Setting up a data repository, what does it entail

Indira Yerramareddy and Nilam Prasai
International Food Policy Research Institute (IFPRI)

April 8, 2019

Data Repository: The What

- IT infrastructure (cloud based/online) set up to manage, share, access, maintain, and archive datasets

- An application database specialized in storing metadata of data files/datasets/databases

- Differs from publication repository mainly in its ability
  - to store metadata at different level/hierarchy
  - to store and ingest data files in various formats for long-term preservation

IFPRI

Data Repository: The Why

- Easy information discovery
- Easy and efficient access
- More exposure and amplify impact
- Persistent access (through persistent URL)
- Long-term storage and preservation

Additionally,

- Enables unprecedented use, analysis, and discovery through interoperability and interlinking with other repositories
- Enhances Open science scholarship

IFPRI

Data Repository: The Software

- Data specific – Dataverse, Comprehensive Knowledge Archive Network (CKAN), HUBzero
- Expanded from text-based institutional repository – DSpace, Fedora, Hydra
- Open source – Dataverse, HUBzero, Dspace, Fedora, CKAN
- Proprietary/commercial – Figshare, Interuniversity Consortium for Political and Social Resource [ICSPR], Digital Commons

IFPRI

Data Repository Software: Limitations

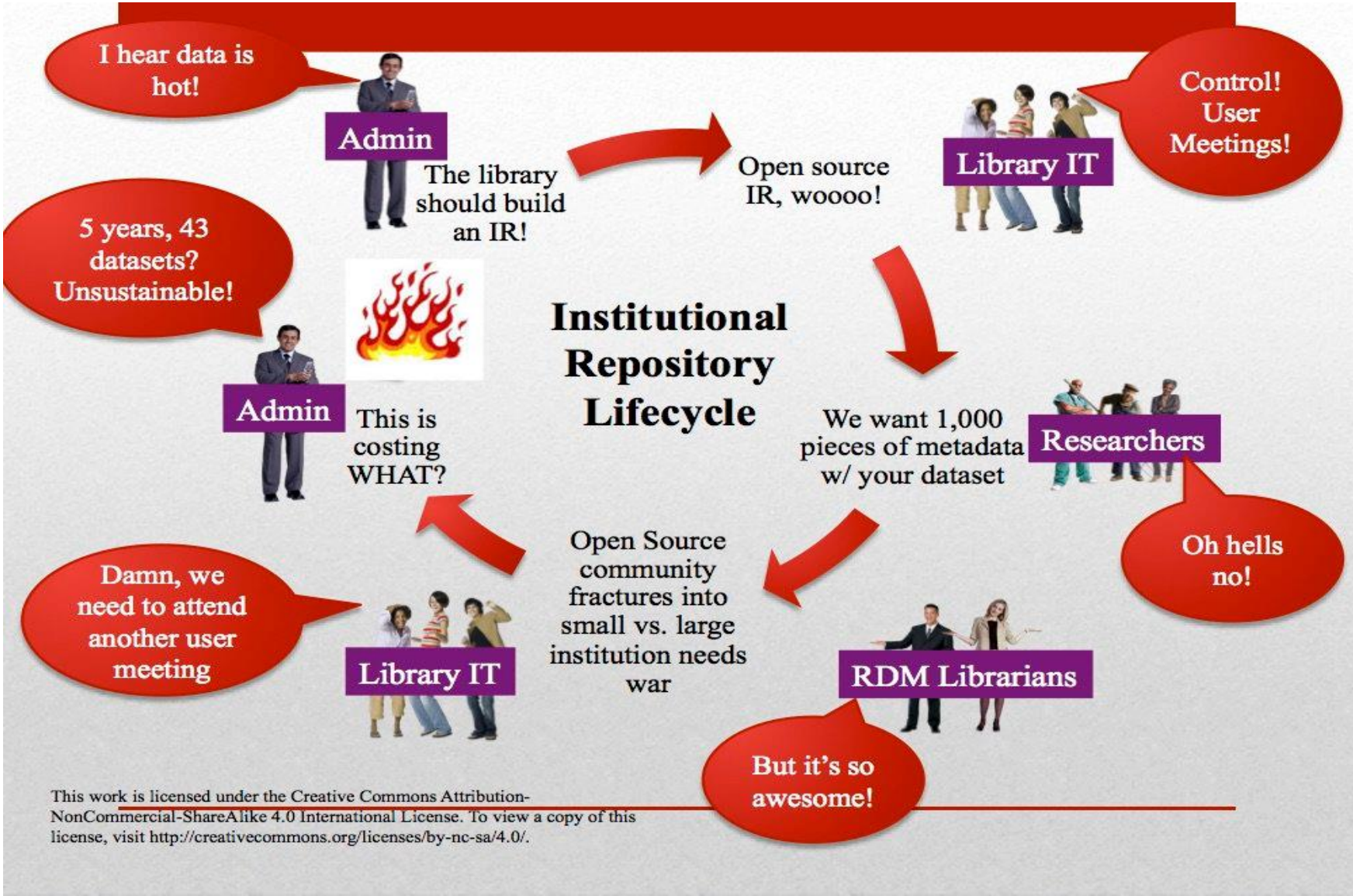
Table 1: Limitations of the identified repository solutions. **Source:** ▽OpenDOAR platform △Corresponding website. †Only available through additional plug-ins. *Only partially.

