



Developing, Packaging, and Sharing Reproducible Research Objects: The Whole Tale Approach

Bertram Ludäscher & Craig Willis

School of Information Sciences & National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

DataONE Webinar, October 8th, 2019





What is **Whole Tale**?

- NSF-funded **Data Infrastructure Building Blocks** (*DIBBs*) project
- **Platform** to create, publish, and execute **tales**
- Simplify process of creating & verifying **reproducible** computational artifacts
- <https://dashboard.wholetale.org>



Why **Whole Tale**?

- Increased reliance on **computation** across domains
 - new skill requirements for researchers
- **Open Science** changing norms and expectations
 - increased emphasis on **sharing data & code**
 - ... with **transparency** and **reproducibility** in mind!
 - => from sharing data to sharing **research objects**
 - **FAIR** principles



Whole Tale: Enables Computational Science

The collage features several distinct panels:

- Chemistry:** Molecular models of CO₂ monomer and dimer with vibrational arrows. Text: "CHEMISTRY Anharmonic vibrational structure... Supplemental information used to produce from 'Anharmonic vibrational structure of the dioxide dimer with a many-body potential energy... The project solves the vibrational Schrodinger equation for the CO₂ monomer and dimer using vibrational theory and a many-body potential energy surface... Olaseni Sode".
- Archaeology:** A grid of 20 heatmaps showing agricultural patterns. Text: "ARCHAEOLOGY Climate change stimulated agriculture... This is a compendium package for d'Alpoim Guedes and Bocinsky (2018). The compendium contains all code associated with the analyses described and presented in the publication. d'Alpoim Guedes, Jade and R. Kyle Bocinsky. Jade d'Alpoim Guedes and R. Kyle Bocinsky".
- LIGO Tutorial:** A visualization of gravitational waves. Text: "SCIENCE LIGO Tutorial LIGO Detected Gravitational Waves from Black Holes. On September 14, 2015 at 5:51 a.m. Eastern Daylight Time (09:51 UTC), the twin Laser Interferometer Gravitational-wave Observatory (LIGO) detectors, located in Livingston and Hanford, California and Louisiana, respectively, made the first direct observation of gravitational waves. Kacper Kowalik".
- Ecology:** Two maps of South America showing species distribution for *bradypus variegatus* and *microryzomys minutus*. Text: "a classical ecological... for species distribution... Anderson and Shapire revisited... probability density functions in order to... for species presence in a geographic region... replicates the logic and gene...".
- Material Science:** A ternary phase diagram for Ni-Al. Text: "SCIENCE Predicting the Properties of Inorganic Materials... This tale describes how to recreate a 2016 paper by Ward et al on using machine learning to predict the properties of materials. The main focus of this paper was the construction of a general purpose method to link the composition of a material (i.e., the fractions of each element) to its properties, which they found can be used for... Logan Ward".
- Machine Learning:** A bokeh-style background of yellow and orange circles. Text: "SCIENCE Machine Learning: An Applied Econometric Approach... Markdown Editor. Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106. URL: https://www.aeaweb.org/articles?id=doi:10.3386/jep.312017. Lili Fang".
- Other:** A satellite image of a coastline with a fish, a spectrum plot, and a star field image.



Whole Tale & the Elements of a ... Reproducible Computational Research Platform

Develop



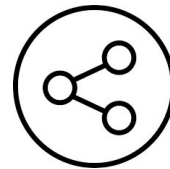
Easy-to-access
cloud-based
computational
environments

