Lifecycle-based Preservation Management of Research Data

Qing Lu[1], Yingdong Zhang, Yinbing Sun, Xiaoying Sun
ShanghaiTech University, Shanghai, China

Abstract: Research data management (RDM) is now an integral part of university management of research, research integrity, and research infrastructure. But RDM is hampered by the segmented treatment and the un-disciplined management of the research lifecycle. RDM fails when looking at it from a repository perspective and when considering it only when the data "is ready". ShanghaiTech University takes a lifecycle-based approach to RDM to ensure the context of, interaction with, and evolution of research data along the research process are managed and preserved. Based on a brief discussion of the challenges to real life RDM, the paper describes, as a work in progress, the underlining framework, and its four components of the lifecycle-based RDM: project-based RDM, active research data management, deposit and sharing of supporting data for research, and distributed preservation of layered research data.

Keywords: Research Data Management, Lifecyle-based, Data Preservation

1. Introduction

Research data management (RDM) is critically important for today's digitally-based and open research[1]. Many institutions have embarked on RDM[2,3] for management of research assets, integrity, public sharing, and a data-centric research infrastructure. In China, the State Council issued the Management Regulations on Research Data[4] in 2018, Chinese Academy of Sciences published its Regulations on Research Data Management and Sharing[5] in 2019, and a national system of 20 research data centers is formally set up as public resource sharing services[6] in 2019.

However, RDM is not only challenged by the velocity, volume, value, variety and veracity of data[7] and by the requirements of FAIR (Findable, Accessible, Interoperable, Reusable) principles[8], it is also troubled by the fact that the lifecycle of research data is often segmented between the research teams, lab data platforms, and institutional or disciplinary data centers, where each may follow different rules and standards and even different codes of conducts or acceptable behaviors. This may lead to proliferation of data silos, lack of consistent quality control, broken relationships among different data and between data and other research resources, difficulty in provenance and verification, difficulty in intellectual rights management, and lack of sustainable preservation, etc. A contextual and lifecycle-based approach is needed to ensure reliable, effective, economical, and sustainable RDM.

2. The lifecycle-based framework of RDM at ShanghaiTech

ShanghaiTech University[9] is a young research intensive university in the heart of Shanghai Pudong's Zhangjiang Hi-Tech Park, with an academic focus on STEM research, spanning from regular small-science labs in physical, life, and information sciences, to big science Cryo-EM clusters and Hard X-ray Free Electron Laser Facility. The types and scales of data are highly varied, researchers are working intensively collaborative, and processing and preservation of data are highly distributed.

To enable a university-wide reliable and efficient RDM, we started out planning with

[1] luqing@shanghaitech.edu.cn

a research-lifecycle-based approach with the lifecycle adopted from that of Princeton University[10], and related management services are designed accordingly (Figure 1).

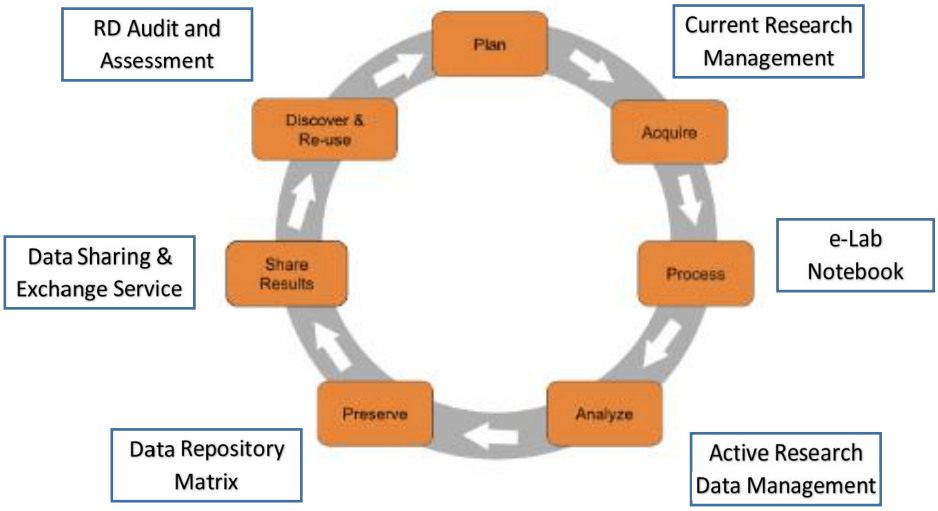

Fig.1 The Lifecycle-based Framework of RDM at ShanghaiTech

The framework is to make sure that all the phases in the lifecycle of research projects are covered. Adding to it, would be a set of policy and services components for a well-disciplined RDM system environment (Fig.2).

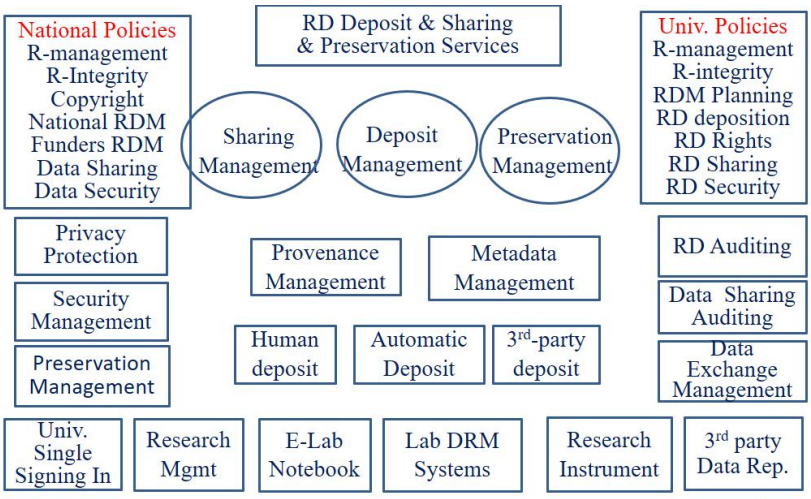

Fig 2. A set of policy and service components for a well-disciplined RDM system

3. Some Current Efforts to Develop the RDM system at ShanghaiTech

In this work-in-progress, current efforts have concentrated to develop policies and standards based on the national and disciplinary practices, to implement the e-Lab Notebook capability, to deploy active research data management, to build the university data deposit and sharing service for supporting data to published research, and to plan a distributed data repository matrix for long term preservation. The later four will be briefly described as follows.

