Security and Privacy in the Era of Big Data

iRODS, a Technological Solution to the Challenge of Implementing Security and Privacy Policies and Procedures

A RENCI/National Consortium for Data Science WHITE PAPER

renci ncds

CONTACT INFORMATION
Charles Schmitt, PhD
Telephone: 919.445.9696
Email: cschmitt@renci.org

Summary

Today's accessibility to "big data" holds much promise to harness the power of massive data sets and translate that power into advances and transformations in science, medicine, health care, education, and economic growth. Nonetheless, many challenges remain in how best to use these massive data sets while ensuring data security and privacy. These challenges include protection against security breaches and data leakage, vulnerabilities in public databases, and third party data sharing. How to implement security and privacy policies poses a major challenge, particularly when managing large-scale, distributed data sets, whereby data access and use needs to be tracked and monitored in a dynamic, decentralized environment. This White Paper discusses a novel technological solution to this challenge; namely, the integrated Rule-Oriented Data System (iRODS).

The Challenges

Issues related to data security and privacy are of paramount concern in today's era of "big data." Governmental agencies, the health care industry, biomedical researchers, and private businesses invest enormous resources into the collection, aggregation, and sharing of large amounts of personal data. The surveillance programs of the National Security Administration (NSA) represent highly publicized examples (Bamford 2012; 2013; Greenwald 2013; Greenwald & MacAskill 2013a,b). Through recent disclosure, we now know that the NSA routinely collects and analyzes massive amounts of personal data derived from heterogenous data sources such as telecommunications, the Internet, and the user databases of large businesses, including Microsoft, Yahoo, Google, Facebook, PalTalk, YouTube, Skype, AOL, and Apple. The opening of NSA's Utah Data Center (Bamford 2012) will enable massive computing and storage capabilities to analyze inconceivable amounts of data, with the goal of managing data on the order of exaflop by 2018 (10^{18} bytes), zettaflop by 2021 (10^{21} bytes), and yottaflop by 2024 (10^{24} bytes).

The health care industry and biomedical researchers also are tapping into rich data sources. As Electronic Medical Records (EMRs) become more widely adopted (Charles, et al. 2013), biomedical researchers, health care providers and insurers increasingly will use patient data. Further, governmental agencies such as the National Institutes of Health are pushing for the sharing and aggregation of patient EMR data as part of new and ongoing biomedical research endeavors.

Researchers also have begun tapping into non-traditional sources of data such as home, environment, and body-based sensors; records of consumption of food, resources, and services; and social media patterns and usage. Health care providers and insurers likely will do the same. These new data sources and applications will provide yet another vector for security and privacy breaches. While access to patient data requires that receivers obtain approval by relevant regulatory bodies and abide by stringent rules and regulations (e.g., the 1996 Health Insurance Portability and Accountability Act [HIPAA, *U.S. Public Law 104–191, 110 Stat. 1936*], the 1998 Children's Online Privacy Protection Act [COPPA, *U.S. Public Law 105–277, 112 Stat. 2581*], and the 2008 Genetic Information

At a Glance

- Implementing security and privacy policies and procedures is a challenge in the era of big data due to the reliance on oral or written contracts, monitoring by IT staff, and the paucity of technologies to provide automated implementation, tracking, and monitoring of policies and procedures.
- iRODS is a major technological advancement in implementing security and privacy policies and procedures across dynamic, decentralized, distributed data systems. It includes automated processes to implement organizational policies and procedures; the ability to tailor or update policies and procedures; advanced encryption technology; and an integrated data archiving system to enable the long-term tracking and monitoring of data access and use, and to ensure compliance and the maintenance of data provenance.
- The iRODS Consortium will ensure future advancements and sustainability of the iRODS technology. Plans are underway to develop security extensions to the iRODS technology. including plug-ins based on RENCI's Secure Medical Workspace and ExoGENI technologies.

The Team

Charles P. Schmitt, PhD, RENCI CTO and Director of Informatics and Data Science; **Kirk C. Wilhelmsen**, MD, PhD, RENCI Chief Domain Scientist for Genomics and Professor of Genetics and Neurology, UNC-Chapel Hill; Brand Fortner, PhD, Executive Director, iRODS Consortium, RENCI; **Ashok Krishnamurthy**, PhD, RENCI Deputy Director and Professor of Computer Science, UNC-Chapel Hill; **Stanley C. Ahalt**, PhD, RENCI Director and Professor of Computer Science, UNC-Chapel Hill; and **Karamarie Fecho**, PhD, Medical and Scientific Writer for RENCI.