Analyze



Transparent
access to
research **data**

Share



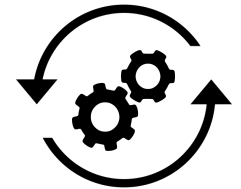
Collaborate
and **share** with
others

Package



Export or publish
executable
research
objects

Reproduce

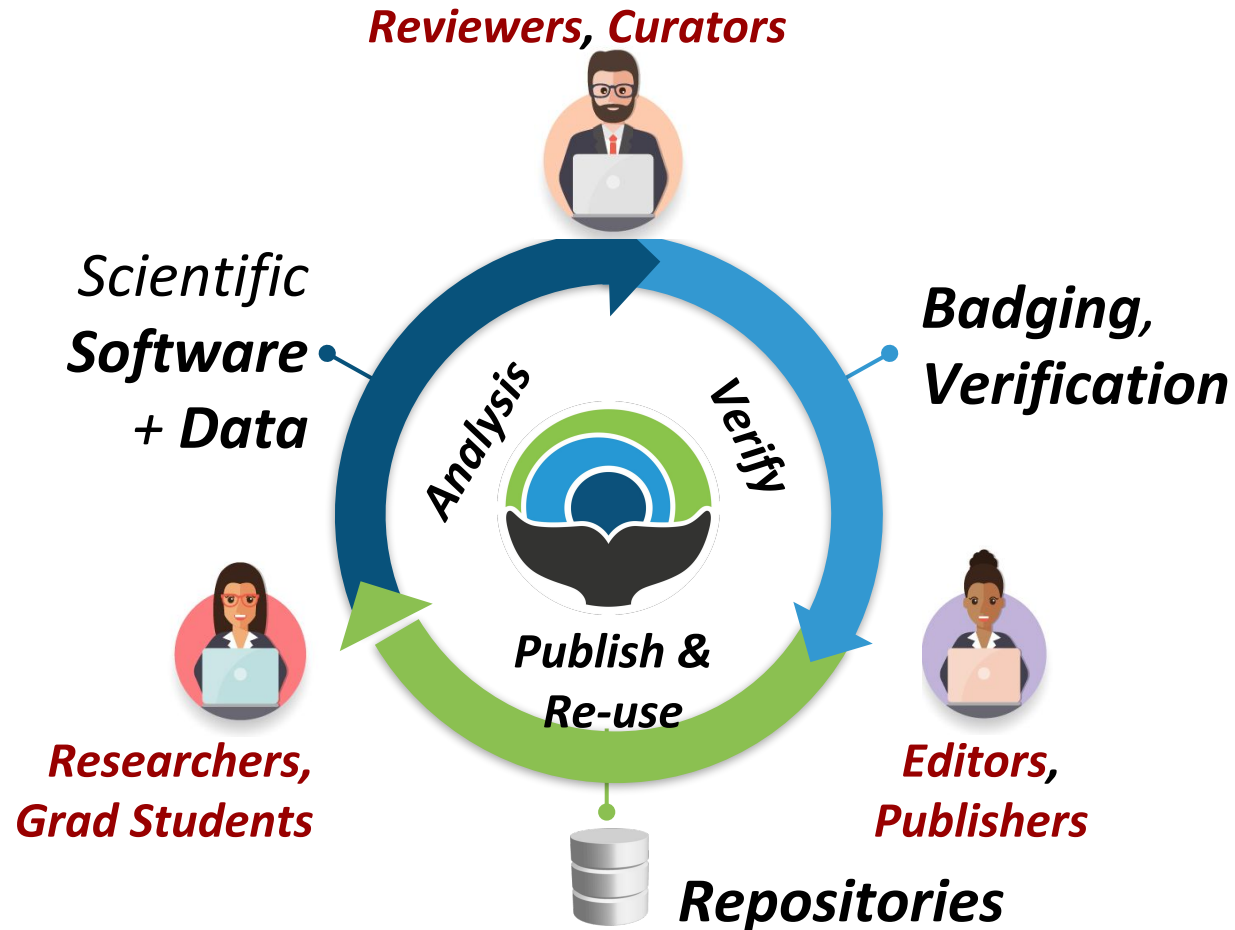


Re-execute
Review
Verify
Re-use

Coming soon



Whole Tale **Roles** and **Stakeholders**





Develop & Analyze with Whole Tale

- Easy to access cloud-based environments
 - Your laptop in the cloud
- Popular tools
 - + ... extensible!
- Work with data & code in **transparent** (*provenance-enabled*) ways
 - Automatic **data citation**
 - Automatic computational **provenance capture** (coming soon)



Package & Reproduce with Whole Tale

- Executable **Research Objects**
- Publish or export to **research archives**
- Compatible with new norms for **reproducibility** and **transparency**
- For **verification** and **re-use**




Whole Tale *and* DataONE

- **Discover & access data** from any DataONE repository
- **Analyze** data in Whole Tale
- **Package & publish** tales to Metacat-based repositories
- **Provenance** support



< Back to search | Search / Metadata

James Duncan, Alexandra Kosiba, Garrett Meigs, and Jennifer Pontius. Standardized Regional Aerial Detection Survey Disturbance Spatial Data. Forest Ecosystem Monitoring Cooperative. p1216.ds2428, version: p1216.ds2428_20191005_0300. 

