Data Sharing in a Complex Computational Study: Easier Said than Done!

Qian Zhang[1,2], Heidi Imker[2], Bertram Ludäscher[1]

1 School of Information Sciences (iSchool), University of Illinois at Urbana-Champaign

2 Research Data Service, University Library, University of Illinois at Urbana-Champaign

INTRODUCTION

This work is a follow-up of the IDCC 2016 data paper "Using a computational study of hydrodynamics in the Wax Lake delta to examine data sharing principles" [1]. In this poster, we will explain the practical considerations and activities that were performed throughout data sharing, from preparing to finally publishing the datasets. Specifically, we will describe our efforts to:

- *explore and evaluate data repositories;*
- *investigate data use policy from both external and internal sources;*
- *decide on granularity of deposit and the associated publication timelines;*
- *and prepare documentation.*

With this example as a use case, we show that data sharing of a complicated computational study for research reproducibility/replication in practice is not as straightforward as stated in "principle", and requires flexible examination and additional care in practice.

Background

Figure 1 describes a complex dataset used to study the circulation and wind-driven flows in the Wax Lake Delta, Louisiana, USA under winter storm conditions. The whole package bundles a large dataset (approximately 74 GB), which includes the numerical model, software and scripts for data analysis and visualization, as well as detailed documentation. The raw data came from multiple external sources (government agencies, community repositories, and deployed field instruments and surveys, etc.), leading to very large datasets with complex data structures. After integrating multiple datasets from diverse data formats from different sources, new data products are obtained which are then used with the numerical model. With a complex algorithm of computation, the model generates a structured output dataset, which is, after post-data analysis, presented as informative scientific figures and tables that allow interpretations and conclusions contributing to the science of coastal physical oceanography.

Data Sharing Motivation

The data can be reused to study reproducibility or as preliminary investigation to explore a new topic. With thorough documentation and well-organized data, both the input and output dataset are ready for sharing in a domain repository or an institutional repository. Furthermore, the data organization and documentation also serves as a guideline for future research data management and the development of workflow protocols. Here we will describe the dataset created by this Wax Lake Delta hydrodynamics study, how sharing the dataset publicly could enable validation of the current study and extension by new studies, and the challenges that arise prior to sharing the dataset.

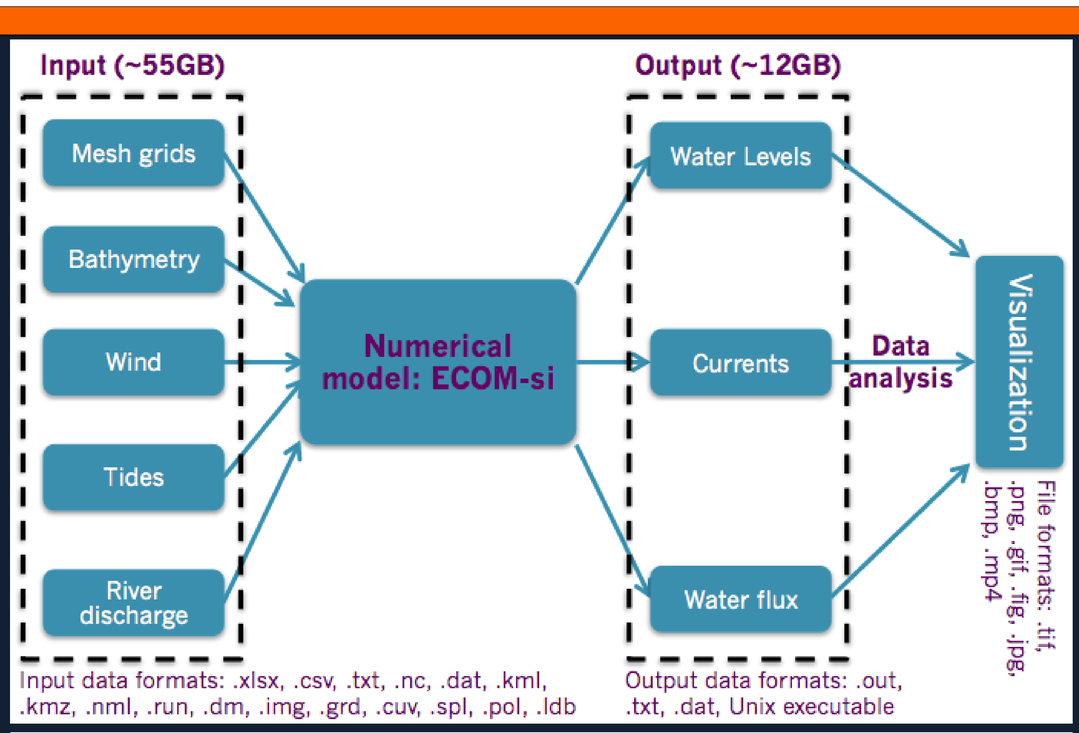

Figure 1. Data flow of the WLD hydrodynamics study: multiple input data are configured for numerical model setup, which initiate the simulation and output user-customized variables with time evolving. Both input and output datasets are highly structured, and must be visualized for further interpretation and analysis.