Nondiscrimination Act [GINA, *U.S. Public Law 110–233, 122 Stat. 881*], and Institutional Review Board, www.hhs.gov/ohrp/assurances/index.html), the amount of data that is available today and the extent to which these data are used and shared between entities increase the possibility of security and privacy breaches. Moreover, approval by regulatory bodies does not ensure compliance with policies and procedures related to access and use of patient data. It is rather ironic that in today's technology-oriented world, oral or written pledges remain the most common method to ensure compliance. But as the recent incident with NSA contractor Edward Snowden demonstrates (Greenwald & MacAskill 2013a,b), oral pledges are useless if the motivation to leak data is stronger than the motivation to protect it.

Private businesses also increasingly aggregate large data sets in an attempt to improve their marketing efforts, and third party or "open" market use is becoming more common. These practices provide even greater opportunity for security and privacy breaches, particularly when policies to ensure data security and privacy are outdated or do not exist and incentives to comply with policies are insufficient.

Consider the website SPOKEO (www.spokeo.com), which reportedly aims to connect people. A search by first and last name will (typically) reveal current and former home addresses, a Google satellite map with an arrow pointing to the current home address (labeled as "✔ Found!"), sex, age range, marital status, adult family members in current residence, a Microsoft Bing map of the geographical area of residence with MLS information on the current home and neighboring homes, and (if one pays for a subscription) all sorts of other personal information, including information on education, occupation, salary, photos, friends, etc.—all scrapped and aggregated from publicly available data files.

"We didn't realize that in the digital world, there are a lot of ways to use the digital technology to control us, to snoop on us. In the old days of mailing letters, you licked it, and when you got an envelope that was still sealed, nobody had seen it. You could have private communication. Now they say because it's email, it cannot be private, anyone can listen."

—Steve Wozniak, co-founder of Apple (as quoted in Franceschi-Bichierai, Mashable, 2013.]

Alarmingly, several groups have demonstrated the ability to de-anonymize large data sets using publicly available information. For example, de-identified genomic data sets can be re-identified using public genealogy databases (Gymrek, et al. 2013). "Anonymous" Internet postings have been de-anonymized using texts of known authorship (Narayanan, et al. 2012). Movie ratings have been used to identify the records of ~500,000 subscribers in the "anonymous" public Netflix Prize database and infer information about a subscriber's political and religious affiliations and, in some cases, sexual orientation (Narayanan & Shmatikov 2008). These privacy breaches are significant in that they affect the "forward privacy" of individuals who have been breached; in other words, once your personal information has been linked or aggregated and basic identifying features have been discovered, it is virtually impossible to re-anonymize one's digital self (Narayanan & Shmatikov 2008).

Of importance, the tools to aggregate publicly available data (or other heterogeneous data sources) have been developed to the point where they can be accessed and used by persons with little or no formal training in software programming or computer science. Few (if any) regulations exist to protect consumers from unauthorized use of personal data, and few (if any) incentives exist for private industries to comply with existing privacy and security policies and procedures, which often contain numerous loopholes and poor compliance rates among employees (Nair 2012). Indeed, hackers capitalize on these inherent weaknesses in business practices (Honan 2012a,b; Nair 2012).

Cloud technology is a relatively new technology that has been defined by the National Institute of Standards and Technology (NIST) as "a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" (Mell & Grance 2011). Amazon Web Services was the first cloud technology to be brought into the public marketplace (aws.amazon.com). Introduced in 2006, it quickly became widely used. Since then, numerous other cloud service providers have appeared and are in use across all sectors of society as a more-or-less public utility.

> ***"If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility…The computer utility could become the basis of a new and important industry."***
>
> ***— John McCarthy, MIT Centennial, 1961 (Garfinkel S.L. "Architects of the Information Society, Thirty-Five Years of the Laboratory for Computer Science at MIT," MIT Centennial, 1961, H. Abelson [Ed.]; as cited in Hewlett-Packard Business, White Paper, 2011)***

The popularity of cloud technology is outpacing improvements in the technology. The Cloud Security Alliance (CSA) cites the greatest cloud-related security concerns in 2013 (ranked in order of severity) as: (1) data breaches; (2) data loss/leakage/transfer; (3) account hijacking; (4) insecure Application Programming Interfaces (APIs); (5) denial of service; (6) malicious insiders; (7) abuse of cloud services; (8) insufficient due diligence; and (9) shared technology vulnerabilities (Cloud Security Alliance 2013). These are essentially the same concerns that the CSA identified in 2010, suggesting the need for dramatic improvements in the security of cloud technology. Indeed, Julian Assange of WikiLeaks exploited the vulnerability of cloud technology for data leakage/transfer (www.wikileaks.org). Nonetheless, the use of cloud services has increased exponentially over the last few years, rendering cloud computing of central importance when developing new approaches to digital security and privacy.

