Cyberinfrastructure Facilitators:

New Approaches to Information Professionals for E-Research

R. David Lankes

Associate Professor

Derrick Cogburn

Associate Professor

Megan Oakleaf

Assistant Professor

Jeffrey Stanton

Associate Professor

Syracuse University's School of Information Studies

Paper Submission to Oxford e-Research Conference 2008

ABSTRACT

This paper introduces the concept of a CI-Facilitator defined as a vital member of the research enterprise who works closely with researchers to identify extant tools, data sets, and other resources that can be integrated into the process of pursuing a research objective. In order to prepare CI-Facilitators to evolve with e-Research endeavors they must be grounded in deep conceptual frameworks that do not go out of date as quickly as any given cyberinfrastructure technology. One such framework, that of participatory librarianship, is presented here and explored in terms of tackling the issue of massive scale data in research. Participatory librarianship is grounded in conversation theory and seeks to organize information as a knowledge process rather than as discreet objects in some taxonomy.

Introducing CI-Facilitators

The Internet has become the informational substrate of most scientific and engineering research enterprises[1]. Few people who experienced the early days of Telnet, Gopher, or even the Web truly anticipated the impact of the Internet on the scientific process. Even the most accurate futurists could not foresee how the Internet would shape the processes involved in creating new knowledge. Databases, statistical datasets, data warehouses, sample libraries, and image collections are just a few of the myriad examples of large scale information collections that scientists and engineers must create, maintain, and share [2,3]. Even these examples only scratch the surface, however, because the most innovative scientific and engineering information use is now in the form of cyberinfrastructure that Facilitates the development of geographically distributed research centers and networked communities of research within and across traditional disciplines [4].

Yet in e-research scholars have three serious problems facing them. First, researchers spend their careers mastering the skills, knowledge, and tools that comprise the core of their respective disciplines [5,6]. Few among them have the capacity to simultaneously become experts in information management, networking, virtual or distributed collaboration, search and retrieval, archiving, user interface development, and all of the other skills of the information professions[7]. Second, advances and convergences in cyberinfrastructure (broadly defined to include the web, wireless grids, parallel processors, laptops, cell phones, mainframes, telecommunication networks, etc.) that have occurred over recent decades have themselves fueled a vast proliferation of information – more findings, more datasets, more papers, more conferences, more journals, more books and so on. Even the brightest and most motivated struggle to keep up with the rapid pace of knowledge creation in their field [8,9,10]. Finally, information infrastructure itself is in the process of an accelerating evolution. Gains in computing power, storage, transmission bandwidth and other fundamental building blocks of cyberinfrastructure create frequent discontinuities in the economics of information technologies, while open source software tools sprawl daily into innovative new application territories [11,12]. The rapid pace of development of information infrastructure implies that only individuals who dedicate their professional lives to it can truly keep up.

One solution to these issues is the preparation of “CI-Facilitators.” These are information professionals able to partner with e-research teams to identify extant data and tools, as well as build new tools in the pursuit of research topics. CI-Facilitators may work in physics, chemistry, biology, neuroscience, sociology, or any of a host of other STEM disciplines [16].

To illustrate the proposed benefits of a CI-Facilitator, imagine a team of geneticists who want to test a new protein folding technique. The team includes individuals from three geographically distributed Facilities in the United States. Large scale supercomputing and/or a distributed grid computing process could substantially speed the research, and yet the scientists have limited knowledge on how to access these resources, the preparation needed to run the experiments, or how to coordinate the efforts of multiple users at the three Facilities. Through needs assessment, the trained CI-Facilitator would work with the scientific team members to identify their research goal, and help design the underlying socio-technical infrastructure needed to support their goals. Working with the research team, the CI-Facilitator would identify a likely pool of simulation technologies, develop contact and partnering information for organizations that can provide access to the technologies, and identify a scientific consortium to provide shared access to simulation tools. The CI-Facilitator would then configure the client Facilities needed to connect to the consortium's servers. Additionally, the CI-Facilitator would deploy collaborative tools to handle administrative meetings, project team meetings, and outreach activities. Once the simulations occur, the CI-Facilitators would archive the results of the study (raw data and synthesis pieces such as articles) and assist with dissemination of the results. Across all of these processes, the CI-Facilitators acts as a catalyst for the interaction of the research problem with the information resources and technologies needed to Facilitate the research process, thereby shortening the developmental curve of accomplishing the research goal.

