xDCI - Accelerating Data Cyberinfrastructure and Research for Community Science Gateways

Ray Idaszak[1,3], Stan Ahalt[1,3], Kira Bradford[1,3], Chris Calloway[1,3], Claris Castillo[1,3], Jason Coposky[1,3], Jon Crabtree[2,3], Sarah Davis[1,3], Yue Guo[1,3], Fan Jiang[1,3], Ashok Krishnamurthy[1,2,3], Howard Lander[1,3], W. Christopher Lenhardt[1,3], Arcot Rajasekar[1,3], Kimberly Robasky[1,3], Terrell Russell[1,3], Erik Scott[1,3], Don Sizemore[2,3], Michael Stealey[1,3], Hao Xu[1,3], Hong Yi[1,3], Wenzhao Zhang[4]

[1]Renaissance Computing Institute (RENCI)

[2]The Odum Institute for Research in Social Science

[3]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27517

[4]North Carolina State University, Raleigh, North Carolina, 27695

Email: (rayi, ahalt, kcbradford, cbc, claris, jasonc, sdavis, yueguo, dcvan, ashok, howard, clenhardt, sekar, krobasky, tgr, escott, stealey, xuhao, hongyi)@renci.org, jonathan_crabtree@unc.edu, dls@email.unc.edu, wzhang27@ncsu.edu

Abstract: *This paper discusses xDCI, a Data Cyberinfrastructure environment that accelerates deployment of Science Gateways. Recognizing the growing importance of Science Gateways, xDCI builds on their elements in making it efficient for individuals and organizations to launch and sustain customizable Science Gateways while growing their respective communities and accelerating resultant science.*

1. Introduction

Science Gateways have proven to be useful to scientific and engineering communities wanting to share data and software, share and run workflows, and generally share and make other resources available [1]. Science Gateways are growing ever popular, firmly establishing them as both useful and necessary for communities to accelerate their respective research. However, establishing a Science Gateway can be time consuming and expensive, often requiring significant funding and personnel resources.

Herein we introduce xDCI as an environment for the efficient and rapid deployment of a Science Gateway that can be readily aligned and branded to a specific science or engineering community. xDCI is an acronym for cross-disciplinary data cyberinfrastructure. xDCI is positioned as a completely customizable solution for scientific communities that simplifies the process of creating and using cyberinfrastructure to support data science [2]. For xDCI herein, we discuss elements that are important to the success of Science Gateways. We then discuss how xDCI positions elements to implement and sustain community Science Gateway cyberinfrastructure, while concomitantly growing the community and accelerating its science and engineering research. We conclude with a recommendation of an xDCI-like approach to pursue a new Science Gateway for one's community.

Proceedings of Gateways 2017, University of Michigan, Ann Arbor, MI; October 23-25, 2017. https://gateways2017.figshare.com/.

2. Elements of Science Gateways

In evolving xDCI, we observed elements common to the success of Science Gateways. xDCI builds on these success elements by adding the additional element of rapid deployment of a Science Gateway that is branded and functionally aligned to a specific community. By rapid deployment, ideally we endeavor to have a one-click deployment of a ready-to-use Science Gateway deployed on, e.g., departmental or cloud virtual machines and based on contents of prepopulated configuration files that specify branding and functionality unique to a given community. At the time of this writing, xDCI is not a one-click deployable Science Gateway, but this stated goal conveys our efforts to continually reduce the effort and expertise expended to deploy a customized Science Gateway.

Following in List 1 are elements of Science Gateways that we highlight in the positioning of xDCI in the community of science and engineering communities as a rapid enabler of Science Gateways. The list, in no particular order, is meant to be representative of common themes that we

have identified and is not intended to be an exhaustive list of elements that comprise successful Science Gateways:

(a) Web portal presence with Graphical User Interface (GUI)
(b) Data sharing supporting FAIR [3] and FAIR-TLC [4] data principles
(c) Computational model and workflow sharing
(d) Scientific reproducibility
(e) Ability to integrate with and operate on national cyberinfrastructure, grid, and public and private cloud at scale
(f) Support for {X}-as-a-Service including Software, Infrastructure, Platform, etc.
(g) Versioning of artifacts, assets with provenance
(h) Comments and ratings of artifacts, assets
(i) Visualization and analysis of artifacts, assets
(j) Virtualization and containerization of artifacts, assets
(k) Citation and persistence of artifacts, assets
(l) Distributed data management supporting federation
(m) Authentication and authorization including support for federated identity and single sign-on; related access control of facilities, assets, networks, and cloud resources
(n) Support for multiple groups or virtual organizations therein
(o) Extensible data model
(p) Support for the full data lifecycle
(q) Support for metadata including existing and extensible user-defined metadata
(r) Search and discovery
(s) Use of open source and standards
(t) Sustainable software practices
(u) Sustainable community practices

List 1. Elements of successful Science Gateways.