As these examples attest, the sharing and aggregation of large-scale data sets across decentralized, distributed data systems provide numerous opportunities for security and privacy breaches, whether accidental or malicious. The risk of breaches increases with third party use of data. The implementation of security and privacy policies and procedures necessitates that data access and use are properly tracked and monitored, in order to ensure compliance and provide an avenue for enforcement; yet, tracking and monitoring of data access and use are extremely challenging when dealing with large-scale, dynamic, distributed data sets.

Existing Solutions to Enforce Security and Privacy Policies Across Dynamic, Distributed Data Systems

As discussed above, ***oral and written pledges*** remain the most common solution to enforce security and privacy policies and procedures; yet, history has shown that this approach is largely flawed. Low-level, but routinely employed, technical solutions to ensure security and privacy when sharing and aggregating data across dynamic, distributed data systems include ***passwords, controlled access,*** and ***two-factor authentication,*** which requires a user to submit two of three authentication factors before gaining access to a resource or service (typically something a user knows [a password], something a user has [a physical bank card], and something a user is [e.g., a biometric characteristic such as a fingerprint]; Federal Financial Institutions Examination Council, http://www.ffiec.gov/pdf/authentication_guidance.pdf). Access permissions such as these can potentially be breached by both the intentional sharing of permissions and the continuation of permissions after they are no longer required or permitted by policy; thus, permission privileges need to be continuously assessed in the absence of automated systems to remove them after they are no longer needed.

More advanced technological solutions include ***cryptography*** and ***encryption.*** Encryption is intended to encode data or information such that access is permitted only to authorized individuals who hold the "key" to unlock the encryption code. The digital application of encryption dates to the 1970s with the introduction of three encryption algorithms: the symmetric cipher, Data Encryption Standard (DES); the asymmetric cipher, RSA (named for developers Ronald Rivest, Adi Shamir, and Leonard Adleman); and the Diffie-Hellman key exchange (Narayanan 2013). The Advanced Encryption Standard (AES) was developed by NIST and introduced in 2001 (FIPS 2001). While AES has been breached, it is expected to remain strong for at least another decade (Bamford 2012). Ironically, although AES has been adopted by NSA for "secret" confidential documents/data (Bamford 2012), the NSA's Utah Data Center and several other federally funded facilities are building the infrastructure for fast, powerful computing capabilities designed to breach any and all encryption schemes, including AES (Bamford 2012;

2013). Indeed, recent revelations indicate that the NSA may have already found ways to breach or circumvent existing Internet encryption schemes (Perlroth, et al. 2013). Once these capabilities are developed, they are bound to become more widely used in the public domain, making AES obsolete. Quantum Computing (e.g., D-Wave Quantum Computer) may one day replace traditional cryptography approaches, but such an approach is purely theoretical at present (Jones & Nature magazine 2013).

Virtual barriers are designed to restrict access to data as it moves across a network or between networks. Firewalls provide virtual barriers and were introduced in the early 1990s (initially as routers with filtering rules) as a means to restrict and authenticate incoming and outgoing network traffic by analyzing data "packets" to determine whether the source of Internet traffic is authorized (Avolio 1999). ***Secure Sockets Layer (SSL) and Transport Layer Security (TLS) technologies*** also were introduced in the 1990s as approaches to authenticate communication across the Internet and reduce eavesdropping by unauthorized third parties; SSL and TLS require both client- and server-side authentication in the form of a "handshake" (Microsoft 2003). Each of these technologies can be breached, however, and thus need to be continuously monitored, with fixes applied as needed. Additionally, firewalls cannot differentiate between data packets containing good intent versus those with malicious intent (virus-infesting), nor can they protect against external connections to a network that bypass the firewall (Avolio 1999).

Tracking/monitoring/auditing software is intended to provide a history of data flow and network access by an individual user in order to ensure compliance with security-related policies and procedures. A major limitation of this technology is that it is difficult and costly to implement on a large scale or with distributed data systems and users because it requires dedicated staff to read and interpret the findings. Another limitation is that the software can be exploited to monitor individual behavior in addition to or instead of protecting data. For example, digital mobility "footprints" are non-random patterns of behavior tracked through mobile devices, GPS, etc. (González, et al. 2008). In the near future, it may very well be harder to hide one's mobility than it is to hide one's identity, and the latter is already extremely difficult.

