Ubiquitous Infrastructure for Deep Accountability

Robert E. McGrath
National Center for Supercomputing Applications
University of Illinois, Urbana-Champaign
mcgrath@ncsa.uiuc.edu

ABSTRACT

The "Web 2.0" has created new capabilities to "mash up" content and "meet" on-line to create new collective knowledge. But to develop, maintain, and sustain serious collective knowledge, serious accountability is needed. The emerging CyberInfrastructure provides ubiquitous mechanisms for accessing shared resources, but serious collective knowledge requires deep accountability for "everything": data, process, and assertions. This will be accomplished by software infrastructure that implements:

- Stable identities for every entity of interest
- Open metadata
- Data and process provenance

These critical issues are addressed by current developments from many sources. Several projects show that these principles are within reach. While emerging from large-scale science and engineering, they provide a toolkit that will be useful for many types of collaborations, and will enhance and expand the deployment of mass collaborations such as Wikipedia.

Keywords: CyberEnvironments, CyberInfrastructure, Collective Knowledge, Collaboration, Accountability

1. INTRODUCTION

The internet provides a ubiquitous decentralized infrastructure for sharing information. This is evolving into a global CyberInfrastructure to support collaboration in virtual organizations [34]. For example, the Grid—e.g., the Open Grid Services Infrastructure [49]—provides strong security, and facilities for managing data and computation. However, it is centralized, complex, and not flexible enough for many uses. For example, the Grid community has struggled to support anonymous community accounts, which are essential for collaborative groups [41].

The so-called "Web 2.0"—blogs, wikis, etc.—has enabled ordinary users to develop complex applications [36, 44, 48]. This has led to the emergence of web-based activities that aim to create collective knowledge—once the domain of scholarship, science, and experts. This is epitomized by wikipedia (http://www.wikipedia.org), a decentralized, volunteer-fueled knowledge base, similar to an encyclopedia, but written and corrected by anyone who donates their effort [35]. Wikipedia employs simple social networking technology and "mass amateurization", and has generated knowledge at a fantastic pace on a broad array of topics [35, 44, 50]. However, in many cases, the longevity and/or quality of the knowledge is suspect (e.g., [52]).

Does the Web 2.0 have "too little accountability", and the Grid "too much accountability", or perhaps both are just not what is needed? What needs to be built? This paper suggests a specific infrastructure that, together with Web, Grid, and emerging national CyberInfrastructure standards, can substantially improve accountability for digital communities of many types. The general principle is to design flexible and reusable middleware that provides the "right" set of services, without "wiring in" a specific set of assumptions about how the systems must be used.

2. COLLABORATION IN DIGITAL COMMUNITIES

While on-line collaboration is scarcely new ([1, 20, 25]), the advent of so-called "Web 2.0" technologies has enabled the Internet to support collective activity of all types. Social computing technologies offer inexpensive and ubiquitous sharing and communication, with capabilities for reuse and border-crossing. These technologies include blogs, RSS feeds, wikis, file sharing, folksonomies, and simple technologies for aggregating this content [7, 36, 48, 51].

Using these technologies, it is comparatively easy for ordinary users to "repurpose" web content to create their own views and aggregate ("mash up") information from many sources. Information is reused outside the borders of its original context, perhaps for radically different purpose.

Serious collaborations have existed long before the Internet, in the form of dictionaries, encyclopedias, scholarship, science, and other knowledge intensive activities. Conventional scholarship places stock in expertise and professional credentials, and heavily filters "content" on the way to publication. Historic standards of publication, citation, and review have carried forward to be applied to digital objects. However, digital environments are too complex and the data too voluminous to manually create all the necessary content and context. In short, digital collaborations cannot rely on manual cataloging, selection, and annotation, as was done in the past.

For example, in scientific fields, instruments can easily generate thousands of results per minute—far too many for any human or group to check or annotate the purpose, quality, and use of each measurement. Furthermore, it is increasingly important to study problems and systems that span phenomena and disciplines, arenas where there is no single discipline or group of experts. These critical studies of complex systems require knowledge "mash ups" across multiple domains of expertise, in which it is not possible for a single person or group to create all the content needed to understand and use the digital artifacts. The producers of software, data, and knowledge do not know who the consumers may be, let alone what they may wish to do.