Electronic Lab Notebook (ELN) capability for all the research teams is considered to be critical since it is the foundation to record the execution of research, to capture research data in real time, and to log in the parameters and processes of methods,

equipment, and operations. ELN comes in different types and shades, but a list of MUST, SHOULD, or MAY requirements is produced based on researcher input, including: the capabilities to model research plans and to adapt when needed; to record, versioning, and manage all the data produced with discipline specific metadata and formats; to track and log-in all the concerned parameters needed for provenance and integrity check; to provide fixity and backup for research process and its data; to interoperate with all the systems in university research infrastructure, including authentication and authorization, CRIS, Instrument Sharing and Data Transferring service, institutional repository of publications, data sharing services; to support FAIRed operation, etc. Currently, there are over 100 ELN software available[11] and ShanghaiTech has begun test use of some of them and the results will be evaluated to support a university-wide implementation decision.

Active Research Data Management (ARDM) is a robust data file storage service for active data in research, especially for those from scientific instruments produced by small science teams. It has a high throughput capacity and serves as a underlining or back-up service to the ELN servers of research teams, or those yet to have an ELN server. It will store larger raw data and processed data or backed-up data but it does so only temporarily during the execution of the research project. Once the project is complete, the raw data will be transmitted to the distribute data repository matrix and the support data to published results will be deposited in the service described below.

Data Repository for Support Data to Published Results (DR4SD) is to enable open sharing of the research data according to the requirements of funder and publisher policies and to support research integrity assessment when needed. It is to store supporting data for published research, including those directly related to figures and tables and other evidence-type data files. It will have enriched metadata and sophisticated right licenses to facilitate sharing, and it will transmit, on behalf of the university and its researchers, needed supporting data to journals and disciplinary or funder data repositories. Customized data portals for research teams, labs, programs, or schools, will be an easily set-up regular service. Auditing of data deposit and sharing will be in place to measure data impact and the policy compliance.

Data Repository Matrix (DRM) is a distributed but interrelated network of data repositories, managed by labs, schools, individual big science programs, and university. They fall under an unified system of division of labor and they all follow the coordinated and consistent sets of standards, including identification, basic metadata and metadata mapping, right licenses, data APIs, privacy protection, data safety and system security. A university-wide data registry, a federated discovery service, and a coherent access control, will be on top of the network. It in essence constitutes a data cloud of the university where data users don't need to know where and who the data is management. A data preservation strategy is in preparation[12], taking into consideration of the interdisciplinary research need and the vast volumes of big science research.

4. Future work

At the moment, the university information center is working closely with several research teams in 3 schools to test the ELN tools, and is cooperating with the Computer Networking Information Center (CNIC)[13] to implement its DR4SD based on a data repository product by CNIC. A guideline for ELN capabilities and a guideline for deposit and sharing supporting data to published research are already in

test use.

Lot of work is still ahead when goes at large scale and as daily & basic functioning. Disciplinary differences and unfamiliarity (unpreparedness or even unwillingness) of researchers with well-disciplined planning and management of research and RD are expected to be the strongest challenges. Politic will and policy dexterity, plus embedded, easy, customizable, and intelligent technical support, will help to realize RDM.

References

[1] Royal Society. Science as an Open Enterprise. https://royalsociety.org/topics-policy/projects/science-public-enterprise/report/

[2] Perrier, L., Erik, B., Patricia Ayala, A., & Dearborn, D. (2017). Research data management in academic institutions: A scoping review. PLoS One, 12(5), e0178261.

[3] The Realities of Research Data Management Part One: A Tour of the Research Data Management (RDM) Service Space https://www.oclc.org/research/publications/2017/oclcresearch-rdm-part-one-service-space-tour.html

[4] State Council. Management Regulations on Research Data. 2018. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm

[5] Chinese Academy of Sciences. Regulations on Research Data Management and Sharing. 2019. https://www.cas.cn/tz/201902/t20190220_4679797.shtml

[6] MoST. Notice on the New List of National Service Platforms of Scientific Resources http://www.gov.cn/xinwen/2019-06/11/content_5399105.htm

[7] The 5V of Big Data. https://www.jigsawacademy.com/blogs/big-data/5v-of-big-data/

[8] Wilkinson, M. D., Dumontier, M., Aalbersberg, *et al.* The FAIR guiding principles for scientific data management and stewardship. Scientific Data, March, 2016, https://www.nature.com/articles/sdata201618/

[9] ShanghaiTech University. https://www.shanghaitech.edu.cn/eng/

[10] Research Lifecycle Guide. https://researchdata.princeton.edu/research-lifecycle-guide/research-lifecycle-guide

[11] Kwok R. How to pick an electronic laboratory notebook. Nature 560, 269-270, 2018

[12] Preserving Research Data: Are you ready for a long-term commitment? https://blog.oclc.org/next/preserving-research-data-are-you-ready-for-a-long-term-commitment/

[13] Computer Networking Information Center, CAS. http://cnic.cas.cn/