DATASET OVERVIEW

Model Input		Model Output
Data I made/generated	External data (public agency data)	
• Mesh grid	• USGS	• Multiple files • Simulation figures, analysis • …
• WAVCIS data (embargo)	• NGDC	
• Field measurements (3 data sets）	• NDBC	
• LIDAR topography image file	• CO-OPS	

Challenges

- Mixture of licences: CC0, CC-BY, etc.
- Mixture of publication timelines: no delay vs. embargo.
- Storage gap: huge datasets (over 70 GB with single data file sizes of over 5 GB) exceed the size limit for most data repositories → no suitable domain repository.
- Difficulty in choosing a trustworthy dissemination platform that can provide a secure, sustainable, and reliable infrastructure.
- Lack of incentives and rewards for data sharing.

POSSIBLE DATA SHARING PATHS

1

1 deposit: All-in-one
(1). model (input + output) dataset: CC-BY, embargoed for 1 year

2

9 separate deposits: Make each separate deposit
(1). mesh grid dataset: CC-BY, no delay
(2). WAVCIS dataset: CC-BY, embargoed for 1 year
(3). Field measurement dataset: CC-BY, no delay
(4). LIDAR dataset: CC-BY, no delay
(5). USGS dataset: CC0, no delay
(6). NGDC dataset: CC0, no delay
(7). NDBC dataset: CC0, no delay
(8). CO-OPS dataset: CC0, no delay
(9). Model output dataset: CC-BY, no delay

3

7 separate deposits: Re-arrange deposit based on granularity
(1). public agency dataset (NGDC + NDBC + USGS + CO-OPS): CC0, no delay
(2). WAVCIS dataset: CC-BY, embargoed for 1 year
(3)-(5). 3 field measurement datasets → 3 deposits: CC-BY, no delay
(6). LIDAR dataset: CC-BY, no delay
(7). model datasets (mesh grid + model output): CC-BY, no delay

RESULTS & CONCLUSION

The granularity of deposit is based not only on the computational infrastructure as a whole but on a mixture of licences applied to different datasets, as well as taking the ease of data access and later reuse into consideration. As a result, our sharing solution is to

- *Option 3;*
- share and publish all the datasets in an *institutional data repository – Illinois Data Bank (https://databank.illinois.edu/);*
- divide the whole data package of this project is into *7 separate dataset deposits;*
- with *no publication delay for 6 [3-8] of the 7 deposits;*
- *and a last deposit with a one year embargo [9] which will make the metadata publicly available but restrict the data files from access until the peer-reviewed journal paper get published.*

Those implementing RDM services at University of Illinois at Urbana-Champaign are working to adapt to similar situations that require flexibility. For example, the data sharing steps introduced in this poster are in good accordance with the Data Management Workshop Series 3: Preparing for Data Sharing [2] provided by the Research Data Service on our campus.

ACKNOWLEDGEMENTS

Special thanks to my PhD supervisor Dr. Chunyan Li at the Louisiana State University, the Data Curation Specialist Elizabeth Wicks and the Senior Information Design Specialist Dena Strong at the Research Data Service as well as the Prairie Research Institute Librarian Susan Braxton, for valuable suggestions during the dataset consultation.

REFERENCES

[1] Qian Zhang, Chunyan Li, Heidi Imker, Bertram Ludäscher, and Megan Senseney. Using a computational study of hydrodynamics in the Wax Lake Delta to examine data sharing principles. IJDC 2016 Data Papers.

[2] Wickes, Elizabeth; Sheehan, Beth (2016). Preparing for Data Sharing Materials, https://www.ideals.illinois.edu/handle/2142/91615.

[3] Zhang, Q. (2016): Public agency data of the Wax Lake delta. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-4871125_V1

[4] Zhang, Q. & Li, C. (2016a): Model dataset for the Wax Lake delta. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-9511904_V1

[5] Zhang, Q. & Li, C. (2016b): Current data of the Wax Lake delta. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-1752285_V1

[6] Zhang, Q. & Li, C. (2016c): Bathymetry data of the Wax Lake delta (2012-12-01). University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-4810873_V1

[7] Zhang, Q. & Li, C. (2016d): Bathymetry data of the Wax Lake delta (late 2012). University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-1001307_V1

[8] Zhang, Q., Li, C., & Braud, D. (2016): LIDAR data for the Wax Lake delta. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-3764213_V1

[9] Zhang, Q., & Li, C. (2017): Meterology and ocean data collected at LSU WAVCIS Lab. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-2436375_V1