ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Rensselaer

UNIVERSITY OF MICHIGAN

INDIANA UNIVERSITY

SEAD Virtual Archive:
Building a Federation of Institutional Repositories for Long-Term Data Preservation in Sustainability Science

Beth Plale, Indiana University, Bloomington, Indiana, USA

Robert H. McDonald, Indiana University, Bloomington, Indiana, USA

Kavitha Chandrasekar, Indiana University, Bloomington, Indiana, USA

Inna Kouper, Indiana University, Bloomington, Indiana, USA

Stacy Konkiel, Indiana University, Bloomington, Indiana, USA

Margaret L. Hedstrom, University of Michigan, Ann Arbor, Michigan, USA

Jim Myers, Rensselaer Polytechnic Institute, Troy, New York, USA

Praveen Kumar, University of Illinois, Urbana, Illinois, USA

SEAD

Sustainable Environment — Actionable Data

SEAD TEAMS

Michigan	Margaret Hedstrom-PI, Marietta Van Buhler, Karen Woollams, George Alter (ICPSR), Bryan Beecher (ICPSR)
Indiana	Beth Plale-Co-PI, Katy Börner, Robert H. McDonald, Robert Light, Kavitha Chandrasekar, Stacy Kowalczyk, Inna Kouper, Stacy Konkiel, Robert Ping, Ryan Cobine
Rensselaear	James Myers-Co-PI, Ram Prasanna Govind Krishnan, Lindsay Todd
Illinois	Praveen Kumar-Co-PI, Terry McLaren (NCSA), Rob Kooper (NCSA), Luigi Marini (NCSA)

SEAD
Sustainable Environment — Actionable Data

Challenge: The Data Deluge

1. Scientific data ingestion must be quick and minimally intrusive on a scientist's time.
2. Ingesting must be flexible enough to handle the varied kinds of data.
           sizes // formats // composition
3. Tools for advertising and serving data from an institutional repository need to be consistent with tools and processes of the scientific community.

