xDCI, a Data Science Cyberinfrastructure for Interdisciplinary Research

Ashok Krishnamurthy*, Kira Bradford*, Chris Calloway*, Claris Castillo*, Mike Conway*,
Jason Coposky*, Yue Guo*, Ray Idaszak*, W. Christopher Lenhardt*, Kimberly Robasky*, Terrell Russell*,
Erik Scott*, Marcin Sliwowski*, Michael Stealey*, Kelsey Urgo*, Hao Xu*, Hong Yi*, and Stan Ahalt*
*Renaissance Computing Institute (RENCI)
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27517
Email: (ashok,kcbradford,cbc,claris,mconway,jasonc,yueguo,rayi,clenhardt,
krobasky,tgr,escott,marcin,stealey,kelseyurgo,xuhao,hongyi,ahalt)@renci.org

Abstract—**This paper introduces xDCI, a Data Science Cyberinfrastructure to support research in a number of scientific domains including genomics, environmental science, biomedical and health science, and social science. xDCI leverages open-source software packages such as the integrated Rule Oriented Data System and the CyVerse Discovery Environment to address significant challenges in data storage, sharing, analysis and visualization. We provide three example applications to evaluate xDCI for different domains: analysis of 3D images of mice brains, videos analysis of neonatal resuscitation, and risk analytics. Finally, we conclude with a discussion of potential improvements to xDCI.**

1. Introduction

With the exponential growth of available data and the rapid advances in algorithms and software for data science, there are significant challenges in data storage, sharing, analysis and visualization for many scientific domains.

For instance, in brain imaging research, recent tissue clearing and modern microscopy techniques [1] enable 3D imaging of a mouse brain at high resolution. This imaging method can result in 30 TB of data for each mice brain, which outstrips our current ability to share and analyze the data, and visualize the results.

Another example is a deep-learning based technique to analyze videos of neonatal resuscitation [2]. The required workflow, that includes several steps and GPU based-computing is cumbersome for many domain science researchers such as medical professional who use the videos to evaluate training methods.

To tackle these challenges, it is essential that researchers have the ability to share data, discover and access the data that has been collected at different institutions, test their analysis methods on a diverse collection of data, and publish and share the results with the community. The size and complexity of the data, as well as the computational requirements of the analysis methods makes conventional methods of data sharing required for such “team science” difficult for most research groups.

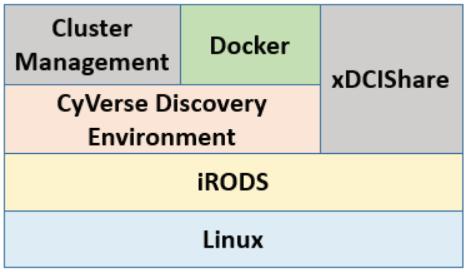


Figure 1. The proposed software stack for xDCI.

We propose a generalized data science cyberinfrastructure called xDCI, to facilitate such interdisciplinary ”big data” research. xDCI is a technology framework designed to run on server class computers. A basic xDCI software stack is shown in Figure 1. The major components of xDCI are:

- *The integrated Rule Oriented Data System (iRODS)*: iRODS [3] is an open source data management software and is used to manage millions of files and tens of petabytes of data worldwide. iRODS rules and microservices provide a number of critical functions for xDCI, namely, owner-specified access control for individual data sets, automatic migration of data between computer systems, automatic backup and replication of data sets, searchable metadata to locate and discover data sets, and creation of workflows for data analysis.
- *The CyVerse Discovery Environment (DE)*: The DE [4] provides a service layer linked to a modern web interface for powerful computing, data, and analysis application resources for scientific analyses. The DE has an Application Programming Interface that provides a way to plug-in new analytical capabilities and can facilitate the execution of resources intensive computations on remote resources. xDCI will use the DE both as a web interface and to facilitate external analytical capabilities.

978-1-5386-3472-1/17/$31.00 ©2017 IEEE

- *xDCIShare*: xDCIShare is derived from HydroShare [5], [6], an online collaborative system to support the open sharing of hydrologic data, analytical tools, and computer models. With HydroShare, scientists can easily discover, access, and analyze hydrologic data and thereby enhance the production and reproducibility of hydrologic scientific results. We are refactoring the HydroShare codebase to remove aspects that are hydrology specific, with the resulting software code being called xDCIShare. xDCIShare is the basis for the Data Discovery and Data Persistence/Identifier functionality of xDCI by retaining these key capabilities of HydroShare in generalized form.

