Making agricultural data
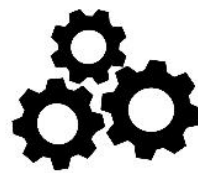
F indable A ccessible I nteroperable R eusable

Cynthia Parr, Ph.D cynthia.parr@usda.gov
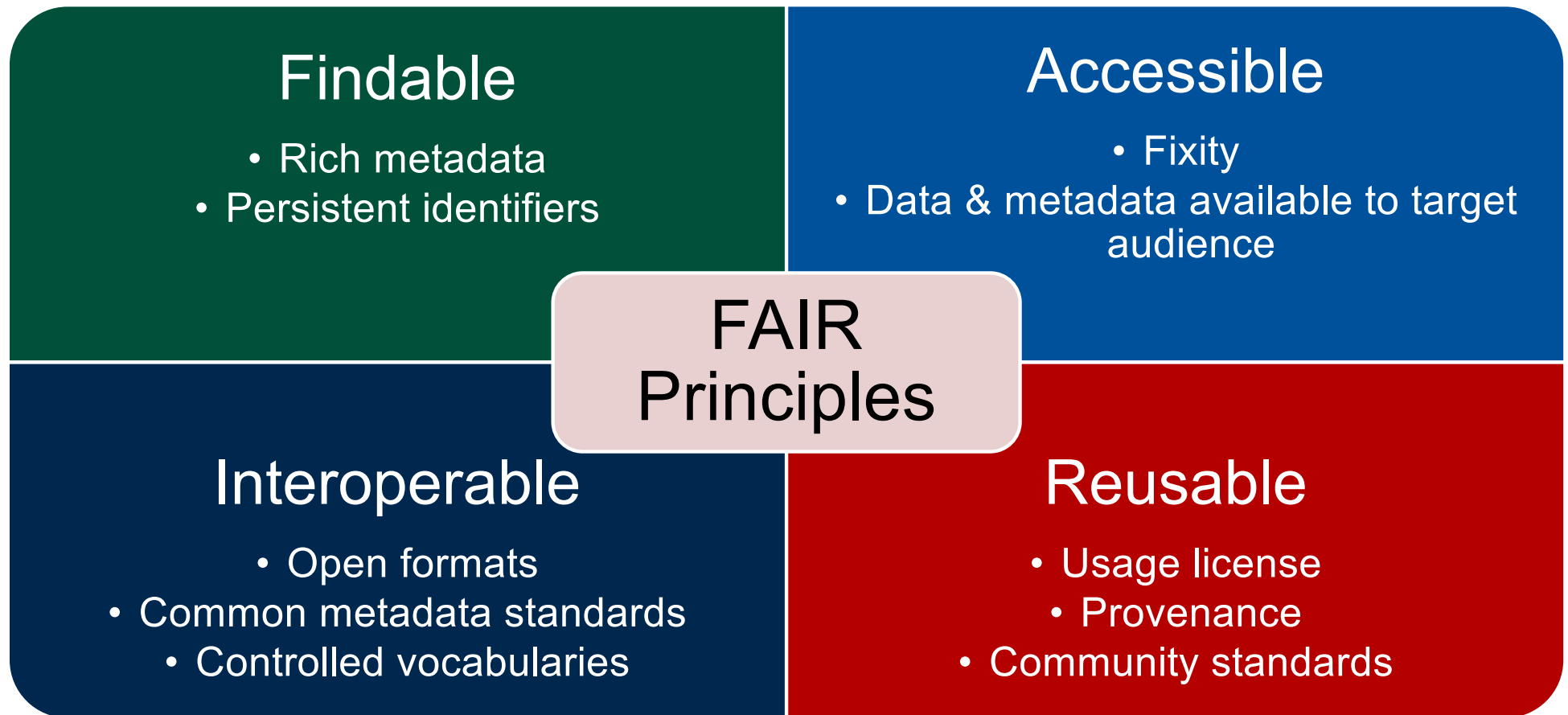
Acting Assistant Chief Data Officer

USDA Research Education and Economics Mission Area

Texas A & M Symposium   July 17, 2020

Agenda

- Why FAIR data is important
- NAL data services
- AI in ARS
- USDA data governance
- Future plans

Findable
- Rich metadata
- Persistent identifiers

Accessible
- Fixity
- Data & metadata available to target audience

FAIR Principles

Interoperable
- Open formats
- Common metadata standards
- Controlled vocabularies

