DuraCloud DfR Narrative

Ver 1.0

February 1, 2012

Introduction

DuraCloud DfR is an open source software solution, with development initially funded by the Alfred P. Sloan Foundation, used to help researchers preserve their project data. The application will be offered as a managed service by DuraSpace, the non-profit organization whose mission is to provide preservation, archiving, and access solutions for scholarly, cultural, and research data by supporting community-driven, open source software projects.

Based on DuraCloud

The DuraCloud DfR project is an extension of DuraCloud designed to meet the needs of researchers and their institutional data managers. DuraCloud is now typically used by libraries to store scholarly content in the cloud for preservation. The DuraCloud service is deployed in the cloud and provides an application environment that serves as a front end in between the user and a number of back-end cloud providers, both public and private. Typically library users interact with DuraCloud, uploading copies of their content and indicating which of the supported cloud providers should be used for storage. Instead of negotiating with multiple cloud providers, users subscribe only once to the DuraCloud service. Users may easily move their content back and forth among cloud providers, view it, stream it, manage and transform it. In the background, DuraCloud runs services to ensure that the data retains integrity and remains valid and durable.

Protecting Research Data

DuraCloud DfR begins with the base DuraCloud service and adds capabilities that are designed to support the specific needs of researchers and data curators. The first goal of DfR is to enable a "safe" cloud storage infrastructure that automatically makes copies of research materials to a remote location, but still under the control of the researcher and/or institution. Remote copies serve the role of backup for research materials, particularly when sometimes only one copy exists on a computer under the desk.

So, for example, DfR will capture data that is created during the research project by PIs, research assistants, and other collaborators and will store it safely and transparently to the cloud, without interfering with the researcher's existing processes. Additionally, DfR will pull basic context and provenance metadata from the captured data and store it in a cloud-based repository for discovery and access.

Data in DfR will be protected through robust access control and optional encryption. Initially DfR will support Shibboleth authentication. DfR will be engineered to support a variety of eventual authentication mechanisms through a pluggable authentication architecture.

Data Management and Visualization

DuraCloud DfR will also integrate with secondary services that allow the researcher to visualize his data, add new metadata to create richer visualizations, and apply compute services to transform, manipulate, and analyze the data. While these secondary services are not part of the core preservation features, they will contribute increasing value to the researcher over time as new services are added.

Data Curation Workflow

Finally, DuraCloud DfR will provide PIs with the ability to pass data on to the institutional data management staff for further curation, archiving, and discovery.

Enterprise Subscriptions

Users will subscribe to DfR at the enterprise or departmental level. Central administrators (e.g., Office of Institutional Research) will allocate DuraCloud accounts to research departments, and/or departments will allocate accounts to individual PIs for their projects.

DuraSpace is working with the Internet2 Net+ Services team in the interest of making DuraCloud for Research available to InCommon and Internet2 members directly, using community cloud providers that are Internet2 connected.

Account Creation

First the PI or his assistant creates a user account for himself using the DfR management tool. Next the PI names and creates the project for which he is producing data. Optionally, the researcher then creates accounts for others working on the project and grants them permission for reading project data and/or contributing their own data to the project.

Automated Backup of Operational Data Sources

Research materials are not just data but include software, papers, notes, calibration records, email — really anything in digital form that the researcher uses. It can also include surrogates, like photos of physical samples. The researcher uses the DfR Monitor/Sync Configuration Tool to identify the working directories in his operational system that will be uploaded transparently to the DuraCloud archive. Each target platform—workstation, laptop, network drive, box.net directory, etc.—will automatically initiate a backup to DuraCloud when data has been added or changed. Collaborators will do the same to sync their working data with the DuraCloud data store for the project. Content will include texts, images, video, audio, and tabular datasets. However, operational storage may also be part of an instrument, software, HPC center or workflow system. In order to accommodate a variety of devices, the DfR project will document a specification making it possible for third party contributors to provide monitor/sync tools for numerous other sources of data.

Researchers will view a log file to confirm the status of their uploads to the DuraCloud archive.

Optional Restore

Data may be restored to the researcher's operational file system in case of loss on the local platform.

Repository in the Cloud

Data uploaded this way to DuraCloud will trigger an "object creation service" that will pull basic metadata from each datastream and package it for ingest into a cloud-based Fedora repository. Contents of the repository will also be backed up to DuraCloud.

Smithsonian Institution Partnership

In an early phase of DfR the researcher will use a prototype under development by the Smithsonian Institution's Office of Research Information Services to manage and visualize project data that is represented in the repository. Using a web-based UI, the researcher will be able to logically tie the digital assets of his project together by defining the relationships among them, adding new metadata as he does so.

In the Smithsonian application, "the 'home page' would be a conceptual object representing the project as a whole. It would include a formal title, descriptive annotations that explain different things about it, variety of metadata fields that classify it, describe people and institutions associated with the project, etc. The research project object could them assert relationships to media objects that contain such things as sets of notes, spreadsheets, texts such as grant proposals, essays, articles that result from the research, and digital information created for the project, such as images, audios, video, and tabular datasets" that have been uploaded.

The more metadata the researcher adds about his project, the easier it will become for him and others to understand and make use of it.

In the future, other services will be plugged in to manage, visualize, and manipulate project data. A variety of user interfaces will be built to act on the data, and a set of repository tools will evolve and be used to analyze and present the data. DuraCloud DfR will be architected to accommodate optional development of full-featured VREs

Researcher/Curator Workflows

The PI will have full control over the project data and will assign permissions to others to access and/or add to it. The PI may choose to partner with data curators during the course of the project. In that case, the PI may give library staff—or other support staff--permission to further document the project data, adding a variety of additional metadata.

Eventually, the PI will have the option to make a copy of the data fully available to the data curation staff, which may move it to an alternate repository, continue to enhance the metadata, or otherwise prepare it for

archiving and future use.

Timeline

The DuraCLoud for Research project is being developed using an iterative, agile methodology. Iterations will proceed throughout 2012, with three interim user-facing releases of working prototypes and a release at the end of the calendar year containing phase one functionality (including a working prototype of the Smithsonian Institution user interface).