



Automating data publication and discovery with Globus

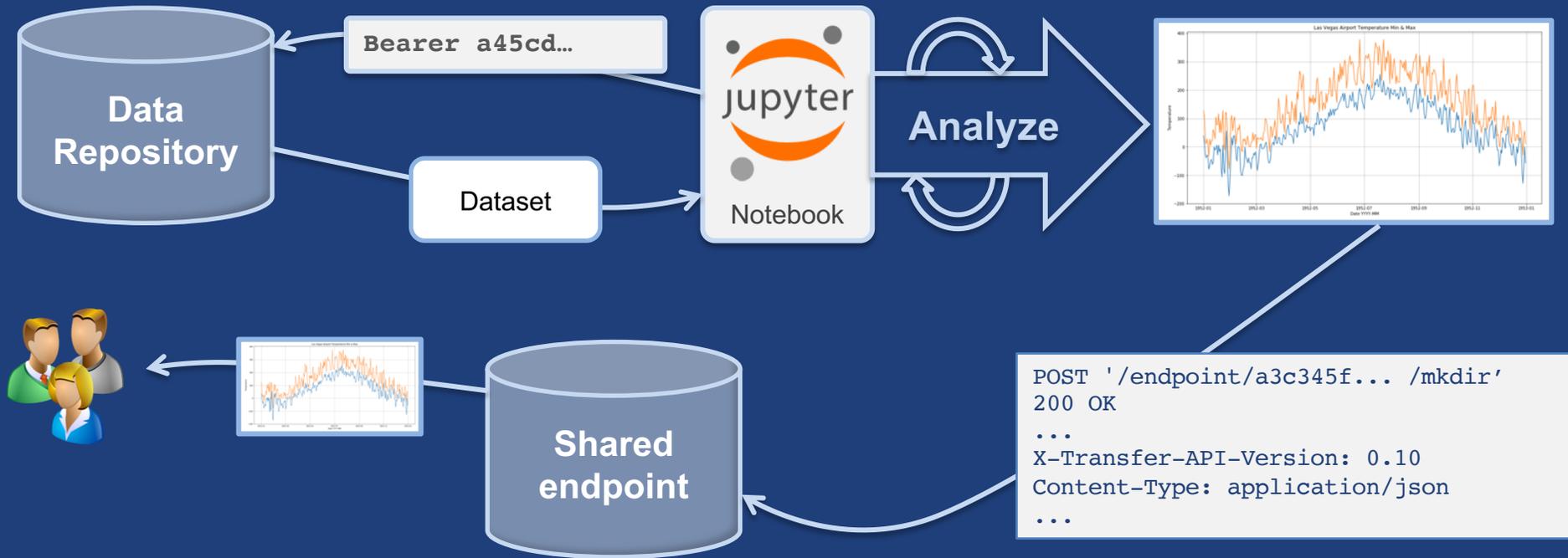
Vas Vasiliadis
vas@uchicago.edu

NCAR – September 5, 2018





Recall our simplistic data flow...



- Adequate for *ad hoc* sharing (implicit knowledge)
- Broader access, reuse requires “formalization”
- Leverage Globus data publication services



Globus Data Publication V1

SaaS publication

BYO Storage & in-place publication

User-managed collections

Arbitrary metadata (with pre-defined schema)

Handle, DOI PIDs

Adoption since 2015:

>1800 users, >600 datasets

The screenshot shows the Globus Data Publication dashboard. At the top, there is a navigation bar with the Globus logo, 'Manage Data', 'Publish', 'Groups', 'Support', and 'Account'. Below this is a secondary navigation bar with 'Browse & Discover', 'Data Publication Dashboard', and 'Communities & Collections'. A search bar is located below the navigation. The main content area features a heading 'Globus Data Publication' and a brief description: 'Globus simplifies the publication and discovery of research data. Use Globus to describe, curate, and preserve data at desired levels of durability. Make your data easily accessible to fellow researchers and other interested parties who can search and browse published datasets.' To the right of the text is an icon representing data storage and discovery. Below the text is a link: 'Click here to learn how to publish data. More information on how Globus data publication works is available here.' A link for 'Try a free trial of Globus data publication.' is also present. The dashboard is divided into four columns: 'Communities', 'Discover', 'Subject', and 'Issue Date'. Each column contains a list of items with associated counts.

Communities	Discover	Subject	Issue Date
<ul style="list-style-type: none"> BD2K Data Repository Big Data to Knowledge Centers Globus Materials Data Facility Community for the Materials Data Facility Collaboration National Data Service RDCEP Center for Robust Decision Making on Climate and Energy Policy 	<ul style="list-style-type: none"> GGCMI 57 McInerney, D. J. 35 Moyer, E. J. 35 Sun, S. 35 Moyer, E.J. 18 Schwarzwald, K. 18 Zhorin, V. 18 Hersam, Mark C. 13 Voorhees, Peter W. 10 	<ul style="list-style-type: none"> Agricultural Impacts 57 Climate Change 57 Climate Impacts 57 CO2 effects 57 Farm system models 57 Food Security 57 Model Intercomparison 57 climate 18 cmip5 18 	<ul style="list-style-type: none"> 2018 33 2017 49 2016 138 2015 2 2014 2 2013 1

Publication V2: A platform for automation

- **Decompose data publication v1 into platform services**
- **Facilitate flexible re-composition, adaptation by customers**
- **Enable extension and enhancement**
- **Initial services**
 - Search, identifiers (and data management)
- **Future services**
 - Description (metadata), flows



Globus Search

- **Scalable service** → billions of entries
- **Schema agnostic**: use standard (e.g. DataCite) or custom metadata
- **Fine grained access control**: only returns results that are visible to user
- **Plain text search**: ranked results
- **Faceted search**: facilitates data discovery
- **Rich query language**: ranges, expressions, regex, etc.

docs.globus.org/api/search



Globus Identifiers



- **Service for issuing persistent identifiers**
 - DOI, ARK, Handle, Globus
 - e.g. <https://identifiers.globus.org/doi:10.1145/2076450.2076468>
- **Within a namespace, e.g. your DataCite namespace**
 - Control which identities/groups can create identifiers
- **Each identifier has...**
 - **Link to data:** one or more https URLs, to file, folder or manifest
 - **Landing page:** provided by service, or by user
 - **Visibility:** identities, groups that can see identifier
 - **Checksum:** of the file or manifest
 - **Metadata:** as required by identifier (e.g., DataCite), extensible
 - **Replaces/replaced-by:** for versioning





Extending the automation flow

- How can we automate a data publication flow using Globus platform services?

