# Supporting Large Datasets with Arch

Contact: Chris Diaz | Digital Publishing Librarian | [chris-diaz@northwestern.edu](mailto:chris-diaz@northwestern.edu)

**Background**  
Arch needs to support the deposit and unmediated access to large datasets to support faculty compliance with public access requirements from research funding organizations and scientific publishers. Arch is currently unequipped to deal with file sizes larger than 500 MB, presenting a barrier for potential users, labs, and research centers. For example, a dataset from [Northwestern Proteomics](http://proteomics.northwestern.edu) is composed of 126 image files that range from 400 to 600 MB each, for total size of 64GB. This project will address the storage, ingest, and delivery problems that prevent Arch from handling large datasets.

This is a change

**Proposed Workflow**

*Note: this workflow assumes that the user is using NUIT’s* [*Research Data Storage*](http://www.it.northwestern.edu/research/user-services/storage/research-data.html) *service for their working datasets. The result is to enable researchers to deposit their final datasets in Arch from NUIT’s data storage environment.*

1. User visits Arch and fills out metadata form to initiate the deposit
2. User is instructed to transfer a copy of the dataset from RDS to a library-managed Isilon share
3. Repository manager retrieves files from the Isilon share and deposits them to Fedora/AWS S3
4. Repository manager updates the deposit record with the location of the public dataset
5. End user requests the files from Arch, using a free [Globus](https://www.globus.org/) account or BitTorrent client that retrieves the dataset from the Fedora/AWS endpoint

**Project Milestones**

1. RDC: Upgrade to [Hyrax 2.0](https://github.com/samvera/hyrax/releases/tag/v2.0.0)
2. RDC: Move Arch to AWS
3. RDC: Create Work Type for Dataset with local metadata and location of large dataset
4. NUIT/ITI: Create an Isilon share for researchers to transfer copies of their data for deposit in Arch
5. RDC: Develop a process for side-loading large datasets into Fedora, accessible via Arch
6. RDC/ITI/NUIT: Implement Globus or BitTorrent to deliver large public datasets