The Web 2.0 deal with this scaling challenge through "mass amateurization" and a "publish, then filter" approach [44]. Wikipedia, in particular, has proved to be a serious provocation to scholarly communities, which have well-established methods for creating cumulative, collective knowledge. Comparing today's popular collectives (e.g., wikipedia) to established scholarship (e.g., a journal such as *Nature* (www.nature.com));

they have similar goals and appearance. However, there are significant differences in practices and expectations about accountability: a substantial difference in culture.

The question of accountability underlies the potential value of a “serious collective”. Is the cumulative knowledge authoritative (reliable, correct, useful, etc.)? What would need to be done to establish a great wiki page as “*the authoritative*” source on that topic, upon which others should rely and build new knowledge? For what audience and for what purposes? If not, what could be done to make it so?

The following sections outline new and achievable infrastructure that will make it possible to achieve deep accountability when needed within Web 2.0 style open networks.

3. COMMUNITIES FOR CREATING COLLECTIVE KNOWLEDGE

A collaborative community is about sharing, presence, and commitment to common goals. Some collaborations—such as scientific communities—have deep and serious goals, such as developing and promulgating reliable knowledge. Creating knowledge requires more than socializing, easy file sharing, and the capability to “mash up” information, it requires clear understanding of the context and quality of the information underlying the knowledge. This understanding is based on accounts of “where it comes from”, “who says so” and “why”.

Supposing that collective knowledge is generated, how can it be propagated and built on by others? A classical approach is to capture knowledge in highly authoritative artifacts (e.g., journal articles), constructed with extreme care toward sources and arguments. Centuries of scholarly and scientific practice have worked from the principle that transparency and accountability are more important than majority votes. Much of the so-called scientific method and real life scientific practice is dedicated to accounting for the sources and destinations of data and arguments.

Digital collectives are exploring different approaches to quality control. These are based on social pressure [35], reputation [15], and voting and cross-references [5] It remains to be seen whether, say, wikipedia’s honor system can achieve high quality, long-term cumulative knowledge [52]. In fact, Wikipedia has evolved norms and processes for managing and coordinating collective action, which resemble those in conventional organizations [50]

There are many different communities, with different practices and cultures of accountability: some communities are informal, requiring limited trust; others have more formal or “serious goals” (such as scholarly or legal authority), which require careful accounts of sources and arguments. Furthermore, some “communities” are composed across multiple contexts, each of which may impose an alternative “view” on knowledge.

Different collaborative communities may require different kinds and levels of accountability. Despite the diversity of ends and means, all communities have common needs, and in any case need to use standard mechanisms to create their own cultures. For communities to interoperate, to “mash up” across multiple domains, it is necessary to have standard automated mechanisms for understanding the sources of knowledge. Infrastructure should provide services that enable communities to implement their own culture of accountability.

As collaborative technologies develop, it will be critical to have standard mechanisms for accountability of several kinds. Accountability cannot be done through a centralized authority, it will follow the philosophy and model of the Web: simple reliable mechanisms that enable users to “mash up” accountability for their own purposes. The general principle is to design flexible and reusable middleware that provides the “right” set of services, without “wiring in” a specific set of assumptions about how the systems must be used.

4. CYBERENVIRONMENTS: INFRASTRUCTURE FOR COLLABORATION

In recent years, a ubiquitous middleware has emerged to enable large-scale, multidiscipline, system-level science and engineering. This *CyberInfrastructure* (CI) is greatly decreasing the costs of sharing data, instrument, and computational resources [34]. For many users, ubiquitous access is necessary but not sufficient. Additional software is needed to enable information-intensive communities to exploit local resources and national CI in their research, development, and teaching activities, which we have termed *Cyberenvironments* [31, 33]. Rather than focusing on universal *access to* resources, Cyberenvironments emphasize the integration of resources into end-to-end scientific processes, integration across Cyberenvironments, and the continuing development and dissemination of new resources and new knowledge.