Citations 0 Downloads 0 Views 38 Copy Citation Analyze Quality report

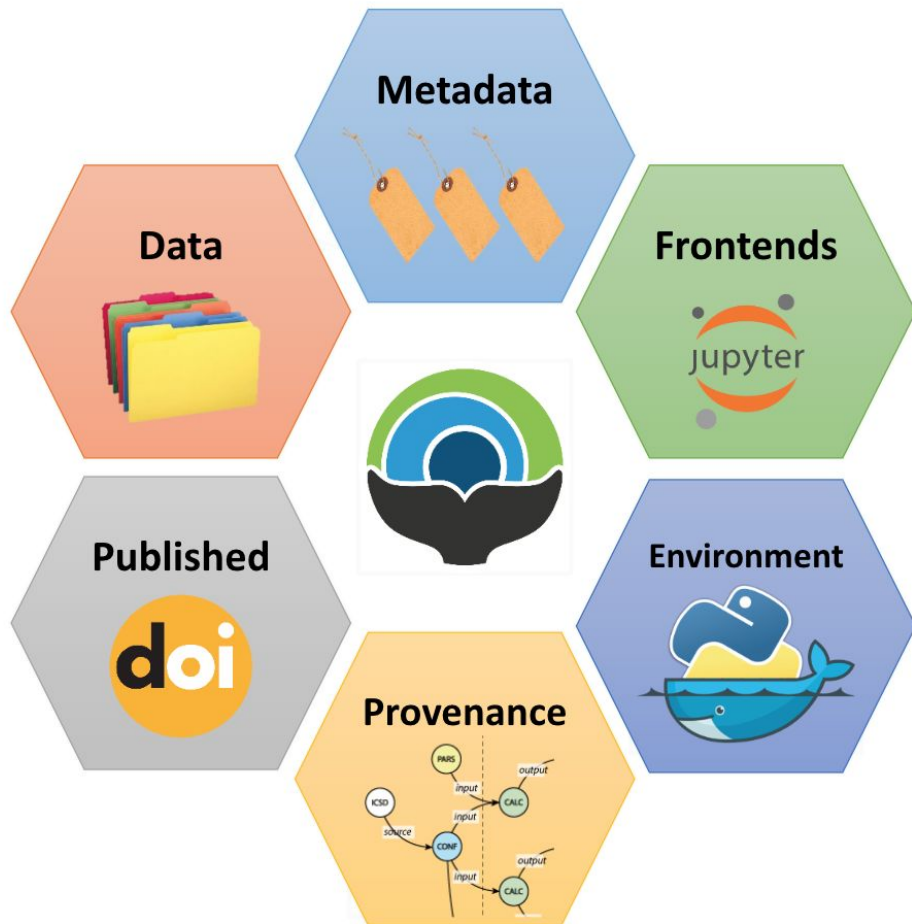
Analyze dropdown menu:
RStudio
Jupyter Notebook

Files in this dataset		
Name	File type	Size
Metadata: Standardized Regional Aerial Detection Survey Disturbance Spatial Data	EML v2.1.1	21 KB