Challenge: Long Tail Scientific Research

- Many research niches
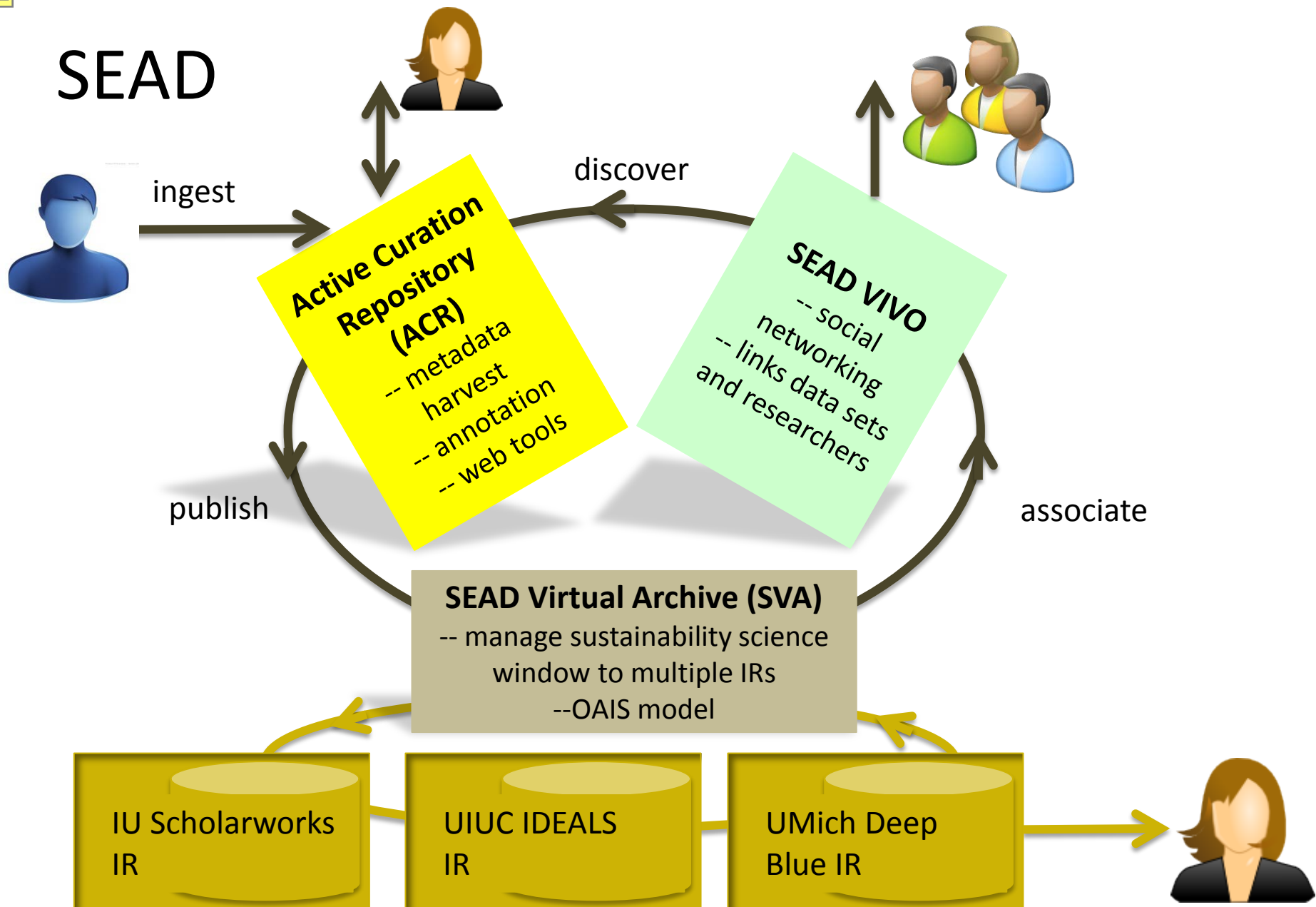 - customized methods & toolsets
  - localized storage

- Less consideration for long-term availability and data reuse

SEAD
Sustainable Environment — Actionable Data

Requirements of Virtual Archive for Sustainability Science

- Must connect multiple IRs
- Must be minimally intrusive on a scientist's time
- Must handle varied data:
  - multi-GB collection,
  - vastly heterogeneous collection of files,
  - small complex database of a thousand variables, or
  - set of files in formats that are unique to the subdiscipline
- Must be consistent with tools and processes of the community

SEAD

ingest

discover

Active Curation Repository (ACR)
-- metadata harvest
-- annotation
-- web tools

SEAD VIVO
-- social networking
-- links data sets and researchers

publish

associate

SEAD Virtual Archive (SVA)
-- manage sustainability science
window to multiple IRs
--OAIS model

IU Scholarworks IR

UIUC IDEALS IR

UMich Deep Blue IR

SEAD Virtual Archive (SVA)

Active Curation Repository (ACR)
-- metadata harvest
-- annotation
-- web tools

SEAD VIVO
-- social networking
-- links data sets and researchers

Design

Policy Decisions

Progress to Date

SEAD Virtual Archive (SVA)
-- manage sustainability science
window to multiple IRs
--OAIS model

[Single view into data] [Easy deposit]

SEAD Virtual Archive Workflow

Accept Repository Agreement

Preview Data → Upload Data to VA → Run Virus Checking → File Charact-erization → Mint DOI → Deposit to IR (& cloud) → Update DOI target → Index Metadata