Background: Defining the Knowledge, Skills, and Tools for the CI-Facilitator

O*Net is the U.S. Department of Labor's successor to the now outmoded Dictionary of Occupational Titles (http://online.onetcenter.org/). O*Net provides an empirically-derived database of job descriptions based on formal job analysis along with a collection of information resources for each job, including skills, abilities, activities, tools, and technology [17]. Not surprisingly, there is no job with cyberinfrastructure in the title or description, let alone with the specific title of cyberinfrastructure Facilitator [16,18]. Nonetheless, the CI-Facilitator roles that we have envisioned do have their roots in existing work roles. In particular, we examined the skills and tools of seven different jobs in order to begin our needs assessment of the CI-Facilitator training regime: computer systems analysts, database administrators, computer support specialists, training and development specialists, natural sciences managers, archivists, and audio-visual collections specialists. The results of this analysis, summarized in Table 1 below, showed a striking degree of commonality in knowledge, skills, and tools across these seven jobs.

Table 1: Knowledge, Skills, and Tools/Technology Shared Across Seven Jobs

Agreement Level	**Knowledge**	**Skills**	**Tools and Technology***
100%	English Language	Reading Comprehension	Desktop Computers
100%	Computers and Electronics	Active Learning	Notebook Computers
100%	Customer and Personal Service	Active Listening	Database management system software
86%	Science/Math	Critical Thinking	Object or component oriented development software
71%	Administration and Management	Instructing/Teaching	Mainframe computers

57%	Education and Training	Written communication	Metadata management software

*Tools and technology data not available across all O*Net jobs; percentages are approximate for this column.

As Table 1 suggests, basic literacy, numeracy, and technology capabilities are fundamental to these jobs. It is interesting to see, however, the importance of a group of interrelated skills and knowledge that pertain to working with the “clients” or “users” that these job roles serve. In particular, active listening, customer service, teaching/training, and active learning appeared prominently in most of these jobs. Given the CI-Facilitator role that we have defined – a vital member of the research enterprise who works closely with researchers to identify extant tools, data sets, and other resources that can be integrated into the process of pursuing a research objective – this finding should not be surprising. To assist scientists and engineers with their cyberinfrastructure needs will require a well-honed capability for eliciting user requirements and translating those requirements into effective systems and services. At a purely intuitive level, it is easy to see how an individual with the knowledge, skills, and mastery of the tools described in Table 1 would be a welcomed member of any science or engineering Facility.

One other aspect of Table 1 to note is that the descriptions of knowledge, skills, and tools are necessarily expressed at a very broad level. This characteristic is intrinsic to O*Net, which must cover a wide range of jobs with as small a collection of categories as possible. To obtain a more nuanced view of the knowledge, skills, and tools required for the CI-Facilitator role, the authors asked subject matter experts (Ph.D. faculty, and students from a variety of STEM disciplines) to brainstorm in these three categories. Table 2 displays the results of that effort.

Table 2: Knowledge, Skills, and Tools/Technology Suggested by SMEs

Importance (1-10 scale)	**Knowledge**	**Skills**	**Tools and Technology***
10	Domain knowledge in one or more areas of science and mathematics	Research skills (data elicitation, data analysis, scientific writing)	Database design tools
9	Architecture and operation of data networks	Communication skills (oral and written)	Content Management System / Website Development Tools
8	Information policy (Access controls, intellectual property rights, licensing, privacy)	Service skills (working with people, determining user needs)	Server Administration

7	Human-computer interaction	Database design skills	Large Scale Computing (Mainframes, Supercomputing, Grids)
6	Scripting, query, and programming languages	Cultural sensitivity, working with people with disabilities	Distributed collaboration tools

Although the subject matter experts focused less on fundamental skills of literacy and numeracy than the O*Net analysis showed, we believe these essential skills were assumed by our subject matter experts. More importantly, these lists of knowledge, skills, and tools provide a much finer level of detail, particularly with respect to prevailing information technologies. Also worthy of comment, the subject matter experts were adamant on the importance of math and science skills for the CI-Facilitator. An individual who solely had in-depth knowledge of information technology would not be as valuable because of the difficulty of knowing the specialized nomenclature and needs in specific domains of science and engineering.

