Trust Threads: Minimal Provenance for Data Publication and Reuse

Beth Plale
Data To Insight Center
Science Director, Pervasive Technology Institute
School of Informatics and Comptuing
Indiana University

May 7, 2015

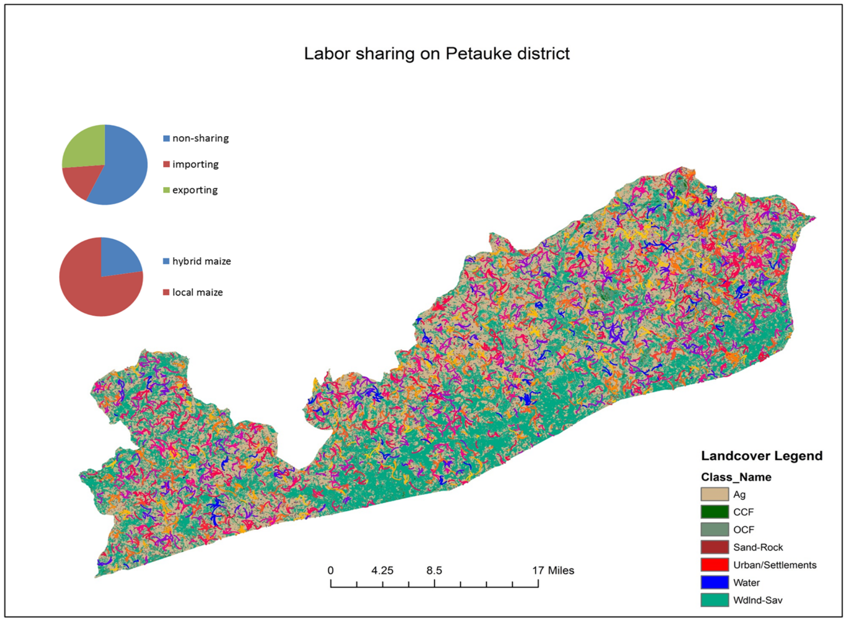
SEAD Sustainable Environment Actionable Data

Cooperative agreement #OCI0940824

The impact that digital data is having on science is increasing because of phenomenal data growth.

Data growth results in

- New and bigger research questions : energy, clean and abundant water, sustainable food, healthy population
- Requires data understood across discipline boundaries



Can technology innovation (tools, data management, knowledge representation) accelerate frequency of reuse and repurpose of scientific digital data?

- Data can carry with it thin threads of information *"trust threads"* that connect data to both its past and its future.
- In carrying this minimal provenance, data becomes more trustworthy.

Provenance suitcase

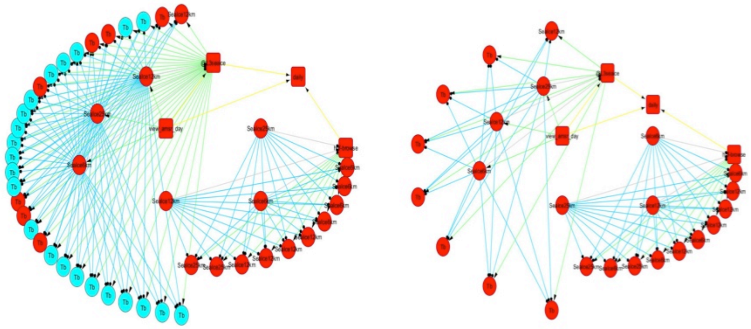
- Trustworthiness critical to successful sharing, reuse of data in science and technology research

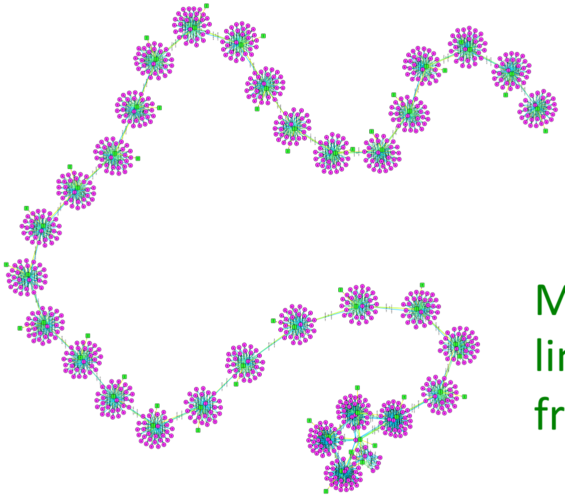
Data provenance

Data Provenance

Lineage of data product. Like piece of art, tells what has happened to data product through its life. Shows relationships between products

Provenance for two polar ice sheet images showing problems with one (on right, missing contributing files)





Month of ice sheet images, linked by template passed from day to day in processing

2012

Story of Publishing

Table of Contents

- The Flood
- The Publishing Exercise

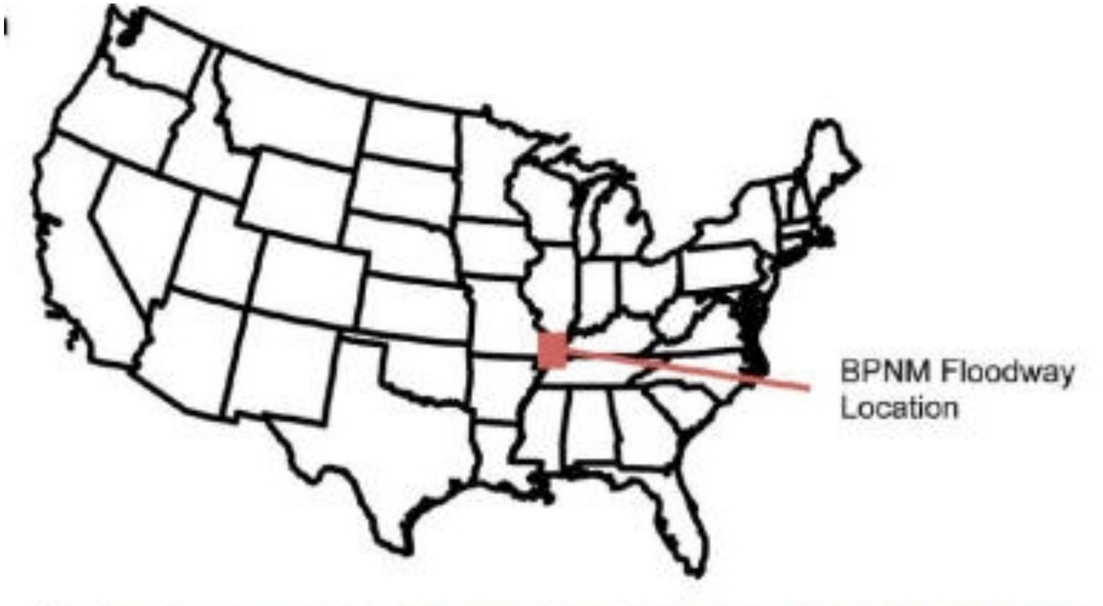
Terminology as used in this story

- **Data publishing** : act of making a data set that is associated with a peer reviewed scientific publication publicly available
- **Data preservation** : steps undertaken to ensure a data set is available and useable by subsequent generations of scientists
- **Data sharing** : reuse and repurpose of a data set for by a third party
- **Long tail science:** science utilizing data from multiple sources; data and models can be manipulated on workstation or small cluster