Version Data

IR Match-maker → Large Dataset Decision

Index Scientific Metadata

Ongoing work

SEAD
Sustainable Environment — Actionable Data
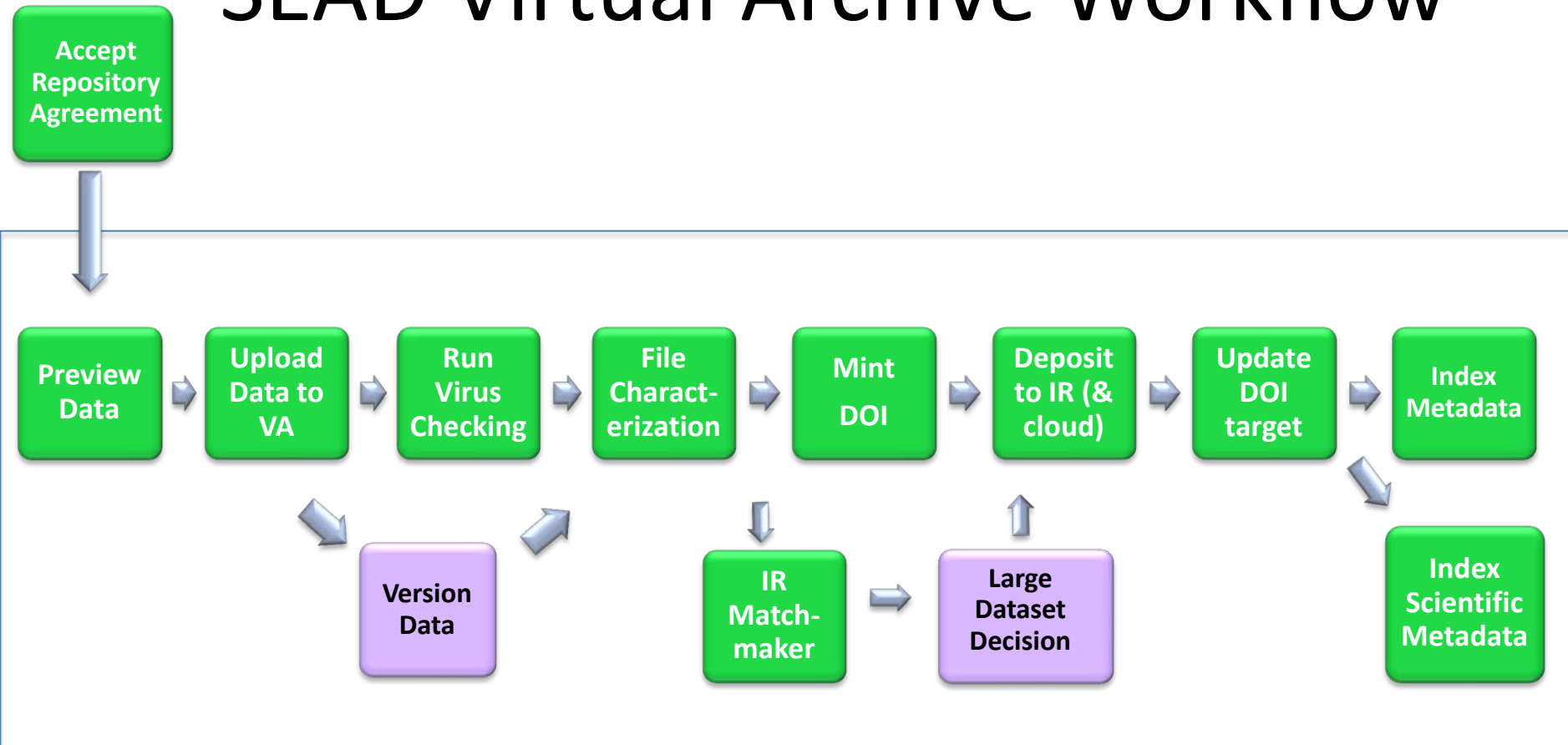