The rest of the paper is organized as follows: Section 2 will introduce the details of xDCI, Section 3 will discuss the three example applications of xDCI, and Section 4 presents future developments envisioned for xDCI.

2. Method

xDCI is a federated and distributed ”big data” framework as depicted in Figure 3. It will include the functionality listed in the subsections below, with all capabilities made available over a web interface.

2.1. Data Ingestion

xDCI implements a flexible, user-controlled, and rule-driven method for data ingest. In several of the applications, the data originates at an instrument computer and is then moved to a laboratory local grid node that is associated with xDCI. This arrangement puts data immediately under the management of the policy based environment. This pipeline for the data is shown in Figure 2 below. The xDCI data ingest process is designed to provide the scientist who collects the data complete flexibility in controlling when the data is moved to xDCI, who has access to the data while the study is underway, and when the data is published and assigned a Data Object Identifier (DOI) and made available to other researchers. One of the described applications, BRAIN-I (described in Section 3.1) accomplishes this by co-locating an iRODS data grid node in close proximity to the instrument computer, connected to BRAIN-I, as shown in Figure 3. Data is transported to this local node using a smart agent via high speed transfer methods. Using the iRODS Rules Engine (iRODS rules are definitions of actions that are to be performed by the server), the investigator can set iRODS rules which will enable “Set it up and Forget about it” disposition of the data at each stage of the data life cycle. We understand that it may not be possible to install and run an agent at all locations, but in these cases a wide range of command line and standard protocols are supported, such as sftp.

2.1.1. Metadata, Standards and Identifiers. In xDCI, we propose to use a flexible and expandable metadata template to capture metadata during the data ingestion process. The metadata template allows fields that are pre-populated, and allows validation and enforcement of metadata standards by the iRODS Rule Engine. This will be a key point of interaction as the metadata captured during data ingestion can ensure that the data meets the defined standards, thus facilitating the work of both researchers using the data and the analysis and visualization tool builders. The combination of a metadata templating system and the iRODS Rules Engine is an effective means of ensuring that the data being ingested meets the needs of being discoverable and reusable in other studies. In addition, the policy driven data grid can be used to automate metadata extraction from ingested data, including through the invocation of workflows automatically and asynchronously. Metadata can be stored and managed within the grid, and collections can be projected into a range of discovery mechanisms, including ElasticSearch, triple-stores, or relational databases.

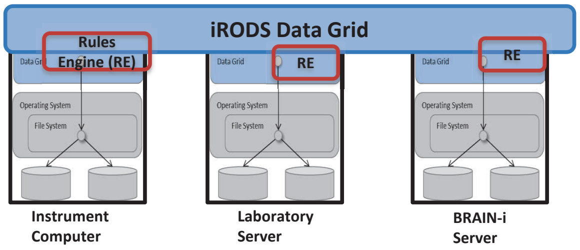


Figure 2. Data ingestion pipeline from instrument computer to the BRAIN-I server.

2.2. Data Accessibility

xDCI data accessibility methods adhere to the following two rules: 1) the researcher that is depositing data in xDCI should have full control over the discovery and accessibility of the data over a reasonable period of time (this period of time forms the embargo time for the data, will be data set specific, and under the control of the data owner); and 2) once the embargo time is over, the data must be discoverable and accessible by other researchers and the public. Note that because of unknown and variable delays in publication, the data owner will be able to adjust the embargo time dynamically; however, xDCI will track these changes and provide reminders to the data owner to set reasonable embargo periods. These data accessibility principles will be enforced using the Rules Engine in iRODS, with activation of the rules being triggered by events in the data life cycle. For example, an initial estimate of the embargo time may be a required parameter when the data is first ingested into xDCI. We expect that the details of the data accessibility policy will be formulated through a community-driven process; the key here is that the xDCI iRODS Rules Engine can be set up to ensure that the community derived rules are effectively enforced.

