PUTTING YOUR USER’S DATA IN THE CLOUD WITH DATAUp

Kristin M. Tolle, Ph.D.
Director, Environmental Science Infrastructure Development
Microsoft Research Connections
TRAINING AGENDA

• What is DataUp? (in brief)
• What’s new? (in brief)
• How do I become and admin and set up my own repository?
• How do I get the DataUp source code deploy my own instance?
• What is the admin/user experience based on selections I make in the UI once I set up my repository

Please ask questions as we go!!
DATAUP OBJECTIVES (CIRCA 2011)

• Create a tool designed to improve data reusability practices within the tool used most* by (environmental) scientists—Microsoft Excel

• Users stepped through a workflow of steps to create the most reusable data and place that in a repository. Designed to:
  • Encourage uploaders to follow “best practices” for data reusability
  • Encourage users to attach appropriate metadata (EML) to aid in reuse
  • Give users a citation so that they can appropriately cite their data
  • Post to a repository occurs only when (required) conditions are met

* Based on several surveys conducted by the California Digital Library, supported by an AGU survey taken by MSR
DATAUp V1

- Collaborative project with University of California, California Digital Library, The Gordon and Betty Moore Foundation, and Microsoft Research.
- Released October 2012 at DataONE Global Meeting—as this was a DataONE-centric project.
- Provided both an Excel Add-in and an Azure web service.
- Source code released on OuterCurve and BitBucket as Apache 2.0 open source license.
- Planned to support all DataONE repositories (via collaborator API). Ended up only supporting ONEShare (Merritt).
- Private funding concluded—application submitted for NSF funding (awarded August 2013—10 month gap)
INVESTIGATOR TOOLKIT SUPPORT

Plan

Collect

Assure

Discover

Describe

Preserve

Integrate

Analyze

MATLAB

Kepler

Digital Library Tools for Data Management

my experiment

ONE Share

zotero

DATAUP

DMP Tool
**What is Sequim??**

- Microsoft Research wanted to repurpose DataUp for internal needs (e.g., downstream data processing).
- We also interacted with many folks at AGU and other events and saw gaps in the DataUp experience for other repositories and organizations.
  - Need UI for administrators to manage their repositories
  - Need flexibility in metadata (EML is too limited—expand to Dublin Core and make dynamic)—support any type of scientist.
  - Need data validation to ensure that data will work in downstream processes and the Environmental Science tools (e.g., MatLab scripts, FetchClimate, Node Atlas, Distribution Modeller, etc.)
  - Need for Data Validation on a more atomic level.
  - Need to set up more repositories—including personal ones like ONEDrive where interim processing steps can be stored with auto-generated metadata
A FAMILY OF TOOLS FOR SCIENCE
SEQUIM => DATAUp V2

• Restarted collaboration with CDL Fall 2013.
• Released February 24th, 2014 in conjunction with IDCC 2014.
• NSF funded programmer started last week to do future updates in BitBucket directly.
• In discussions with DataONE to develop mechanisms for creating a “vibrant” open source contribution community.
• Replaced V1 on http://dataup.org site.
• Add-in deprecated (no planned future release).
BRASS TACKS

• DataUp was/is developed:
  • HTML5 frontend
  • C# backend developed and optimized to run on .NET and Windows Azure

• http://dataup.org is the California Digital Libraries (CDL) Azure instance of DataUp—not Microsoft Research

• https://bitbucket.org/dataup/dataup2 BitBucket repository is also managed by CDL

• But... you can stand up your own version ...more on this later
WHAT CAN I DO WITH THIS VERSION?

http://www.dataup.org/
REQUEST ADMIN ACCESS

DataUp is a data documenting, archiving and sharing tool that allows you to:

- Upload tabular data to your own repository in the cloud
- Perform a best practices check on your data
- Add metadata and descriptions to your data
- Share your data with others

Sign in with...