Architecture: SEAD VA Matchmaker
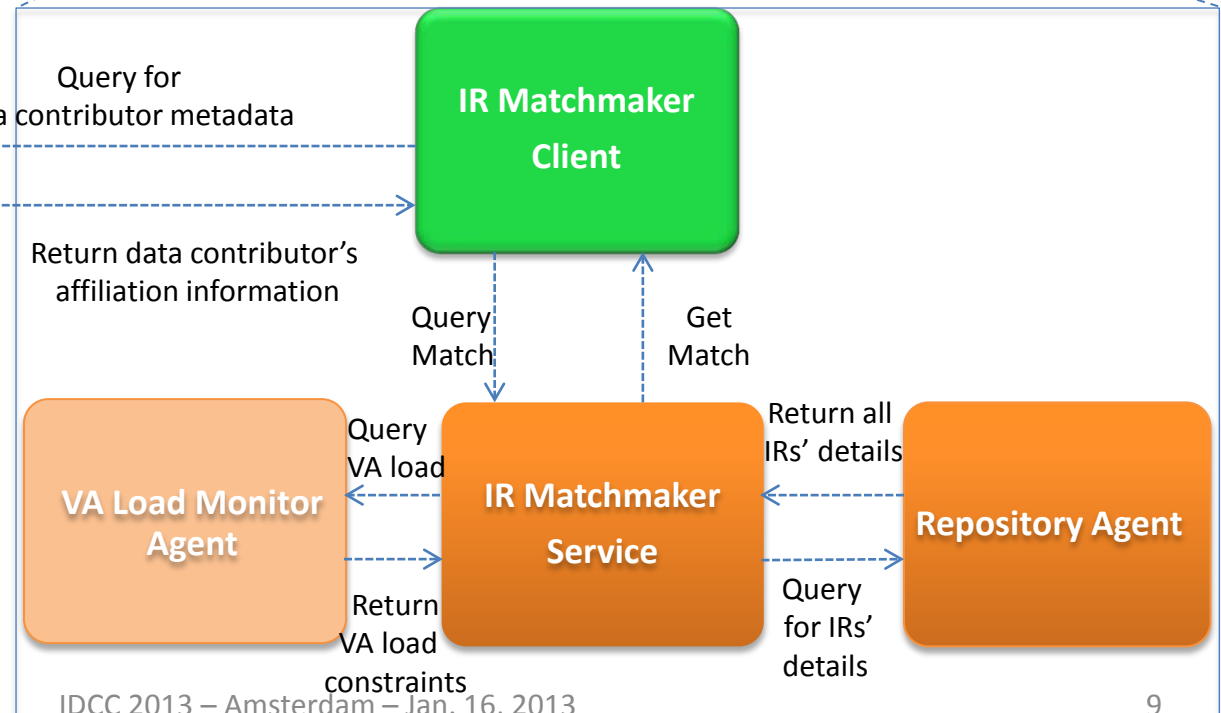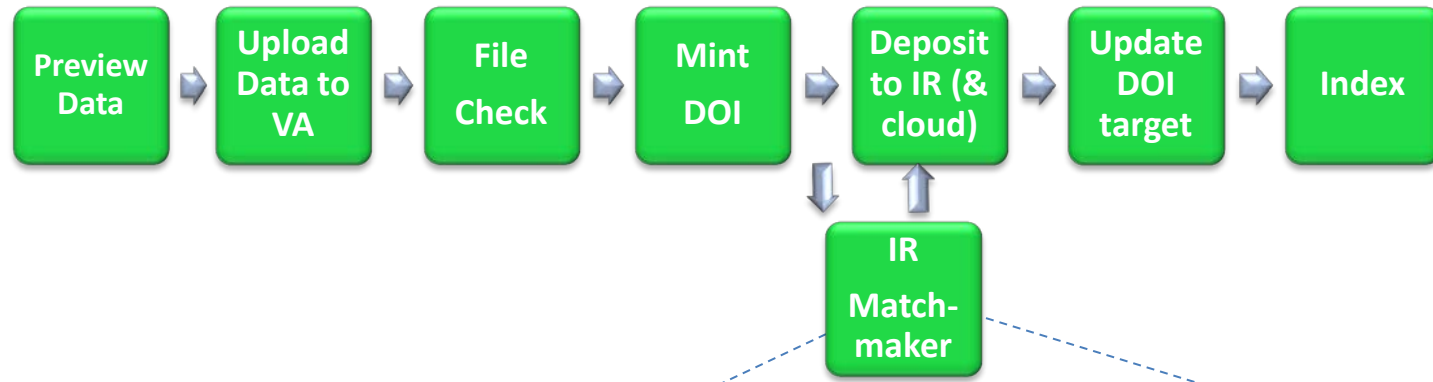
Policy: Licensing Agreements