	Registered repositories▽	Closed source	No API	No unique identifiers	Complex installation or setup	No OAI-PMH compliance
CKAN	139△			×†		×*
ContentDM	53	×				
Dataverse	2					
Digital Commons	141	×	×			
DSpace	1305					
ePrints	407			×†		
EUDAT	—	×*				
Fedora	41				×	
Figshare	—	×				
Greenstone	51		×	×	×	
Invenio	20					
Omeka	4			×		×†
SciELO	18	×				
WEKO	40			No data		
Zenodo	—					

Source: Amorim R.C., Castro J.A., da Silva J.R., Ribeiro C. (2015) A Comparative Study of Platforms for Research Data Management: Interoperability, Metadata Capabilities and Integration Potential. https://doi.org/10.1007/978-3-319-16486-1_10

IFPRI

Data Repository: Hosting??



IFPRI

Data Repository: Hosting or Partnering

- Cost
  - Free ??
  - Direct and ongoing cost of purchasing and keeping sever or cloud space
  - Ongoing cost for maintenance
- Inhouse technology infrastructure and expertise
- Number of datasets
- Size/volume of data
- Sustainability

IFPRI

Data Repository: Selecting partners

- Reputation
  - Is the repository service provider reputable
  - Does the repository demonstrate acceptance and usage in the field of its scholarship
- Visibility
  - How will other discover datasets
  - Is the repository indexed by google and other databases
- Sustainability
  - Does the repository have long-term data management plan
  - Does the repository have contingency plans
- Cost
  - Free
  - Direct cost for depositing datasets
  - Indirect cost for maintaining and preserving datasets
- Persistence identifier
  - Does the repository assigns persistence identifier (PID) for datasets/datafiles

IFPRI

Data Repository: Selecting partners cont…

- Open Archives Initiatives Protocol for Metadata Harvesting (OAI-PMH)
  - Is the repository OAI-PMH compliant
- Policies and licenses
  - Does the repository allow data use agreements and licensing to state explicitly what uses the data owners will allow
  - Does the repository allow to set open, closed or restricted access
  - Does the repository allow creative commons licensing
- Scholarly impact
  - Does the repository track data citation and impact
  - How does the repository track downloads and usage
  - Does the repository allow integration with third party impact tracking software (Altmetrics, Plum etc.)
- Support services
  - Is the support service of the repository well-established and quick

