THE NEED FOR FORMALIZED TRUST IN DIGITAL REPOSITORY COLLABORATIVE INFRASTRUCTURE

Fran Berman
Robert H. McDonald
San Diego Supercomputer Center

Brian E. C. Schottlaender
Ardys Kozbial
UC San Diego Libraries

Section 1: Introduction

A recent IDC report posits that 161 exabytes (10^{18} bytes) of digital information existed in the world in 2006 (Gantz, 2007). Given the unrelenting increase in digital data; its value to modern life, work, entertainment, and scholarship; and the challenge of developing and supporting adequate infrastructure for its management, stewardship, and preservation, it is clear that new approaches will be needed to meet the needs of digital data stewardship and preservation in the information age.

Today, a broad spectrum of institutions, communities, and individuals in both the public and private sectors are concerned with digital preservation, including universities, libraries, government agencies, researchers, and educators. At a recent workshop sponsored by the National Science Foundation (NSF) and the Association of Research Libraries (ARL), Dr. Chris Greer, Senior Advisor for Digital Data in the NSF Office of Cyberinfrastructure, presented an illustration of potential participants in digital preservation (Figure 1) and discussed the importance of crossing institutional and sector boundaries to craft a comprehensive solution to the data preservation challenge.

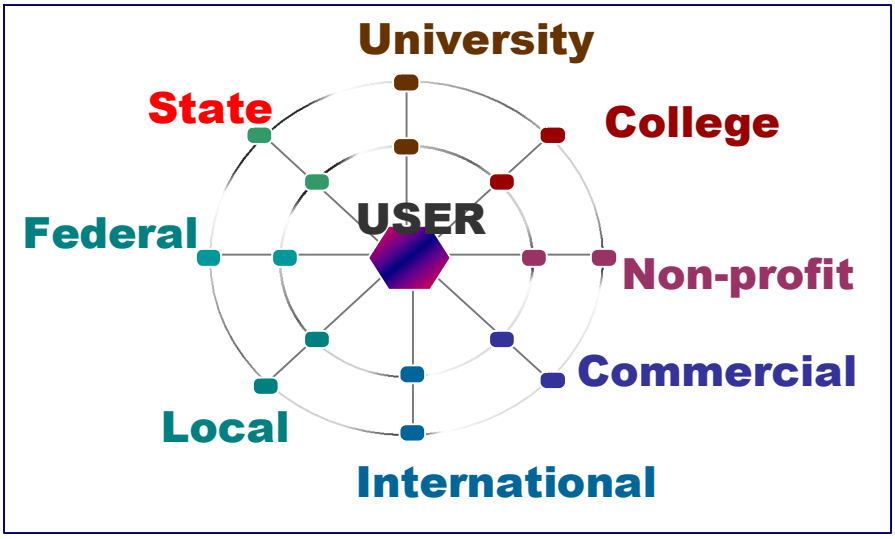

Figure 1. National Digital Data Framework. Source: ARL, 2006.

The increased number and diversity of those concerned with digital preservation—coupled with the current general scarcity of resources for preservation infrastructure—suggests that new collaborative relationships that cross institutional and sector boundaries could provide important and promising ways to deal with the data preservation challenge. These collaborations could potentially help spread the burden of preservation, create economies of scale needed to support it, and mitigate the risks of data loss.

One of the requirements of successful preservation partnership collaborations is that the roles and responsibilities of the partners be made clear. In this paper, we describe a collaboration to develop and deploy Chronopolis™, a model for preservation predicated on data grid infrastructure and replication (Section 2). One of the key components of the Chronopolis™ model is the formalization of the notion of trust between Chronopolis™ participants. In Section 3, we discuss our experience with and thoughts on formalizing trust. We conclude (Section 4) with key challenges for trust relationships that must be addressed to ensure their success.

Section 2: The Chronopolis™ Pilot

The Chronopolis™ model (Moore et. al, 2005) describes a datagrid configured for the purpose of replicated data preservation. The intent of the datagrid is to aggregate participants into a distributed, trusted repository that contains multiple copies of valued data collections and that provides varying degrees of access to those collections at each of the partner sites. In the model, each site can play any or all of several different roles for each collection, and can serve different roles for different collections. The multiple instances of Chronopolis™ collections serve to provide access to the relevant user community and sufficient "backup" copies to protect the data. The Chronopolis™ model is technology-independent and its pilot instantiation seeks to use the best suited and most appropriate software for each component available.