This vision of Cyberenvironments assumes that research results such as papers, processes, and data can be conveyed with enough information about themselves to be incorporated into further research work. This capability is essential to establishing accountability, and requires infrastructure for managing metadata (i.e. what units are the data in) and provenance (which data was discussed in a paper, what analysis was applied to it) [31].

Cyberenvironments draw inspiration from reflective software [22]. Adaptive systems employ brokers to dynamically compose software components using several critical design principles [33]:

- Decomposition into abstract interfaces
- Exposed metadata
- Instrumentation

The next sections present an abstract architecture and some implementations of the concepts.

5. TECHNOLOGIES FOR ACCOUNTABILITY

This section presents three related and mutually supporting concepts which provide the key patterns for flexible and adaptive, yet potentially thoroughly accountable systems. This will be accomplished by software infrastructure that implements:

- Stable identities for every entity of interest
- Open metadata
- Data and process provenance

These concepts are designed to enable individual communities can implement their own practices, yet still exchange processes, data, and knowledge. These abstract services can be implemented with current technology.

5.1. Borders and Sharing: Contexts

Knowledge is created and used in a context. Data cannot “speak for itself”; it is meaningful in the context of a community, termed here a *Virtual Organization*. A Virtual Organization is a container for a shared view of the universe of entities and activities of interest [34]. This context has many facets: it is a filter (for what is "relevant"), a reputation and recommendation system (for "best practice" and "expertise"), and a social network ("who knows what"), as well as a collection of specific techniques and artifacts (such as standard datasets and programs).

Virtual Organization contexts can be created whenever people wish to collaborate; from buddy lists, to ad hoc task-oriented teams, through more formal organizations. For example, consider a group of scientists that wish to study an ecological anomaly, such as a suspected sudden geographical shift in insect species. To study this problem will require experts with knowledge of entomology, botany, hydrology, statistics, and more. Furthermore, the experts need to be brought together quickly and efficiently, possibly from around the world. This ad hoc team should be able to form a virtual organization, building on common infrastructure to combine data and computation from multiple domains to address the question. This sort of boundary crossing is always difficult: even if such data and software can be located and accessed, interpreting and integrating it requires a level of "meta-explanation" to assure valid results.

Virtual Organizations are more than a social networking tool, they can form an important part of problem solving. The contexts of a Virtual Organization can provide critical short cuts into complex knowledge spaces, through which a person can find relevant practice, knowledge, and high quality data for solving a current problem. A problem solver can first, find the right contexts (Virtual Organizations), and then, exploit the pre-built views to attack the new problem. Interestingly, this process parallels fundamental principles of creativity [46] (and possibly the function of the human brain [18]).

5.2. Identity and Credentials

Collaborations require people to have a stable identity *within the relevant social context of the collaboration*—that is, an identity based on community defined tokens of achievement and contribution. A person might have many digital "personalities," the precise set of credentials of interest depends on the goals of the collective activity.

However, for a collaboration across communities, it will be necessary to utilize "reputations" of people, data, and techniques from several fields. For example, a multidiscipline group studying ecology cannot rely on a single discipline to define the reputation or credentials of all the experts, or to define what or how datasets or methods should be employed.

While collaborations are contextual, all digital objects need permanent, universal identities. Web standard Universal Resource Identifiers (URIs) must be extended to provide "actionable" tags, which will reliably identify an object even as it may migrate over time [9, 23].

In addition to global IDs, current technology provides mechanisms for implementing stable identity and reputation, as well as certain kinds of anonymity ([3, 4]). Recommendation based on social network analysis helps navigate these identities to discover expertise and mutual interests (e.g., [11]).

Ultimately, reputation will be used to help assess the evidential value of data and metadata and people. Since there may be many alternative judgments of a particular claim based on different views, languages for reasoning about evidence may be needed. Logical foundations for notions of reputation and trust have been proposed (e.g., [14]), though these must be extended to consider context, and to reason about a body of evidence.