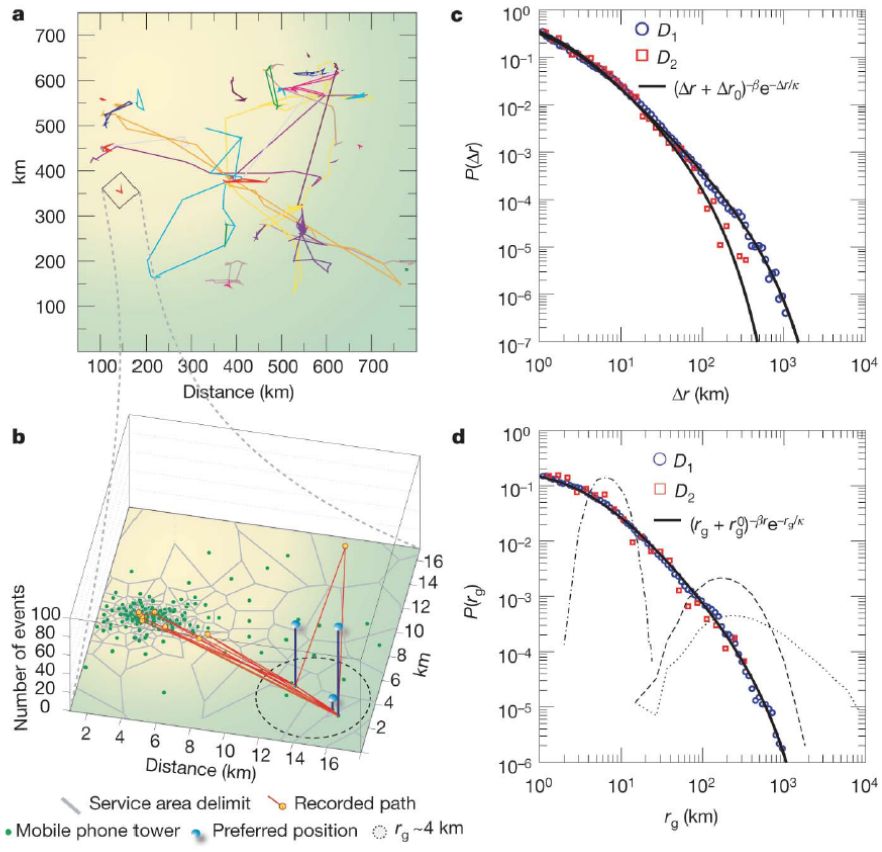

Figure 1. Basic human mobility patterns. a) Spatial trajectory of 40 mobile phone users over the course of a week indicates that most users travel only short distances; b) Detailed trajectory of a single mobile phone user over 186 two-hour reports, with cell towers in green dots and the approximate reception area of each tower in grey marks, indicates that the user visited a total of 12 different locations (cell tower vicinities), with two preferred locations displayed as blue lines (one was visited on 96 occasions; the other on 67 occasions), with the circle indicating the radius of gyration; c) Probability density function, P(Δr), for the travel distances derived from the two studied data sets, D1 and D2, fitted using a truncated power law; d) The distribution of the radius of gyration, P(rg), for the two studied data sets, where rgT was measured after T = 6 months of observation, fitted using a truncated power law. *(From González, et al. Nature, 453, 779-782, 2008. Reprinted with permission from Nature Publishing Group.)*

Ideas Into Action: The integrated Rule-Oriented Data System (iRODS)

iRODS was developed by the Data Intensive Cyber Environments group at UNC-Chapel Hill and the University of California at San Diego and RENCI, with contributions from around the world, as an open source, policy-based solution to the security challenges involved in allowing communities to access, share, publish, preserve, and manage needed data and associated metadata regardless of the physical location of the data or the technology used to store it (Rajasekar, et al. 2010a,b; Liu, et al. 2011). iRODS was architected and designed to address these challenges across a broad spectrum of communities, with differing institutional goals and security and privacy concerns, by providing each adopter community the ability to develop and deploy solutions for data management and sharing that are specific to organizational needs.

Key technological features include: (1) federated data grids or "intelligent clouds" to share data across the iRODS user base, with logical namespaces in place that are managed independently of the storage location, in order to ensure consistent user access despite evolution in Information Technology (IT) storage solutions and locations; (2) a distributed rules engine to automate administrative tasks, enforce management policies, and evaluate data attributes (e.g., zipped files, metadata tags) through system- and user-defined rules and microservices; (3) an "iCAT" metadata catalog to store data and associated metadata in a single database; (4) a storage access layer that allows common access to data stored in traditional local and network attached file systems, object and block storage systems, newer systems such as HDFS and Amazon S3, as well as dynamically instantiated data such as data generated through a SQL query, a web service call, or a Hadoop Map-Reduce job; and (5) a rich combination of graphical user interface (GUI) and command-line–based clients and APIs for interaction with an iRODS data grid.