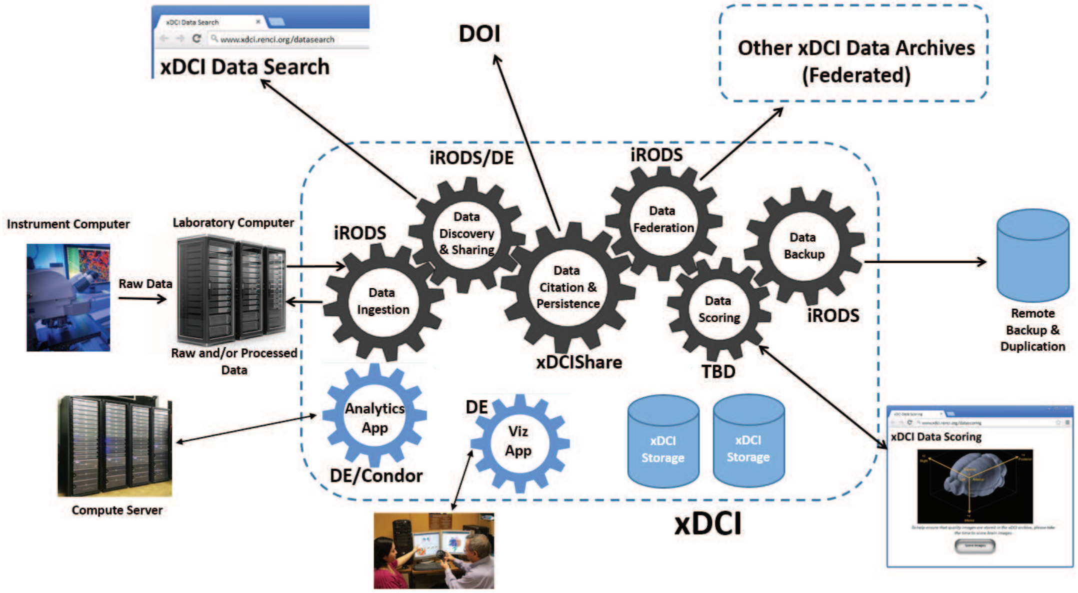


Figure 3. This figure shows the functions that will be provided by xDCI and the software components that are involved in providing the functionality using the Brain-I application described in Section 3.1 as an xDCI exemplar.

2.3. Data Discovery and Sharing

Data Sharing in xDCI can follow two paths. The first path is that the owner of the data invites another user or group of users to access and use the data. Since the data owner controls access during the data embargo period, this is easily done. The second path to data sharing occurs once the data is published, thus opening it up for general use. Data sharing after publication will require the ability of other users and scientists to discover the data based on different search criteria. The key to creating a successful data discovery system is that there is rich metadata that is attached to the data sets. xDCIShare has developed an effective data discovery strategy, and xDCI will implement the data discovery methods used in xDCIShare. The policy and analysis frameworks can be employed to run metadata extraction workflows. The policy framework allows event-driven workflows to maintain near-real-time indexes, allowing rich query and discovery using a range of query languages and discovery methods.

Following the philosophy espoused in xDCIShare, xDCI will define two levels of metadata that can be used for data discovery - system metadata and science metadata. System metadata refers to system information about the data files and other objects that are maintained in xDCI; for example, ownership of a data file, or the date it was last accessed represents system level metadata. iRODS uses a relational database called iCAT to maintain such metadata. We propose to use iCATS for the system metadata for xDCI also. The iCAT server stores metadata in the form of “triples” to its relational database. The triples consist of an attribute field, a value field, and a unit field. The content of each of these fields can be independently defined and applied. Metadata in iCAT can be used in various ways, for example to trigger actions, based on rules defined in the iRODS rule engine. The system metadata can be searched, using for example the iRODS command imeta for simple searches, while more complex queries can be generated using a subset of SQL operations issued through the iquest command.

Science metadata, on the other hand, provides information about the data resource itself. Science metadata, for example, could be an ID number used to identify a specimen, or information about the tissue clearing process for brain images that was used. Science metadata is typically created by the data owner and will be used in xDCI for data discovery. Following the approach of xDCIShare, xDCI will use the standard Dublin Core Metadata element set to describe xDCI data sets [7], with specific Dublin Core Application Profiles for different data types and experimental protocols [8]. We propose to use metadata templates to manage structured metadata, and to use the policy engine to project metadata into a range of discovery mechanisms.

xDCI will allow for the search and discovery of available raw and derived data sets based on both system and science metadata. The searchable interface will be made available through a web interface, through off-the-shelf search technology. In addition, publishing mechanisms will allow the packaging of data sets and metadata into formats suitable for target repositories and allow publishing of disseminated data to external platforms and meta-catalogs.

2.4. Data Citation and Persistence

Once data is published, we need an effective method for citing the data and a persistent identifier that points to the data. xDCI will adopt the Digital Object Identifier (DOI) system for providing an actionable, persistent link to the archived data. The xDCI DOI system will be based on a similar system in xDCIShare.

All data in xDCI, including processed data, will be assigned an internal identifier. Each subsequent version of a data set will be further assigned its own unique internal identifier. The internal identifiers will be used for tracking usage, access patterns of data and for system administration purposes. In xDCIShare, the internal identifiers are also used for emerging social media functionality to enable users to enhance information about and collaboration around xDCIShare resources.

Only when a data set is published is it assigned a DOI. xDCI will maintain a mapping between the internal identifiers and the assigned DOI. New versions of published data set will get their own unique DOI, which ensures that users may cite a particular version of a resource using its DOI, which will be a persistent pointer to the content for that version. With different DOIs for different versions, users can easily cite the correct version they used for their research.

2.5. Data Federation, Replication, and Backup

xDCI provides a means for federation with other research archives. The underlying iRODS data grid provides a mechanism that allows an emergent connected network of collaborating institutions, while allowing each institution to maintain control of their own data according to their own management policies. This ability to expand the nodes to other institutions, or to collaborate across other science domains is a unique, defining quality of the xDCI architecture. Within an organization, the abstraction layer allows a heterogeneous mix of storage, including Lustre, object stores, high-performance disk arrays, hierarchical storage, and cloud storage to be represented in a unified logical name-space as one coherent data repository. xDCI will provide for remote replication and backup of all data in the archive. This will be achieved through iRODS, as data replication, archiving, and backup is one of the most widely prevalent use case for iRODS.

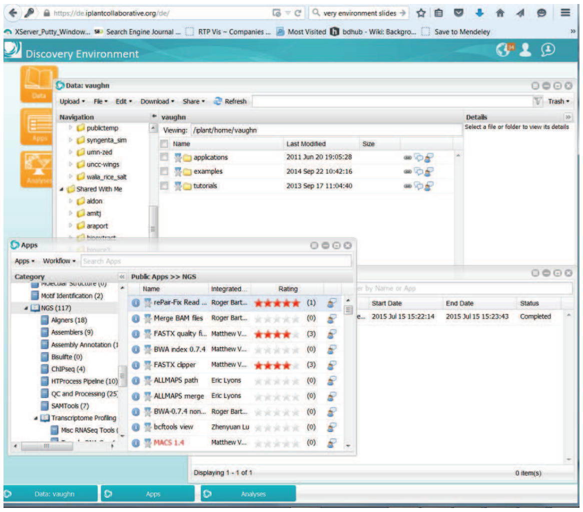
2.6. Data Scoring

There are a number of different criteria that can be used to score data: for example, based on the completeness of the data set, the completeness of the metadata associated with the data, the popularity of the data derived from access patterns, and ratings of the data from users. At another level, data can be scored based on properties of the data itself: for example, number of resolved cell bodies in a given region. Automated analysis codes may be suitable for providing scores based on the image properties. The combination of such criteria to derive a composite score for the data will be a community-led effort in xDCI, since the scientists using the data are best suited to determine the scoring method.

2.7. Analysis and Visualization Tools

The CyVerse Discovery Environment (DE) can be used to create and package workflows as Apps that can be launched from the user's virtual desktop when they log into xDCI see Figure 4. One method of creating these Apps is to encapsulate the workflows into a Docker container, and the Docker image is then launched on a suitable compute resource. The xDCI system will take care of any necessary data movement between xDCI and the compute resource, so that the analysis process itself will be a seamless experience for the user. Note that this can also be a mechanism to provide "Bring your own analysis tool" for the xDCI community [9].

The Cyverse/DE presently uses Condor to launch compute jobs on high performance compute (HPC) resources, including specialized systems such as General Purpose Graphical Processing Units (GPUs). This can be useful for compute-heavy analysis methods that are better suited to such specialized resources. DE also integrates the Agave Science API, allowing access to a wide range of computational platforms [10].



Viewing: /iplant/home/vaughn

Name	Last Modified	Size
applications	2011 Jun 20 19:05:28	
examples	2014 Sep 22 10:42:16	
tutorials	2013 Sep 17 11:04:40	

Name	Integrated...	Rating	
rePair-Fix Read ...	Roger Bart...	★★★★★	(1)
Merge BAM files	Roger Bart...	☆☆☆☆☆	(0)
FASTX quality fi...	Matthew V...	★★★★☆	(3)
BWA index 0.7.4	Matthew V...	☆☆☆☆☆	(0)
FASTX clipper	Matthew V...	★★★★☆	(3)
ALLMAPS path	Eric Lyons	☆☆☆☆☆	(0)
ALLMAPS merge	Eric Lyons	☆☆☆☆☆	(0)
BWA-0.7.4 non...	Roger Bart...	☆☆☆☆☆	(0)
bcftools view	Zhenyuan Lu	☆☆☆☆☆	(0)
MACS 1.4	Matthew V...	☆☆☆☆☆	(0)

	Start Date	End Date	Status
e...	2015 Jul 15 15:22:14	2015 Jul 15 15:23:43	Completed

Figure 4. Screenshot of virtual desktop showing the Discovery Environment for an xDCI user. Note that the open window in the foreground shows the Apps available to the user, while the open background window shows the data sets available.

3. Applications

In this section, we explore how xDCI is currently being implemented in three applications, 3D Brain Imaging Analysis, Neonatal Resuscitation Video Analysis, and Risk Analytics.

3.1. 3D Brain Imaging Analysis

The study of nervous system function requires an understanding of the structure of the brain at the cellular or sub-cellular level [11]. This in turn drives the development

978-1-5386-3472-1/17/$31.00 ©2017 IEEE

of methods to image the brain at sufficient resolution to see single neurons and trace the patterns of synaptic connections. Traditional two-dimensional microscopy methods are insufficient for this purpose, because the sectioning leads to difficulty in registration across sections resulting in distortions in the reconstructed image. 3D imaging methods such as MRI do not have the micron level spatial resolution that is required for cellular imaging [12]. Recently, a number of 3D brain imaging techniques have been developed that overcome many of these limitations and can produce 3D micron resolution images of fluorescently labeled brain samples. The key innovations are in novel tissue clearing methods coupled with fast microscopy methods.

Tissue clearing techniques such as CLARITY can preserve the structure of the brain, maintain mechanical stability of the tissue and at the same time render it transparent by removing the lipids responsible for light scattering [13], [14]. In light sheet microscopy, a specimen is illuminated along an entire plane with a light sheet [15], [16]. Separate lenses are used for the excitation of fluorophores and the detection of a fluorescence signal. Light sheet microscopy has been used for the imaging of intact brains from organism such as fish, mouse, marmoset, and human [14], [17], [18].

The goal of this application is to create a robust data archive called BRAIN-I for 3D images of intact brains that will enable the neuroscience research community to rapidly advance brain research. The archive will initially be developed for light sheet images of CLARITY-cleared brains. BRAIN-I will adhere to, and enforce standards developed for such 3D images allowing for data from multiple research centers to be easily accessible for research. BRAIN-I will also integrate analytical and visualization tools to allow researchers to easily conduct repeated analyze-evaluate-re-analyze workflows. We believe that BRAIN-I will be an important component of the envisioned BRAIN network for advancing our understanding of the brain.

3.2. Neonatal Resuscitation Video Analysis

Another example application of xDCI is for the analysis of video sequences of neonatal resuscitation. Neonatal resuscitation, e.g. stimulation, airway cleaning, bag-and-mask ventilation, is critical to infants who need additional help to initiate breathing. Assessing the performance of such procedures via videos is necessary for training purposes, but requires labor-intensive manual inspections. Guo *et al.* [2] reported a deep learning based approach to partially automatic this process. Their approach first selected motion salient regions by using a pre-trained Faster Region-based Convolutional Neural Network (RCNN), and then extracted motion and spatial deep features in the resulted regions. Next, the extracted features were fed into a linear SVM for prediction. Lastly, they applied a pairwise model for classification consistency.

One major drawback of this deep learning method is that the workflow is somewhat complex for domain scientists such as medical professional, thus limiting the use of the method. The deep learning component was implemented by Caffe [19] in Linux and MATLAB was also utilized for other computational tasks and the system does not support multiuser concurrent usage.

In order to achieve easy-access and high-scalability, we extend the method to allow users to set several computational parameters and then submit the analysis workflow in a web interface. In the server, each task is executed in a Docker container, which is built via a Dockerfile, and the result is returned to the web interface once completed. This simplified system does not require any environment configuration or software installation, and thus increases the availability of the original method.

3.3. Risk Analytics Discovery Environment (RADE)

The RADE pilot project, part of the North Carolina Data Science and Analytics Data Science Initiative (NC DSAI) is used for studying risk analytics. RADE uses xDCI, including iRODS, data federation, and DE with a goal to create a community platform for risk analytics research. For the purposes of the RADE project, risk analytics are defined as research using computationally and/or data intensive methods to assess and quantify risk and related impacts. The RADE pilot focuses on two exemplar use cases, one on the application of natural language processing (NLP) and one on coastal natural hazards impacts.

RADE - Deriving Business Trends Insights from Text-Based NLP Natural language processing (NLP) is generally run against large volumes of unstructured text in order to develop or refine the learning algorithm. In this use case, researchers at RENCI tested the efficacy of neural networks NLP to uncover trends in the local business climate by processing online sources of local business news to potentially identify events such as layoffs and IPOs. This particular use case was operationalized as a working proof of concept to test the viability of the approach [20].

RADE - Coastal Hazards Coastal hazards research is necessarily interdisciplinary and multi-methodological. In this example, researchers at North Carolina State University are working to better understand both the dynamics of hurricane impacts on coastal landscapes and how predicted changes in those landscapes may influence decision-makers at the local level such as town managers and property owners. This research combines high resolution geospatial data, hurricane impact forecasts, and an agent-based model in the analysis. In addition, the methodology requires access to compute, and an environment that can easily handle tens of thousands of discrete files.

Reusable Analytics Workflows

One of the goals for creating the risk analytics platform was to develop reusable tools and apps to support risk analytics research. The Discovery Environment component of xDCI is ideally suited to support this type of capability. In both cases, the researchers operationalized their research methodology as tools and apps in the Discovery Environment. See figure 6 In the context of the natural language processing use case, the primary processing algorithm was

978-1-5386-3472-1/17/$31.00 ©2017 IEEE

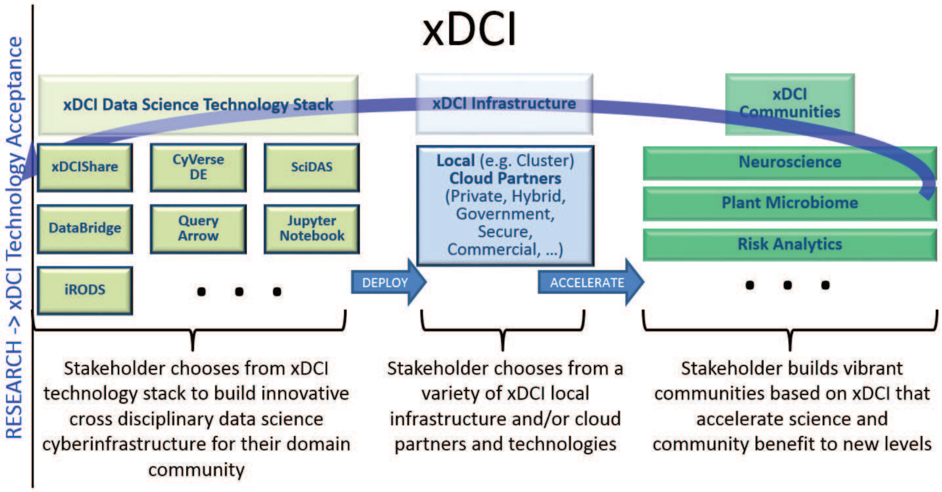

xDCI
RESEARCH -> xDCI Technology Acceptance
xDCI Data Science Technology Stack
xDCI Infrastructure
xDCI Communities
xDCIShare
CyVerse DE
SciDAS
DataBridge
Query Arrow
Jupyter Notebook
iRODS
• • •
Local (e.g. Cluster) Cloud Partners (Private, Hybrid, Government, Secure, Commercial, ...)
DEPLOY
ACCELERATE
Neuroscience
Plant Microbiome
Risk Analytics
• • •
Stakeholder chooses from xDCI technology stack to build innovative cross disciplinary data science cyberinfrastructure for their domain community
Stakeholder chooses from a variety of xDCI local infrastructure and/or cloud partners and technologies
Stakeholder builds vibrant communities based on xDCI that accelerate science and community benefit to new levels

Figure 5. We envision the future of xDCI as enabling stakeholders to readily deploy data science cyberinfrastructure for their community in a manner that will accelerate their community's science.

made into a reusable app. The analysis supported by the app could easily be used against different sets of input data. The output data is also easily staged for further processing for additional research insights. For the coastal research use case these tools and apps included data processing elements, GIS tools, model runs, and creation of custom workflows. Commonly needed tools were put into Docker containers and configured to be accessed as part of building the research specific catalog of tools, including GRASS, Python, R, and GNU Octave.