5.3. Deep Content and Annotation

One important aspect of serious collective work is creation of rich content that is linked to other content. This content may be any type of data—including "workflows," which capture practices [12]—along with metadata, annotations, and cross links. Content may be created in one context, used and reused many times in other contexts, and potentially accessed for years or decades. Given the scale and complexity of the digital world, and the cross-boundary nature of collaborative computing, it is unrealistic to expect a single repository will manage all data of interest, or create all the metadata needed. On the contrary, the typical case will be to use many sources.

Mechanisms for managing of content must be general and flexible enough to manage many forms of data, including software and very complex objects, as well as treating metadata, annotations, and provenance as "first class content". Furthermore, the physical representation of digital objects evolves over time, ultimately migrating into persistent archives. During this process, identity and annotations must "move" with the objects. For example, the same dataset may move from local copies into shared repositories, and then be preserved in an archive. Over the life of the dataset, annotations or cross-references must automatically follow the object.

Abstract models such as Java Content Repository [21] combined with reliable identifiers [9, 23] illustrate the features needed. The "Data Grid" is a step in this direction (e.g., [8, 38]), but needs to be pushed to even greater abstraction, as may be seen in current research projects such as Tupelo (http://tupeloproject.org), and the Data Format Definition Language [47].

Discovery and search rely on the availability of attributes about objects, which may come from metadata, user annotations, automated feature extraction, or other sources. Creating metadata is difficult and time consuming, so it is unrealistic to expect all metadata to be created manually either by "authorities" or by users. Furthermore, the attributes of interest depend on the purposes of the user. Annotations from one point of view may or may not be valuable to another, so an object cannot be annotated once and for all by its creator.

Existing technology provides much of the infrastructure needed for manually linking and annotating. Del.icos.us (http://del.icis.us), flikr (http://www.flickr.com), YouTube (http://youtibe.com), and similar services provide a simple and flexible scaffolding for sharing many kinds of annotations (termed "folksonomies") [6, 15-17]. This approach has been used for scholarly activities as well (e.g., http://citeulike.org, http://arxiv.org/, http://www.merlot.org/merlot/index.htm).

In addition to manual annotation, Automated Learning techniques may provide yet more metadata, through automated feature extraction (e.g., [37]). For example, techniques similar to spam filtering (e.g., [40, 43]) could be employed to detect data and events of interest, as defined by individuals or communities. These techniques could embody part of the "context" of the collective—which, in turn, could be annotated and shared.

Collective production of knowledge needs a reliable account of not only the identity of an artifact, but also its *provenance*—how data was generated, what software was used, what was the input data, and so on [28]. This is needed for many purposes. For example, if an anomaly is detected in a dataset, it will be necessary to discover the software versions that created the data, the input datasets, and parameter settings. In his case, it will also be important to know what processes used the flawed data, and how.

Provenance can be automatically collected and aggregated by the software infrastructure (e.g., [2, 10, 27-29]). This provenance can be combined with other metadata, and reasoned with to answer questions and infer relations, such as what data was used to reach a conclusion, or what conclusions depend on a particular data or process [26].

6. IMPLEMENTATION IN E-SCIENCE

This paper has described abstract services which can be built on current technologies. These ideas may be realized in many concrete implementations. This section discusses some projects that are building end-to-end Cyberenvironments tailored to specific scientific and engineering disciplines. These illustrate the use of the abstract mechanisms described above, and show they are in reach.

The Collaboratory for Multi-scale Chemical Sciences (CMCS) (http:/cmcs.org), funded by the U.S. Department of Energy funded effort, led by Sandia National Laboratories, is designed to enhance information transfer between chemistry sub-disciplines, connecting quantum chemistry, thermochemistry, kinetics, and modeling of combustion devices, e.g. diesel engines [29, 42].

The CMCS implements versions of the abstractions discussed above. CMCS is built on services that implement content management (via augmented WEBDAV [13, 30] service) and provenance capture (e.g., from a workflow from Extensible Computational Chemistry Environment , http://ecce.pnl.gov/index.shtml), which have enabled the exchange of data and development of new understanding of complex phenomena. The CMCS supports multiple contexts, international groups in a number of sub-fields are coordinating their research efforts and acting as expert groups, publishing new data and models backed by rich information about their creation and ranges of validity. Through automated metadata generation and provenance capture, the CMCS provides deep accountability required for science: what was done, what data was used, and who did what.