iRODS is used in a number of data management applications; examples include the provision of a digital library solution to support large-scale publication, an archival environment to support management over the full data lifecycle, and data-oriented workflows. The system is highly scalable; to date iRODS installations have been used to manage hundreds of millions of files, >64 petabytes of data, and >10,000 users. iRODS has been adopted by numerous institutions around the world, including RENCI, UNC-Chapel Hill, and other leaders in data science such as the National Aeronautics and Space Administration, National Science Foundation, National Optical Astronomy Observatory, National Archives and Records Administration, Broad Institute of MIT and Harvard, Welcome Sanger Trust Institute, Bejing Genome Institute, Merck, and the United Kingdom's e-Science Data Grid. Many publications describe the myriad ways in which iRODS technology has been adapted and applied to solve of variety of challenges in policy-based, large-scale data management (e.g., Hedges, et al. 2007; Rajasekar, et al. 2010a,b; Barg, et al. 2011; Chiang, et al. 2011; Schnase, et al. 2011).

The iRODS technology (www.irods.org) provides improvements in common approaches to securing data and ensuring privacy, including:

Comprehensive set of security controls: iRODS supports multiple authentication methods, including its own secure password system, global security infrastructure (GSI), Kerberos, pluggable authentication module (PAM)/LDAP), and/or OS authentication. iRODS also supports role- and group-based policy settings for access to digital objects, similar to most file systems and database management systems. Metadata and rules are stored in an access-controlled database. All operations that deal with data stored in an iRODS data grid invoke policy enforcement points to allow the data grid to tailor security requirements. Policy enforcement points are also used to provide audit trails on all instances of data access. Finally, iRODS provides support for SSL and TLS.

Improved control of data access and use through metadata: Most file systems are limited in their ability to control access and use of large data collections because policies are based on files and directories, which means the ability to keep policies current as users join and leave projects and as files are added and moved requires tremendous effort and resources and often leads to situations where access policies become outdated over time. iRODS provides the ability to use metadata to control access and use policies, thus allowing policies to be enforced despite changes in the location of the data and/or user base.

Storage virtualization and data security lifecycle: The security requirements around data used by a community typically evolve and become more stringent as awareness and use of a data collection grows and

as increasingly diverse data are added to a collection. Enforcement of evolving policies, such as moving data to a more secured physical environment or adopting encryption technologies, disrupts access to data and leads to lost productivity and financial costs—factors which may lead to avoidance or delay in important security upgrades. The iRODS middleware, however, allows data to be moved to more secure systems and permits new security policies to be enacted with minimal or no impact on users.

Persistent identifiers: iRODS supports the use of persistent identifiers for data and the integration of data management operations with persistent identifiers. Examples include: Globally Unique Identifiers (GUIDs), which provide a persistent identifier but do not provide information about location or access controls; Handles or Object Identifiers, which provide a persistent identifier and associated location but not access controls; and Tickets, which provide a persistent identifier, location, and access controls and can be restricted by specific time period, number of accesses, or amount of data. Policy-encoded objects can be sent to a remote site, with subsequent access triggering verification that the environment is authorized and compliant and the encoded policies have been deployed. The encryption of a policy-encoded object enables that object to be sent anywhere, while preserving the enforcement of policies to prevent arbitrary access and manipulation.

The international iRODS Consortium has recently been established to provide the policies and technologies to ensure sustainability of iRODS and safeguard ongoing improvements to the open source iRODS technology such as those discussed above (RENCI Press Release 2012; Brieger 2013). Current members of the iRODS Consortium include RENCI, DICE, and the Max Planck Institute.