Cast of Characters

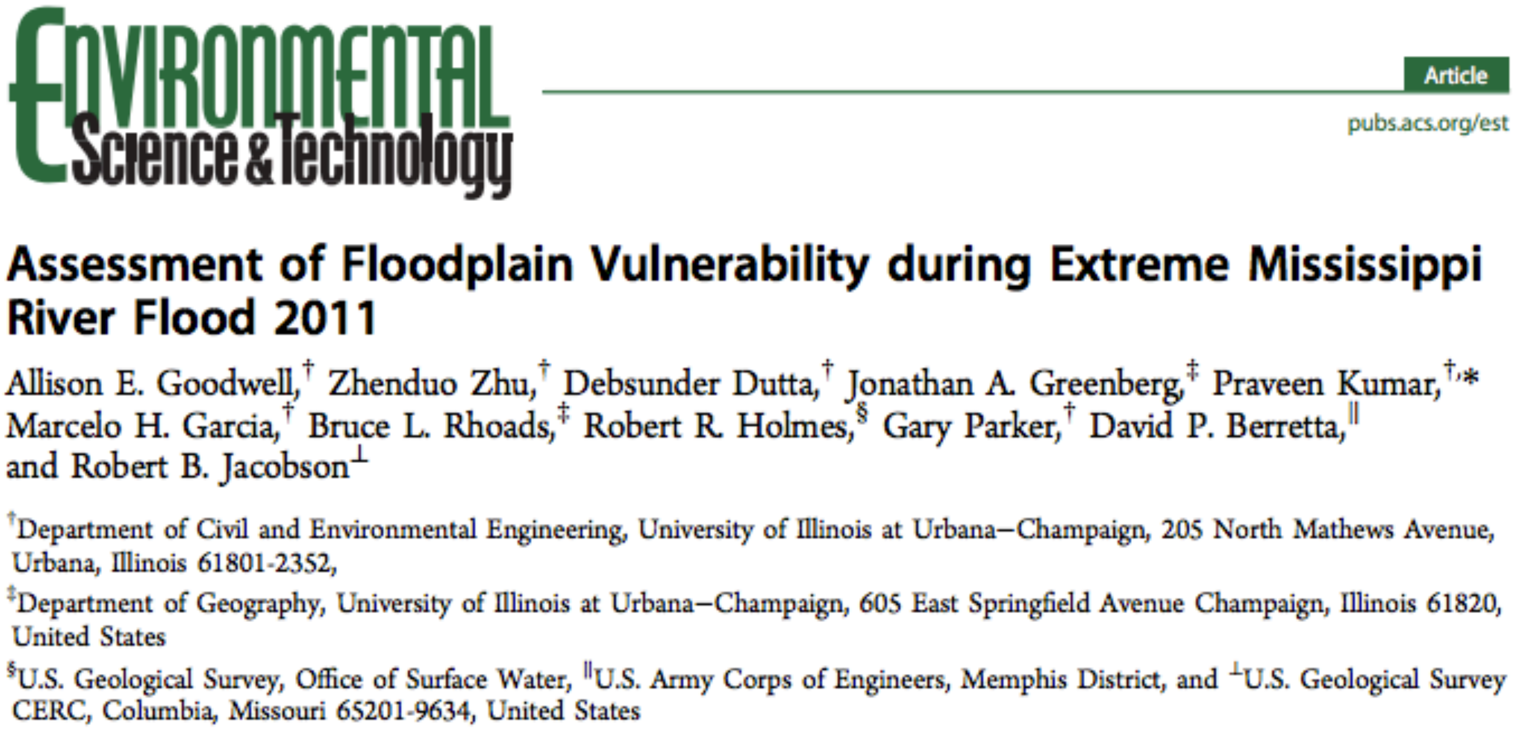
- Hydrologist (PI level)
- Hydrologist Postdoc
- Data scientist (PI level)
- Research programmer
- Data curator (CLIR Fellow)
- University library (data management services)

Extreme Mississippi Flood 2011



Thanks to

-- Praveen Kumar and his team at UIUC



Environmental Science & Technology

Article

pubs.acs.org/est

Assessment of Floodplain Vulnerability during Extreme Mississippi River Flood 2011

Allison E. Goodwell,[†] Zhenduo Zhu,[†] Debsunder Dutta,[†] Jonathan A. Greenberg,[‡] Praveen Kumar,[†,*] Marcelo H. Garcia,[†] Bruce L. Rhoads,[‡] Robert R. Holmes,[§] Gary Parker,[†] David P. Berretta,[‖] and Robert B. Jacobson[⊥]

[†]Department of Civil and Environmental Engineering, University of Illinois at Urbana–Champaign, 205 North Mathews Avenue, Urbana, Illinois 61801-2352,

[‡]Department of Geography, University of Illinois at Urbana–Champaign, 605 East Springfield Avenue Champaign, Illinois 61820, United States

[§]U.S. Geological Survey, Office of Surface Water, [‖]U.S. Army Corps of Engineers, Memphis District, and [⊥]U.S. Geological Survey CERC, Columbia, Missouri 65201-9634, United States

SEAD | Sustainable Environment Actionable Data

SEAD funded by National Science Foundation. B. Plale and P. Kumar, co-PIs.