What is important to emphasize about the contents of List 1 in the context of xDCI is that xDCI enables rapid customization and deployment of those elements of the list unique to a particular science or engineering community's needs. In other words, not all Science Gateways support all of the elements identified in the list, nor do they need to. Rather, what is important is that the xDCI environment collectively support these elements and then enable a subset of those elements to be uniquely and rapidly customizable and deployable for a given science or engineering community.

3. Constituting xDCI

In constituting a technology stack for xDCI that services elements of successful Science Gateways identified in List 1, we took the approach that individual xDCI technology stack components will service multiple elements, and that there is some overlap and redundancy. The reason for expecting overlap and redundancy is that we chose pre-existing components initially designed and developed for functionality other than the stated intent of xDCI, and thus some Science Gateway elements overlapping that collective functionality are addressed redundantly by multiple xDCI components.

The chosen xDCI technology stack components, shown in Figure 1, necessarily bias those funded through or pursued by our host organization for a straightforward reason: they are components that our staff has most experience with and–as a consequence of their external funding–that external stakeholders have deemed to be of value to a broader community. We welcome the reader to use the specific xDCI technology stack components we have selected. However, other groups can populate their respective technology stack with components from within their respective organizations that their staff may have most experience with and that still align with elements of successful Science Gateways, complemented with xDCI-derived sustainability practices.

3.1 HydroShare and xDCIShare

An important consideration in constituting the xDCI technology stack is the initial and ongoing sustainability of its elements. In this sense, List 1 elements (s,t,u) are important taken together. To demonstrate this, we discuss our selection of HydroShare [5] to support the xDCI web portal with graphical user interface (a) while also supporting all the rest of the List 1 elements (b through r). HydroShare is a hydrology community

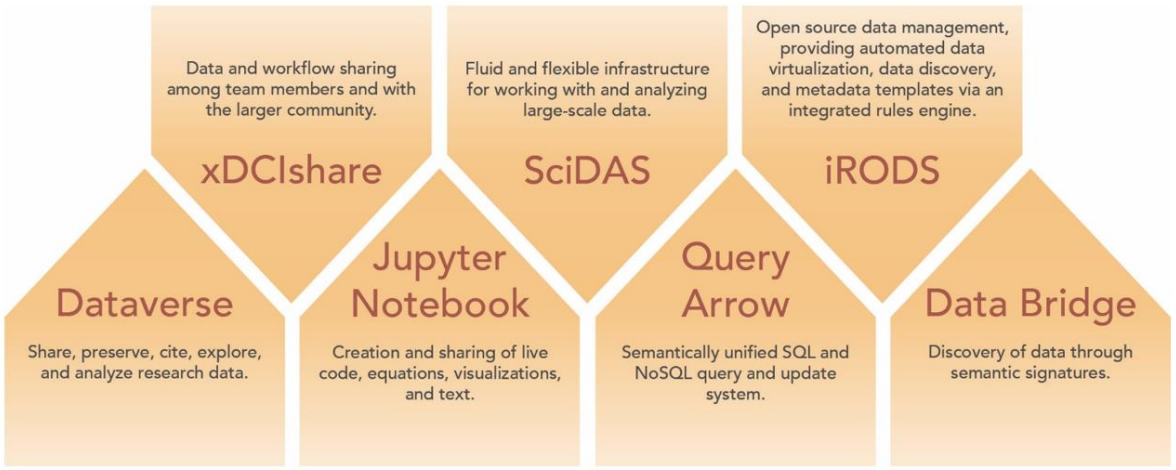

Fig. 1. Components of the xDCI technology stack.

