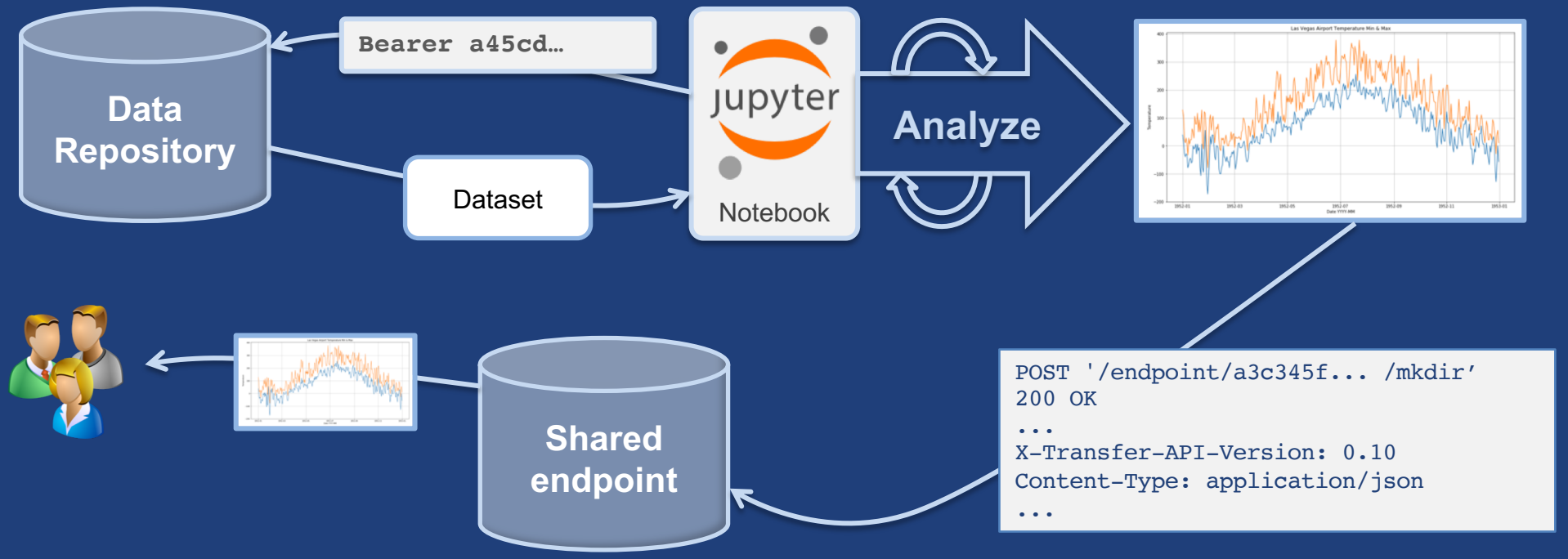
Automating data publication and discovery with Globus

Vas Vasiliadis
vas@uchicago.edu

NCAR – September 5, 2018

globus

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

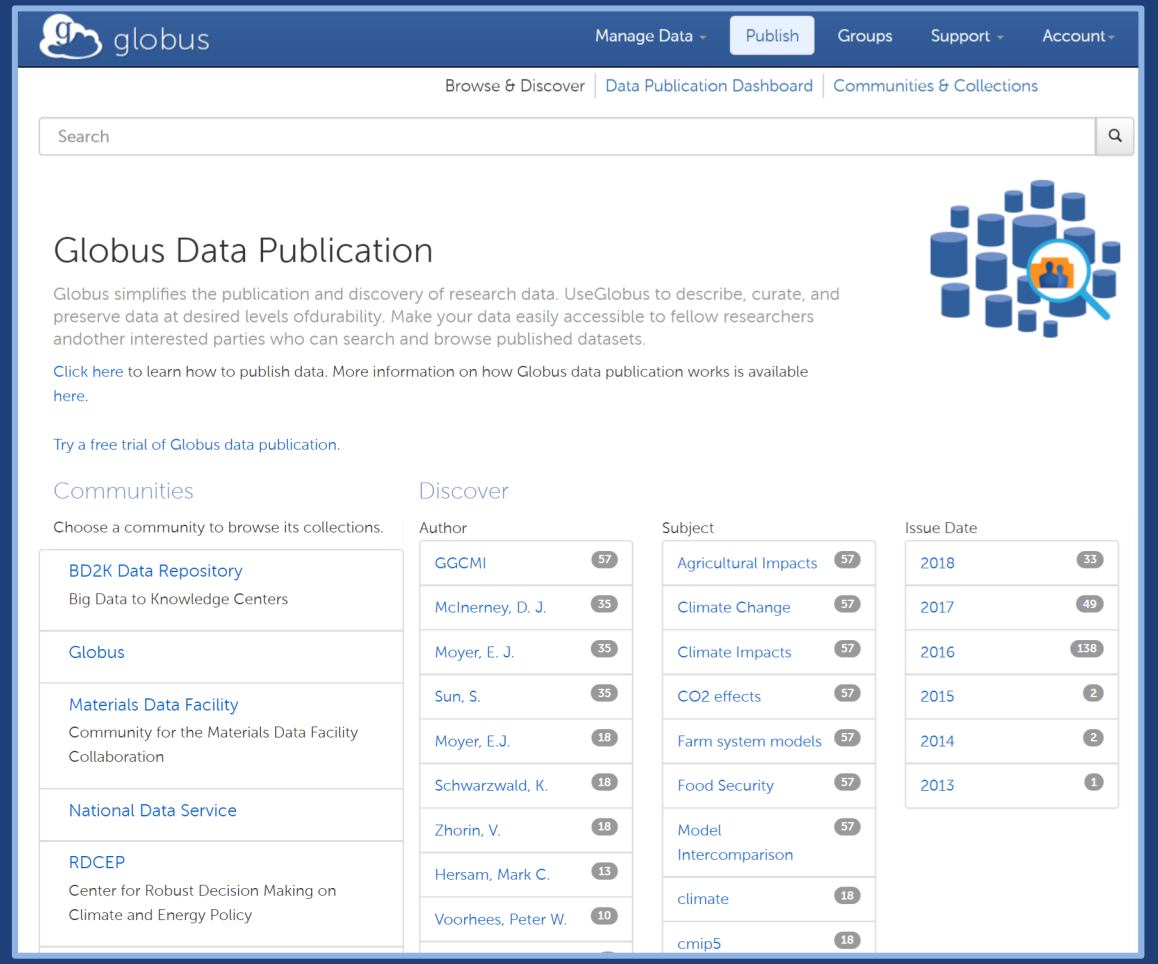
User-managed collections

Arbitrary metadata (with pre-defined schema)

Handle, DOI PIDs

Adoption since 2015:

>1800 users, >600 datasets



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?**