

Assistant Editor: Joanne Kaczmarek, University of Illinois. Contact Joanne at jkaczmar@illinois.edu if you would like to guest author a column or have a good idea to share.

Beyond Floppy Disks: Collecting Digital Content for Archives Today

by Bertram Lyons, AVPreserve

For the past decade, archives have focused on developing the internal capacity to store and manage digital records. Many methodologies exist for describing, managing, storing, and providing access to digital content in archives today.¹ With tools such as BitCurator, we also have methods to extract and analyze digital content from external storage devices acquired with archival collections such as floppy disks and hard drives. Building from these advances, and working with the Louie B. Nunn Center for Oral History at the University of Kentucky Library,² AVPreserve saw the need for a tool that could engage records creators in the activity of preparing and sending digital content directly to an archives. Together we turned our attention to addressing digital collections made available by donors and records creators by creating a tool called “Exactly.”

Enabling Content Creators and Donors

One of the main drivers behind the project to build Exactly was to enable donors to help archivists by preparing their content for acquisition in advance of depositing it with the archives. Sometimes living donors can work with archivists to prepare digital content for delivery and ingest into the archives, either directly or indirectly. But contemporary donors sometimes do not wish to part with their physical devices or they have little more to give to the archives than a set of files they select and deliver themselves. Exactly was built to enable donors and records creators to get their materials to an archives. The other central impetus to build Exactly was that there is simply not enough time for archivists to acquire everything by hand. Because of the growing mountain of digital content, we wanted to start thinking about (building) tools that can incorporate preservation techniques to ensure accurate documentation and acquisition, while simultaneously enlisting the help of the content creators.

The primary-user stories that led to the development of Exactly include the various needs of archivists who collect digital content from donors remotely. Archivists expressed these needs as follows:

- I want to be able to gather and prepare provenance information and establish fixity for files as early in the process as possible.
- I do not want donors to copy content to external drives or optical disks to send content to the archives. Instead,

I want them to send files directly via FTP or other file-sharing methods.

- I want to receive an e-mail with a list of files/checksums and a note about the details of the transfer once the donor sends the content to the archives.

Leveraging Existing Technologies

Exactly is a simple, free, and open source desktop application that can be used by archivists as part of their suite of tools during acquisition to address the above-mentioned needs. Exactly makes use of the following technologies:

1. BagIt Java library—The underlying packaging specification we use is the BagIt File Packaging Format.³
2. ftp4j—Exactly has a built-in FTP client to allow the tool to connect to servers using FTP to transfer files directly from the client computing environment to a remote server.
3. JavaMail—Exactly uses JavaMail to communicate to users and to designated recipients about the results of delivery activities.
4. Derby database—Internally, Exactly uses a Derby database to store profiles, e-mail credentials (encrypted), FTP credentials (encrypted), delivery metadata, and configuration information for the tool.

Exactly’s current central services include a broad array of functionality. The tool creates valid BagIt bags with additional event documentation and xml/csv baginfo versions. This documentation supports the requirements of establishing fixity information for digital files early in the acquisition process. It also generates a manifest that enables an archives to answer questions about file attendance, or the presence and/or absence of files within the acquisition. Last, this feature generates event documentation that assists with questions about the provenance of the acquired digital files.

Exactly also features user-defined metadata fields allowing archives to custom create information for donors to populate before acquisition. This feature assists with questions about provenance by enabling donors and records creators to document metadata about the content they are sending to the archives. It also provides the seed for descriptive metadata about the acquisition.

(Continued on page 28)

(Continued from page 27)

By default, Exactly creates a local copy of each acquisition package at a specified location. Because of this, Exactly is usable in internal networks (e.g., if a corporate archives wants divisions to deliver files directly through shared internal network drives) and locally (e.g., if an archives wants a donor to package files and move them to a local dropbox folder). Exactly also offers optional FTP delivery to remote servers. This allows donors and records creators to connect directly with the archives to deliver files remotely from anywhere in the world. It eliminates double-hop dropbox functionalities and encourages timely delivery from remote depositors, reducing the practice of copying files to CDs, flash drives, or hard drives to be shipped in the mail.

Exactly supports optional e-mail notification to specified recipients upon completion of successful transfer. The e-mail that is sent includes a manifest of files in the acquisition and a checksum for each file. This e-mail exchange communicates completion of the delivery from the donor to the archives, and it supports independent verification by the archives of the expected contents of the acquisition and the fixity of each file.

One important design element of Exactly is that it can be deployed on a user's local computer. Because it is not web-based, it can work more efficiently toward packaging and delivering large numbers and sizes of files. Exactly was built to be easy for donors to use (does not require complex installation) and includes builds for Windows, Mac OS, and a Java .jar package to make it as widely compatible as possible.

Another important functionality for Exactly is that it allows users (ideally the archives) to create, save, and share application configurations that can be exported as .xml files from one instance and imported by users into their own instances of the application. This means that an archives can send a donor a configuration file that will automatically load FTP information and credentials and e-mail authentication information and a required metadata set from the donor. Doing so makes it easier for the donor to point to the files to be transferred, fill out the required metadata fields, and initiate the transfer without having to do anything else. An archives can also create a variety of templates for separate use cases and donors.

Exactly features an Archives section to support quick validation of BagIt packages upon receipt from a donor.

This provides a quick check to assure the validity of the delivery before moving on to next steps. Exactly reports BagIt Java library results in the application's on-screen log. For archives that do not want to keep the BagIt structure upon receipt, Exactly has a simple unbagging option. Unbagging moves a bag to a designated local destination, validates the bag after the move, and then unpacks the bag.

The newest feature of Exactly is the ability to capture file system information from the original computer environment of the donor before the package is created. This feature acquires original directory paths, create/modify/access dates, file sizes, and file ownership information and stores the data as a text document with one row of information for each file in the package.

Future Development

Of course, there is always room for improvements, and we have already begun to receive requests. On the horizon is an upgraded FTP library to be more compatible with cloud service providers, as well as an addition to support SFTP for secure transfers. Information about Exactly can be found on the AVPreserve website.⁴ There is also an open user group where Exactly users discuss workflows and report bugs.⁵ Please be in touch if you have ideas for improving Exactly's features. Since Exactly is an open source tool, if you are so inclined you can get access to the GitHub repository.⁶

Notes (All websites accessed on June 2, 2016.)

1. For long lists of relevant tools and methodologies, see the COPTR website, coptr.digipres.org/Main_Page and see the "matrix" on the Digital POWRR project website at digipres.org/tools/ubergrid.
2. With support from colleagues at Gates Archive and StoryCorps, and some last minute feedback from MIT
3. The BagIt File Packaging Format is an Internet Engineering Task-Force standard originally developed by the Library of Congress and the California Digital Library with current support from George Washington University and the University of Maryland.
4. AVPreserve website, avpreserve.com/tools/exactly.
5. Exactly user group, groups.google.com/d/forum/exactly-users.
6. GitHub repository, github.com/avpreserve/uk-exactly.