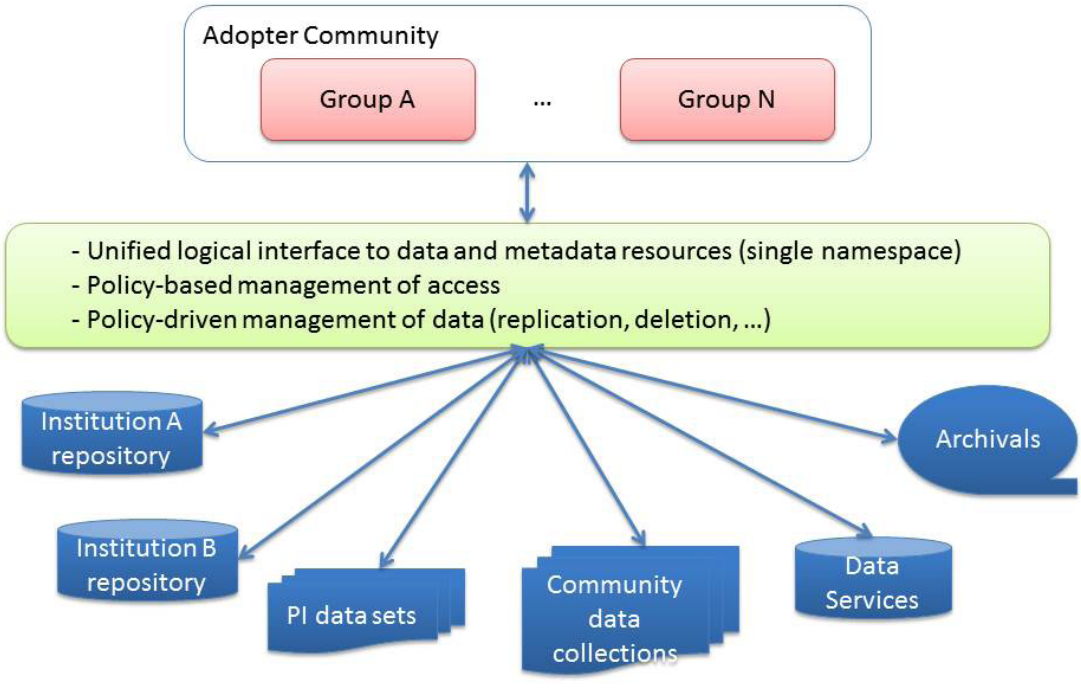

Figure 2. Overview of the iRODS technology. (Image courtesy of RENCI)

The Upshot

The iRODS technology addresses many of the current security and privacy challenges involved with the implementation of security and privacy policies and procedures across dynamic, decentralized, distributed data systems:

1. Security and privacy policies are automatically implemented through technological solutions rather than reliance on monitoring by IT staff.
2. Security and privacy policies can be updated as needed and tailored to organizational needs.
3. Advanced encryption technology is employed to secure access to data and prevent tampering with embedded policies.
4. Breaches trigger automatic "red flags" and block access to data.
5. The technology minimizes the risk of unintentional or malicious third party use of data.
6. The technology includes a data archiving system that enables long-term tracking and monitoring of data access and use and ensures data provenance.
7. The technology is scalable and can be implemented in virtually any environment.
8. The technology can be adapted for a variety of uses.
9. The technology is open source and thus available for use by any group, regardless of budgetary concerns.
10. Advances in the technology will be supported by the iRODS Consortium, thereby ensuring long-term sustainability.

The Big Picture

The iRODS technology represents a novel and innovative technological solution to the challenge of implementing security and privacy policies and procedures as large-scale data are shared and aggregated across decentralized, distributed data systems. The technology will certainly advance as new challenges and vulnerabilities in security and privacy are recognized, and the newly established iRODS Consortium will ensure improvements and sustainability of the iRODS technology.

Of significance, the iRODS technology has been and can be adopted by governmental agencies, the health care industry, biomedical researchers, private businesses, and any other type of organization with a need to protect data in a dynamic, decentralized, distributed data system. Widespread adoption of iRODS is anticipated, along with future integration of iRODS with other technologies that address data security and privacy, such as RENCI's Secure Medical Workspace and ExoGENI technologies.

Acknowledgments

Karen Green provided editorial and design support for the preparation of this white paper, and Tasha Wilhelmsen provided proofreading and assistance with the references. RENCI provided funding for this support.

About RENCI

RENCI, an institute of UNC Chapel Hill, develops and deploys advanced technologies to enable research discoveries and practical innovations. RENCI partners with scientists, policy makers, and industry to engage and solve the problems that affect North Carolina, the U.S., and the world. RENCI is a collaboration involving UNC Chapel Hill, Duke University and North Carolina State University. For more information, see www.renci.org.

Vol. 1, No. 3 in the RENCI White Paper Series, November 2013. Created in collaboration with the **National Consortium for Data Science** (www.data2discovery.org).

About the NCDS

The National Consortium for Data Science (NCDS) is a public-private partnership that offers a foundation for advancing data science research, educating the next generation of data scientists, and translating data innovations into economic opportunity. The NCDS formed in 2012 in the Research Triangle area of North Carolina. For more information, see www.data2discovery.org

How to reference this paper:
Schmitt, C., Wilhelmsen, K., Krishnamurthy, A., Ahalt, S. & Fecho, K. (2013): Security and Privacy in the Era of Big Data: iRODS, a Technological Solution to the Challenge of Implementing Security and Privacy Policies and Procedures. RENCI, University of North Carolina at Chapel Hill. http://dx.doi.org/10.7921/G0H41PBR

References

Avolio, F. (1999). Firewalls and Internet security. *The Internet Protocol Journal*, 2 (2). http://www.cisco.com/web/about/ac123/ac147/ac174/ac200/about_cisco_ipj_archive_article09186a00800c85ae.html. [Accessed July 8, 2013]