- Right to store and re-format files (preservation)
- Allow editing to protect human subjects, sensitive data (protection)
- Make metadata public (discoverability)
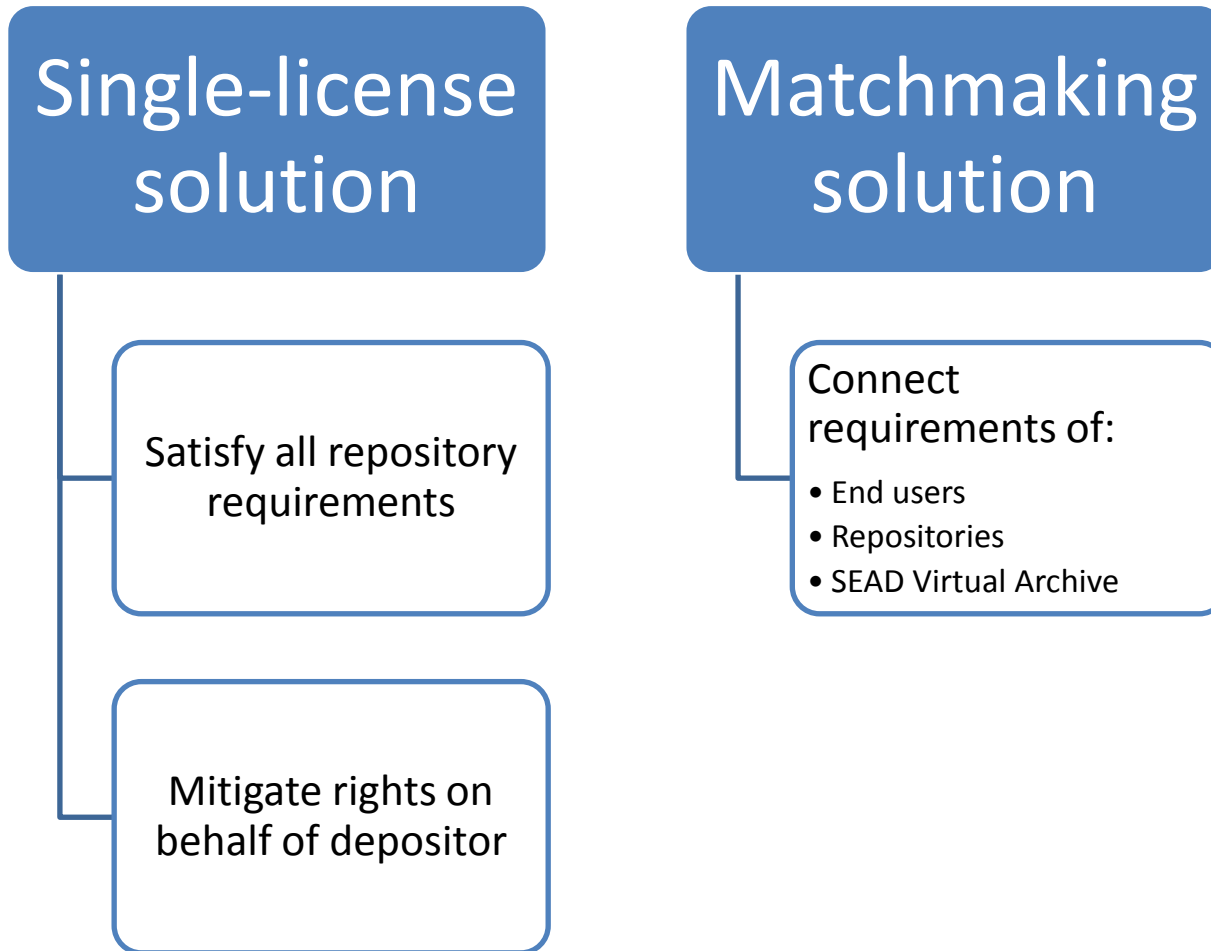- Ensure sponsor compliance (liability)