CMCS has been used by internationally distributed groups in a number of sub-fields who are coordinating their research efforts and acting as community ‘expert groups’ to publish new reference data and models backed by rich information about their creation and ranges of validity. These groups use the base content management and service integration capabilities of the system to assemble and curate computational and experimental data, to transparently exchange data between modelers who use different software with different file formats, and to stage data for use with newly developed tools. This ability to gather data from researchers around the globe and statistically analyze it together has resulted in a factor of ten improvement in the precision of knowledge of the properties of important chemical species [39]. Further, any group wishing to understand this work, or to repeat or extend it, can (with permission) access the data, tools, and discussions that occurred, see the specific analyses that were performed, and create their own space to work in CMCS and perform “what-if” analyses that include new data. [32].

The myGrid project provides similar middleware for bioinformatics, enabling *in silico* experiments. Taverna workflows are created and shared through a general content management system, which captures annotations and provenance. Notably, the workflows are available to the communities as another kind of content, capturing the processes used, enabling users to share and discover “best practices”, and assisting in the evaluation and improvement of methodology [53].

MyGrid has been used by scientific communities, for example in drug discovery [45]. Taverna services have enabled tools to (securely) integrate data and information from many sources, create, discover, and reuse workflows, and identify objects. Provenance is tracked automatically, providing critical accountability. In this environment (biomedical investigation), it is critical to develop a deep understanding of a potential finding (the underlying data, theory, and assumptions), and to compare alternative analyses (again, “what-if” scenarios), as well as discovery of non-obvious relations (e.g., from apparently unrelated studies). MyGrid illustrates that implementation of the abstractions described in this paper enable the construction of rich web of knowledge, to improve understanding of who, what, when, where, and how.

Other projects such as CI-Shell [19], Nanohub (http://www.nanohub.org), Comb-e-Chem (http://www.combechem.org), and the WATer and Environmental Research Systems (WATERS) Network [24] are building specific environments using these principles. These projects suggest that the abstractions are both feasible and useful, enabling communities to create their own collaborative contexts.

7. CONCLUSION

This paper has described an intertwined set of abstract services that, on top of ubiquitous Cyberinfrastructure, provide flexible mechanisms for accountability in mass collaborations at many scales. These Cyberenvironments can be implemented with technology that is in reach. While emerging from large-scale science and engineering, they provide a toolkit that will be useful for many types of collaborations, and will enhance and expand the deployment of mass collaborations such as Wikipedia.

Social computing opens the door to creating collective knowledge through new capabilities to “mash up” content and “meet” on-line to create new collective artifacts of many kinds. But, in order to develop, maintain, and sustain serious collective knowledge, serious accountability is needed. This accountability cannot be done through a centralized authority, it will follow the philosophy and model of the Web: simple reliable mechanisms that enable users to “mash up” the required accountability. This paper outlines important developments, which are in reach today.

The emerging CyberInfrastructure provides ubiquitous mechanisms for accessing shared resources. This ubiquitous CyberInfrastructure is a necessary but not sufficient foundation for creating collective knowledge. In particular, serious collective knowledge requires not only ubiquitous sharing and open software, it needs deep accountability for “everything”: data, process, and assertions.

This will be accomplished by software infrastructure that implements:

- Stable identities for every entity of interest
- Open metadata

- Data and process provenance

These critical issues are addressed by current developments from many sources.

Several projects show that these capabilities are within reach, and that the concepts enable creation of Cyberenvironments that can be used by professional science and engineers, and by teachers and learners; in large projects, individual labs or classes, and in ad hoc work groups. These techniques are not limited to technical or scholarly activity, including collaborations for collective commerce, socializing, and entertainment.