Bamford, J. (2013). Five myths about the National Security Agency. *The Washington Post*. http://articles.washingtonpost.com/2013-06-21/opinions/40114085_1_national-security-agency-foreign-intelligence-surveillance-court-guardian. [Accessed June 25, 2013]

Bamford, J. (2012). The NSA is building the country's biggest spy center (watch what you say). *WIRED*. http://www.wired.com/threatlevel/2012/03/ff_nsadatacenter. [Accessed June 25, 2013]

Barg, I., Scott, D., & Timmermann, E. (2011). NOAO E2E integrated data cache initiative using iRODS. In I. N. Evans, A. Accomazzi, D. J. Mink, & A. H. Rots (Eds.), Astronomical Data Analysis Software and Systems XX. ASP Conference Proceedings, Vol. 442, pp. 497-500. San Francisco, CA, USA: Astronomical Society of the Pacific.

Brieger, L. (2013). iRODS Technologies at UNC. E-iRODS: Enterprise iRODS at RENCI. Presented at SC12: Supercomputing 2012, Salt Lake City, UT, USA. http://eirods.org/dev/wp-content/uploads/2012/04/e-irods-slides-sc-bof.pdf. [Accessed August 6, 2013]

Charles, D., King, J., Patel, V., & Furukawa, M. F. (2013) Adoption of Electronic Health Record systems among U.S. non-federal acute care hospitals: 2008-2012. *ONC Data Brief No. 9.* Washington, DC, USA: Office of the National Coordinator for Health Information Technology. http://www.healthit.gov/sites/default/files/oncdatabrief9final.pdf. [Accessed July 15, 2013]

Chiang, G.T., Clapham, P., Qi, G., Sale, K., & Coates, G. (2011). Implementing a genomic data management system using iRODS in the Wellcome Trust Sanger Institute. *BMC Bioinformatics*, 12 12 (1), 361. http://www.biomedcentral.com/1471-2105/12/361. [Accessed July 31, 2013]

Cloud Security Alliance. (2010). Top threats to cloud computing v1.0. https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf. [Accessed July 3, 2013]

Cloud Security Alliance, Top Threats Working Group. (2013). The notorious nine. Cloud computing top threats in 2013. https://downloads.cloudsecurityalliance.org/initiatives/top_threats/The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf. [Accessed July 3, 2013]

Federal Information Processing Standards. (2001). Announcing the ADVANCED ENCRYPTION STANDARD (AES). Publication 197. Washington, DC, USA: National Institute of Standard and Technology. http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf. [Accessed July 18, 2013]

Franceschi-Bicchierai, L. (2013). Steve Wozniak on NSA snooping: 'I feel a little guilty.' *Mashable*. http://mashable.com/2013/06/21/steve-wozniak-on-nsa. [Accessed August 2, 2013]

González, M. C., Hidalgo, C. A., & Barabási, A. L. (2008). Understanding individual human mobility patterns. *Nature*. 453, 779-782.

Greenwald, G. (2013). XKeyscore: NSA tool collects 'nearly everything a user does on the Internet.' *The Guardian*. http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data. [Accessed July 31, 2013]

Greenwald, G., & MacAskill, E. (2013a). NSA Prism program taps in to user data of Apple, Google and others. *The Guardian*. http://www.theguardian.com/world/2013/jun/06/us-tech-giants-nsa-data. [Accessed July 3, 2013]

Greenwald, G., MacAskill, E., & Poitras, L. (2013b). Edward Snowden: The whistleblower behind the NSA surveillance revelations. The 29-year-old source behind the biggest intelligence leak in the NSA's history explains his motives, his uncertain future and why he never intended on hiding in the shadows. *The Guardian*. http://www.theguardian.com/world/2013/jun/09/edward-snowden-nsa-whistleblower-surveillance. [Accessed August 7, 2013]

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, 339 (6117), 321-324.

Hedges, M., Hasan, A., & Blanke, T. (2007). Management and preservation of research data with iRODS. Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience, pp. 17-22. New York, NY, USA: ACM.

Hewlett-Packard. (2011). Five myths of cloud computing. Business white paper. http://www.hp.com/hpinfo/newsroom/press_kits/2011/HPDiscover2011/DISCOVER_5_Myths_of_Cloud_Computing.pdf. [Accessed July 3, 2013]

Honan, M. (2012a). How Apple and Amazon security flaws led to my epic hacking. *WIRED*. http://www.wired.com/gadgetlab/2012/08/apple-amazon-mat-honan-hacking. [Accessed July 8, 2013]

Honan, M. (2012b). Mat Honan: How I resurrected my digital life after an epic hacking. *WIRED*. http://www.wired.com/gadgetlab/2012/08/mat-honan-data-recovery. [Accessed July 8, 2012]

Jones, N., & *Nature* magazine. (2013). D-Wave's quantum computer courts controversy. *Scientific American*. http://www.scientificamerican.com/article.cfm?id=d-waves-quantum-computer-courts-controversy. [Accessed June 25, 2013]

Liu, S., Schulze, J. P., Herr, L., Weekley, J. D., Zhu, B., VanOsdol, N., Plepys, D.M., & Wan, M. (2011). CineGrid Exchange: A workflow-based peta-scale distributed storage platform on a high-speed network. *Future Generation Computer Systems*, 27 (7), 966-976.