Reusable
- Usage license
- Provenance
- Community standards

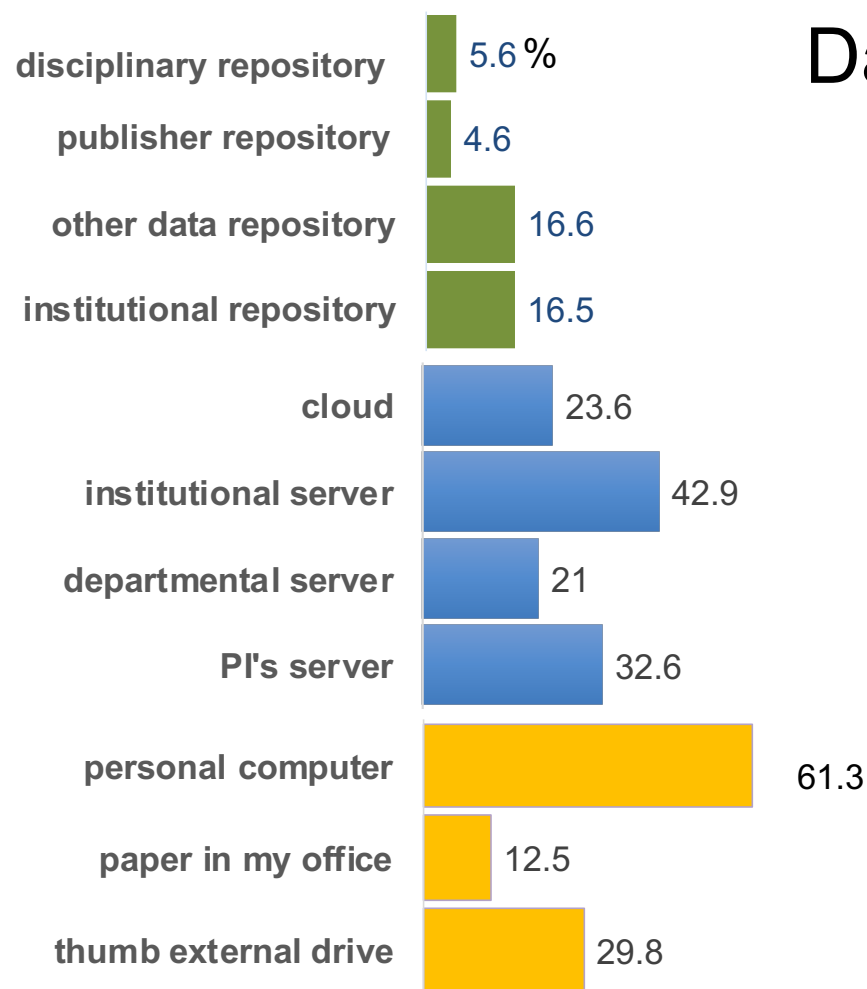https://www.force11.org/group/fairgroup/fairprinciples

3

Why Good Data Management is Important

Retraction: "Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis"

The authors "can no longer vouch for the veracity of the primary data sources"

https://www.thelancet.com/lancet/article/s0140673620313246

USDA National Agricultural Library
United States Department of Agriculture

disciplinary repository 5.6 %

publisher repository 4.6

other data repository 16.6

institutional repository 16.5

cloud 23.6

institutional server 42.9

departmental server 21

PI's server 32.6

personal computer 61.3

paper in my office 12.5

thumb external drive 29.8

Data storage practices need improvement
2017 - 2018
n = 2,184 researchers

5

Benefits to researchers

Jones, et al. 2019.
http://doi.org/10.1629/uksg.463

More top tier journals will strongly encourage or require data availability statements.

A

Number of articles (0 – 32000)

BMC
- no DAS
- Upon request
- In paper & SI
- In repository

Publication date (2000, 2003, 2006, 2009, 2012, 2015, 2018)

B

Number of articles (0 – 32000)