Download



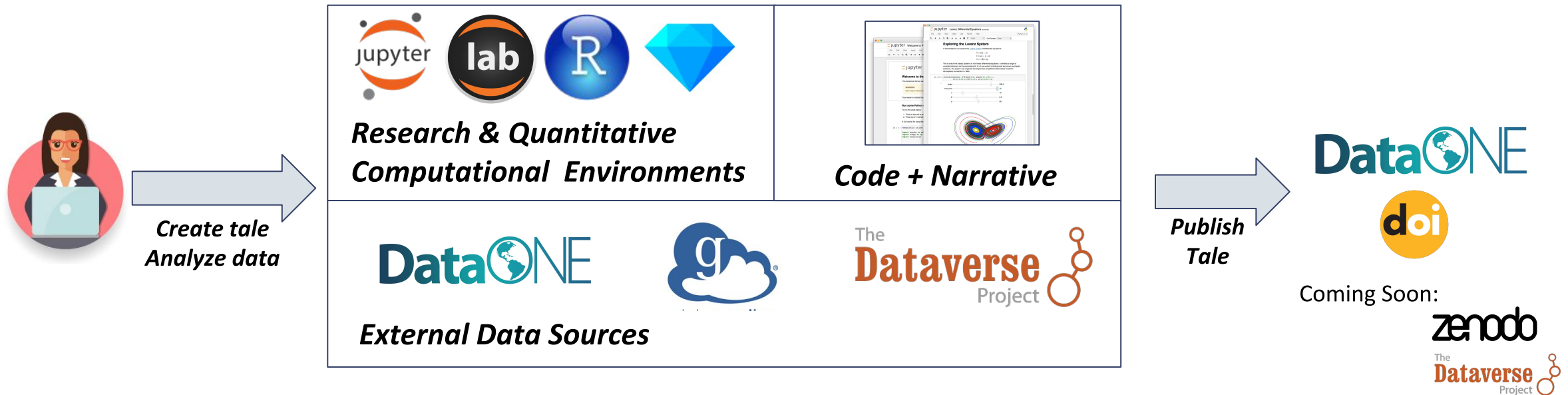
What exactly is (in) a **Tale**?



- ✓ **Tale: Research object**
 - data, code, narrative, compute environment
- ✓ **Executable**
- ✓ **Transparent**
- ✓ **Publishable**
- **Verifiable**
- **Remixable**
- **Standards-based**