Example Challenge: Massive Scale Information

Clearly the challenge of preparing CI-Facilitators for the work they must do is a vast undertaking. A great deal of foundational work must be done to define conceptual and technical approaches to facilitation. Without such deep foundations, every new tool and cyberinfrastructure development will necessitate a new learning curve. Just as STEM scientists depend on both a theoretical orientation and methodological base in their work, so to must a CI-Facilitator. There are many approaches to this conceptual foundation that must be explored throughout the development of the CI-Facilitator concept and educational program. The following presents one such approach to be considered – one based on conversation theory [15] and participatory librarianship [19]. These concepts are explored in relationship to one of the challenges of the CI-Facilitator: massive scale information.

The ideas presented here on massive scale and facilitation began in conjunction with a U.S. National Academies of Science study on information management in the transportation industry. Several study panel members observed that soon every mile of U.S. highway will generate a gigabyte of data a day [13]. This data will come from road sensors embedded into asphalt to detect temperature for winter salting, real-time traffic data from roadway cameras, weather information, toll data from RFID (Radio-Frequency Identification) expressway systems, car black boxes, and a myriad of other data sources. It is assumed that this will become a gigabyte an hour as more and more technology finds its way into our vehicles and management systems (GPS data, real time environment monitoring, etc.). As there are 3.5 million miles of highways in the U.S. that would be 3.3 petabytes of data per hour, or 28 exabytes per year. Such data can be of immense value to the scientific and research communities; however, it can just as certainly overwhelm these communities and mask important findings.

Some readers may not be familiar with an exabyte. It is the name for a very large volume of storage like megabytes, gigabytes (1024 megabytes) and terabytes (1024 gigabytes); technically 2^60 bytes. Table 3 (derived from http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm) will give the reader some sense of the scale involved.

Byte	1 byte: a single character
Kilobyte	2 Kilobytes: A typewritten page
Megabyte	2 Megabytes: A high resolution photograph
Gigabyte	2 Gigabytes: 20 meters of shelved books
Terabyte	2 Terabytes: An academic research library
Petabyte	2 Petabytes: All US academic research libraries
Exabyte	5 Exabytes: All words ever spoken by humans
Zettabyte	
Yottabyte	

What the reader needs to know is that each succeeding row in the table, from megabyte to gigabyte to terabyte and so forth, is an exponential increase.

E-science environments have demonstrated that the trends of the transportation industry in the utilization of massive-scale data sets will be wide spread across STEM disciplines and beyond. Given the proclivity of the sciences towards larger data sets the issues of archiving and retrieving these data sets will be of major important. The problems go beyond simply storing materials, or the application of information retrieval technologies. While important aspects, neither retrieval nor storage directly addresses intellectual access issues. These issues deal with questions of taxonomy development and the utilization of metadata to more precisely represent the contexts of an information object not simply the content.

The areas of classification and ontology development have created a rich empirical and theoretical understanding of the issues of intellectual access. One of the clear findings of these fields is that contextual clues, classifications, and ontologies can greatly aid access to materials, at least for those familiar with the classification system. Another key finding is the applications of descriptive metadata, or contextual information, is a process of applying external data to an object, not simply highlighting inherent aspects of an object. An example might clarify this point.

Imagine a large-scale meteorological simulation. That simulation contains some intrinsic data elements (variables of the simulation, run date, computing environment, etc). However, it will not contain useful extrinsic data such as who developed the simulation, the purpose of the simulation, links to articles that emerged from the simulation, etc. These will be added (authorship may well be contained in resulting files, but are considered extrinsic because this data is not necessary for the simulation to run). What’s more, the uses of this simulation are extrinsic. So, this simulation may be useful in educational settings, it may also be useful in legal settings, policy settings, etc. To describe the utility of the simulation to these alternate settings requires some descriptive process. Further, this process will be applying descriptions more about the context of use than the simulation itself.

The use of intellectual access has a long tradition in libraries. For centuries libraries have developed indexes, classification systems, and metadata schemas to describe, organize and locate information. Yet, the rise of massive scale information repositories is complicating traditional systems of intellectual access. By and large libraries have relied upon human cataloging. In the early days of the Web, there were several attempts to apply these human descriptive processes to web sites and pages. Such approaches were quickly abandoned or scaled-back given the realities of the number of pages involved (this does not even take into account added complexities such as dynamic and shifting data). The question for CI-Facilitators is how can the benefits of intellectual access be gained without the inherent scalability problems of a librarian-centered descriptive process?