PLOS
- no DAS
- Upon request
- In paper & SI
- In repository

Publication date (2000, 2003, 2006, 2009, 2012, 2015, 2018)
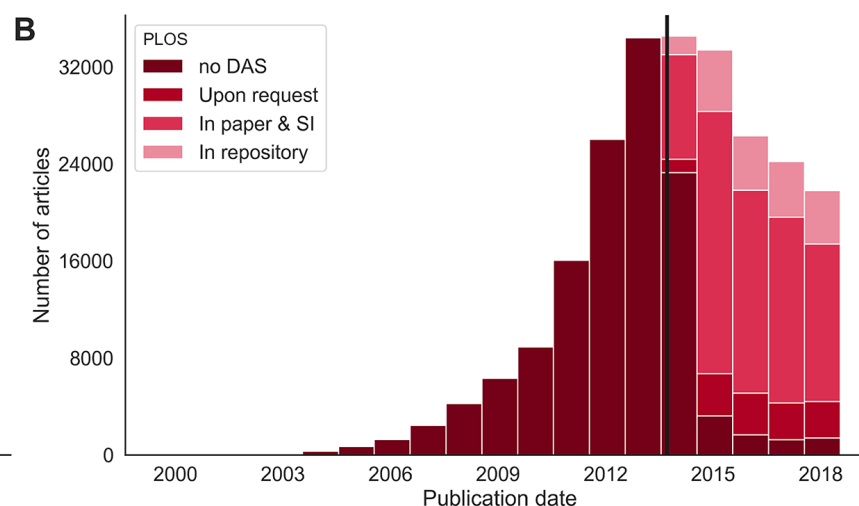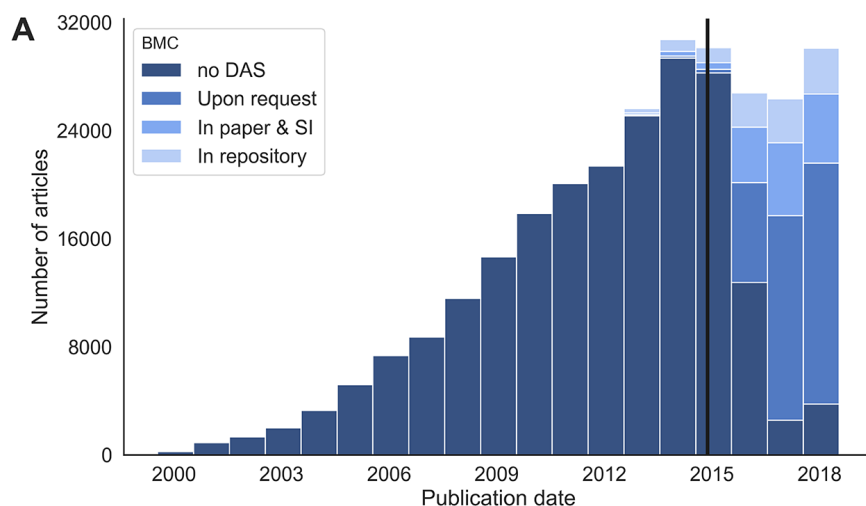
25% Citation Rate Advantage if Data Availability Statement Linked to a Repository