- Microsoft Account
- Facebook Account

Contact is the place to request access
Contact Us

For information on the DataUp project partners, visit About the Project.

We welcome your questions, comments, and/or feedback!

By email: uc3@ucop.edu
Via Facebook: www.facebook.com/DataUpCDL

Or use this feedback form:

Name: 

Email: 

Your questions, comments and feedback:

Presently this points to the CDL website set up for DataUp—which is appropriate since you are asking to use their instance of DataUp
SOME TERMINOLOGY

• **Repository type**: new repository you want to integrate into DataUp. DataUp v2 Feb 2014 release supports Merritt ONEShare and OneDrive out of the box.

• **Repository Adapter**: Software component that needs to be implemented for every repository type to facilitate bidirectional data transfer. This is how you make your repository available to your users.
HOW DO I STAND UP MY OWN INSTANCE?
**INITIAL STEPS**

• Sign up for a BitBucket account.

• Browse to the DataUp v2 repository on BitBucket located at [https://bitbucket.org/dataup/dataup2](https://bitbucket.org/dataup/dataup2)

• Fork the DataUp v2 repository under your account.

• Clone the forked repository from the DataUp v2 repository onto your local disk.

• Now you can start developing or deploying (as is) your own instance.
AUTHENTICATION CONSIDERATIONS

• What security scheme does the repository type conform to?

• DataUp v2 has two reference implementations:
  • Basic authentication (user name + password combo). ONEShare repositories are secured by this scheme.
  • Windows live authentication (token based auth). OneDrive repositories are secured by this scheme.

• If the repository leverages one of the above schemes, you can reuse the code to facilitate the new repository type.

• If not extend the existing AuthN scheme to implement the new mechanism.
DATA TRANSFER PROTOCOLS

• What data transfer protocol do I want the repository type conform to?

• All repositories in DataUp v2 follow HTTP/HTTPS protocol to transmit files and data.
**Next Steps**

- Add a new repository adapter by extending the repository adapter façade. (Requires implementing `IRepositoryAdapter` interface and extend `RepositoryAdapterFactory` class)

- Create (or reuse/customize if you support HTTP/HTTPS) the administrative user interface for admins to configure a new repository instance in DataUp (example, UI to provide uri supporting HTTP verbs GET, POST, PUT and DELETE).

- Re-use existing authentication scheme or implement new one to secure data transfer.
SHARING WITH OTHERS

• If you create a repository adapter that is specific to your repository you may wish to share that with others.

• After testing the changes on your local box, submit a “Pull” request to the DataUp v2 repository that you forked from.

• Wait for the repo owner (CDL) to accept or reject your pull request.
I’ve got the code, I’ve got my adapter. How do I stand up an instance that I own?
SHORTEST PATH

• DataUp source on BitBucket is optimized for .NET/Azure the easiest thing is to set up an Azure instance of your own and push the bits out there.

• First Step—get an Azure account:
  • (if academic) Apply for an Azure4Research one-year grant
  • (if not) Request a free 1-month subscription from
LONGER PATH

• Porting the code to another platform (e.g., Linux)
• ...Then standing it up on Azure 😊
• We’ll teach you how to set up a Linux VM at an Azure training near you http://research.microsoft.com/en-us/projects/azure/training.aspx
WHAT’S NEXT?

1. You must be an admin on the Azure instance
2. (recommended) HTTPS Certificate to secure web application and the API
3. Deployment script templates are checked into BitBucket
   https://bitbucket.org/dataup/dataup2/src/e26c62b12c6f935190fc5db47076f905b13d3f81/Deployment/?at=master
4. Publish to Azure
A CLOSER LOOK AT STEP #3

• Create a new ACS namespace under Azure Active Directory Services (or use existing)

• Configure Identity providers. DataUp2 (presently) supports Microsoft Account (Windows Live Id) and Facebook:
  • Configuring Facebook requires registering the deployment with Facebook developer network.
  • No additional configuration required for Windows Live