Whole Tale Platform Overview

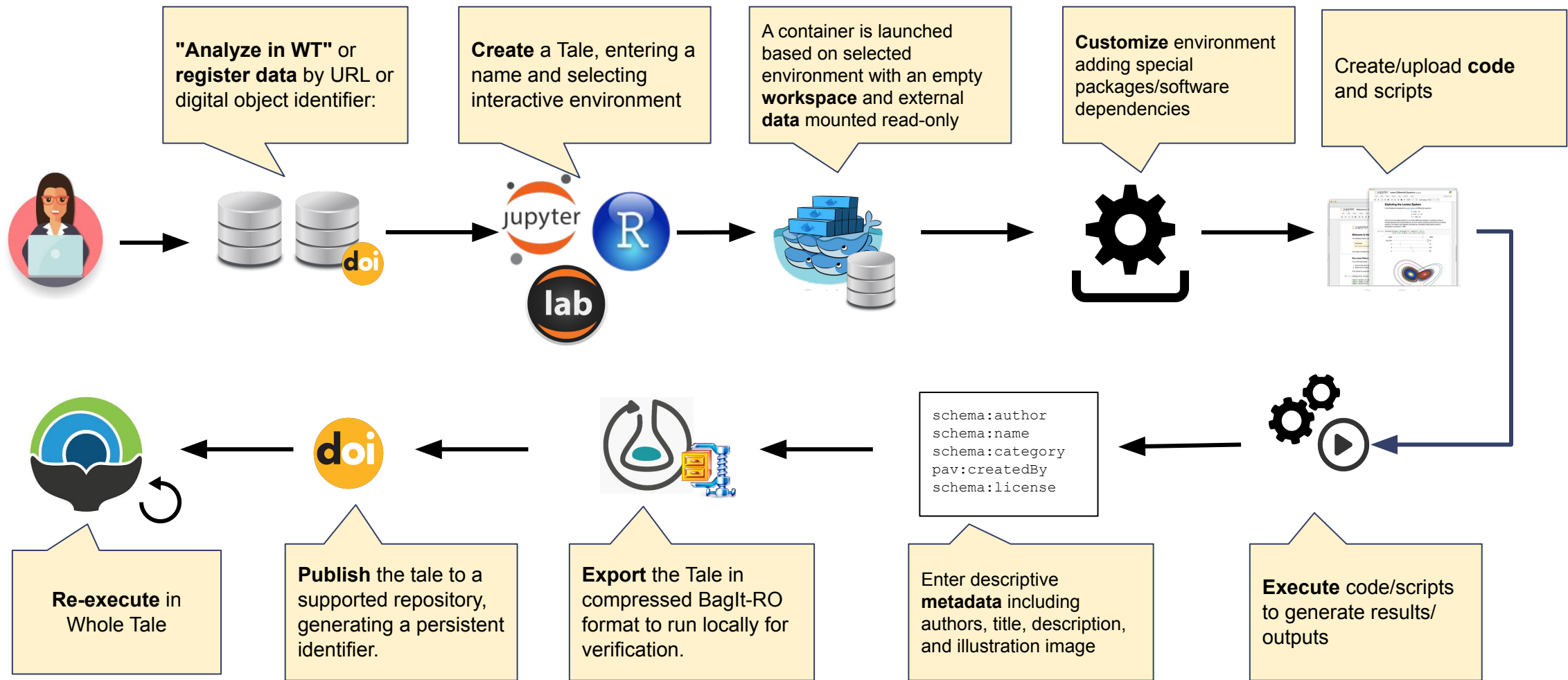


- **Authenticate** using your institutional identity
- **Access** commonly-used **computational environments**
- Easily **customize** your environment (via repo2docker)
- Reference and access externally **registered data**

- Create or upload **your data and code**
- Add **metadata** (including **provenance** information)
- Submit code, data, and environment to **archival repository**
- Get a **persistent identifier**
- **Share** for **verification** and **re-use**



Tale Creation Workflow



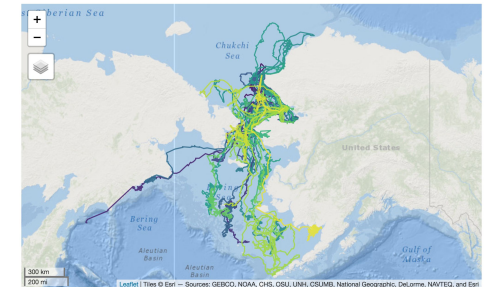


Demo: Analyzing Seal Migration Patterns

A research team is preparing to publish a manuscript describing a computational model for estimating animal movement paths from telemetry data:

- **Telemetry data** published in **Research Workspace**
- **Analysis** and **visualization** in **RStudio**
- Existing routines stored in **local R files**
- Analysis requires **specialized R packages**
- **Publish results** for the community in **DataONE**

[Live Demo](#) or [Demo Video](#)



Based on: J.M. London and D.S. Johnson. Alaska bearded and spotted seal example dataset and analysis. <https://github.com/jmlondon/crwexampleakbs>, 2019



Key features

Supported **environments**

- Extension to Binder's **repo2docker**
 - Jupyter, JupyterLab
 - RStudio (based on Rocker Project)
 - OpenRefine
- Coming soon:
 - Matlab, Stata



jupyter-repo2docker



Key features

Supported **data repositories**

- **Register data** from supported research data repositories
- Referenced data is **cited**
 - Ideally eventually contributing to citation counts
- **Publish tales** back to research repositories

DataONE



The
Dataverse
Project

Coming Soon:

zenodo



Key features

Export to BagIt-RO

- BagIt: archival format
- Re-runnable in WT
- BagIt-RO
 - **Open archival** format
 - **Research Object** support
 - Extended for **Big Data**

```
tale/  
  bagit.txt  
  bag-info.txt  
  data/  
    workspace/  
      run.py  
      LICENSE  
      requirements.txt  
      output.csv  
  LICENSE  
  metadata/  
    manifest.json  
  manifest-sha1.txt  
  start-here/  
    README.md  
  tagmanifest-sha1.txt
```




Key features Export and **Run Locally**

- Natural outcome of Tale **export** and **repo2docker**
- Download a zip file (BagIt-RO)
- run-local.sh
 - Build image (**repo2docker**)
 - Fetch external data (**bdbag**)
 - Execute (**Docker**)



jupyter-repo2docker



BIG DATA *for* DISCOVERY SCIENCE



Coming soon

- Publish to Zenodo, Dataverse
- Tapis/Agave data sources
- Sharing/collaboration
- Create tale from Git repository
- Image preservation
- System provenance capture
- Better user experience



Thank you! Questions?