Chronopolis™ participant roles are evolving as follows:

- "Users" will utilize the Chronopolis™ environment and services for data management and preservation of their collections.
- "Partners" will support the installation of servers (e.g., SRB, DSpace, or Fedora) at their sites, register their collections into Chronopolis™, and use the Chronopolis™ environment to replicate their collections.
- "Providers" will constitute the federated Chronopolis™ environment, and will serve as a Core Center (CC), a Replication Center (RC), or a Deep (Write-Once) Archive (DA), including deploying distributed storage infrastructure at their sites and working as a team to provide research and development infrastructure for preservation tools and services.

Chronopolis™ is currently being piloted by a consortium of partner/providers: the San Diego Supercomputer Center (SDSC), the UC San Diego Libraries (UCSDL), the National Center for Atmospheric Research (NCAR), and the University of Maryland (UMD). The partnership provides an opportunity to explore collaboration across institutional boundaries and to establish expectations for users, partners, and providers.

Section 3: Building Formalized Trust

In order for Chronopolis™ partner institutions to work together with clear expectations of outcomes, generally vague notions of trust must be embodied formally.

Ring and Van der Ven (1994) define trust as confidence in the, "goodwill of others which is produced through interpersonal interactions …dealing with matters of uncertainty," or risk. Maister, Green, and Galford (2001) posit four components of trust:

- Credibility
- Reliability
- Intimacy
- Self-Interest

Of these four components, self-interest is weighted over the others because the authors believe this is where the greatest risk lies in the equation of trust. If partnerships are credible and reliable then there can exist enough intimacy to share information. It is what is done with this shared information that reflects self-interest.

In the case of Chronopolis™, self-interest is shared among partners because many of the collections that are stored in the Chronopolis™ environment are too large for any one institution's infrastructure to store more than one copy. Thus any single institution's important information becomes its partner's important information and vice-versa. The identification of collections to be ingested within the Chronopolis™ grid is, thus, driven both by enlightened self-interest and by an interest in preserving one's partners' digital information.

In the business world, trust is usually enforced by contractual agreement tied to monetary incentives (or penalties, as the case may be). In the higher education domain, trust is more informal and generally the product of personal relationships, rather than formalized agreements. The federated preservation environment, however, demands more: namely,

Terms Used in this Paper

Data-Cyberinfrastructure

- Cyberinfrastructure is defined by the NSF (Atkins, 2007) as "infrastructure based upon distributed computer, information and communication technologies," that enable modern research.

- Data-cyberinfrastructure comprises the data storage, access, discovery, and preservation components of cyberinfrastructure.

Data-cyberinfrastructure offers a host of opportunities for testing theories and proposed toolsets in support of long-term digital preservation depending on the user group being supported. Considering the general domains of science and engineering, social sciences, and cultural heritage institutions, we see varied needs for both infrastructure and curation. It is helpful to look at concrete examples in these disciplinary areas to see the challenges that currently exist.

Memorandum of Understanding (MOU)
"Document that expresses mutual-accord on an issue between two or more parties." (Memorandum of Understanding, 2007)

Service Level Agreement (SLA)
"Contract between a service provider and a customer, it details the nature, quality, and scope of the service to be provided." (Service Level Agreement, 2007)

formalization using policy-based trust mechanisms.

The federated entity that is Chronopolis™ can be described as a virtual organization, in the sense used by Holland and Lockett (1998). Virtual organizations (federations) can exhibit trust both from dispositional (the natural tendency of an individual to trust other people) and situational (dispositional trust combined with structural and situational factors) perspectives.

Viewed through the lens of Holland and Lockett, the Chronopolis™ pilot comprehends both dispositional and situational trust as member organizations have successful prior working relationships built on collaborative data-cyberinfrastructure and high- performance computing projects . Holland and Lockett's trust model, as instantiated in the Chronopolis™ pilot project is shown in Figure 2.