Repository rights

SEAD
Sustainable Environment — Actionable Data

Policy: Licensing Agreements

Depositor rights

- Retain copyright/moral rights
- Deposits will not be changed from original intent
- Embargoes will be honored

SEAD
Sustainable Environment — Actionable Data

Policy: Licensing Agreements

Single-license solution

- Satisfy all repository requirements

- Mitigate rights on behalf of depositor

Matchmaking solution

Connect requirements of:
- End users
- Repositories
- SEAD Virtual Archive

SEAD
Sustainable Environment — Actionable Data

Policy: Permanent Identifiers

Author IDs

- VIVO identifiers

Dataset IDs

- Digital Object Identifiers (DOIs)

SEAD
Sustainable Environment — Actionable Data

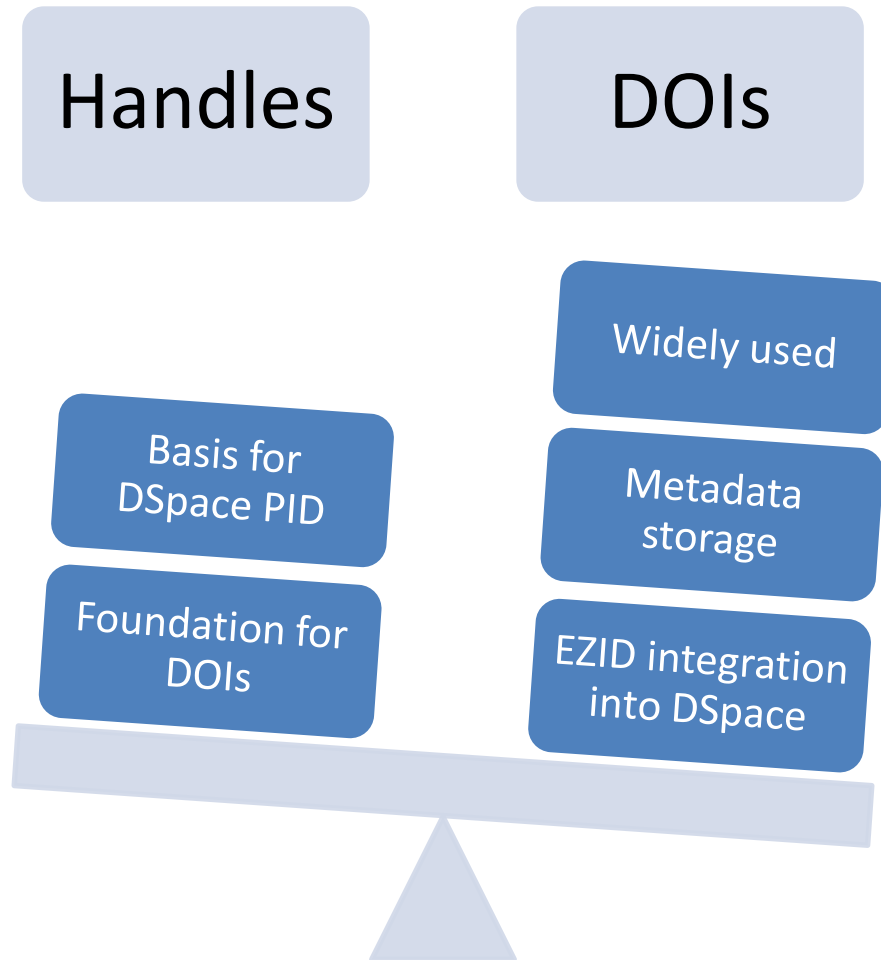# Policy: Author IDs

- Used primarily at domain/institution al level
- Supports many researcher ID systems, including ORCID