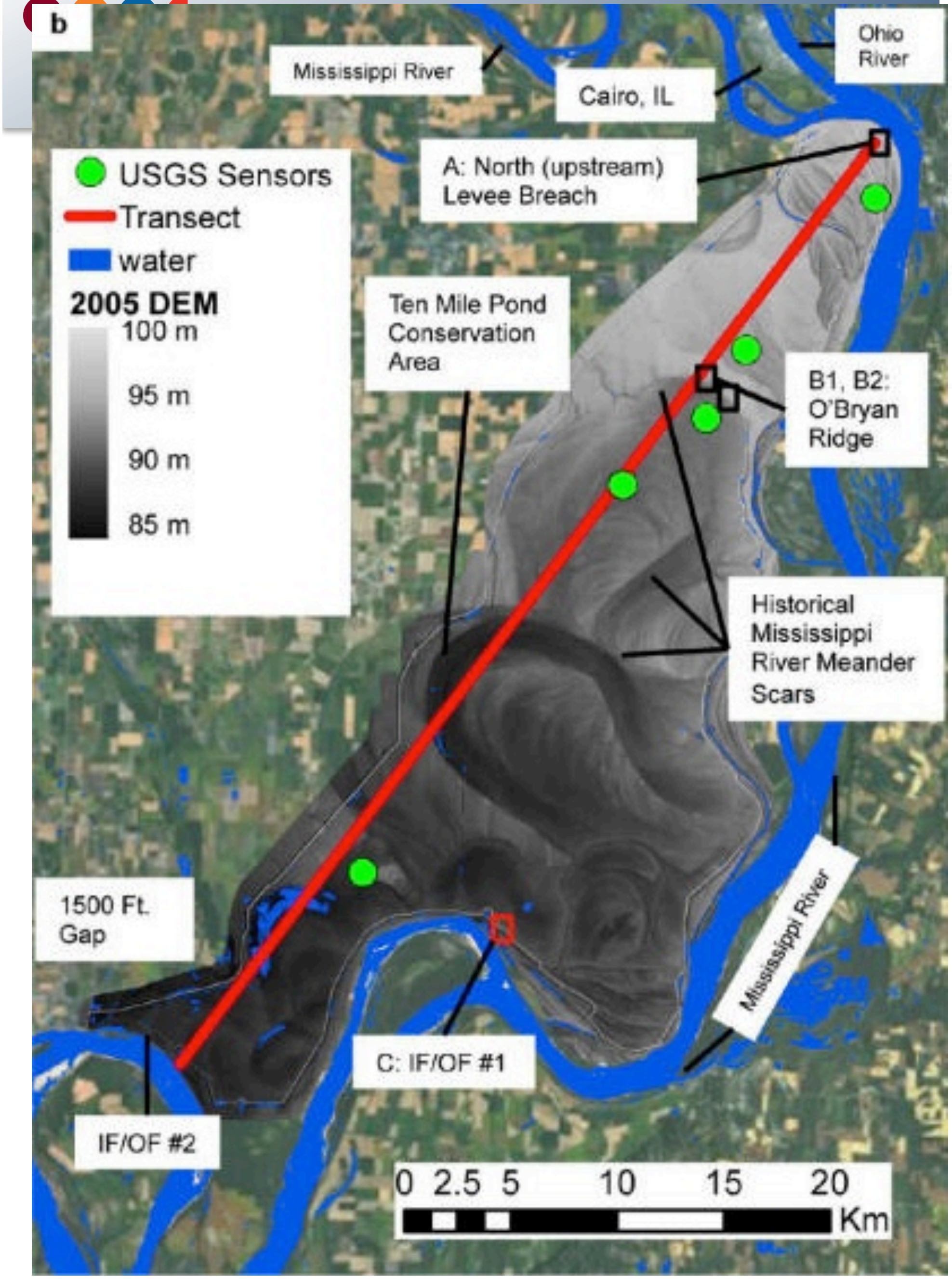
Extreme Mississippi Flood 2011*

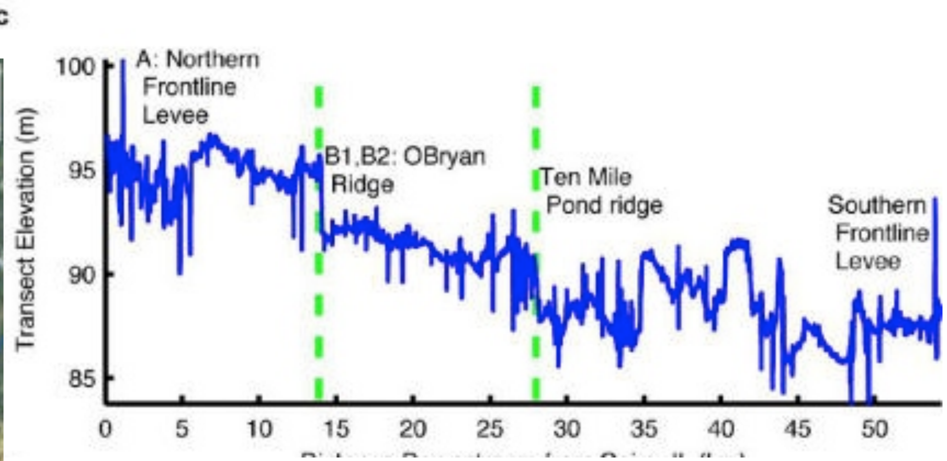
- Birds Point New Madrid (BPNM) Floodway is an agricultural region located west of Mississippi River just south of its confluence with the Ohio River, near city of Cairo, Illinois.
- Policy established following disastrous 1927 Flood permits levees surrounding BPNM Floodway to be intentionally breached during extreme floods.
- During historic flood stages of May 2011, Floodway was activated for first time since 1937. U.S. Army Corps of Engineers used blasting agents on May 2, 3 and 5, 2011 to create artificial breaches.

* A. E. Goodwell, Z. Zhu, D. Dutta, J. A. Greenberg, P. Kumar, M. H. Garcia, B. L. Rhoads, R. R. Holmes, G. Parker, D. P. Berretta, R. B. Jacobson (2014). Assessment of Floodplain Vulnerability during Extreme River Flood 2011, Environmental Science and Technology.

Why study breach?

Impact was dramatic. Parts of agricultural floodplain were inundated for over a month.