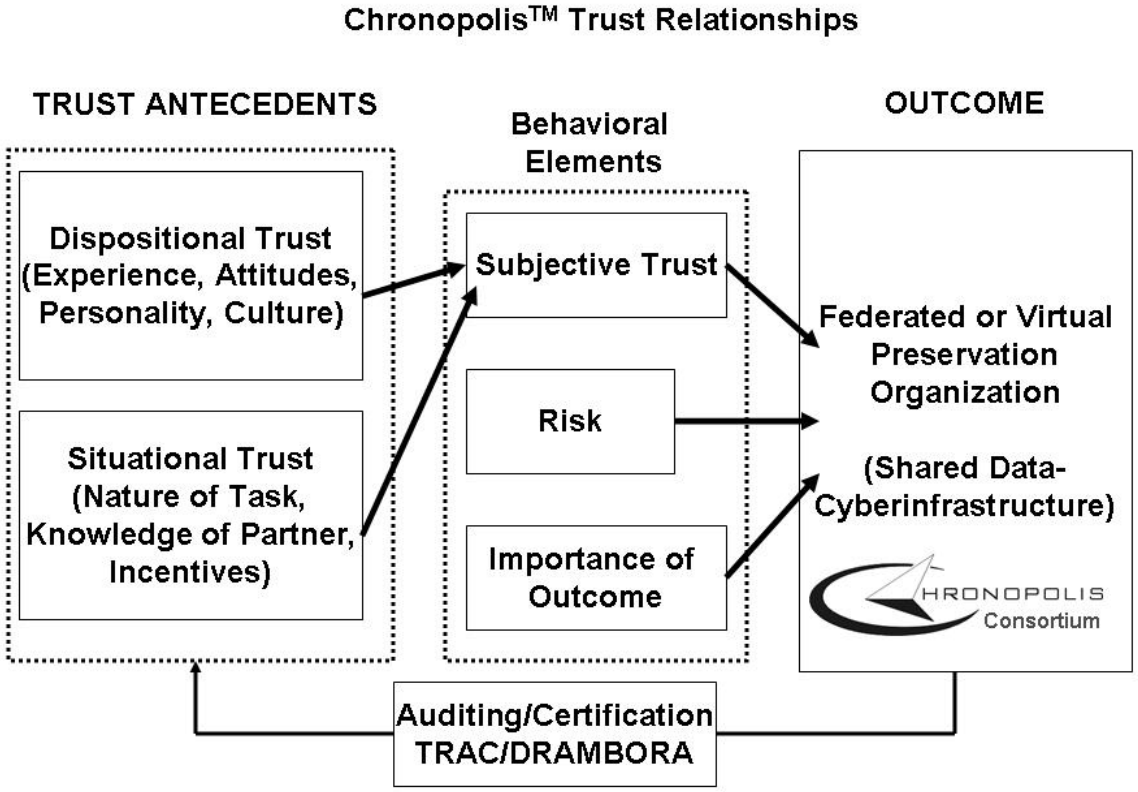

Figure 2. Model of Chronopolis™ Trust: adapted from Holland and Lockett, 2006.

In the Chronopolis™ pilot, each provider/partner institution must have a formal trust relationship with the others. The nature of these relationships depends upon the roles the partners play with respect to one another on the datagrid; if they play multiple roles with respect to one another, they have multiple relationships. General trust relationships amongst the partner institutions are formalized via Memoranda of Understanding (MOUs); service-oriented trust relationships, via Service Level Agreements (SLAs). These types of agreements are useful as vehicles for

"implementing" trust relationships between entities, and as specifications of the expectations and commitments of self-interest and goodwill required to work together closely and successfully.

Formalizing trust can create the foundation upon which certification for "trusted digital repository" status will be built. In both the CRL/OCLC/NARA Trustworthy Repositories Audit and Certification (TRAC) document and the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) Toolkit, sections describe appropriate institutional governance and formalized trust of the parties involved with the sustainability and governance of the trusted digital repository. Chronopolis™ is currently undertaking an audit of pilot participants in order to formalize the credibility and reliability of preservation providers. This will set the stage for formalizing trust among the Chronopolis™ pilot institutions.