• Configure DataUp2 API as a “Relying party” application

• Create new SQL Azure Database

• Create new Azure blob storage account (use an existing)

• Update the configuration in DataUp2 deployment scripts with the keys provided in the Azure subscription (requires some technical knowledge and working with XML)

• Package the deployment scripts
CREATE A NEW ACS NAMESPACE UNDER AZURE ACTIVE DIRECTORY SERVICES
CONFIGURE IDENTITY PROVIDERS
# Relying Entity Partner Setup

## Identity Provider Settings

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Display name</td>
<td>Enter a display name for this Facebook application. This name is used in the ACS Management Portal only.</td>
</tr>
<tr>
<td>Application ID</td>
<td>Enter your Facebook Application ID.</td>
</tr>
<tr>
<td>Application secret</td>
<td>Enter your Facebook application secret.</td>
</tr>
</tbody>
</table>

## Application Permissions

Enter a comma-separated list of permissions to request from Facebook.

## Login Page Settings

<table>
<thead>
<tr>
<th>Setting</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Login link text</td>
<td>Enter the text to display for the login link for this Facebook application.</td>
</tr>
</tbody>
</table>

## Used By

Select any existing relying party applications that you want to associate with this identity provider.

- dat eup.org
- dat eup-client
- dat eup-client dev
- localhost:54238

[Image of the setup page with the ACS Identity Provider settings and options]
CREATE A SQL AZURE DATABASE
SETUP AZURE BLOB STORAGE
**What are SQL and the Blob Storage For?**

- As you know, DataUp doesn’t hold files. We want people to put them in the repository and we do not want to pay for storing them twice.
- SQL holds the file properties (e.g., owner, location, status, etc.) and is used to populate the UI that users see.
- Blob storage (used ‘as needed’) is the temporary storage location for the files users upload.
New Admin UI Experience
WHAT CAN I DO AS AN ADMIN?

- Files—same as users
- Add/edit repositories
- Create data validation rules—repository independent!!
# Managing Repositories

## Repositories

<table>
<thead>
<tr>
<th>Repository Name</th>
<th>File Type</th>
<th>Visibility</th>
<th>Date Created</th>
<th>Created By</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Repository</td>
<td></td>
<td></td>
<td>2/25/2014 11:27:18 AM</td>
<td>Susan Borda</td>
<td></td>
</tr>
<tr>
<td>deptest_merritt.imp02</td>
<td></td>
<td></td>
<td>2/13/2014 2:50:08 PM</td>
<td>Sequim Test</td>
<td></td>
</tr>
<tr>
<td>OneShare</td>
<td></td>
<td></td>
<td>2/11/2014 4:15:39 PM</td>
<td>Susan Borda</td>
<td></td>
</tr>
<tr>
<td>deptest_sd_imp</td>
<td></td>
<td></td>
<td>2/11/2014 1:32:32 PM</td>
<td>Sequim Test</td>
<td></td>
</tr>
</tbody>
</table>
THINGS TO NOTE

• Admins of *an instance* of DataUp can see all of the “visible” repositories in that instance—even if they are not visible to users.

• Three major functions on the page that an admin can do:
  • Add repository
  • Edit an existing repository
  • Remove a repository
# Visibility of a Repository

<table>
<thead>
<tr>
<th>Repository Name</th>
<th>File Type</th>
<th>Visibility</th>
<th>Date Created</th>
<th>Created By</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Repository</td>
<td></td>
<td></td>
<td>2/25/2014 11:27:18 AM</td>
<td>Susan Borda</td>
<td></td>
</tr>
<tr>
<td>depetest_merritt_imp02</td>
<td></td>
<td></td>
<td>2/13/2014 2:50:08 PM</td>
<td>Sequim Test</td>
<td></td>
</tr>
<tr>
<td>OneShare</td>
<td></td>
<td></td>
<td>2/11/2014 4:15:39 PM</td>
<td>Susan Borda</td>
<td></td>
</tr>
<tr>
<td>depetest_sd_imp</td>
<td></td>
<td></td>
<td>2/11/2014 1:32:32 PM</td>
<td>Sequim Test</td>
<td></td>
</tr>
</tbody>
</table>
Users can upload single, multiple or zip files.
# Editing a Repository