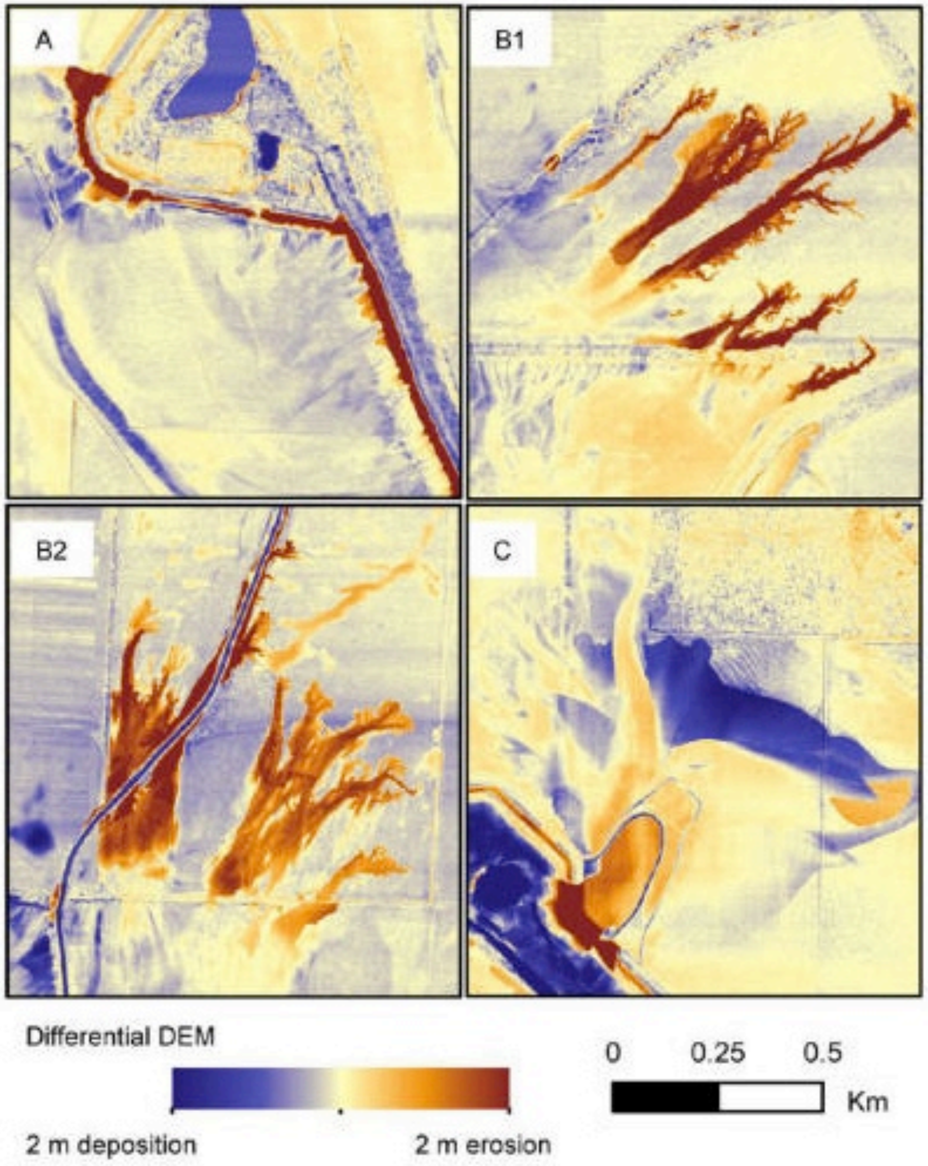


Grey area is floodplain, blue is Mississippi River. Green circles are USGS installed sensors. Levee breach at point A.

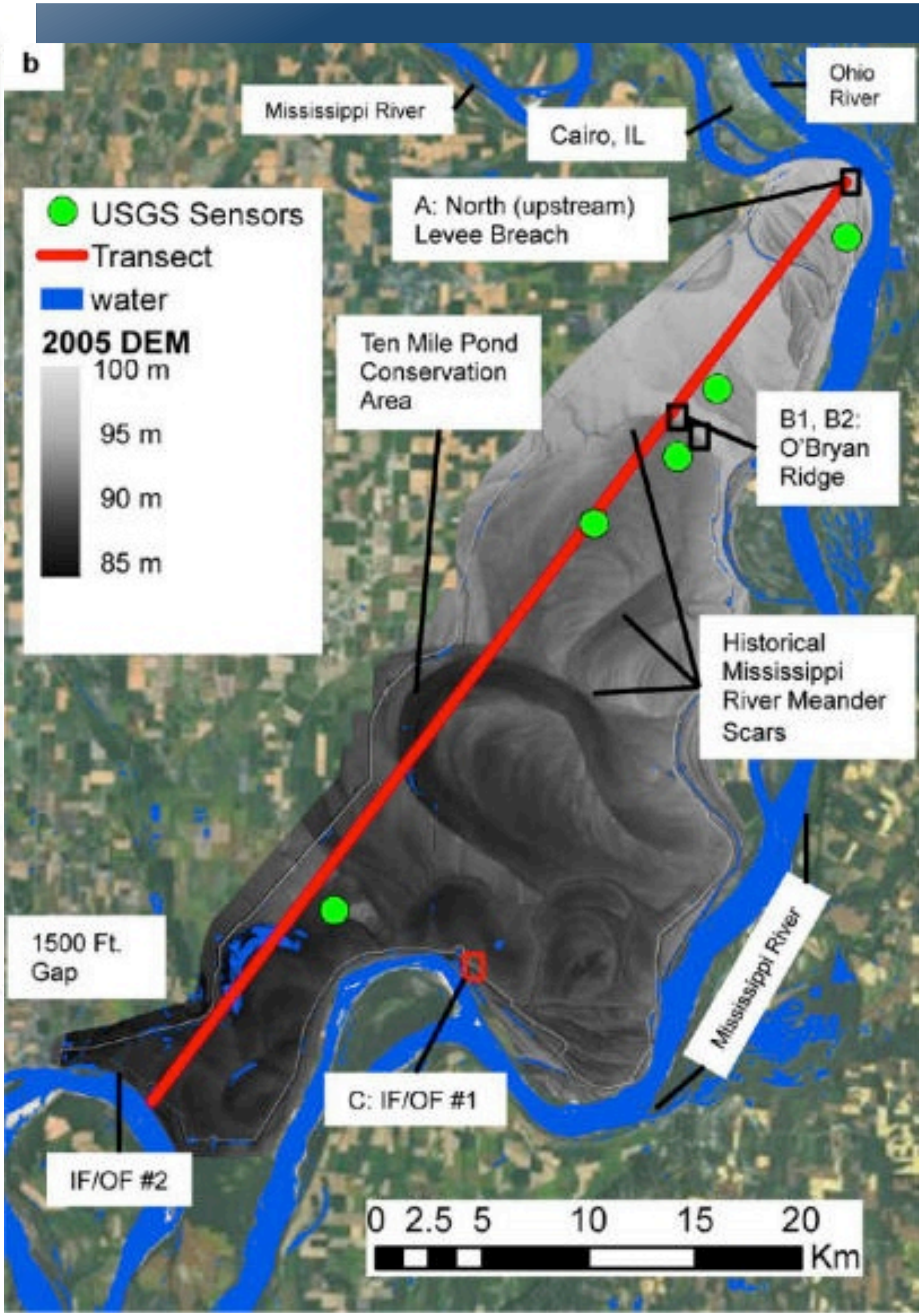
Sudden elevation drop at O'Bryan Ridge(B1, B2) resulted in substantial erosion there.

Oxbow of old river bed (Ten Mile Pond) should have seen substantial erosion but didn't because trees had filled into area since 1937.

Credit for images: A. E. Goodwell, et al. (2014). Assessment of Floodplain Vulnerability during Extreme River Flood 2011, Environmental Science and Technology.

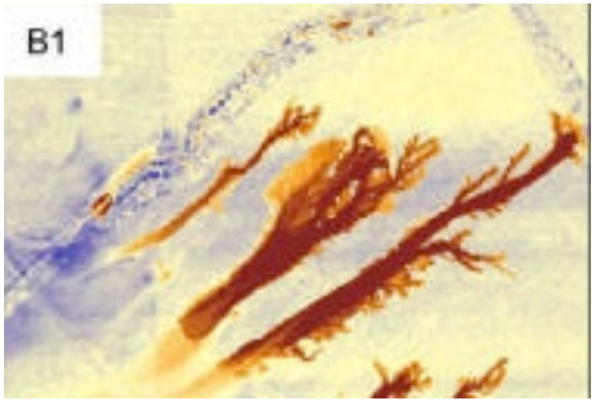


Sudden elevation drop at O'Bryan Ridge resulted in substantial erosion - see B1, B2.



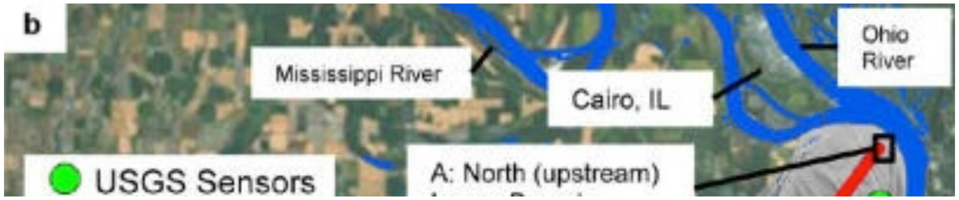
Part II: Publishing Exercise

- Fall 2014 Postdoctoral Fellow asked to publish datasets associated with 2014 ES&T paper.
- SEAD services used to ingest into institutional repository at her university
- Questions Fellow had to address:
  - *What do I publish?*
  - *Why do I publish?*
  - *How do I publish?*



Kinds of Data Used

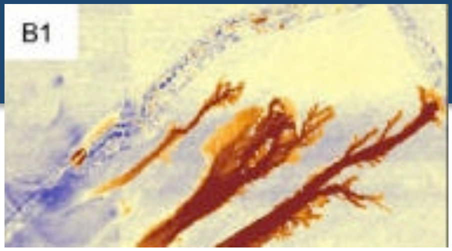
- Map of land elevation (Digital Elevation Model). LIDAR data from 2005 and 2011. US Army Corps of Engineers. *Images*
- Max velocity of water. Hydrosed 2D hydrology model. *Modeled data*
- Woody vegetation mapping. AVIRIS – airborne visible/infrared imaging. *Images*
- Water level sensors. USGS Water level sensors. *Sensor data*
- Soil/landscape sensitivity to erosion.   K/T function. *Computed values*



Published: single Research Object

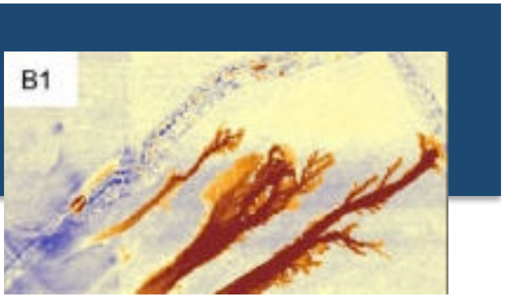
- 5 **Hydrosed model** max velocity files (whole Birds Point New Madrid floodway (BPNM), Case 1 and 2: O'Bryan, Case 1 and 2 Ten Mile Pond)
  - Whole BPNM: `Umax_BPNM.tif` 1.47 MB
  - Case 1 O’Bryan: `Umax_OBR_Case1.tif` 3.23 MB
  - Case 2 O’Bryan: `Umax_OBR_Case2.tif` 3.23 MB
  - Case 1 Ten Mile Pond: `Umax_TMP_Case1.tif` 6.59 MB
  - Case 2 Ten Mile Pond: `Umax_TMP_Case2.tif` 6.59 MB
- K/T Factor **soil map**: `KT_BPNMShape.tif` 7.34 MB
- Classified AVIRIS map (**woody vegetation**) `AVIRIS_50M_BPNMshape.tif` 919.48 KB
- Original **differential DEM** (1.5 m resolution) `Sub_Original_5ft.tif` 2.84 GB
- Final **differential DEM** (10 m resolution) `Sub_Corrected_10m.tif` 72.75 MB
- Final **differential DEM** (3 m resolution) `Sub_Corrected_3m.tif` 798.12 MB

Publishing Exercise



- *Why do I publish?*
  - ***A: publish data necessary to reproduce images in published paper***
- *What do I publish?*
  - ***A: publish data necessary to reproduce images in published paper***