Preferred,

- Recognized by the FAIRsharing.org

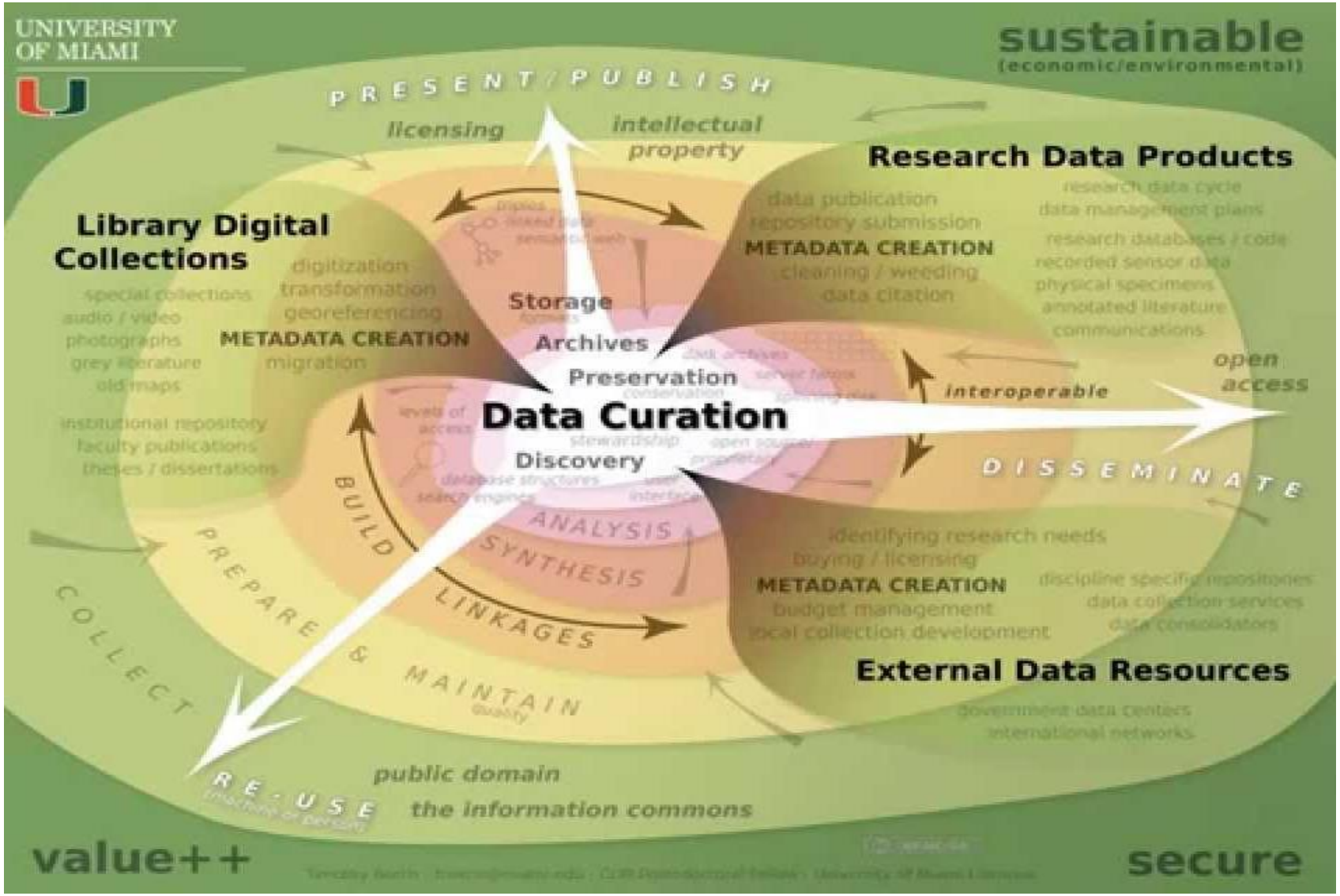
IFPRI

Data Curation: A Crucial Function for Successful Repository

- What is data curation
  - active and ongoing management of data through lifecycle of interest and usefulness
  - processes and activities related to organizing, **describing**, cleaning, **annotating**, enhancing and **preserving** data for public use.
  - centered around maintaining and managing the metadata rather than the data itself [documenting]
  - directly linked with preservation and maintenance, interoperability, and reuse
- Begins with the research lifecycle
- Requires change in culture/practice

IFPRI

Data Curation Mountain: Understanding the concept



Dataverse: The What

- Harvard's open source research data repository software

Features

Data Citation
automatically generated

Multiple Publishing Workflows
dataset in draft, in review, and then published

Terms of Use + Guestbook
CC0 waiver default, custom terms of use, and download metrics

Account + Data Notifications
access request, roles granted, and when data is published to name a few

Faceted Search
metadata fields based facets

Pull header metadata from Astronomy (FITS) files

APIs for interoperability
search API, data deposit API

Shibboleth
single sign on using your institution's credentials

Three Levels of Metadata
description/citation, domain-specific or custom fields, file metadata

Access Control Support
pre-defined and custom roles

Restricted Files + Ability to request access to restricted files
allow anyone, certain people, or no one to be able to download files

Customization of dataverses
branding, metadata based facets, sub-dataverses, featured dataverses