Mell, P., Grance, T.; on behalf of the National Institute of Standards and Technology, U.S. Department of Commerce. (2011). The NIST definition of cloud computing: Recommendations of the National Institute of Standards and Technology. SP 800-145. Gaithersburg, MD, USA: National Institute of Standards and Technology. http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf. [Accessed July 31, 2013]

Microsoft. (2003). Introduction (SSL/TLS in Windows Server 2003). http://technet.microsoft.com/en-us/library/cc757054%28v=ws.10%29.aspx. [Accessed July 18, 2013]

Nair, L. V. J. (2012). How you are helping hackers steal your data. *hongkiat.com*. http://www.hongkiat.com/blog/keeping-online-data-safe. [Accessed July 8, 2013]

Narayanan, A. (2013). What happened to the crypto dream?, Part 1. Security & Privacy, *IEEE*, 11 (2), 75-76. http://randomwalker.info/publications/crypto-dream-part1.pdf. [Accessed July 30, 2013]

Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., & Song, D. (2012). On the feasibility of internet-scale author identification. SP-12 Proceedings of the 2012 IEEE Symposium on Security and Privacy, pp. 300-314. Washington, DC, USA: IEEE Computer Society. http://www.cs.berkeley.edu/~dawnsong/papers/2012%20On%20the%20Feasibility%20of%20Internet-Scale%20Author%20Identification.pdf. [Accessed July 22, 2013]

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. SP-08 Proceedings of the 2008 IEEE Symposium on Security and Privacy, pp. 111-125. Washington, DC: IEEE Computer Society. http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf. [Accessed July 22, 2013]

Perlroth, N., Larson, J., & Shane, S. (2013). N.S.A. able to foil basic safeguards of privacy on web. *The New York Times*. http://www.nytimes.com/2013/09/06/us/nsa-foils-much-internet-encryption.html?pagewanted=all&_r=0. [Accessed September 9, 2013]

Rajasekar, A., Moore, R., Hou, C. Y., Lee, C. A., Marciano, R., De Torcy, A., Wan, M., Schroeder, W., Chen, S. Y., Gilbert, L., Tooby, P., & Zhu, B. (2010a). iRODS Primer: Integrated Rule-Oriented Data System. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2 (1), 1-143.

Rajasekar, A., Moore, R., Wan, M., Schroeder, W., & Hasan, A. (2010b). Applying rules as policies for large-scale data sharing. Proceedings of the UKSim/AM SS First International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 322-327. Washington, DC, USA: IEEE Computer Society Washington.

RENCI Press Release. (2012). RENCI Announces E-iRODS Consortium at SC12. *HPC Wire*. http://www.hpcwire.com/hpcwire/2012-11-05/renci_announces_e-irods_consortium_at_sc12.html. [Accessed July 22, 2013]

Schnase, J. L., Webster, W. P., Parnell, L. A., & Duffy, D. Q. (2011). The NASA Center for Climate Simulation Data Management System. 2011 IEEE 27th Symposium on Mass Storage Systems and Technologies MSST, pp. 1-6. Washington, DC, USA: IEEE Computer Society Washington.

U.S. Law

1996 Health Insurance Portability and Accountability Act (HIPAA) *(U.S. Public Law 104–191, 110 Stat. 1936)*

1998 Children's Online Privacy Protection Act (COPPA) *(U.S. Public Law 105–277, 112 Stat. 2581)*

2008 Genetic Information Nondiscrimination Act (GINA) *(U.S. Public Law 110–233, 122 Stat. 881)*

Websites

Federal Financial Institutions Examination Council, www.ffiec.gov/pdf/authentication_guidance.pdf.

Institutional Review Board, www.hhs.gov/ohrp/assurances/index.html.

National Consortium for Data Science, www.data2discovery.org.

SPOKEO, www.spokeo.com.

TRUSTe, www.truste.com.

WikiLeaks, www.wikileaks.org.

renci RESEARCH \ ENGAGEMENT \ INNOVATION ncds THE NATIONAL CONSORTIUM for DATA SCIENCE THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL