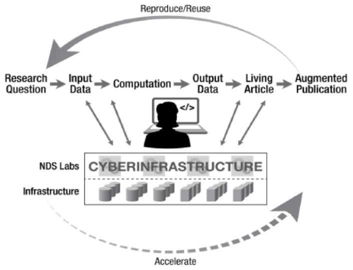
#1541450: CC*DNI DIBBS: Merging Science and Cyberinfrastructure Pathways: **The Whole Tale**

B. Ludäscher, K. Chard, N. Gaffney, M. Jones, J. Nabrzyski, V. Stodden, M. Turk, K. Turner

1. The Challenge

Despite an increasing recognition of the need to share all aspects of the research process, scholarly publications today are often disconnected from the underlying data and code that was used to produce the findings. There is no shortage of tools and cyberinfrastructure (CI) addressing specific aspects of this challenge, yet scientists find it difficult to utilize these different pieces and building blocks in a seamless way that spans the "whole story", i.e., from conducting the computational science to the publication of a "living" or executable paper. These new types of publications include not only the science narrative, but also (references to) all the relevant data, code, and provenance information needed to reproduce and experience the computational and research processes described by the paper.

2. Whole Tale Vision and Approach



The Whole Tale project aims to link existing tools and CI and thus provide support for the entire computational process that underlies discovery using popular frontends, e.g., Jupyter and RStudio, thereby simplifying the ability for researchers to conduct, share, and publish their research. To facilitate reproducible computational science, workflow modeling, provenance capture, and advanced provenance querying capabilities will be provided through the Whole Tale: we envision an environment where data providers, scientists, and application developers can collaborate and create reproducible workflows using software they are already familiar with. Our aim is to support scientific investigation at all computational scales, from HPC environments to single-user endeavors (i.e., the "long tail" of science). We will provide a research environment that captures and, at the time of publication, exposes salient details of the research process via access to persistent versions of the data and code used, the underlying workflow, and relevant data lineage including parameter settings, and intermediate data.

Rather than developing new CI and tools from scratch, Whole Tale will leverage and link existing CI and tools. The underlying CI will be exposed to users through well-known applications such as Jupyter Notebooks that support commonly used data analysis languages. Data repositories and Whole Tale storage will be unified and visible to users through several open source interfaces. By building data repository access into modules that present filesystem-like interfaces, we further lower the barrier to access to remote data stores. The Whole Tale system will also incorporate Globus Auth, a unified identity management system for users to leverage their own campus-, ORCID-, or other existing identities. Whole Tale will enable deployment of Dockerfile-based environments to support extensible and customizable research workflows. Close collaboration with science communities is essential to success in development and uptake. We have initiated community-driven scientific working groups to identify and pilot use cases, and to provide feedback from the early stages of the project.

3. Progress to Date and Next Steps

We have completed an initial system architecture, selected key technologies and started CI integration, and designed a mockup user dashboard. We have designed a data model for managing disparate data and metadata, and developed an initial REST API for interacting with the system. We have adopted Globus authentication which facilitates federated login with a range of identities, and we are working towards extending this system with support for ORCID. Finally, we have deployed a federated storage system between TACC and NCSA. We also have formed the Whole Tale External Advisory Board, started the first science working groups, and presented Whole Tale at stakeholder meetings, looking for feedback and support from potential users.

Our next steps focus on extending the range of systems supported by Whole Tale by developing data ingestion wrappers for Globus and DataONE. We will integrate NextCloud with our data management fabric to provide an accessible user-oriented home directory model. Subsequently we will integrate Jupyter to provide a personalized analysis environment with interfaces to access supported data sources. Finally, we will implement and refine a user dashboard that ties together the many microservices provided by the Whole Tale.