Existing collaborations suggest that these mechanisms are both needed and would be better than improvised solutions. Wikipedia provides an interesting case. A wikipedia page is a snapshot of knowledge, with an associated history of discussion and edits that occurred within the wikipedia environment. Thus, Wikipedia implements a context for collaboration (i.e. Wikipedia is Virtual Organization), with shared processes, norms and tools, which have emerged through discussions, and are publicly documented and enforced by user surveillance [44, 50]. Wikipedia's process depends heavily on a complete provenance for each page, which is needed to allow public scrutiny and corrections, as well as to undo mistakes or vandalism. Wikipedia also has strong norms about representation of expertise and credentials. Thus, Wikipedia has created its own versions of the key concepts discussed above. Ubiquitous deployment of the services described in this paper would make it much easier to create new Wikipedia-style collaborations.

Furthermore, ubiquitous Cyberenvironments envisioned here would enable a "better Wikipedia"; with a broader and deeper account of the purported knowledge presented, across the boundaries of a single virtual organization. Rather than a single snapshot of knowledge ("the current version of an article"), the artifact can be a complex web of knowledge including data, computation, and visualizations, and the history of the current artifact (compare, for example, myGrid [45] to a Wikipedia page). Furthermore, drilling down from the "article" leads to representations of the history, sources, and processes underlying the claims, including the data and software used, as well as citations and who did what. Standard representations of provenance and other metadata, with stable identities enables evaluation of the knowledge (is it credible?), and comparison of alternative accounts (e.g., using different data or assumptions).

8. ACKNOWLEGMENTS

The author wishes to acknowledge the contribution of many CI researchers to the concepts and systems discussed here with specific recognition of members of NCSA's Cyberenvironments Directorate. The National Center for Supercomputing Applications is funded by the US National Science Foundation under Grant No. SCI-0438712. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Abdel-Wahab, H.M., Guan, S.-U. and Nievergelt, J. Shared workspaces for group collaboration: an experiment using Internet and UNIX interprocess communications. *IEEE Communications Magazine*, *26* (11). 10-16.
2. Altintas, I., Barney, O. and Jaeger-Frank, E., Provenance Collection Support in the Kepler Scientific Workflow System. in *IPAW'06 International Provenance and Annotation Workshop* (Heidelberg, Germany, 2006), Springer-Verlag Berlin, 118-132.
3. Baron, D.P. Private Ordering on the Internet: The EBay Community of Traders. *Business and Politics*, *4* (3). Article 1.
4. Basney, J., Humphrey, M. and Welch, V. The MyProxy online credential repository. *Software Practice and Experience*, *35* (9). 801-916.
5. Brin, S. and Page, L., The Anatomy of a Large-Scale Hypertextual Web Search Engine. in *Seventh International Conference on World Wide Web*, (Brisbane 1998), 107 - 117.
6. Chi, E.H. and Mytkowicz, T. Understanding Navigability of Social Tagging Systems, Palo Alto Research Center, 2007.
7. Feiler, J. *How To Do Everything with Web 2.0 Mashups*. McGraw Hill, New York, 2008.
8. Foster, I.T., Vöckler, J.-S., Wilde, M. and Zhao, Y., Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation. in *14th International Conference on Scientific and Statistical Database Management*, (2002), 37 - 46
9. Futrelle, J. Actionable resource tags for virtual organizations, NCSA, 2006.
10. Futrelle, J. and Myers, J. Tracking Provenance in Heterogeneous Execution Contexts. *Concurrency and Computation: Practice and Experience*, *20* (5). 555-564.
11. Futrelle, J., Myers, J., Minsker, B. and Bajcsy, P., Community-based Metadata Integration for Environmental Research. in *Proceedings of the 7th International Conference on Hydro-science an Engineering (ICHE)*, (Philadelphia, PA, 2006).
12. Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L. and Myers, J. Examining the Challenges of Scientific Workflows *IEEE Computer*, *40* (12). 24-32.
13. Goland, Y., Whitehead, E., Faizi, A., Carter, S. and Jensen, D. HTTP Extensions for Distributed Authoring -- WEBDAV, IETF, 1999.
14. Golbeck, J., Combining Provenance with Trust in Social networks for Semantic Web content filtering. in *International provenance and Annotation Workshop*, (2006), 101-108.
15. Golbeck, J. and Hendler, J., Accuracy of Metrics for Inferring Trust and Reputation. in *14th International Conference on Knowledge Engineering and Knowledge Management*, (Northamptonshire, UK., 2004).
16. Golder, S.A. and Huberman, B.A. The Structure of Collaborative Tagging Systems., Information Dynamics Lab, HP Labs, 2005.
17. Hammon, T., Hannay, T., Lund, T. and Scott, J. Social Bookmarking Tools (I). *D-Lib Magazine*, *11* (4).
18. Hawkins, J. *On Intelligence*. Henry Holt and Company, New York, 2005.
19. Herr, B.W., Huang, W., Penumarthy, S. and Börner, K. Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. *Annals of the New York Academy of Sciences*, *1093*. 161-179.
20. Jackson, L.S. and Grossman, E. Integration of synchronous and asynchronous collaboration activities. *Computing Surveys*, *31* (2es). article 12.

