World Data System

Domain Specific *versus* Generalist Repositories

Alex de Sherbinin, PhD

Associate Director, CIESIN, Columbia University

Deputy Manager, NASA SEDAC

Chair, Scientific Committee of the World Data System

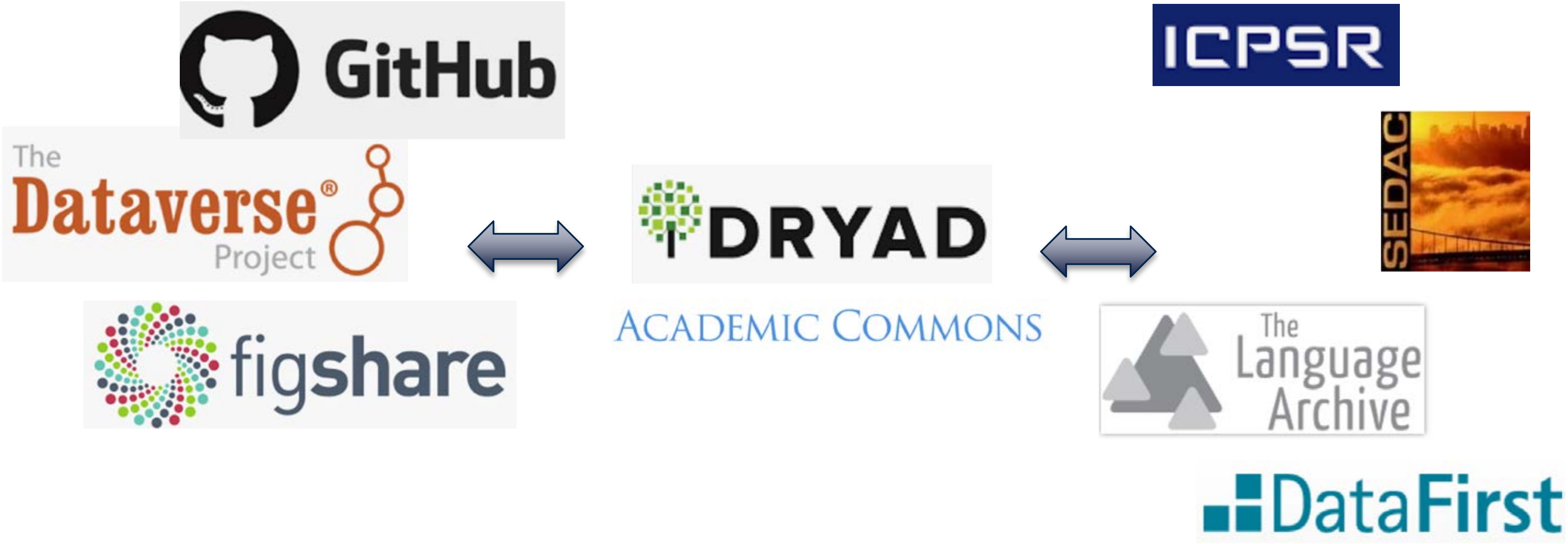
WORLD DATA SYSTEM

Domain repositories

WDS membership and CoreTrustSeal accreditation are targeted to domain specific repositories

Repositories are expected to have a mission to serve a **specific scientific community**, to engage representatives of those communities in **advisory groups**, and to use **domain knowledge** to **curate and serve** data

Spectrum from generalist to domain



- Many generalist repositories also run underlying infrastructure for domain repositories (e.g. Dataverse, CKAN)

Generalist repository risks

- Data could be lost (depending on their sustainability/commitment)
- For users:
  - Reduced discoverability
  - Relevant data are scattered
  - Self-archived data documentation are often insufficient for end users
  - Potential lack of QA/QC of data
- For domain repositories:
  - Propagate perception that domain expertise and peer review of data are not needed
  - Potentially undercut sustainability of domain repositories

Case Study: NASA Socioeconomic Data and Applications Center (SEDAC)

- The open science movement – in itself an important advance that SEDAC supports – has made the publication of data along with journal articles an increasingly accepted norm
- DOIs are unique, global, persistent identifiers for a given version of a data set
- In the past, SEDAC would approach authors to discuss dissemination of those data; today, by the time an article is published the data are often made available via generalist repositories with DOIs
- We even find that many of the submissions through SEDAC's submissions page are for data sets that already have DOIs
- This prevents SEDAC from disseminating the data under its own DOI
- SEDAC response:
  - Educate authors and work with them upstream in the process
  - Learn from authors and explore domain repositories to understand better why they are attractive, and potentially adopt some of their approaches
  - Have SEDAC listed by journals as a preferred repository

Potential responses

- CoreTrustSeal is being approached by generalist repositories for accreditation; it may create a separate category for them
- Domain repositories can work with journals to be listed as repositories of choice in given fields
- Domain repositories can make the case to authors that if they deposit with them:
  - There will be higher levels of data (and journal article) citation owing to enhanced discoverability or prestige
  - Reputation of the author is likely to be enhanced
  - Data reuse will increase (including by researchers outside their domain)
  - Having the data in a larger ecosystem of similar data makes it easier to develop services based on the data
- Domain repositories could just acknowledge that there are more data being published than can be archived by any one repository, and that this trend is okay

WORLD DATA SYSTEM

Discussion questions

- Do you see this as an issue – and if not, why not?
- If you do see it as an issue, has it affected your repository?
- If so, what has been your response so far?
- Do you have thoughts on future responses?
- How might WDS and CoreTrustSeal address the issue / help?