The Smithsonian Institution is the largest museum complex in the world. Its 19 museums, 9 research centers and the National Zoo contain over 130 million artifacts, specimens and/or works of art. Results of research at the Institution are published by hundreds of scientists, historians and other scholars every year. Recently the Smithsonian Institution Libraries (SIL) has begun collecting the digital publications of Institution researchers to ensure both the long-term care and public availability of these objects via the Smithsonian Digital Repository (http://si-pddr.si.edu).

Currently the Repository contains published reprints from scientific, peer-reviewed journals, including the Institution’s Smithsonian Contributions series. A limited amount of additional (ephemeral) material produced by the Institution is also archived in the system.

The open source software, DSpace was chosen because of it’s widespread adoption and user community. Among the specific requirements for the system were the use of persistent URLs and the ability to batch-ingest multiple items at once. Because of security concerns the system was designed so that content ingest is done on a server which is maintained inside the Institution’s firewall. A second, mirrored server is available for public search and data harvest. It is updated nightly via file synchronization and database dump/restore routines from the secure (internal-only) server. The public-facing server does not permit user login, edit or upload.

Development of the Repository was somewhat challenging at first due to a shortage of IT staff and support. Because the program was undertaken by the SI Libraries, the lack of technical knowledge of library staff required that a contractor install the initial hardware and software. A librarian then took computer-based training course(s) in UNIX, SQL and XML, among other modules in order to develop the skills necessary to maintain and manage the data. Many universities which begin digital repositories have a ready supply of student help who may be studying or have had training in server management, network architecture and the Unix/Linux environment. The Smithsonian does not have a similar pool of talent to draw from.

For both the internal and public servers, the Smithsonian’s Office of the Chief Information Officer maintains the operating system, updates and backups. All other applications (database, repository, utilities, etc.) are installed, maintained and updated by library staff. Due to this shortage of technical expertise devoted to the project, it has not been possible for the Smithsonian Digital Repository to migrate to latest version of DSpace or to employ many plug-ins such as statistics enhancements, which might make the service more appealing.

Prior to the development of the Repository, many museum scholars kept copies of their own publications on their office workstations, CDs or other servers. These items were not
adequately described or backed up and were at risk of loss. Most authors welcomed the idea of a centrally managed repository. However SIL staff soon realized from talking to librarians at other repository-hosting institutions that recruiting content from creators is among the biggest hurdles to populating a repository. It was clear that many scholars, while interested in the idea of a digital repository, do not yet seem prepared to begin the habit of regularly inputting their digital content. Learning from the experience of others, the SIL approached content recruitment with a different strategy.

Although there have been over 3000 items submitted in the first year, the vast majority of that content has been ingested to the system by Smithsonian Institution Library staff rather than by authors or creators of the digital objects. Typically authors send their electronic reprints to the SIL on CDs or via some other file transfer. Library staff search for and download corresponding metadata from several available resources including an Institution-wide research bibliography also managed by the library. Metadata and digital object(s) are matched and later uploaded in bulk. This approach imposes somewhat tedious labor on the library staff but it eliminates time that would have been spent on training (and re-training) scholars on the ingest process, etc.

Despite the IT and widely anticipated user-adoption challenges, the Institution did have some advantages over typical repository experiences. For example, most authors who contribute their reprints to the Repository are federal employees and the papers which they write are for the most part in the public domain. This has relieved the SIL staff from the task of investigating rights, licenses and other property restrictions which can influence content curation in many university-based repositories.

Normally success would be measured by the degree of staff participation and contribution in the repository by creators of digital content. But because authors are not required to enter digital objects themselves, this is not something that can be measured. It was hoped also that after a “critical mass” of content was entered into the system by library staff and a streamlined procedure for entering data could be developed, we would begin to see authors or their designees enter their own content as it is published or created. However, anecdotally the service is widely appreciated by Institution research staff. A more immediate measure of success—sustainable Institution funding—is yet to be seen.

The Smithsonian Institution Libraries, while lacking resources to develop a robust repository service, has nonetheless learned from the larger repository community regarding user adoption and decided to concentrate on content ingest as a primary service. Additionally, the Repository is closely related to the Institution-wide research bibliography, in which many citations are linked to their full text counterpart in the Repository using the item’s persistent URL. The Smithsonian Digital Repository serves as the copy-of-record for many Institutional publications where webmasters and other staff can be assured that the item will be preserved and the URL will remain persistent over time.

While the SIL’s approach may not work for many other institutions, the Smithsonian has realized modest success in exposing scholarly works to the general public and raising the
profile of research conducted at the Institution. As far as long term usability of these objects is concerned, only time will tell. However, given the wide dispersal of digital content prior to creation of the Repository and the fact that accidental erasure remains one of the most common threats to loss of digital content, it will undoubtedly become clear that the Repository merits the support of the Institution.