21. Java Content Repository. Content Repository API for Java™ Technology Specification 2005.
22. Kon, F., Costa, F., Blair, G. and Campbell, R.H. The Case for Reflective Middleware. *Communications of the ACM, 45* (6). 33-38.
23. Kunze, J., Towards electronic persistence using ARK identifiers. in *3rd ECDL Workshop on Web Archives*, (2003).
24. Liu, Y., McGrath, R.E., Myers, J.D. and Futrelle, J., Towards a Rich-context Participatory Cyberenvironment. in *Workshop on Grid Computing Environments (GCE07)*, (Reno, 2007).
25. McGrath, J.E., Groups interacting with technology: the complex and dynamic fit of group, task, technology, and time. in *ACM conference on Computer-supported cooperative work*, (Toronto, Ontario, Canada 1992), 4.
26. McGrath, R.E. and Futrelle, J., Reasoning about Provenance with OWL and SWRL. in *AAAI 2008 Spring Symposium "AI Meets Business Rules and Process Management"*, (Palo Alto, 2008).
27. Moreau, L., Freire, J., McGrath, R.E., Myers, J., Futrelle, J. and Paulson, P. The Open Provenance Model, 2007.
28. Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., Munroe, S., Rana, O., Schreiber, A., Tan, V. and Varga, L. The provenance of electronic data. *Communications of the ACM, 51* (4). 52-58.
29. Myers, J.D., Allison, T.C., Bittner, S., Didier, B., Frenklach, M., Green, W.H., Ho, Y.-l., Hewson, J., Koegler, W.S., Lansing, C., Leahy, D., Lee, M., McCoy, R., Minkoff, M., Nijsure, S., von Laszewski, G., Montoya, D., Oluwole, L., Pancerella, C., Pinzon, R., Pitz, W., Rahn, L.A., Ruscic, B., Schuchardt, K., Stephan, E., Wagner, A., Windus, T. and Yang, C. A Collaborative Informatics Infrastructure for Multi-scale Science. *Cluster Computing, 8* (4). 243-253.
30. Myers, J.D., Chappell, A.R., Elder, M., Geist, A. and Schwidder, J. Re-Integrating The Research Record. *Computing in Science and Engineering, 5* (3). 44-50.
31. Myers, J.D. and Dunning, T.H., Cyberenvironments and Cyberinfrastructure: Powering Cyber-research in the 21st Century. in *Foundations of Molecular Modeling and Simulation (FOMMS 2006)*, (2006).
32. Myers, J.D., Dunning, T.H. and McGrath, R.E. Cyberenvironments: Ubiquitous Research and Learning (in press). in Cope, B. and Kalantzis, M. eds. *Ubiquitous Learning*, 2008.
33. Myers, J.D. and McGrath, R.E., Cyberenvironments: Adaptive Middleware for Scientific Cyberinfrastructure. in *Adaptive and Reflective Middleware*, (Newport Beach, 2007).
34. National Science Foundation Cyberinfrastructure Council. NSF'S Cyberinfrastructure Vision for 21st Century Discovery, NSF, 2007.
35. Nov, O. What Motivates Wikipedians? *Communications of the ACM, 50* (11). 6064.
36. Pierce, M.E., Fox, G., Yuan, H. and Deng, Y., Cyberinfrastructure and Web 2.0. in *International Advanced Research Workshop on High Performance Computing and Grids*, (Cetraro (Italy), 2007).
37. Plale, B., Vijayakumar, N., Ramachandran, R., Li, X. and Baltzer, T., Real time Filtering and Mining of NEXRAD Streams for Mesoscale Forecast and Prediction. in *23rd AMS Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology}*, (San Antonio, 2007).