Replacing traditional approaches to information organization based on artifacts and metadata with a new approach based on use and context, CI-Facilitators can more effectively aid large-scale e-research initiatives. Such a participatory approach, grounded in both complexity [14] and conversation theory [15], seeks to utilize knowledge creation processes to interconnect e-research projects, aiding in the dissemination and impact of ongoing projects.

Conversation theory, developed by Gordon Pask, is a macro-theory that seeks a general approach to the questions of learning and knowledge. It has at its core the premise that knowledge is created through conversation. A conversation consists of:

- Conversants: two or more cognitive agents. A conversant is a scalable concept where a person can be a conversant, or a group of people. Also, multiple agents can reside within a single person as in metacognition [19] and critical thinking.

- Language: conversants exchange language. Language can be at a very high level where all participants share a great deal of domain knowledge, or at a low level where one or more of the parties has little pre-existing knowledge.

- Agreements: Conversants seek agreements through language interchange. These agreements form a common context and can be scaffolded to seek greater domain understanding.

- Memory: is the storage of agreements in a relational way that integrates them into a domain-wide knowledge.

One of the implications of this theory as applied to the question of information organization is that the artifacts created from a knowledge process - be it a simulation, article, or data set – are artifacts of true knowledge, and therefore “secondary objects.” To be clear, this does not mean unimportant, simply derivative of what should be the focus of information organization: the conversation.

A focus on use and conversation over taxonomies changes the approach to intellectual access. First, access is to conversations and relationships not simply artifacts. The second implication of conversations is the distribution of the descriptive process as a social and behavioral process. Each of these implications is discussed below.

Cataloging Conversations

Where did the meteorological simulation presented in an earlier example come from? Certainly it was the result of data analysis and programming. However, it was also the result of previous research, discussions with colleagues, and more than likely a complex series of interactions between scientists, graduate students, funders, system administrators, and others. In essence, the simulation is the final result of an ongoing conversation. Further, each aspect of that conversation were themselves the result of other conversations (the best way to program, optimal data visualizations, grant negations, etc). The end result is that the actual simulation can be seen in a variety of contexts. To the scientist the simulation is the result of an ongoing research agenda. For the programmers it is a demonstration of their skill. For another researcher it might be a confirming piece of evidence. Each of these contexts is equally "true." Yet the current approach to intellectual access is to try and match this complex series of interactions through a single object to a point in a unified classification scheme. Why not map the conversations themselves?

This approach is probably best seen in the shift from traditional IR approaches to page rank like algorithms in web searching. Traditional IR techniques, before Google, used the text of a given document and some limited domain knowledge to rank and present material. Therefore the rank of the document was most influenced by attributes of the document itself. Google added behavioral data to dramatically improve the results. The behavior utilized was that of linking. The more sites that linked to a given page, it was reasoned, the more important that page must be. No longer were search engines limited to textual data within a document, they could now take advantage of the unique nature of networked corpora.

The same can be seen in approaching conversations over documents. Here the artifacts take on context from how they are used and created. This is far from a new concept. After all, scholarly communications have long mapped conversations through citations. Citation analysis and other bibliometric techniques map conversations and reflect how artifacts are part of a much larger discourse. This citation approach must be extended to link not only highly polish and synthesized artifacts (such as articles) in a computationally intelligible fashion, but to data sets, simulations, emails, and other associated artifacts in an e-science environment as well.

This kind of conversational linking goes beyond increasing the "browsability" of scientific data (read the article, click on the table to see the raw data; click on the raw data and examine the sensor that created the data; see who else uses that type of data; find their analysis and resulting articles), but also the ways in which data is located. For example, patterns can be identified. So that most of the scientists who do this type of analysis use this type of tool. Now a system can be built that automatically suggests that tool to a scientist or CI-Facilitator. It is a sort of meta-analysis of whole domains that includes not only articles, but the entire research process.

E-science environments provide the optimal environment for this approach. Unlike traditional approaches to information organization where the focus is on noise reduction to create an optimal system, a conversational system should actually benefit from larger and more diverse data. Blogs, lab notes, run time reports, datasets and citation data only increase the richness of the patterns and conversations. “Seminal information” may exist in the foment of scholarly online debate and trials long before it ever makes it to article writing. By focusing on conversations, these patterns can emerge more quickly and be turned into system features. It is important, however, that e-science environments are built to capture items beyond artifacts. E-science platforms must allow for conversation, debate, hypotheses, and all of the work of scientists, not simply computing resources and data sets.