Colavizza, et al. 2020. https://doi.org/10.1371/journal.pone.0230416
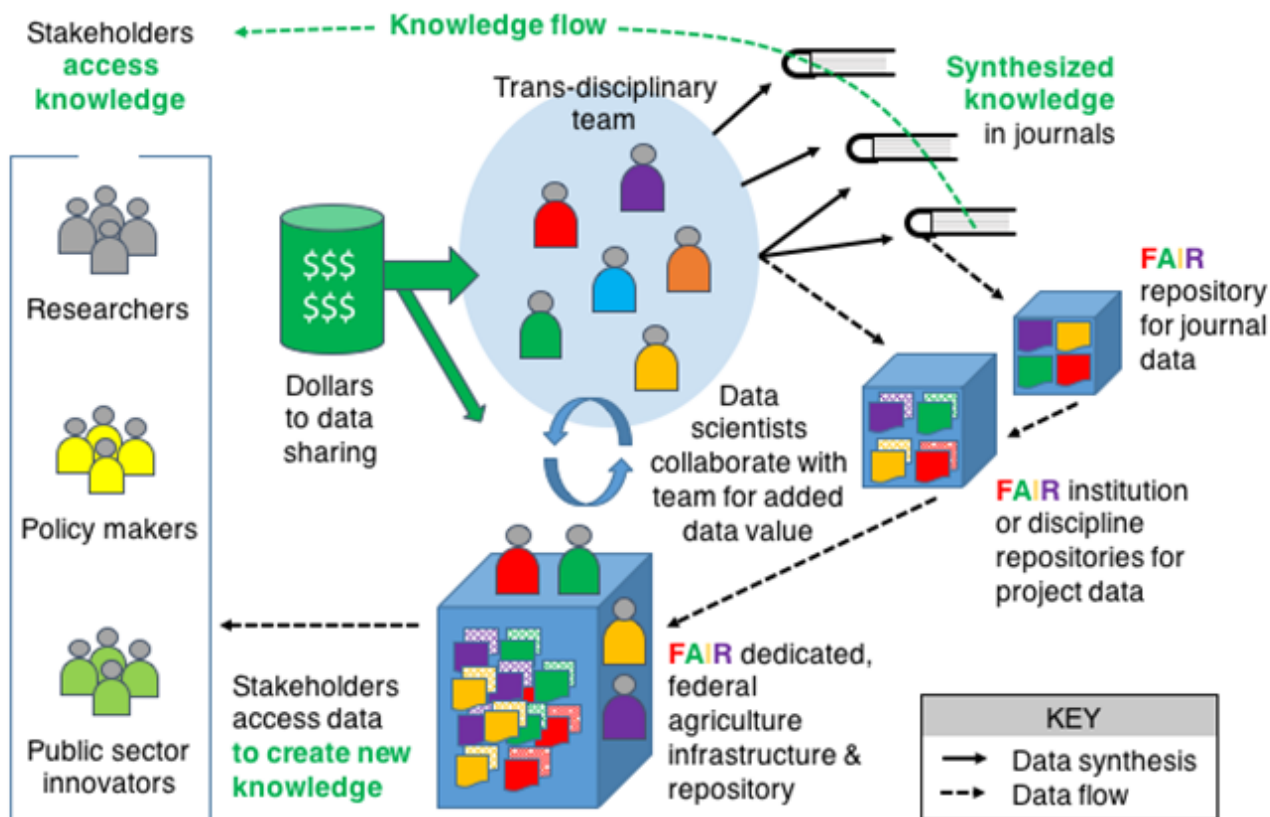
Opportunities

- Funder policy and infrastructure
- Computational Power
- Natural Language Processing
- Machine learning

From Brouder, et al. 2019 CAST Commentary. QYA2019-1

The Ag Data Commons…

- Is a catalog and data repository for USDA-funded research data

- Provides expert services for creating, curating, and enabling access to complete and machine-readable scientific metadata (FAIR data)

- Creates infrastructure for linking information, data, publications, people,…

- Helps the USDA-funded research community meet public access requirements

https://data.nal.usda.gov/

Ag Data Commons Use Cases

Data Manager

Use Case:
I want DOIs assigned to datasets in my data management platform.

Publishing

Bench Scientist

Use Case:
I need a persistent identifier and to provide public access to comply with publisher and Federal open access policies.

Publishing

Policymaker / Administrator

Use Case:
I need to assess trends and understand the impact of Federal research.
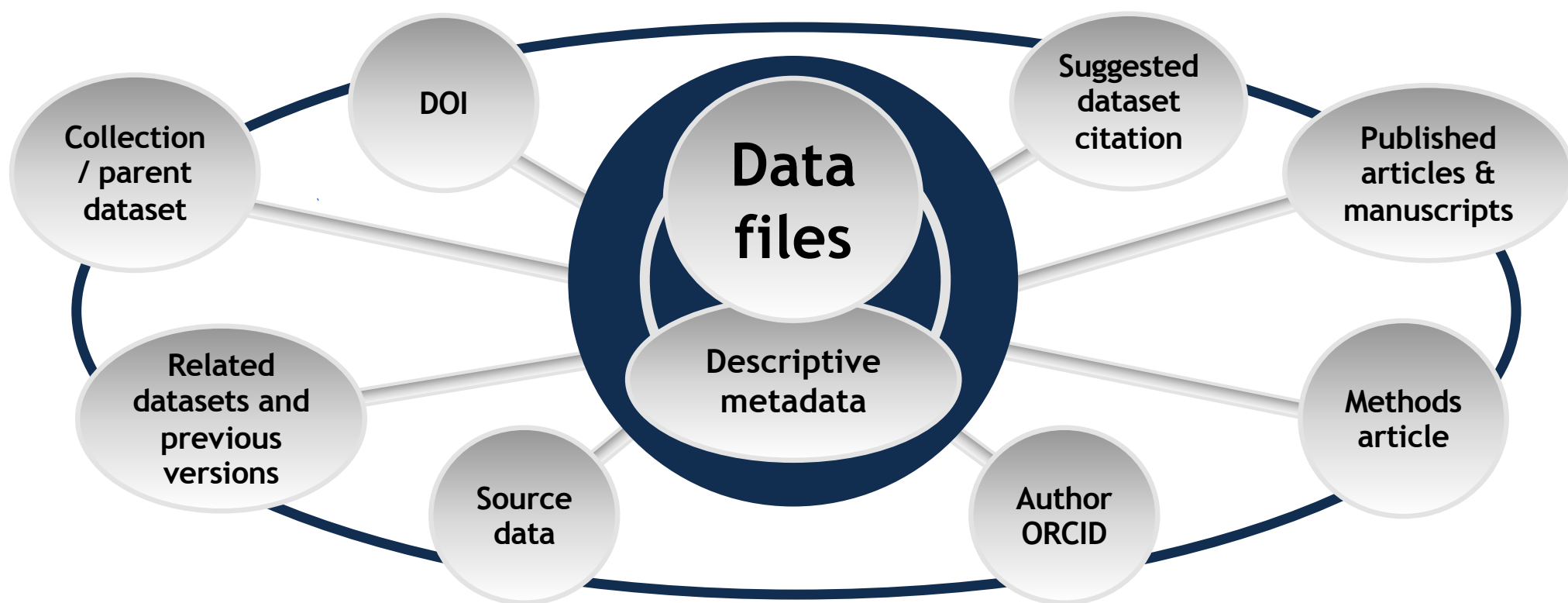
Analytics