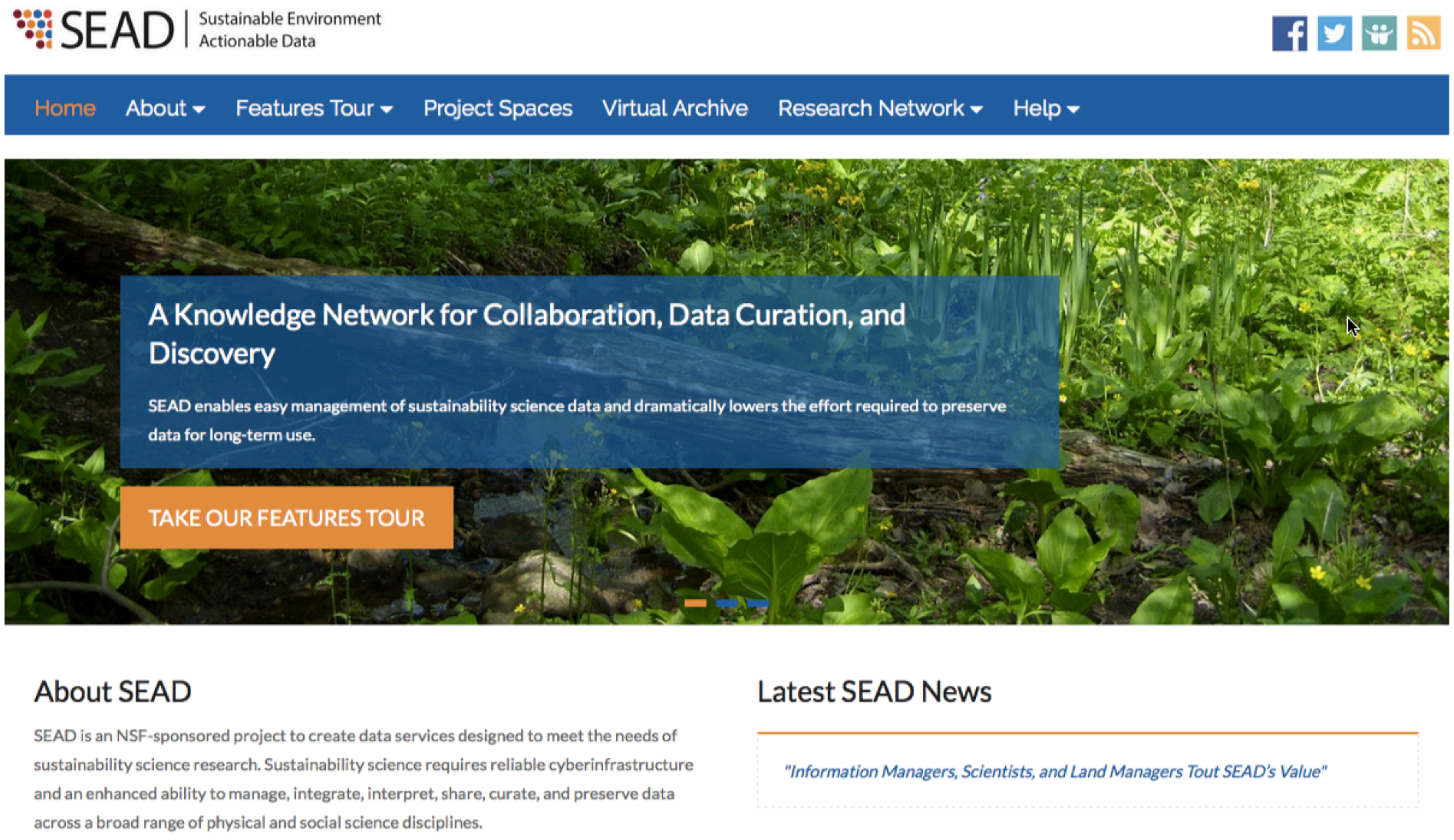
Implications



- TIF files (geoTIF) not sufficient for data reuse
- 3.74 GB Research Object too large for typical repositories
- Data to support images is 3.74 GB, data sufficient for new research substantially larger
- Libraries need partnerships with large scale storage providers

- Umax_BPNM.tif
- Umax_OBR_Case1.t
- Umax_OBR_Case2.t
- Umax_TMP_Case1.t
- Umax_TMP_Case2.t
- KT_BPNMShape.ti
- AVIRIS_50M_BPNM
- Sub_Original_5ft

Technology frame of reference: SEAD



A Knowledge Network for Collaboration, Data Curation, and Discovery

SEAD enables easy management of sustainability science data and dramatically lowers the effort required to preserve data for long-term use.

About SEAD

SEAD is an NSF-sponsored project to create data services designed to meet the needs of sustainability science research. Sustainability science requires reliable cyberinfrastructure and an enhanced ability to manage, integrate, interpret, share, curate, and preserve data across a broad range of physical and social science disciplines.

Latest SEAD News

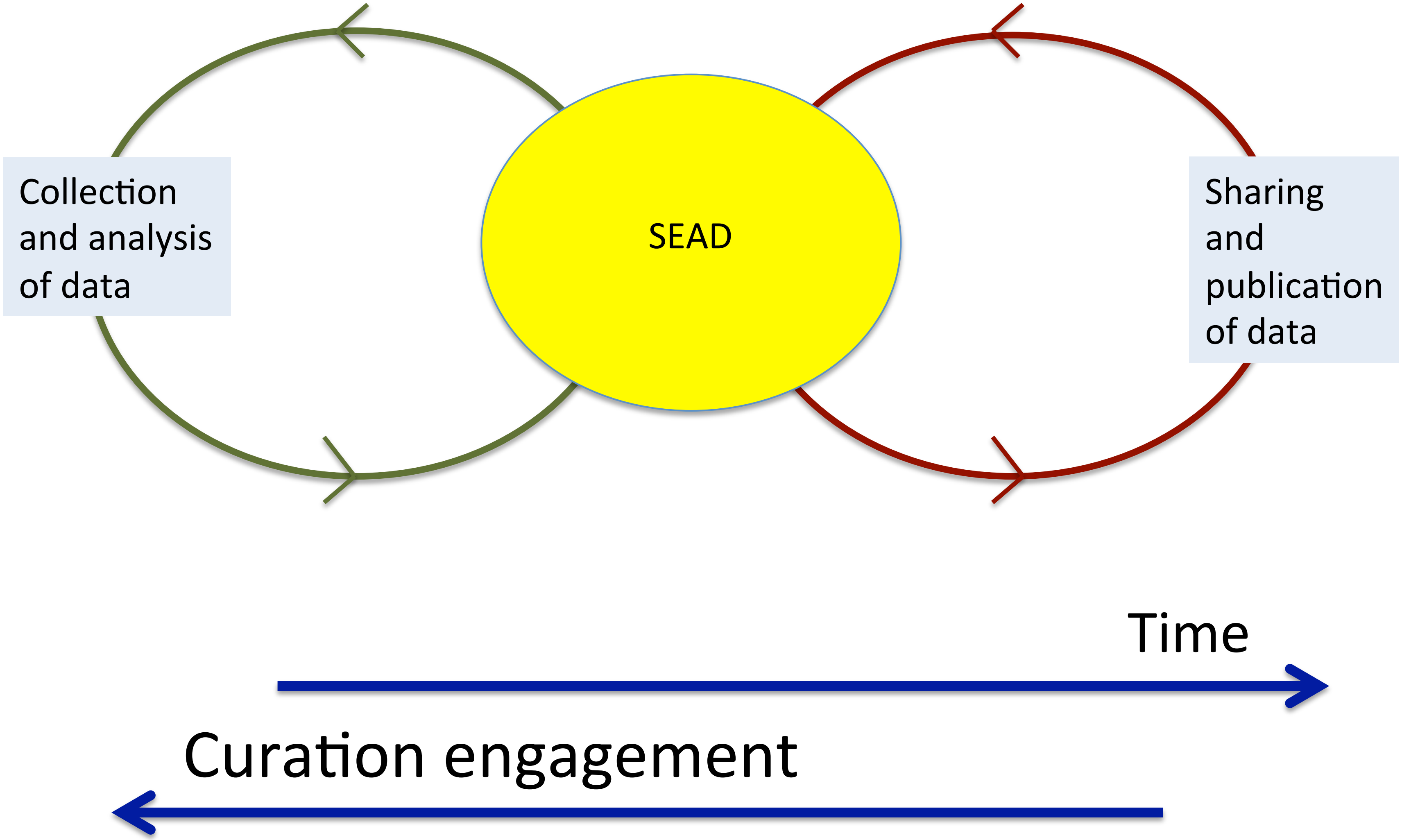
"Information Managers, Scientists, and Land Managers Tout SEAD's Value"