## Repositories

<table>
<thead>
<tr>
<th>Repository Name</th>
<th>File Type</th>
<th>Visibility</th>
<th>Date Created</th>
<th>Created By</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Repository</td>
<td></td>
<td></td>
<td>2/25/2014 11:27:18 AM</td>
<td>Susan Borda</td>
<td></td>
</tr>
<tr>
<td>deptest_merritt_imp02</td>
<td></td>
<td></td>
<td>2/13/2014 2:50:08 PM</td>
<td>Sequim Test</td>
<td></td>
</tr>
<tr>
<td>OneShare</td>
<td></td>
<td></td>
<td>2/11/2014 4:15:39 PM</td>
<td>Susan Borda</td>
<td></td>
</tr>
<tr>
<td>deptest_sd_imp</td>
<td></td>
<td></td>
<td>2/11/2014 1:32:32 PM</td>
<td>Sequim Test</td>
<td></td>
</tr>
</tbody>
</table>
Name, Type, User Login

Edit Repository

Repository Name: KrisTest_sd_imp
Repository Type: Merrill
      SkyDrive

User Agreement

File Type: xskcevshpnc.cdf
Visibility: Visible to Admin

Create Metadata

Add Item
Get Data from XML
### File Types, Visibility

**Repository Name**: depres3_kup

**Repository Type**: SkyDrive

**Impersonate**: 

**User Agreement**: 

**File Type**: xlsxlxtxtxtxtxtxtxtxtxtxtxtxtxtxtxtxtxtxtxtxtxtcsvcdflncshhp

**Visibility**: Visible to All

---

#### Create Metadata

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Field</th>
<th>Description</th>
<th>IS-Required</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Name</td>
<td>First Name</td>
<td>enter your First Name</td>
<td>x</td>
<td>Text</td>
</tr>
<tr>
<td>Last Name</td>
<td>Last Name</td>
<td>enter your Last Name</td>
<td></td>
<td>Text</td>
</tr>
</tbody>
</table>
Create Metadata

Click on +Add Item or Get By XML button to provide Meta Data Fields
### Create Metadata

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Field</th>
<th>Description</th>
<th>IS-Required</th>
<th>Type</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Name</td>
<td>First Name</td>
<td>enter First Name</td>
<td><strong>✓</strong></td>
<td>Text</td>
<td></td>
</tr>
<tr>
<td>Last Name</td>
<td>Last Name</td>
<td>enter Last Name</td>
<td><strong>✓</strong></td>
<td>Text</td>
<td></td>
</tr>
<tr>
<td>DateTime</td>
<td>Date and Time</td>
<td>enter the Date/Time</td>
<td><strong>✓</strong></td>
<td>DateTime</td>
<td></td>
</tr>
<tr>
<td>Latitude</td>
<td>Latitude</td>
<td>if applicable enter the Latitude</td>
<td></td>
<td>Numeric</td>
<td>0 to 90</td>
</tr>
<tr>
<td>Longitude</td>
<td>Longitude</td>
<td>applicable enter the Longitude</td>
<td></td>
<td>Numeric</td>
<td>0 to 180</td>
</tr>
</tbody>
</table>

[Select options and submit]
WHAT AN END USER SEES
**METADATA IS REPOSITORY SPECIFIC**

Some repositories may require additional metadata to be persisted along with the data. Following section(s) cover required and optional fields that need to be filled in before submitting the file to the selected repository.

- First Name
- Last Name
- Date and Time
- Latitude
- Longitude

Values between 1.90

Values between 1.180
Basic Type Validation

Some repositories may require additional metadata to be persisted along with the data. Following section(s) cover required and optional fields that need to be filled in before submitting the file to the selected repository.

