Title	Data distribution
Author	*John Bradley*, Erich S. Huang, Jonathan Turner
Affiliation	https://www.genome.duke.edu
Contact	john.bradley@duke.edu
URL	https://github.com/Duke-Translational-Bioinformatics/duke-data-service
URL	https://github.com/Duke-GCB/DukeDSClient
License	GPLv3, MIT

The large amount of data created by modern genome sequencing has outstripped the practicality of file/block storage methods of data distribution. Issues related to data transfer, provenance and scaling have grown into costly problems. At Duke we have been working on a object store-based data distribution system to address these issues. The core of the system is a REST server that works with a swift object store.

Digital distribution of datasets from a producer to a consumer can be an imprecise process. Transferring ownership may result in copying the data from one storage location to another without any clear method of determining if the copy/transfer has completed successfully. This leads to maintaining multiple copies of the data and inefficient large scale comparisons. To this end we created functionality allowing researchers to transfer a project requiring the recipient to accept the project before granting access/ownership.

Lack of permanent unique identifiers can result in broken or wrong provenance chains when files are moved around. Using an object store can provide unique ids but still requires software to maintain the provenance metadata. To enable accurate provenance data we store the unique IDs and their provenance relationships in a relational database.

Using an object store instead of NAS provides incremental scaling of storage and provides for simpler administration. Instead of a few costly/complex pieces hardware you end up with a larger number of commodity computers.