SEAD Leadership

Clockwise from top: Margaret Hedstrom, UMich; Beth Plale, IU; Praveen Kumar, UIUC; Jim Myers, UMich; Sandy Payette, UMich

Cooperative agreement #OCI0940824

SEAD – Researcher's Perspective



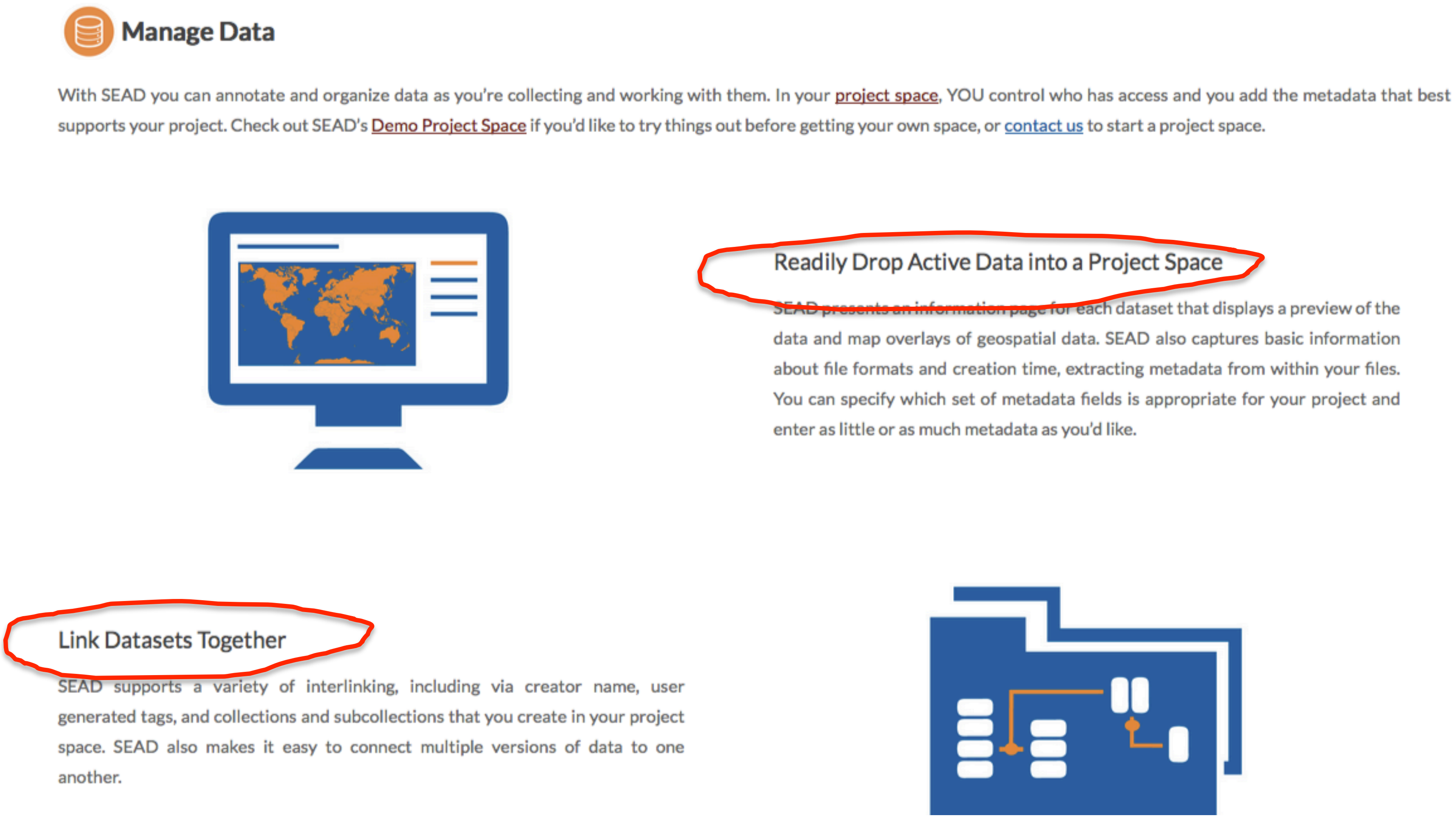
Collection and analysis of data

SEAD

Sharing and publication of data

Time

Curation engagement



Manage Data

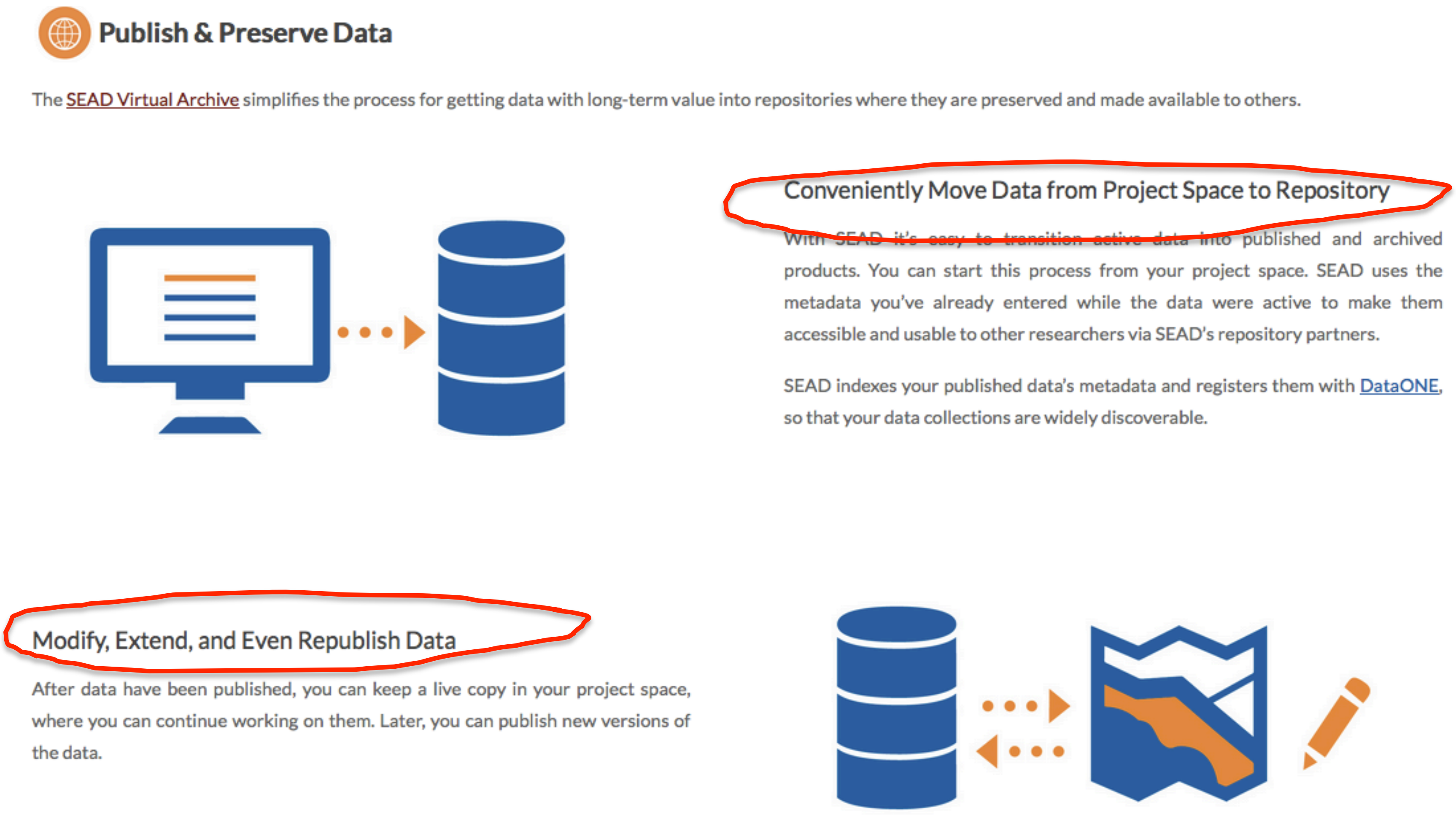
With SEAD you can annotate and organize data as you're collecting and working with them. In your project space, YOU control who has access and you add the metadata that best supports your project. Check out SEAD's Demo Project Space if you'd like to try things out before getting your own space, or contact us to start a project space.

Readily Drop Active Data into a Project Space

SEAD presents an information page for each dataset that displays a preview of the data and map overlays of geospatial data. SEAD also captures basic information about file formats and creation time, extracting metadata from within your files. You can specify which set of metadata fields is appropriate for your project and enter as little or as much metadata as you'd like.

Link Datasets Together

SEAD supports a variety of interlinking, including via creator name, user generated tags, and collections and subcollections that you create in your project space. SEAD also makes it easy to connect multiple versions of data to one another.