open-source cyberinfrastructure project supported by the National Science Foundation (NSF) through its Software Infrastructure for Sustained Innovation program (SI2) [6] and currently has over 1,400 registered users. Recognizing in part how well HydroShare supports the elements of a Science Gateway for its hydrology community, our team chose to generalize the open source HydroShare codebase to apply it to other science and engineering communities. Contributing to this decision is our team's funded participation on and direct experience with the codebase since the HydroShare's project onset. We have named this derivative codebase xDCIShare and are endeavoring to keep it synchronized with the main HydroShare codebase, made possible by the HydroShare team's use of advanced software engineering [7]. Eight teams distributed across the United States work together to produce a new release of the approximately half-million line codebase of HydroShare every 3-5 weeks, supported by continuous integration, formal testing, and code review. For example, a foundation-funded gateway for advanced health care directives uses a derivative xDCIShare codebase, demonstrating the span of diverse domains that can now be addressed. Furthermore, derivative projects tend to have their own development teams. Thus, innovations across the synchronized project codebases will amplify as the number of derivative projects grow, and sustainability will strengthen. Sustainable software practices promoted by forward-thinking programs like the NSF SI2 program make this possible, noting also that the NSF SI2 program is also the primary funding for the overarching NSF Science Gateways Community Institute (SGCI) [8].

3.2 SciDAS

The scientific discovery performed via Science Gateways can be dependent on big datasets that require computing at scale. Thus, our team chose the NSF-funded SciDAS project [9] for inclusion as part of the xDCI technology stack to provide a fluid and flexible infrastructure for working with and analyzing large-scale data. SciDAS, or Scientific Data Analysis at Scale, federates access to multiple cyberinfrastructure resources including NSF Cloud, Open Science Grid, XSEDE v2.0 and campus resources as well as commercial cloud resources. SciDAS relies on the integrated-Rule-Oriented-Data-System (see Section 3.4), enhanced with software-defined-networking (SDN) capabilities, to support network-aware data management decisions and efficient use of network resources. The distributed and scalable nature of both data-sharing and compute infrastructure are exploited to optimize for computer and data locality boosting the performance of workflows and scientific productivity.

3.3 Dataverse

Another component of the xDCI technology stack that addresses all the Science Gateway elements in List 1 is Dataverse [10]. Dataverse is an open source web application that enables its users to share, preserve, cite, explore, and analyze

research data. Generally, each xDCIShare instance is deployed and aligned to a given community, but Dataverse offers the notion of a single Dataverse repository capable of hosting multiple Dataverses therein, with each of these Dataverses aligned to a given community. While it does take some time to instantiate the Dataverse repository, subsequent creation of a Dataverse therein is a very straightforward and simple operation accomplished by a member of that community. In this respect, Dataverse achieves our xDCI objective of rapid deployment, approaching the one-click deployment of a ready-to-use Science Gateway objective discussed in Section 2. Dataverse is also formally supported across our team's University campus [11], thus also addressing our objective discussed in Section 3 of using Science Gateway elements that our team is familiar with. Having multiple xDCI offerings that span the Science Gateway elements in List 1 helps xDCI address a broad range of stakeholder communities.

3.4 iRODS

Our team selected the integrated Rule-Oriented Data System (iRODS) [12] for xDCI as middleware, comprising a data grid that is used to organize distributed data into a shareable collection with associated metadata, while enforcing management policies across distributed storage systems. The iRODS data grid supports registration of files from existing storage systems into a logical name space and access to the remote files. Thus, an iRODS data grid can be used to build a logical collection that spans multiple existing systems without modification to the existing infrastructure. The iRODS data grid also supports a metadata framework for associating additional information with each data file or collection. Application-specific metadata (such as latitude-longitude coordinates), provenance metadata (such as parameters used to derive a dataset), and referential metadata (such as semantic relationships among data files) can be maintained by the iRODS metadata service. The iRODS data grid implements policies as computer actionable rules that control the execution of procedures at each storage location. Rules can be created that enforce a specific preservation policy, that automate an administrative function, or that validate an assessment criterion.

3.5 Specialized xDCI Components

Rounding out our xDCI technology stack are components that address specialized needs. These include QueryArrow [13] that offers a system of semantically unified SQL and NoSQL query and update; Data Bridge [14] that enables discovery of data through semantic signatures; and Jupyter Notebooks [15] that offer creation and sharing of live code, equations, visualizations, and text. Specialized components enable specialized functionality that is often required by Science Gateway communities. Our selection of these particular components for xDCI fits recurrent needs we have observed and exploits internal expertise with these components, as discussed in the introduction to Section 3 herein.

