NGDA Phase II Data Transfer Report
Alpana Pande and Katherine Kott for the
Stanford Digital Repository

In mid-August 2009, Alpana Pande launched the data transfer process for geospatial data ingested into the Stanford Digital Repository (SDR) during the second phase of the National Geospatial Digital Archive (NGDA) project. The transfers were planned to occur from August to December 2009 and would use the BagIt transfer protocol created by the Library of Congress in collaboration with NDIIPP partners, to move the data from the Stanford environment to the Library of Congress (LoC). The Bagit protocol had changed somewhat since phase I data had been transferred in June 2008, and the person who had been responsible for the phase I transfer was no longer working at Stanford. The first task was for Alpana to become familiar with the BagIt protocol and establish contacts at LoC. She also worked with the systems administrator who had facilitated the phase I transfer to understand the network transfer protocols and with the NGDA project manager and collections processing staff to inventory the collections that would ultimately be ingested, reconstructed, bagged, and retrieved by LoC.

By mid-October, Alpana had the first phase II bag; content from the Stanford Geological Survey (SGS) ready for LoC to pick up and by the end of the month the Stanford to LoC transfer process was in operation. The process allowed all the functions of the Stanford Digital Repository to be exercised. Lessons learned are informing the architectural design for the next phase of SDR and provide useful information on workflow, interoperability, and repository design for the preservation community.

As mentioned above, the BagIt protocol was improved from phase I of the NGDA project to phase II, based on phase I transfer experience. The changes made the BagIt process much more modular, which enabled BagIt to be de-coupled from the internal SDR reconstruction process. The decoupling made adjustments to the workflow much easier than phase I adjustments had been. Integration of BagIt within reconstruction during phase I meant code adjustments within reconstruction, complicating the process. Reconstruction is a predecessor process to BagIt that involves pulling files from storage and reassembling them in their original deposit form, checksumming, and comparing the checksum generated to the original checksum. Improved BagIt tools for validation, including the "verifyvalid" operation, which checks for fixity failures in the bag make the BagIt transfer process easy to use. Performing the "verifyvalid" check when the bag is created on the Stanford side and re-running the check on the LoC side once the bag is pulled from the Stanford server using rsynch indicates a successful transfer. Transfer details for two sample bags are included in Appendix I.

The bulk of the phase II content has been bagged and transferred to LoC. One collection remains to be ingested in the SDR, bagged and transferred. One other bag that contains content from a collection that was ingested early in the project may duplicate content that was transferred to LoC in phase I. Once the bulk of the phase II transfers have completed, Alpana will work with LoC to exercise the tools designed to detect duplicate content, determine what still needs to be sent, "BagIt" and send it.

This project has provided an excellent opportunity for SDR to exercise the complete ingest, reconstruction, and transfer workflow. Improvements to BagIt between phase I and phase II made the transfer process itself completely seamless. Separation of BagIt from reconstruction enabled us to see the SDR workflow clearly and identify improvements to implement in the next version of the repository.

Appendix I: Statistics Recorded during Reconstruction and Transfer

Object	Size	Online/Offline	Process	Elapsed time	Comments
SGS : ***NGDA-SULAIR2LOC-Bag-03***	38.9 GB	Online	Reconstruction	1hr 12m 14s	Log level set to "TRACE" increases processing time
			Bagging	3m 52s	
			Transfer (internal)	1h 14m	rsync
Rumsey : ***NGDA-SULAIR2LOC-Bag-09-1***	143.1 GB	Online	Reconstruction	3h 9m 9s	Log level set to "TRACE" increases processing time
			Bagging	21m 24s	
			Transfer (internal)	TBD	rsync