ORCID

VIVO ID

ResearcherID
Scopus
Author ID
Pivot ID
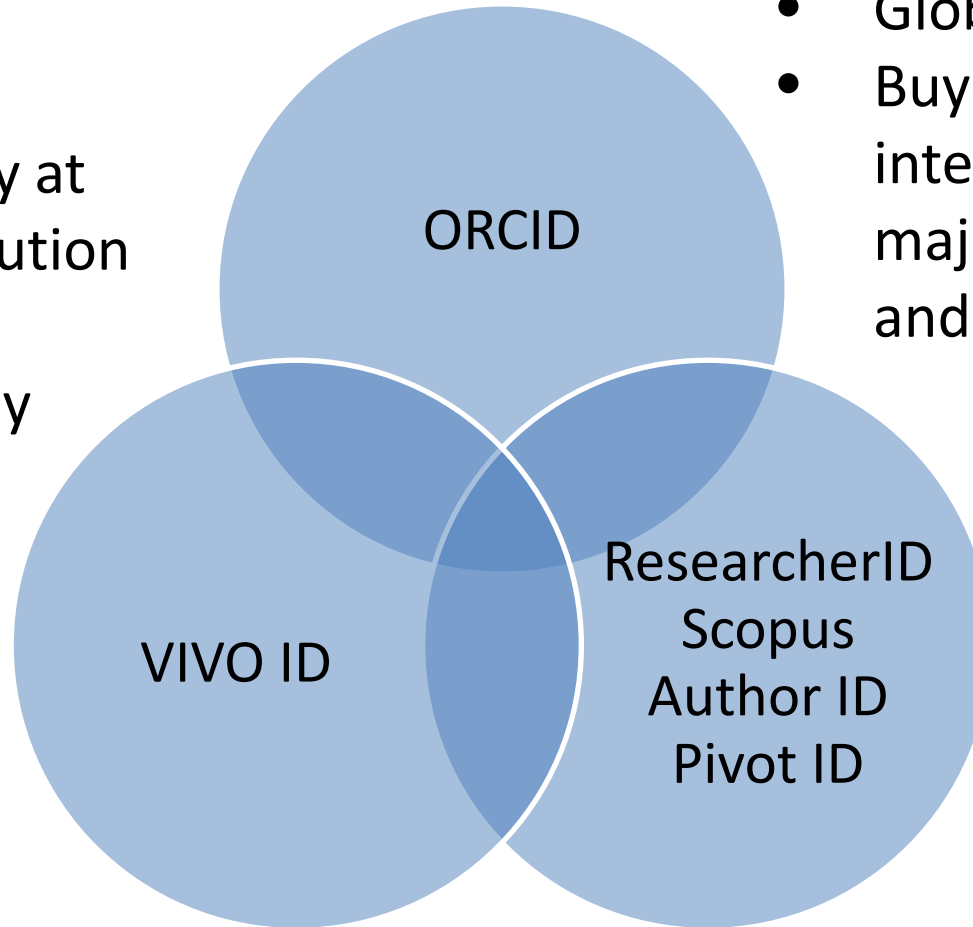
- Global system
- Buy-in from and integration with major publishers and institutions

SEAD
Sustainable Environment — Actionable Data

Policy: Dataset IDs

Progress to Date

- Ingested all NCED data
  - Small-sized collection (overall < 150 Mb)
  - File organization for heterogeneous collection of related files with flat or hierarchical structure
- Tested deposit between the VA, UIUC IDEALS, and IUScholarWorks

SEAD
Sustainable Environment — Actionable Data

Future Work

- Address other use cases
  - Large size collections (overall > 1 Gb)
  - Relational database / interconnected variables
  - Unique formats (to project, discipline, community)
- Interoperability with other DataNets
- Support for API access
- Determine how prototype fits researcher workflows

SEAD
Sustainable Environment — Actionable Data

Thank you

http://www.sead-data.net

@SEADdatanet

Download this presentation at

http://slidesha.re/11vqeN9