- **First Name**: [Input Field]
- **Last Name**: [Input Field]
- **Date and Time**: 123-123-1234
  
  Please specify a valid date in the YYYY-MM-DD format.

- **Latitude**: 60
  
  Please enter a value between 1 and 90.

Values between 1-90
Some of the potential issues listed here must be manually altered to comply with best practices. Download modified files to fix errors. You can ignore these errors and continue posting.

<table>
<thead>
<tr>
<th>Potential Problem</th>
<th>Why this may be a problem</th>
<th>Suggested Remedy</th>
<th>Location(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merged cells</td>
<td>Merged cells will not be maintained when data are exported as a .csv file. Information may be lost when cells are unmerged upon export.</td>
<td>Unmerge cells and annotate appropriately so information is not lost.</td>
<td>B24:C24,</td>
</tr>
<tr>
<td>Commas</td>
<td>Commas are often used to separate multiple piece of information/data (e.g., City, State). Cells should have only one piece of information.</td>
<td>Split pieces of information into multiple columns (e.g., City column and State column)</td>
<td>C8:B24</td>
</tr>
</tbody>
</table>
Some of the potential issues listed here must be manually altered to comply with best practices. Download modified files to fix errors. You can ignore these errors and continue posting.

<table>
<thead>
<tr>
<th>Potential Problem</th>
<th>Why this may be a problem</th>
<th>Suggested Remedy</th>
<th>Location(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commas</td>
<td>Commas are often used to separate multiple piece of information/data (e.g., City, State). Cells should have only one piece of information.</td>
<td>Split pieces of information into multiple columns (e.g., City column and State column)</td>
<td>C8,824</td>
</tr>
<tr>
<td>Special characters</td>
<td>Special characters may cause problems for other programs or may be modified upon export</td>
<td>Use alpha-numeric characters only. If needed, describe the symbol in a new column</td>
<td>B2:C4,G8:C13,C20,824</td>
</tr>
</tbody>
</table>
Creating Data Validation Rules

Add New Rule

Rule Name: Node Atlas

Description:
Files that are processed in Node Atlas need a minimum of DateTime, Latitude and Longitude in any order.

Header Names

<table>
<thead>
<tr>
<th>Header Name</th>
<th>Type</th>
<th>Range Start</th>
<th>Range End</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>Numeric</td>
<td>0</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>Longitude</td>
<td>Numeric</td>
<td>0</td>
<td>180</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>None</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Add Header Name
Manage Quality Check Rules

<table>
<thead>
<tr>
<th>Rule Name</th>
<th>Description</th>
<th>Visibility</th>
<th>Date Created</th>
<th>Created By</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node Atlas</td>
<td>Files that are processed in Node Atlas need a minimum of DateTime Latitude</td>
<td></td>
<td>2/26/2014 2:08:47 PM</td>
<td>Sequim Test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and Longitude in any order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QARule0212</td>
<td>QARule0212</td>
<td></td>
<td>2/11/2014 9:50:23 PM</td>
<td>Sequim Test</td>
<td></td>
</tr>
<tr>
<td>masspec</td>
<td></td>
<td></td>
<td>2/6/2014 11:06:13 AM</td>
<td>Sequim Test</td>
<td></td>
</tr>
</tbody>
</table>
QUALITY CHECK USER EXPERIENCE (ALL FILE TYPES)
Citation

Title: excel_good_example2
Version: v1
Publisher: Tolle, Kristin
Publication Year: 2014
TERMS AND CONDITIONS
WHAT HAPPENS NEXT?

- Posted or not, the actual file is deleted from the Azure instance to reduce storage costs.
- Timing is shown so users can see where they are in the “timeline” to delete.
- 24 hours to delete, the expiration time is bolder and in red (unless posted).
- If posted the file remains visible in the file store and users can still download that version (from the repository) including the citation and metadata.
QUESTIONS??