By building intellectual access around conversations the scalability of intellectual access begins to be addressed as well. Where the current focus on artifacts waits for the development of a highly polished information object, and then creates a separate process for artifact description, a conversational approach would use behavior to describe the materials. It would utilize inherent metadata, and capture descriptive metadata at the point of creation (authorship, creation date). It could also infer additional metadata such as topics, and project affinity. Note, this would not be done by constantly presenting a scientist with some sort of cataloging screen to fill in, rather it would be captured automatically and often unobtrusively. In essence intellectual access emerges from use of a system, and as an aggregation of individual efforts. While certainly there is room for applying traditional ontological approaches to the conversations, these high effort high reward activities (such as peer review and pathfinder construction) is no longer a necessary first step. It can be targeted and prioritized.

There are other implications for CI-Facilitators in utilizing a conversational approach. It reaffirms the opening comments about the importance of communication and collaborative skills for example. It also aligns the work of in developing CI-Facilitators with existing and well-explored areas in system design, information organization, cognitive psychology, and discourse analysis in addition to STEM disciplines. As the development of CI-Facilitators programs evolve it shall be interesting to see how the underlying conceptual frameworks develop alongside cyberinfrastructure itself.

References:

1. Blythe, E. (2004). Déjà Vu. EDUCAUSE REVIEW, 39(3), 60-61.

2. Buetow, K. H. (2005). Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research (Vol. 308, pp. 821-824): American Association for the Advancement of Science.

3. Hey, T., & Trefethen, A. E. (2005). Cyberinfrastructure for e-Science (Vol. 308, pp. 817-821): American Association for the Advancement of Science.

4. Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., et al. (2003). Revolutionizing Science and Engineering Through

Cyberinfrastructure. National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, January.

5. Klein, J. T. (1990). Interdisciplinarity: History, Theory, and Practice: Wayne State University Press.

6. Seymour, E. (2002). Tracking the processes of change in US undergraduate education in science, mathematics, engineering, and technology. Science Education, 86(1), 79-105.

7. Fox, M. A., & Hackerman, N. (2003). Evaluating and Improving Undergraduate Teaching in Science, Technology, Engineering, and Mathematics: National Academies Press.

8. Gibbons, M. (1994). The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies: Sage Publications.

9. Hicks, D. M., & Katz, J. S. (1996). Where Is Science Going? Science, Technology & Human Values, 21(4), 379.

10. Odlyzko, A. (2002). The rapid evolution of scholarly communication. Learned Publishing, 15(1), 7-19.

11. Tuomi, I. (2002). The Lives and Death of Moore’s Law. First Monday, 7(11), 4.

12. Varian, H. R., Farrell, J. V., & Shapiro, C. (2004). The Economics of Information Technology: An Introduction: Cambridge University Press.

13. Transportation Research Board, Committee for a Future Strategy for Transportation Information Management, Transportation Knowledge Networks: A Management Strategy for the 21st Century, TRB Special Report 284 (Washington D.C.: 2006), www.trb.org/news/blurb_detail.asp?id=5789 (accessed Sept. 15, 2007).

14. Holland, J. H. (1995). Hidden order: How adaptation builds complexity. New York: Addison Wesley.

15. Pask, G. 1976. Conversation Theory : Applications in Education and Epistemology. New York: Elsevier.

16. Berman, F. D., & Brady, H. E. (2005). Final Report NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences: National Science Foundation.

17. McCloy, R., Waugh, G., Medsker, G., Wall, J., Rivkin, D., & Lewis, P. (1999). Development of the O* NET™ Computerized Work Importance Profiler. Raleigh, NC: National Center for O* NET Development.

18. Clery, D. (2006). INFRASTRUCTURE: Can Grid Computing Help Us Work Together? (Vol. 313, pp. 433-434).

19. Hartman, H. J. (2001). Metacognition in Learning and Instruction: Theory, Research and Practice. Dordrecht: Kluwer Academic Publishers.

19. Lankes, R. David, Silverstein, J. L., Nicholson, S. “Participatory Networks: The Library as Conversation.” 52(2). Information Technology and Libraries.