An example of Chronopolis™ pilot efforts to formalize trust is the evolving relationship between SDSC and the UCSD Libraries. Although both reside on the same campus (UCSD), they are separate organizational entities. SDSC and UCSDL have developed a joint MOU that describes each institution's intention to fulfill common goals in building the Chronopolis™ pilot preservation environment, and to share the wealth of experience to build mutual grid-based storage architecture and metadata for digital preservation. This agreement is complemented by a more specific SLA to provide for support in running the Libraries' production instance of the Storage Resource Broker (SRB) at SDSC. The SLA outlines the participation of each entity and specifically states the requirements necessary for operating such a production storage environment. This agreement has been in effect since 2003, with options for three year renewals.

Within the Chronopolis™ pilot, separate MOUs have been developed between SDSC and NCAR, and are being drawn up for SDSC and UMD. At this juncture, these "joint" agreements are being extended to cover all of the Chronopolis™ pilot partnerships and will provide the basis for a more permanent set of Chronopolis™ formalized trust agreements.

Section 4: Final Thoughts

It is clear that the successful preservation of our most valuable digital information will necessarily involve groups of partners and providers who can help craft the highly reliable, economically sustainable, and trusted environments needed to house our most valued digital assets. Working across institutional and organizational boundaries is one step toward developing the necessary shared data-cyberinfrastructure. Solidifying this process using formalized trust mechanisms is crucial to long-term sustainability.

The development of the trust relationships necessary to ensure successful data preservation and access is our ultimate objective, and the deployment of structural mechanisms like MOUs and SLAs is a means by which we hope to achieve this goal. In formalizing the trust relationships between preservation providers, partners, and users, many issues are left unresolved.

Key questions include:

- How can accountability be built into trust agreements?
- What are reasonable expectations from providers, partners, and users?
- What vehicles are appropriate for testing trust?
- No system is 100% reliable. What kinds of system failures break trust; what kinds of system failures maintain trust?
- What happens when trust relationships are broken?

Answering these and other questions will prove critical as the community works to ensure preservation of its most critical digital information assets.

Acknowledgements: We are grateful to our colleagues at UCSD and in the Chronopolis™ pilot team for their hard work, useful discussions, and commitment to data preservation.

Section 5: References

Association of Research Libraries (2006). To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering" (Wash., DC: ARL). < http://www.arl.org/info/events/digdatarpt.pdf>.

Atkins, D. E., et. al. (2003). Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure. (Wash., DC: NSF). <http://www.nsf.gov/cise/sci/reports/atkins.pdf>.

DRAMBORA Consortium (2007). DRAMBORA Toolkit. < http://www.repositoryaudit.eu/download/>.

Gantz, J.F. et. al. (2007). "The Expanding Digital Universe: A Forecast of Worldwide Information Growth through 2010" (IDC Whitepaper). <http://www.emc.com/about/destination/digitaluniverse/pdf/Expanding_Digital_Universe_IDC WhitePaper_022507.pdf>.

Holland, C.P. and A.G. Lockett (1998). "Business Trust and the Formation of Virtual Organizations." Proceedings of the Thirty-First Hawaii International Conference on System Sciences, v. 6: 602-10.

Maister, D.H., C.H. Green, and R.M. Galford (2001). The Trusted Advisor. New York: The Free Press.

Memorandum of Understanding (2007). In www.businessdictionary.com, Retrieved April 15, 2007, from http://www.businessdictionary.com/definition/memorandum-of-understanding-MOU.html.

Moore, R.W., F. Berman, B. Schottlaender, A. Rajasekar, D. Middleton, and J. JaJa (2005). "Chronopolis: Federated Digital Preservation Across Time and Space." Proceedings of the IEEE-CS International Symposium on Global Data Interoperability: 171-76. <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1612488>.

OCLC, CRL, and NARA. (2007). Trusted Repositories Audit and Certification: Criteria and Checklist. <http://www.crl.edu/PDF/trac.pdf>.

Ring, P.S. and A. Van de Ven (1994), "Development Processes of Cooperative Interorganizational Relationships," Academy of Management Review, Vol. 19, No. 1: 90-118.

Service Level Agreement (2007). In www.businessdictionary.com, Retrieved April 15, 2007, from http://www.businessdictionary.com/definition/service-level-agreement.html.