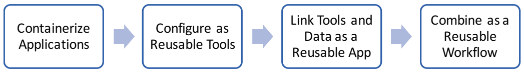

Containerize Applications
Configure as Reusable Tools
Link Tools and Data as a Reusable App
Combine as a Reusable Workflow

Figure 6. App Creation Steps in Discovery Environment

Utilizing the xDCI-enabled RADE framework facilitated 1) the development of reusable tools and apps that allow for customization and hypothesis testing, 2) connectable workflows to further enhance the research, 3) straightforward data input and output management, 4) collaborative research across institutions, and 5) a platform for community-specific analytical capabilities, data and vocabulary. In addition to the more general benefits for a community-based, interdisciplinary research field, the use cases also accrued more specific benefits. The coastal hazards researchers cited the potential for research apps to become publishable products in and of themselves (supports reproducibility), and more efficient analysis (C. Dietrich and A. Karance, personal communication, 11 May 2017). In the case of the NLP research, using RADE provided a foundation for moving forward practical tools to automate the collection of business events for academic and industry research purposes (C. Schmitt, personal communication, 25 May 2017).

4. Future Work

Figure 5 illustrates how we envision the future of xDCI in operation. Shown on the left of Figure 5 is where the multidisciplinary research community at large continually produces technology artifacts that are vetted for inclusion in the xDCI Data Science Technology Stack. Vetting criteria could include e.g. criticality, value, sustainability, usability, reusability, performance, functionality, capability, availability, scientific impact, usefulness, reliability, and so forth [21]. A representative xDCI Data Science Technology Stack is shown. We envision an xDCI stakeholder working with cyberinfrastructure (CI) staff proficient with xDCI to select those xDCI technologies best suited for their community's needs. Certain xDCI technologies such as Discovery Environment, xDCIShare, iRODS, will be made available as a live instance so that users can experiment with the technologies prior to finalizing selection. Shown in the middle of Figure 5 is xDCI Infrastructure that could include local departmental cluster or cloud provider or technology. Finally, on the right are shown example communities constructed using xDCI where, by virtue of using the xDCI approach discussed herein, accelerate their respective community's science by using and reusing community-established and -proven xDCI elements. Invariably, these communities will produce additional technology artifacts that can be vetted for inclusion in xDCI as depicted by the right-to-left feedback arrow. Over time we see this approach as both building the quantity, capability, and interoperability of xDCI elements and leading to novel additional elements, for example we envision the notion of an ”App Store” that can transcend elements such that an App that is discovered in for example xDCIShare, can equally be discovered and run within the Discovery Environment. These xDCI constructs will encourage increased stakeholder participation thus building ever-increasing vibrant communities that are able to accelerate science and community benefit to new levels.

References

[1] D. S. Richardson and J. W. Lichtman, "Clarifying tissue clearing," *Cell*, vol. 162, no. 2, pp. 246–257, 2015.

[2] Y. Guo, J. Wrammert, K. Singh, K. Ashish, K. Bradford, and A. Krishnamurthy, "Automatic analysis of neonatal video data to evaluate resuscitation performance," in *Computational Advances in Bio and Medical Sciences (ICCABS), 2016 IEEE 6th International Conference on*. IEEE, 2016, pp. 1–6.

[3] "irods," https://irods.org.

[4] "Cyverse discovery environment," http://www.cyverse.org/discovery-environment.

[5] D. Tarboton, R. Idaszak, J. Horsburgh, J. Heard, D. Ames, J. Goodall, L. Band, and V. Merwade, "A resource centric approach for advancing collaboration through hydrologic data and model sharing," in *Proceedings of the 11th International Conference on Hydroinformatics*, New York City, USA, 2014.

[6] R. Idaszak, D. Tarboton, H. Yi, L. Christopherson, M. Stealey, B. Miles, P. Dash, A. Couch, C. Spealman, D. Ames, and J. Horsburgh, *Software Engineering for Science. HydroShare – A case study of the application of modern software engineering to a large distributed federally-funded scientific software development project*, J. Carver, N. Chue Hong, and G. Thiruvathukal, Eds. CRC Press, Chap. 10, pp. 219–236, 2016.

[7] "Dcmi metadata terms 2012," https://dublincore.org/documents/dcmi-terms/.