4. xDCI Concierges

To ensure Science Gateway sustainability and accelerated research, the xDCI approach requires expert hand-holding at various stages of Science Gateway deployment and operation. We have determined success here requires that xDCI staff proactively "own" these stages, unlike a more traditional IT support model, which is reactive to inquiries primarily when there are problems. We have deemed these xDCI staff "concierges" [16]. The first stage is the xDCI Technology Concierge, who works with the Science Gateway community stakeholders to select relevant xDCI components, create an architecture, and rapidly move from concept to implementation. The second stage is the xDCI Software Concierge, who instills sustainable software best practices as promoted by and discussed in NSF SI2 literature [6,7]. The third stage is the xDCI Data Science Concierge, who offers data science expertise on how to extract maximum scientific and community benefit from one's Science Gateway deployment via xDCI. The final stage is the xDCI Sustainability Concierge, who assists in the identification of requirements necessary to support future scenarios, including migration to new cloud infrastructure, business plans for identifying future funding models, and identification of technology resulting from use of xDCI which may feed back into xDCI itself.

5. Conclusion

In constructing a sustainable data science cyberinfrastructure ecosystem employing Science Gateways, there is no "one size fits all." Rather, it is only through sustainable practices that the community of science and engineering communities can concomitantly grow their respective communities and accelerate resulting research. Herein, we have positioned xDCI as an approach that allows these communities to more rapidly deploy Science Gateways that achieve these goals while using sustainable practices. The xDCI approach discussed herein proposes reuse of our thought processes in selecting components for one's technology stack elements in making it efficient for individuals and organizations to launch and sustain customizable Science Gateways while growing their respective communities and accelerating resultant science.

6. Acknowledgments

The authors wish to especially thank Stan Ahalt, Director of RENCI, and Ashok Krishnamurthy, Deputy Director of RENCI, for their initiation of xDCI and continued organizational and financial support of xDCI. This material is additionally based upon work supported by the NSF under awards 1148453, 1148090, 1247602, 1247663, 1247652, 1560625, 1649397, 1659300, 1664018, 1664061, and 1664119; a multitude of entities that funded the development of Dataverse [17]; and a multitude of funded projects that supported the development of iRODS technology [18]. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the additional funders.

7. References

[1] https://www.nsf.gov/news/news_summ.jsp?cntn_id=189347

[2] A. Krishnamurthy, K. Bradford, C. Calloway, C. Castillo, M. Conway, J. Coposky, Y. Guo, R. Idaszak, W.C. Lenhardt, K. Robasky, T. Russell, E. Scott, M. Sliwowski, M. Stealey, K. Urgo, H. Xu, H. Yi, S. Ahalt, "xDCI, a Data Science Cyberinfrastructure for Interdisciplinary Research," 2017 IEEE HPEC, Waltham, MA, September 2017.

[3] https://www.ncbi.nlm.nih.gov/pubmed/26978244

[4] http://dx.doi.org/10.5281/zenodo.203295

[5] D.G. Tarboton, R. Idaszak, J.S. Horsburgh, J. Heard, D. Ames, J.L. Goodall, L. Band, V. Merwade, A. Couch, J. Arrigo, R. Hooper, D. Valentine and D. Maidment, (2014), "HydroShare: Advancing Collaboration through Hydrologic Data and Model Sharing," in D. P. Ames, N. W. T. Quinn and A. E. Rizzoli (eds.), Proceedings of the 7th International Congress on Environmental Modelling and Software, San Diego, California, USA, International Environmental Modelling and Software Society (iEMSs), ISBN: 978-88-9035-744-2.

[6] http://www.nsf.gov/si2/

[7] R. Idaszak, D.G. Tarboton, H. Yi, L. Christopherson, M.J. Stealey, B. Miles P. Dash, A. Couch, C. Spealman, D.P. Ames, J.S. Horsburgh. HydroShare – A case study of the application of modern software engineering to a large distributed federally-funded scientific software development project. In: J. Carver, N.P.C. Hong, and G.K. Thiruvathukal (eds.) Software Engineering for Science, ISBN 9781498743853, Taylor&Francis CRC Press, November 2016.

[8] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1547611

[9] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1659300

[10] https://dataverse.org/

[11] https://dataverse.unc.edu/

[12] https://irods.org/

[13] https://irods.org/uploads/2015/01/xu-queryarrow-2016.pdf

[14] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1649397

[15] http://jupyter.org/

[16] https://www.hpcwire.com/2017/09/06/need-data-science-cyberinfrastructure-check-rencis-xdci-concierge/

[17] https://dataverse.org/about

[18] https://irods.org/history/