer than has been outlined here and is well worth proper documentation. It also includes significant contributions by individuals and institutions not supported by NSF, although in the area of scientific computing and communications infrastructure for the open science community, NSF has long been the leader.

Before looking specifically at the etymology of the term "cyberinfrastructure," several things should be clear from this history:

- Scientific infrastructure that includes computers and communications gear may often lead similar infrastructure in other domains (*e.g.* business), but eventually (and rather quickly) becomes the commodity infrastructure for a wide variety of usages (something we must expect will continue);
- The infrastructure for one group of scientists and engineers (*e.g.* computer scientists) may be different in many respects with that needed by another group (*e.g.* particle physicists), both in composition and capability;
- Advanced infrastructure for a given set of users, changes over time as new technologies become available (*e.g.* early 56kB connections between supercomputers have been replaced today by dedicated fibers).

The author's predecessor at NSF, Prof. Ruzena Bajcsy, used⁵ the term 'cyberinfrastructure' in setting up the Blue Ribbon Panel on cyberinfrastructure⁶, chaired by Dan Atkins of Michigan (now at NSF where he heads the Office of Cyberinfrastructure). The context in which she was working was the result of the history outlined above and clearly led to the creation of a term for

⁴ See www.nsf.gov/cise/cns/geni/ and www.geni.net .

⁵ In a similar, but slightly different context, the term was apparently first used in a press briefing on PDD-63 on May 22, 1998 with Richard Clarke. See <http://en.wikipedia.org/wiki/Cyberinfrastructure> for references.

⁶ *Atkins Report, op cit.*, "Charge to the Committee."

infrastructure that attempts to capture the integration of computing, communications, and information for the support of other activities (especially scientific in the case of NSF)⁷.

The definition that the Atkins Report starts with is given by analogy to industrial infrastructure such as transportation or power systems: “The new term, cyberinfrastructure, refers to infrastructure based upon distributed computer, information and communication technology⁸.” This was clearly consistent with the computing and communications infrastructure that had grown up in the scientific community over the past forty years. Much of the existing scientific cyberinfrastructure is based on NSF research in computing and communications, and, at least in the open scientific community, significant amounts of the infrastructure were originally deployed with NSF support. This made the general acceptance by NSF of the recommendations of the Report a logical consequence, and use of this definition a natural starting point.

The Atkins Report definition (and the report itself) contains the seed of a much broader scope, however, when it includes “information.” Indeed, much of the discussion in the NSF community has already appropriately shifted from whether particular pieces of hardware or communications gear are to be included in a cyberinfrastructure to what information is included and how is it to be interfaced. A constructive way to look at this is that it is yet another instance of yesterday’s cyberinfrastructure definition (which basically just included gear) is evolving to include software, database, and more.

As John King has argued⁹, it is a good thing at this stage that we don’t have a single, one-size fits all definition because each discipline must sort out what it needs while at the same time those of us helping build cyberinfrastructure need to explore alternative technical approaches. It is clear that this evolutionary process must, perforce, continue unabated. At the same time, at appropriate points we can snapshot a current definition for the purpose of undertaking specific developments, acquisitions, and deployments. This understanding places certain strong requirements on the design of cyberinfrastructure that we will address below.

While understanding that ambiguity in the definition of cyberinfrastructure is to be expected at this and, very likely, any future stage, NSF and other agencies need a working definition that will permit them to move forward with funding programs aimed at serving the scientific community. The working definition has been, very simply, that cyberinfrastructure is the *integration of hardware, middleware, software, data bases, sensors, and human resources, all interconnected by a network (the Internet in almost all cases)*. When you pull all of these together, you get a comprehensive cyberinfrastructure that is already revolutionizing discovery, learning and innovation across the science and engineering frontier. At the same time, the expansion of this definition to include information, virtual organizations, and other entities, is already underway while at the same time the early operational versions of an integrated infrastructure are just beginning to appear.

The intended take-away from this discussion then is:

⁷ Ruzena Bajczyk, private communication, 2007.

⁸ *Atkins Report, op cit.*, p. 1.

⁹ Private communication to NSF management, 2003.

Cyberinfrastructure can have many definitions and, to some extent, the definition is in the eye of the beholder.

Designing Cyberinfrastructure

If we can't define cyberinfrastructure unambiguously, how can we possibly design it?

Actually, that is quite possible once you take the position that a cyberinfrastructure is a substrate that is fairly stable and upon which other constructions are built. Call those other constructions what you like – specialized cyberinfrastructures, knowledge ecologies, etc.; remember, this author is presenting the viewpoint of a builder of what is intended to be the lowest level and most general cyberinfrastructure. Design of complex systems that are open-ended has become a common task in the systems design world, even though how to do it is not yet a science.

The issue then is to insure that what one designs (for the lowest level cyberinfrastructure) admits the widest possible constructions to be built on top of it. We don't and can't know what all of those constructions will be in the future, but we at least know to consider how every design decision may effect future constructions that will use the cyberinfrastructure. This is not the place to go into the details of how one should approach the issue of insuring generality, but there are three fundamental principles that are worth noting.

First, remember that all the laws and policies in the world cannot *expand* what the underlying technology can do¹⁰, while at the same time that technology is guaranteed to change almost continuously and dramatically (consider the changes in the transport technology and speeds over the lifetime of the Internet). So, the technical design of cyberinfrastructure is very important, not just so that it works for today but so that it will enable continual change underneath while still presenting a stable platform on which the higher-level constructions can be built. At the same time, the definition of the stable platform will need to evolve in several ways and this, too, must be planned for. This is an old, but not completely solved, problem in building computer and networking systems; nonetheless, a good bit is known about how to proceed. The future possibilities for innovative constructions on top of the cyberinfrastructure will be largely shaped by the underlying networking – which is why the GENI Project is so important.

Second, it is very important for the designers of the cyberinfrastructure to keep in mind that once built, laws, policies, and practices will be imposed that will shape (and by definition, limit) the possibilities for later expansion and use. To enable wide and stable usage, laws, policies, and practices are needed, of course, but the technical design itself can shape the possibilities of what transpires in that realm. For example, if the basic design incorporates technical elements that are not in the public domain, then the later possibilities for policies and laws will be very different than if public domain elements were used.

Third, the designers of the basic cyberinfrastructure should also consider the processes by which any modifications or expansions to it will be effected. As those processes are also designed, it is

¹⁰ This is the fundamental thesis of Larry Lessig's book *Code Laws*.

imperative that experts in the social and organizational aspects of change work closely with the technical designers who are the experts in how technical changes can be made. The success of the current Internet, for example, was clearly enabled by the change processes (primarily the IETF process) initially set up by the technical designers and their funding supporters (the commercial sector was uninterested in networking at that point in time). That success, however, may well have been serendipitous and, in any event, we now have a lot of experience in what works and what doesn't.

Let's consider some consequences of these points:

Multiple cyberinfrastructures will certainly exist. This is fine and appropriate, but the critical issue is how can they relate? In technical terms: How can we interface, for example, a cyberinfrastructure being used by microbiologists and a (different) cyberinfrastructure being used by climatologists? The technical answer that makes the most sense to this author lies in the nature of the network architectures and the data structures and ontological methods that are used. This is one reason why NSF is investing in the research proposed by the GENI Project¹¹ along with significant work in multiple ways on dealing with large collections of data (curation, organization, ontological methods, searching, and so on).

The technical architecture of a cyberinfrastructure, while being the initial and most fundamental factor affecting the design of the cyberinfrastructure, is not the only factor. Indeed, if done properly, then the technical architecture will enable and permit the widest possible space of cyberinfrastructures of a more specific nature to be built and everything else (policies, rules, laws, etc.) will only serve to limit the space of possibilities. The astounding success of the Internet to enable a very wide range of innovations in many different fields is generally attributed to two things: the fact that the fundamental architecture of the Internet (the TCP/IP protocols) enables such a wide range of eventual and unknown (to the original Internet designers) applications and that, at least until now, there have been a minimum of laws and governmental policies that unduly restricted what can be carved out of the very large design space enabled by the underlying technical architecture¹².

An extremely important corollary is that there must be much more communication between the technical and other communities. The technical community must be aware of the broader policy, legal, and usage possibilities as a specific cyberinfrastructure is designed and, to the extent possible, involve those communities at the earliest possible stages in the design processes. At the same time, the policy, legal, and usage communities need to assist the technical community and understand that effective and reasonable laws, policies, and usage procedures cannot be developed in ignorance of the technology.

¹¹ *GENI, op. cit.*

¹² The generality of the TCP/IP protocol that has enabled so much expansion is starting to show its limits, however. IPv6 was created to address some of the limitations of number of nodes, and newer services and security requirements may not be realizable in a practical manner. The lesson is that even fantastic generality has its limits.

A final comment concerns understanding how cyberinfrastructure is used (and who the users really are). While there are several thousand users of today's most advanced scientific infrastructure, very few, if any, policy or legal experts have actual experience with using it. Likewise, for the more common cyberinfrastructure that is behind e-government, a lot of current business activity, and, increasingly, personal and entertainment uses, very little organized and accurate characterization of uses and users is available on which policy and legal experts can base constructive and effective policies and laws. It will be essential as we move forward not only to develop such systematic understandings, but to pay attention to them in fashioning policies and laws.

Summary

Returning to the questions posed by the title of this short note:

It is possible to define cyberinfrastructure.

In doing so, however, one must remember that the definition will evolve and the nature of the definition depends very much on the use of definition.

Likewise,

It is possible to design cyberinfrastructure.

The design process must constructively involve the technical, policy, legal, and usage communities and aim for a design that admits of open-ended change both above and below the defined platform.

Working on infrastructure has never been considered a "sexy" or high-status undertaking, especially for technical people. Yet, the definition, design, and deployment of cyberinfrastructure offers the opportunity of enormous leverage and impact of one's professional work. If done properly, the result will far outlive any of us!