[8] "Dcmi guidelines," http://dublincore.org/documents/profile-guidelines/.

[9] K. J. Gorgolewski, F. Alfaro-Almagro, T. Auer, P. Bellec, M. Capotă, M. M. Chakravarty, N. W. Churchill, A. L. Cohen, R. C. Craddock, G. A. Devenyi *et al.*, "Bids apps: Improving ease of use, accessibility, and reproducibility of neuroimaging data analysis methods," *PLoS Computational Biology*, vol. 13, no. 3:e1005209, 2017.

[10] "Using hpc apps in the de - discovery environment manual - cyverse wiki," https://pods.iplantcollaborative.org/wiki/display/DEmanual/Using+HPC+Apps+in+the+DE.

[11] N. Kasthuri, K. Hayworth, D. Berger, R. Schalek, J. Conchello, S. Knowles-Barley, D. Lee, A. Vazquez-Reina, V. Kaynig, T. Jones, M. Roberts, J. Morgan, J. Tapia, H. Seung, W. Roncal, J. Vogelstein, R. Burns, D. Sussman, C. Priebe, H. Pfister, and J. Lichtman, "Saturated reconstruction of a volume of neocortex," *Cell*, vol. 162, no. 3, pp. 648–661, 2015.

[12] H. Dodt, U. Leischner, A. Schierloh, N. Jahrling, C. Mauch, K. Deininger, J. Deussing, M. Eder, W. Zieglgansberger, and K. Becker, "Ultramicroscopy: three-dimensional visualization of neuronal netowrks in the whole mouse brain," *Nature Methods*, vol. 4, no. 4, pp. 331–336, 2007.

[13] R. Tomer, L. Ye, B. Hsueh, and K. Deisseroth, "Advanced clarity for rapid and high-resolution imaging of intact tissues," *Nature Protocols*, vol. 9, no. 7, pp. 1682–1697, 2014.

[14] K. Chung, J. Wallace, S. Kim, S. Kalyanasundaram, A. Andalman, T. Davidson, J. Mirzabekov, K. Zalocusky, J. Mattis, A. Denisin, S. Pak, H. Bernstein, C. Ramakrishnan, L. Grosenick, V. Gradinaru, and K. Deisseroth, "Structural and molecular interrogation of intact biological systems," *Nature*, vol. 497, no. 7449, pp. 332–337, 2013.

[15] E. Reynaud, J. Peychl, J. Huisken, and P. Tomancak, "Guide to light-sheet microscopy for adventurous biologists," *Nature Methods*, vol. 12, no. 1, pp. 30–34, 2015.

[16] R. Tomer, K. Khairy, and P. Keller, "Light sheet microscopy in cell biology," *Methods Molecular Biology*, vol. 931, pp. 123–137, 2013.

[17] E. Susaki, K. Tainaka, D. Perrin, F. Kishino, T. Tawara, T. Watanabe, C. Yokoyama, H. Onoe, M. Eguchi, S. Yamaguchi, T. Abe, H. Kiyonari, Y. Shimizu, A. Miyawaki, H. Yokota, and H. Ueda, "Whole-brain imaging with single-cell resolution using chemical cocktails and computational analysis," *Cell*, vol. 157, no. 3, pp. 726–739, 2014.

[18] R. Tomer, M. Lovett-Barron, I. Kauvar, A. Andalman, V. Burns, S. Sankaran, L. Grosenick, M. Broxton, S. Yang, and K. Deisseroth, "Sped light sheet microscopy: Fast mapping of biological system structure and function," *Cell*, vol. 163, no. 7, pp. 1796–1806, 2015.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[20] Y. Wang and C. Schmitt, "Extraction of Business Events from Text for Industrial Forensics," in *Proceedings of KDD conference, August 2017 (KDD2017), unpublished*, Halifax, Nova Scotia, 2017.

[21] S. Ahalt, B. Berriman, M. Brown, J. Carver, N. C. Hong, A. Fish, R. Idaszak, G. Newman, D. Panda, A. Patra, E. G. Puckett, C. Roland, D. Thain, S. Uluagac, and B. Zhang, "Toward a Framework for Evaluating Software Success: A Proposed First Step," 10 2015. [Online]. Available: https://figshare.com/articles/Toward_a_Framework_for_Evaluating_Software_Success_A_Proposed_First_Step/1561451

978-1-5386-3472-1/17/$31.00 ©2017 IEEE