Bertram Ludäscher
ludaesch@illinois.edu

Craig Willis
willis8@illinois.edu



References

- Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M.B., Kowalik, K., Kulasekaran, S., Ludäscher, B., Mecum, B.D., Nabrzyski, J. and Stodden, V., 2019. [Computing environments for reproducibility: Capturing the “Whole Tale”](#). *Future Generation Computer Systems*, 94, pp.854-867.
- Chard, K., Gaffney, N., Jones, M.B., Kowalik, K., Ludäscher, B., McPhillips, T., Nabrzyski, J., Stodden, V., Taylor, I., Thelen, T., Turk, M.J. and Willis, C., 2019. [Application of BagIt-Serialized Research Object Bundles for Packaging and Re-execution of Computational Analyses](#). In *2019 IEEE 15th International Conference on e-Science (e-Science)*. IEEE.
- Chard, K., Gaffney, N., Jones, M.B., Kowalik, K., Ludäscher, B., Nabrzyski, J., Stodden, V., Taylor, I., Turk, M.J. and Willis, C., 2019, June. [Implementing Computational Reproducibility in the Whole Tale Environment](#). In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems* (pp. 17-22). ACM.
- McPhillips, T., Willis, C., Gryk, M., Nunez-Corrales, S., Ludäscher, B. 2019. [Reproducibility by Other Means: Transparent Research Objects](#). In *2019 IEEE 15th International Conference on e-Science (e-Science)*. IEEE.
- Mecum, B., Wyngaard, S., Willis, C., Turk, M., Thelen, T., Taylor, I., Stodden, V., Perez, D., Nabrzyski, J., Ludaescher, B. and Kulasekaran, S., 2018, December. [Science, containerized: Integrating provenance and compute environments with the Whole Tale](#). In *AGU Fall Meeting Abstracts*.
- Mecum, B., Jones, M.B., Vieglais, D. and Willis, C., 2018, October. [Preserving Reproducibility: Provenance and Executable Containers in DataONE Data Packages](#). In *2018 IEEE 14th International Conference on e-Science (e-Science)*. IEEE.



Whole Tale **Collaboration** (PI Team)

- **U Illinois** (NCSA) **Bertram Ludäscher, Victoria Stodden, Matt Turk**
 - overall lead (co-operative agreement)
 - reproducibility; provenance; open source software development; outreach
- **U Chicago** (Globus) **Kyle Chard**
 - data transfer & storage; compute; infrastructure
- **UC Santa Barbara** (NCEAS) **Matt Jones**
 - (meta-)data publishing; provenance; repositories
- **U Texas, Austin** (TACC) **Niall Gaffney**
 - compute; HTC; “big tale”; Science Gateways
- **U Notre Dame** (CRC) **Jarek Nabrzyski**
 - UX design; UI design





The Whole Team

- **Adam** Brinckman (Notre Dame, former Dev)
- **Bertram** Ludäscher (UIUC, PI)
- **Bryce** Mecum (UCSB, former Dev)
- **Craig** Willis (UIUC, Dev, tech project manager)
- **Damian** Perez (Notre Dame, former Dev)
- **Ian** Taylor (Notre Dame, SP, Dev)
- **Jarek** Nabrzyski (Notre Dame, co-PI)
- **Joe** Stubbs (U Texas, Dev)
- **Kacper** Kowalik (UIUC, Dev, Senior Architect)
- **Kandace** Turner (UIUC, former project mgr)
- **Kristina** Davis (Notre Dame, UI, UX)
- **Kyle** Chard (U Chicago, co-PI)
- **MT** Campbell (UIUC, project manager)
- **Matt** Jones (UCSB, co-PI)
- **Matt** Turk (UIUC, co-PI)
- **Michael** Lambert (UIUC, Dev)
- **Mihael** Hategan (U Chicago, Dev)
- **Niall** Gaffney (U Texas, co-PI)
- **Rachel** Volentine (UTK, UX)
- **Sebastian** Wyngaard (Notre Dame, Dev)
- **Sivakumar** Kulasekaran (U Texas, former Dev)
- **Thomas** Thelen (UCSB, Dev)
- **Timothy** McPhillips (UIUC, Dev)
- **Victoria** Stodden (UIUC, co-PI)

+ *WT Summer Interns (7); WT/RDA Fellows (4+4); WG Leads (5); other collaborators*