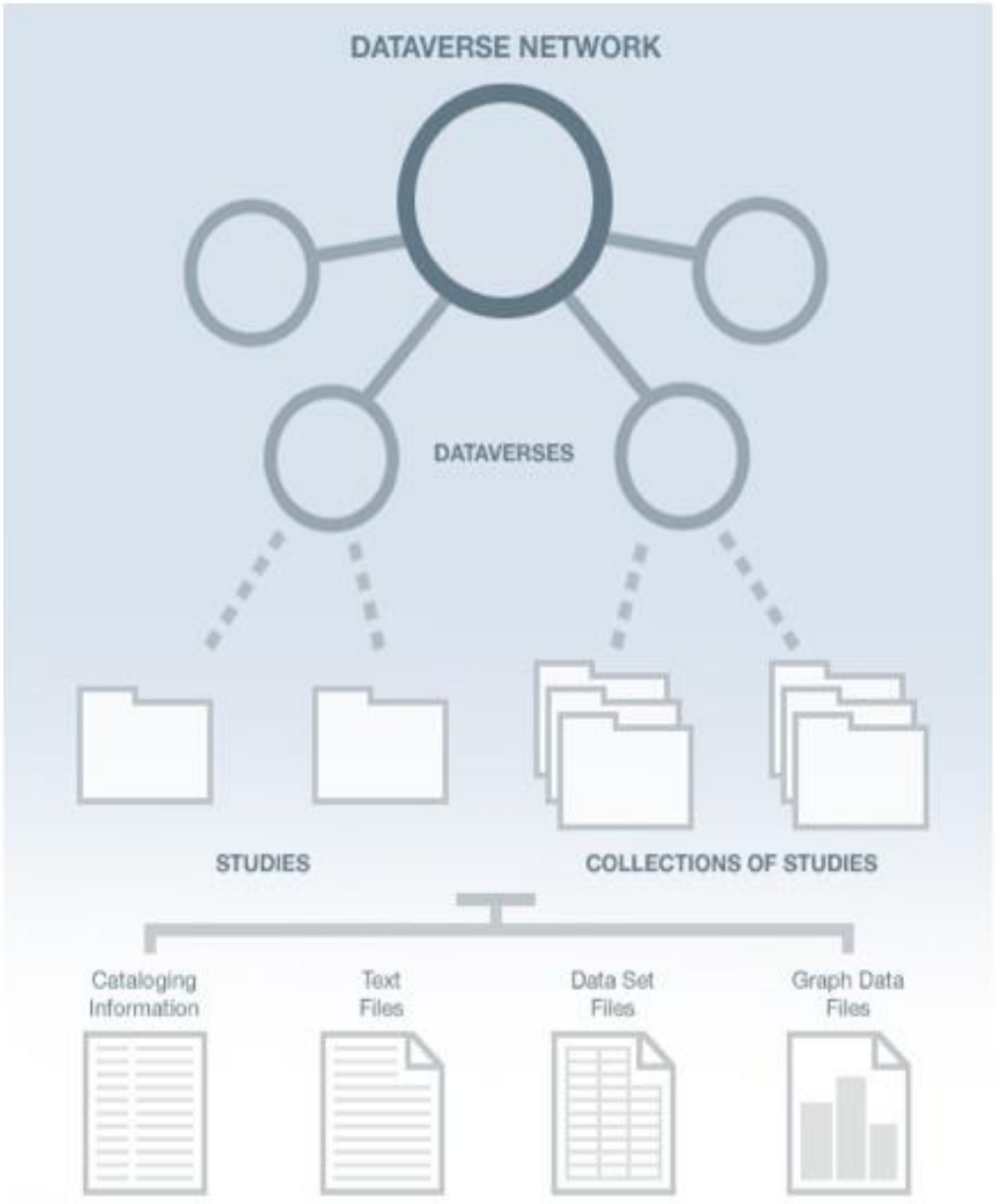
Re-format, Summary Statistics, and Analysis for Tabular Files
integration with TwoRavens

Mapping of Geospatial files
integration with WorldMap

IFPRI

Dataverse: The What

- 40 installations around the world
- Harvard Dataverse, Odum Dataverse, Abacus, INRA, Bostwana Harvard Data
- CGIAR Centers with local installation: CIFOR, CIMMYT, ICRISAT