38. Rajasekar, A., Wan, M., Moore, R., Schroeder, W., Kremenek, G., Jagatheesan, A., Cowart, C., Zhu, B., Chen, S.-Y. and Olschanowsky, R. Storage Resource Broker - Managing Distributed Data in a Grid. *Computer Society of India Journal, Special Issue on SAN, 33* (4). 42-54.
39. Ruscic, B., Pinzon, R.E., Laszewski, G.v., Kodeboyina, D., Burcat, A., Leahy, D., Montoy, D. and Wagner, A.F. Active Thermochemical Tables: thermochemistry for the 21st century. *Journal of Physics: Conference Series,, 16* (1). 561-570.
40. Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E., A Bayesian approach to filtering junk e-mail. in *AAAI'98 Workshop on Learning for Text Categorization.*, (1998).
41. Scavo, T. and Welch, V., A Grid Authorization Model for Science Gateways. in *Grid Computing Environments (GCE) workshop*, (Reno, 2007).
42. Schuchardt, K., Pancerella, C., Rahn, L.A., Didier, B., Kodeboyina, D., Leahy, D., Myers, J.D., Oluwole, O., Pitz, W., Ruscic, B., Song, J., Laszewski, G.v. and Yang, C., Portal-based Knowledge Environment for Collaborative Science. in *GCE 2005: Workshop on Grid Computing Portals*, (Seattle, 2005).
43. Segaran, T. *Programming Collective Intelligence*. O'Reilley, Sebastopol, CA, 2007.
44. Shirky, C. *Here Comes Everybody: The Power of Organizing without Organizations*. Penguin, New York, 2008.
45. Stevens, R., McEntire, R., Goble, C., Greenwood, M., Zhao, J., Wipat, A. and Li, P. myGrid and the drug discovery process. *Drug Discovery Today: BIOSILICO, 2* (4). 140-148.
46. Sunstein, C.R. *Infotopia: How Many Minds Produce Knowledge*. oxford University press, New York, 2006.
47. Talbott, T.D., Schuchardt, K.L., Stephan, E.G. and Myers, J.D., Mapping Physical Formats to Logical Models to Extract Data and metadata: The Defuddle Parsing Engine. in *International Provenance and Annotation Workshop*, (Heidelberg, 2006), Springer, 73-81.
48. Tapscott, D. and Williams, A.D. *Wikinomics: How Mass Collaboration Changes Everyhing*. Penguin, New York, 2006.
49. Tuecke, S., Czajkowski, K., Foster, I., Frey, J., Graham, S., Kesselman, C., Maquire, T., Sandholm, T., Snelling, D. and Vanderbilt, P. Open Grid Service Infrastructure (OGSI), Global Grid Forum, 2003.
50. Viégas, F.B., Wattenberg, M. and McKeon, M.M., The Hidden Order of Wikipedia. in *Online Communities and Social Computing*, (Beijing, 2007), 445-454.
51. Viégas, F.B., Wattenberg, M., van Ham, F., Kriss, J. and McKeon, M., *Many Eyes*: A Site for Visualization at Internet Scale. in *IEEE Information Visualization*, (Sacremento, 2007), 1121-1128.
52. Waters, N.L. Why You Can't Cite Wikipedia in My Class. *Communications of the ACM, 50* (9). 15-17.
53. Wroe, C., Goble, C., Goderis, A., Lord, P., Miles, S., Papay, J., Alper, P. and Moreau, L. Recycling Workflows and Services Through Discovery and Reuse. *Concurrency and Computation: Practice and Experience, 19* (2). 181 - 194.