Researcher

Use Case:
I need to find data related to my research and stay current on the state of the science in my field.

Search & Discovery

Personas developed by Emily Marsh

Ag Data Commons dataset

Create datasets manually

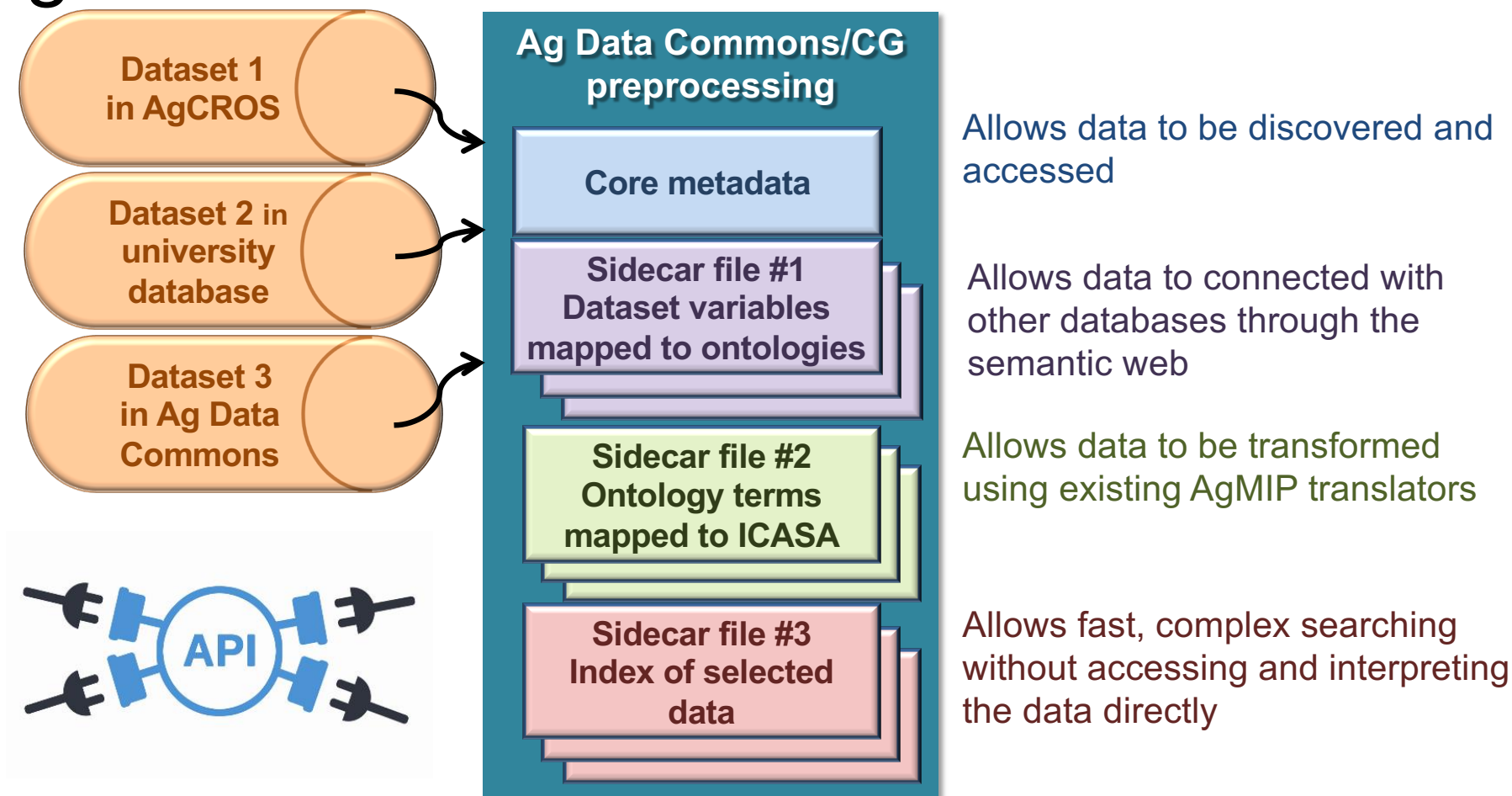| Register for an account (or log in)	Create a dataset record	Link related content	Upload or link resource files	Submit record for review