IFPRI

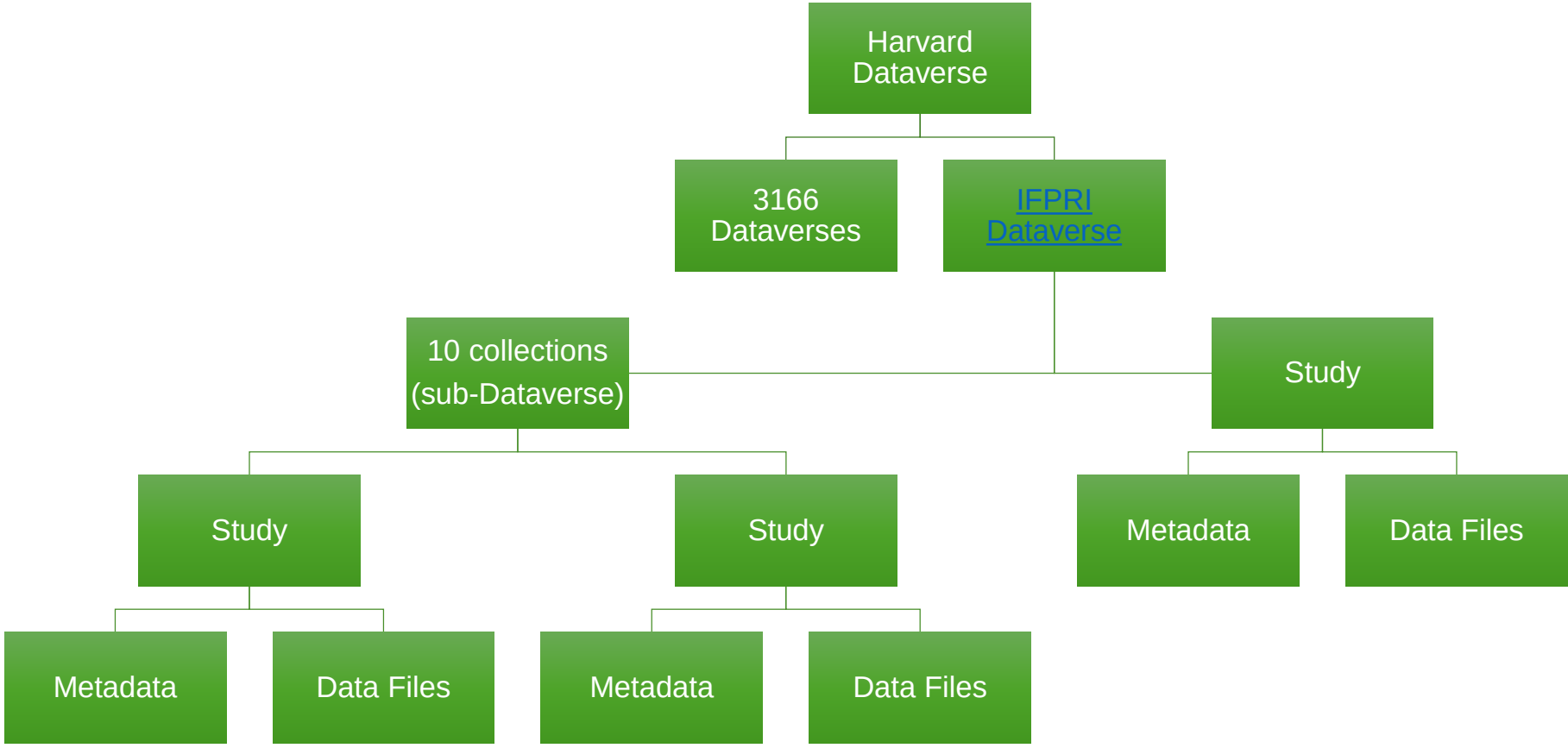
Dataverse: Pros and Cons

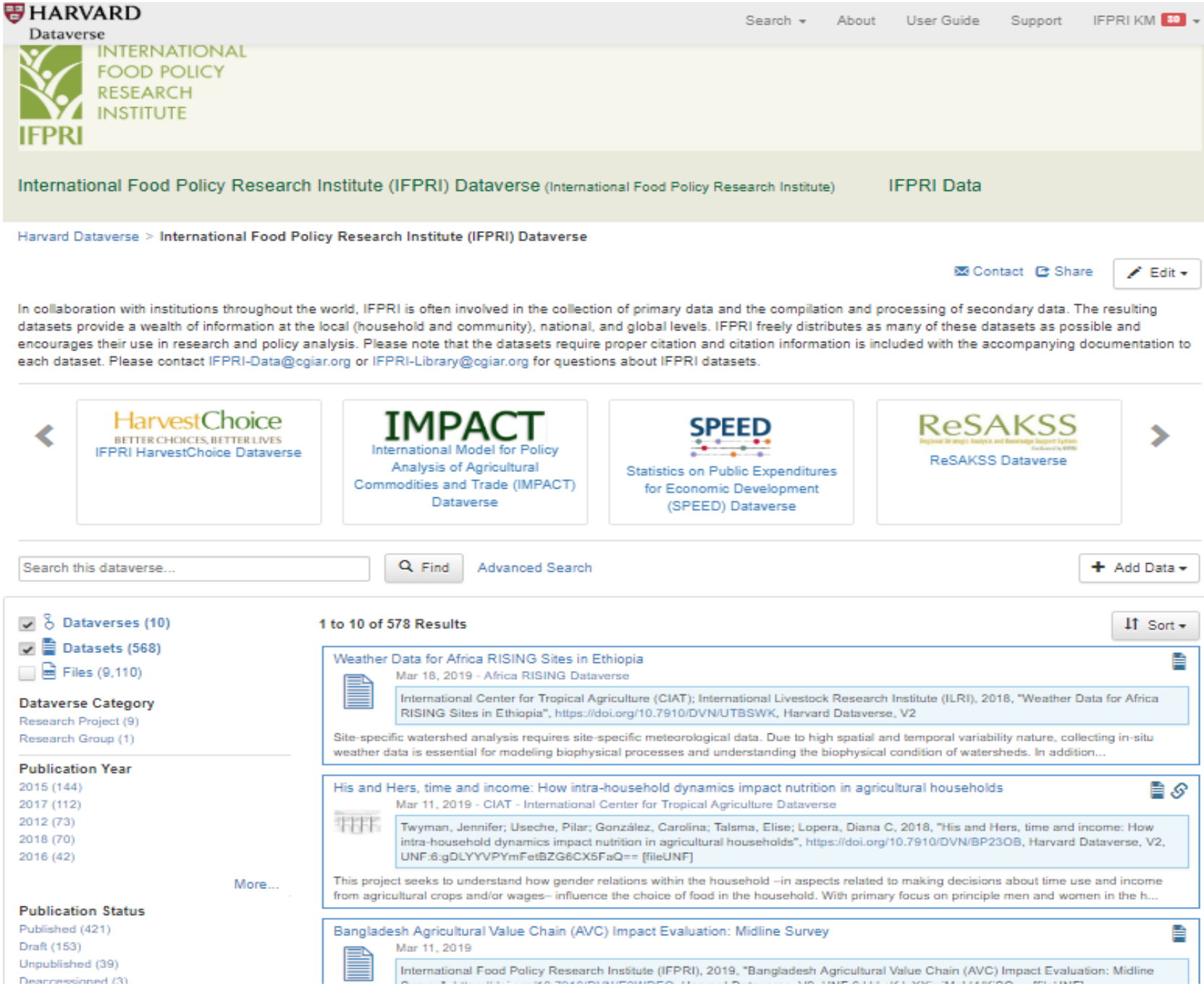
Pros	Cons
Open source software	Metadata uniformity across domains (e.g. Social sciences, geology, computer sciences, etc.)
Evolving and addresses users concern	Not effective in handling sensitive data
DOI minting; Handles	Can't delete datasets; only deaccessioned them
OAI-PMH compliant	Custom version control
API calls	