Ag Data Commons synthesizes metadata

Ag Data Commons metadata fields may be output into a number of schemas to achieve a variety of functions

- Project Open Data
  - Data.gov
- ISO 19115
  - Geospatial data
- DataCite
  - DOIs

Tidy data and data dictionaries

Making Data Machine-Readable

https://www.data.nal.usda.gov

https://go.usa.gov/xfjg6

Wickham, Hadley. (2014). Tidy data. The American Statistician. 14. 10.18637/jss.v059.i10.

Frictionlessdata.io

Agricultural Research Data Network ARDN

National Agricultural Library
United States Department of Agriculture

Dataset 1 in AgCROS

Dataset 2 in university database

Dataset 3 in Ag Data Commons

API

Ag Data Commons/CG preprocessing

Core metadata

Sidecar file #1 Dataset variables mapped to ontologies

Sidecar file #2 Ontology terms mapped to ICASA

Sidecar file #3 Index of selected data

Allows data to be discovered and accessed

Allows data to connected with other databases through the semantic web

Allows data to be transformed using existing AgMIP translators

Allows fast, complex searching without accessing and interpreting the data directly

AI in Agricultural Research Service

www.computer.org/itpro
May/June 2020

Multimodal AI to Improve Agriculture

Parr et al. In press. IT Professional
- – Computer Vision
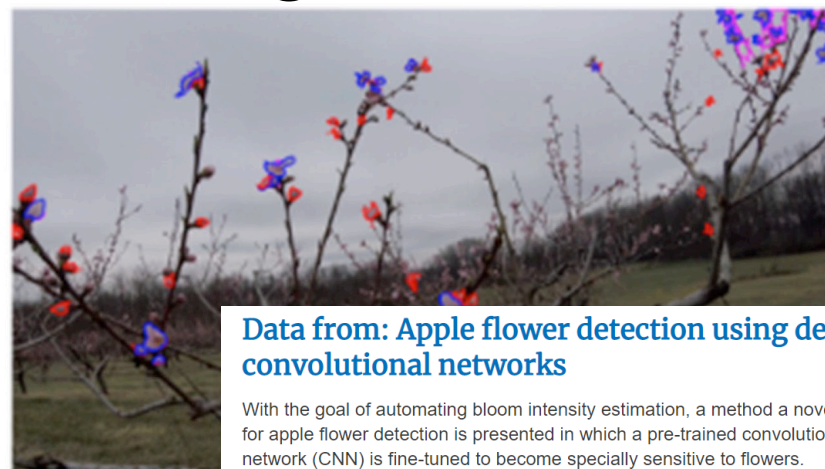- – Natural Language Processing

Plans to apply in combination for:
- – Personalized Nutrition
- – Invasive pest detection and modeling

NAL is helping:

Training data for ML: Ag Data Commons
Developing skilled workforce via NAL-UMD data science fellowships

Data from: Apple flower detection using deep convolutional networks

With the goal of automating bloom intensity estimation, a method a novel method for apple flower detection is presented in which a pre-trained convolutional neural network (CNN) is fine-tuned to become specially sensitive to flowers.