IFPRI

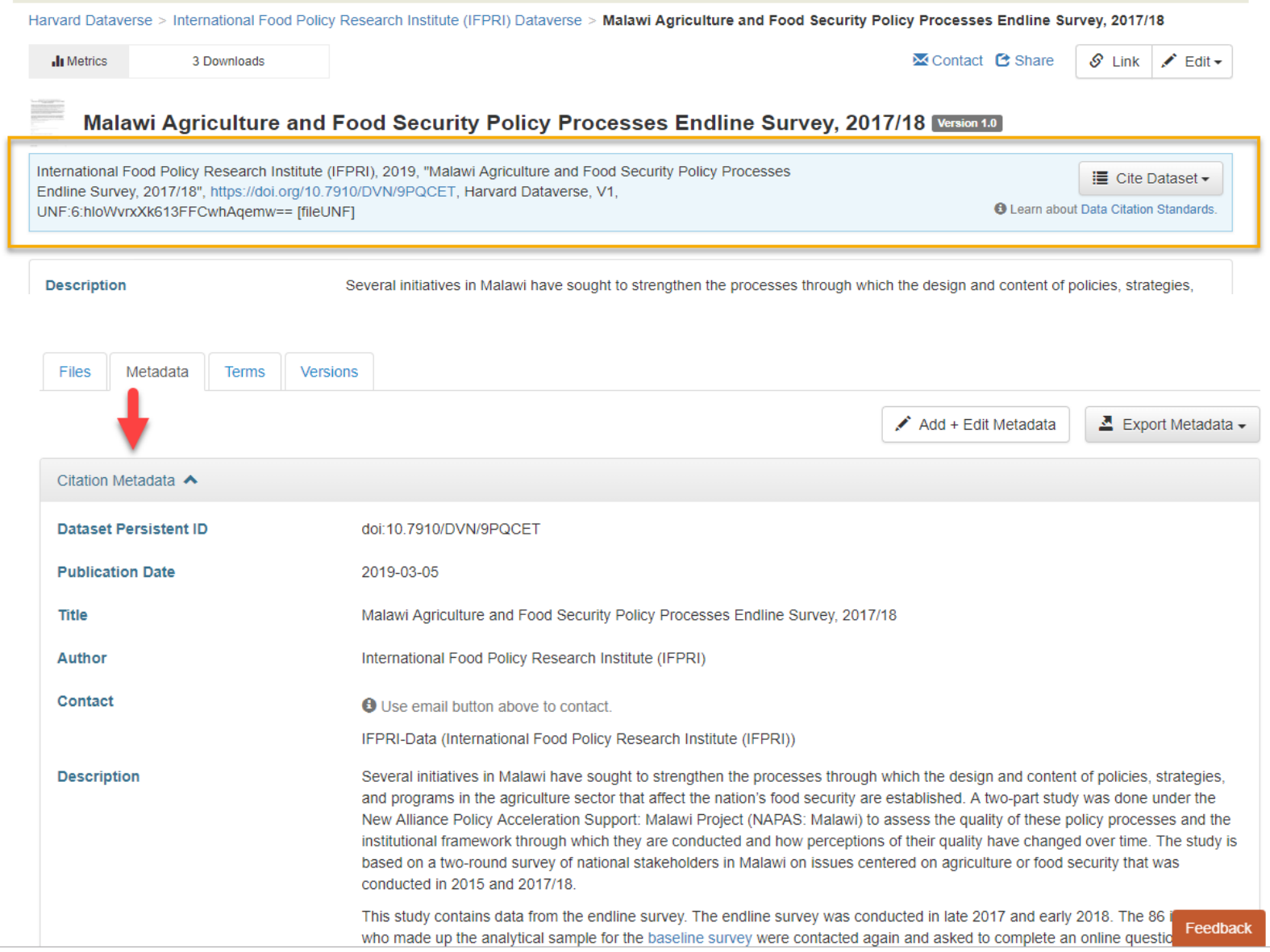
Harvard Dataverse

- One installation of Dataverse software hosted and managed by Harvard University




HARVARD Dataverse
Search
About
User Guide
Support
IFPRI KM
INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE
IFPRI
International Food Policy Research Institute (IFPRI) Dataverse (International Food Policy Research Institute)
IFPRI Data
Harvard Dataverse > International Food Policy Research Institute (IFPRI) Dataverse
Contact
Share
Edit
In collaboration with institutions throughout the world, IFPRI is often involved in the collection of primary data and the compilation and processing of secondary data. The resulting datasets provide a wealth of information at the local (household and community), national, and global levels. IFPRI freely distributes as many of these datasets as possible and encourages their use in research and policy analysis. Please note that the datasets require proper citation and citation information is included with the accompanying documentation to each dataset. Please contact IFPRI-Data@cgiar.org or IFPRI-Library@cgiar.org for questions about IFPRI datasets.
HarvestChoice
BETTER CHOICES, BETTER LIVES
IFPRI HarvestChoice Dataverse
IMPACT
International Model for Policy Analysis of Agricultural Commodities and Trade (IMPACT) Dataverse
SPEED
Statistics on Public Expenditures for Economic Development (SPEED) Dataverse
ReSAKSS
ReSAKSS Dataverse
Search this dataverse...
Find
Advanced Search
Add Data
Dataverses (10)
Datasets (568)
Files (9,110)
Dataverse Category
Research Project (9)
Research Group (1)
Publication Year
2015 (144)
2017 (112)
2012 (73)
2018 (70)
2016 (42)
More...
Publication Status
Published (421)
Draft (153)
Unpublished (39)
1 to 10 of 578 Results
Sort
Weather Data for Africa RISING Sites in Ethiopia
Mar 18, 2019 - Africa RISING Dataverse
International Center for Tropical Agriculture (CIAT); International Livestock Research Institute (ILRI), 2018, "Weather Data for Africa RISING Sites in Ethiopia", https://doi.org/10.7910/DVN/UTBSWK, Harvard Dataverse, V2
Site-specific watershed analysis requires site-specific meteorological data. Due to high spatial and temporal variability nature, collecting in-situ weather data is essential for modeling biophysical processes and understanding the biophysical condition of watersheds. In addition...
His and Hers, time and income: How intra-household dynamics impact nutrition in agricultural households
Mar 11, 2019 - CIAT - International Center for Tropical Agriculture Dataverse
Twyman, Jennifer; Useche, Pilar; González, Carolina; Talsma, Elise; Lopera, Diana C, 2018, "His and Hers, time and income: How intra-household dynamics impact nutrition in agricultural households", https://doi.org/10.7910/DVN/BP23OB, Harvard Dataverse, V2, UNF:6:gDLYYVPYmFetBZG6CX5FaQ== [fileUNF]
This project seeks to understand how gender relations within the household –in aspects related to making decisions about time use and income from agricultural crops and/or wages– influence the choice of food in the household. With primary focus on principle men and women in the h...
Bangladesh Agricultural Value Chain (AVC) Impact Evaluation: Midline Survey
Mar 11, 2019
International Food Policy Research Institute (IFPRI), 2019, "Bangladesh Agricultural Value Chain (AVC) Impact Evaluation: Midline

Data citation and Metadata Example from IFPRI Dataverse



Data Sharing: IFPRI's History

CD-ROMs/IFPRI.Org

- Began in 1999
- Shared through IFPRI.org as a zipped file
- Data files in proprietary format
- No guarantee for long term preservation
- No unique identifier for dataset/datafile
- Cart system check out imposed barrier to access
- No licensing mechanism for dataset/datafiles

Harvard Dataverse

- Started migrating datasets to Dataverse in 2008
- All datasets are in Dataverse since 2013

IFPRI

Harvard Dataverse Benefits: IFPRI's Perspectives

- Apparently free of cost than staff time for curation
- No stress of maintaining server and backups
- Always evolving and open to customer's feedback
- Quick customer support service
- More sustainable for IFPRI in resource constrained environment
- Complies with the standards of data sharing and FAIR principles
- Easy access and data discovery
- Recognized by FAIRsharing.org, open data registry, and many journal publishers
- Indexed by google datasets

IFPRI

Harvard Dataverse Hiccups: IFPRI's Perspectives

- Difficult to fit-in IFPRI specific feature requests unless it serves the wider community
- Slow rendering because of growing data volume
- Not designed for handling sensitive data
- Batch modification a challenge if you don't have programming/API expertise in-house
- Limited options for customizing interface
- Free service – can't enforce to execute the new features quickly

IFPRI

Data Services at IFPRI: What and How

- Data policy awareness
  - IFPRI Data Policy
  - CGIAR Open Access and Open Data Policy
  - Donor Policies (USA, BMGF, EU)
- Data management plan
  - Assist in developing data management plan
  - Review data management plan
- Data publishing
  - Review datasets (direct/indirect personally identifiable information)
  - Create metadata records (using controlled vocabulary and data publishing standards)
  - Create codebooks
  - Develop guidelines for researchers (preparing data for publishing, data anonymization etc)
  - Registering datasets in donor specific library (USAID data development library

IFPRI

Questions and Comments ??

IFPRI