6x mp4

Data from: Multi-species fruit flower detection using a refined semantic segmentation network

This dataset consists of four sets of flower images, from three different fruit tree species: apple, peach, and pear, and accompanying ground truth images. This data is provided to support a paper as well as to provide labeled data to the community for the development of new algorithms and models for object detection.

9x zip   2x txt

DATA AND ANALYTICS AT USDA
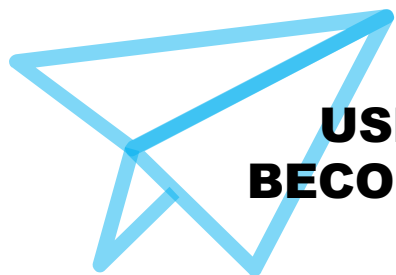
USDA

USDA IS COMMITTED TO BECOMING A DATA-DRIVEN ORGANIZATION...

Agency Priority Goal: Modernize information technology and data analytics capabilities across the Department, resulting in a USDA that is customer-focused, evidence-based, and efficient in the use of American taxpayer dollars.
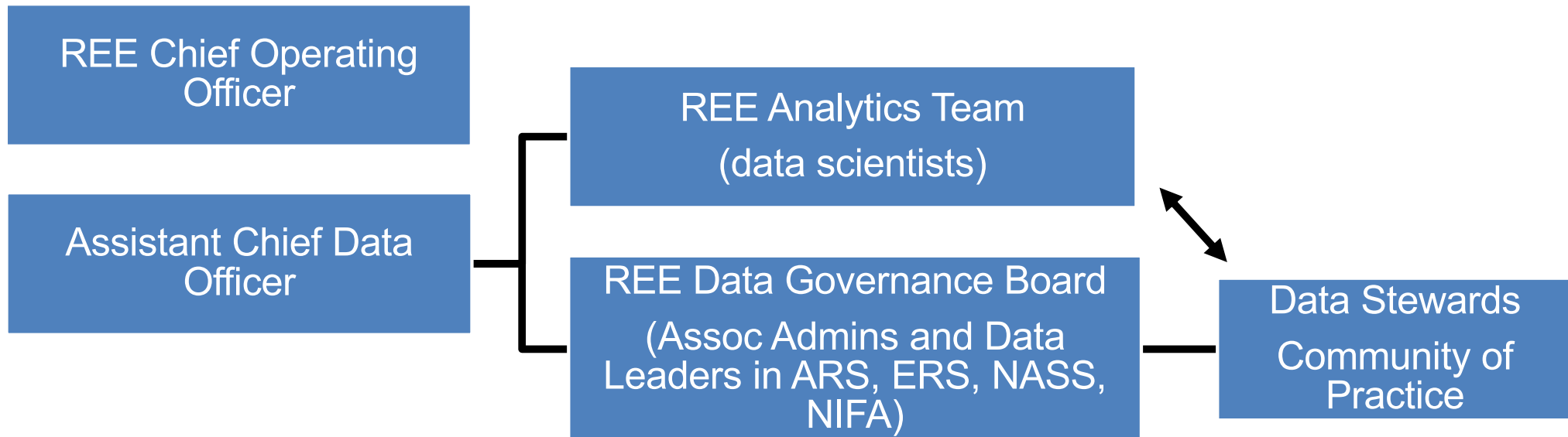
Foundations for Evidence-Based Policymaking Act of 2018 ("Evidence Act") Enacted into law on January 14, 2019

...AND IS POISED TO OVERCOME KEY CHALLENGES.

- Siloed and fragmented data

- Undeveloped data driven culture

- USDA role with on-farm data revolution

- Communicate the use and value of USDA's open data

Data Governance in REE Mission Area

USDA

The Future

- Increase access to FAIR data from REE-funded Research
  - Policies
  - Engagement with researchers & institutional repositories
  - Ag Data Commons collaboration with AI institutes

- Respond to data needs from stakeholders
  - Prioritize USDA datasets based on demand
  - Hold hackathon with USDA data

- Improve workforce skills

Thank you! **cynthia.parr@usda.gov**
@cydparr

Ag Data Commons

https://data.nal.usda.gov/

MORE ABOUT NAL DATA SERVICES

https://nal.usda.gov

Home | Topics | Publications | Collections | Data | Services | About Us

Scientific Data Services
Data Management Glossary
Data Management Plan Guidance
Data Management Planning

Notice

To help protect the public from the spread of
public starting Sunday, March 15th until furth

Online services will continue to be available a