Publish & Preserve Data

The SEAD Virtual Archive simplifies the process for getting data with long-term value into repositories where they are preserved and made available to others.

Conveniently Move Data from Project Space to Repository

With SEAD it's easy to transition active data into published and archived products. You can start this process from your project space. SEAD uses the metadata you've already entered while the data were active to make them accessible and usable to other researchers via SEAD's repository partners.

SEAD indexes your published data's metadata and registers them with DataONE, so that your data collections are widely discoverable.

Modify, Extend, and Even Republish Data

After data have been published, you can keep a live copy in your project space, where you can continue working on them. Later, you can publish new versions of the data.

SEAD Technology

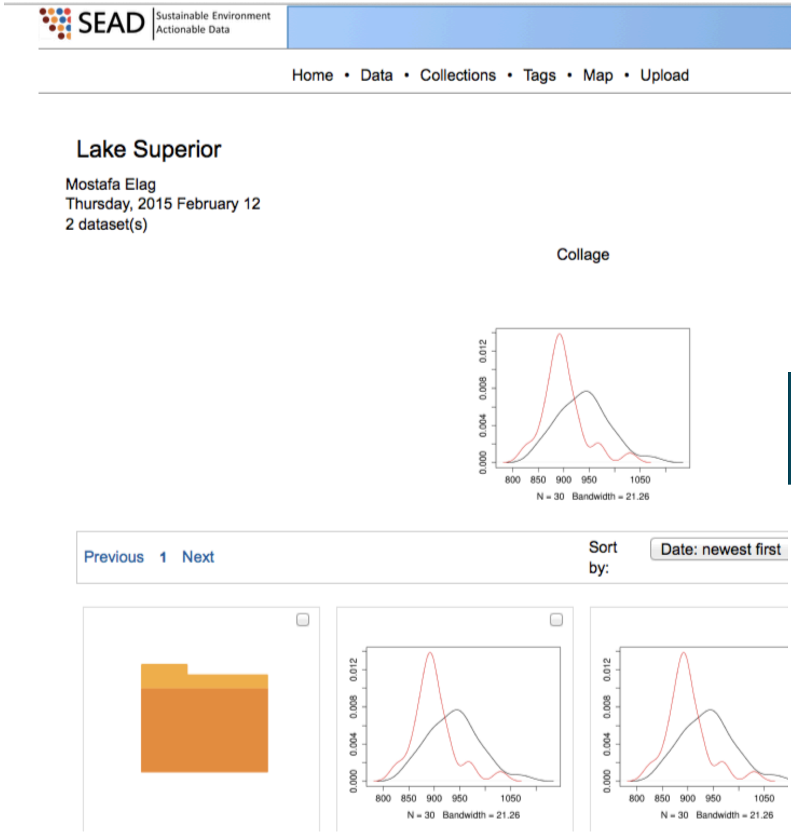
- **Active Project Spaces** : Collaboration / file sharing space for research projects and staging area for data curation prior to publishing
- **Active Curation and Publishing Services**

  i. Publish workflow : push publishable data (as Research Objects) through publish lifecycle

  ii. People, Data, Things service

  iii. Matchmaker : select appropriate preservation and discovery environment

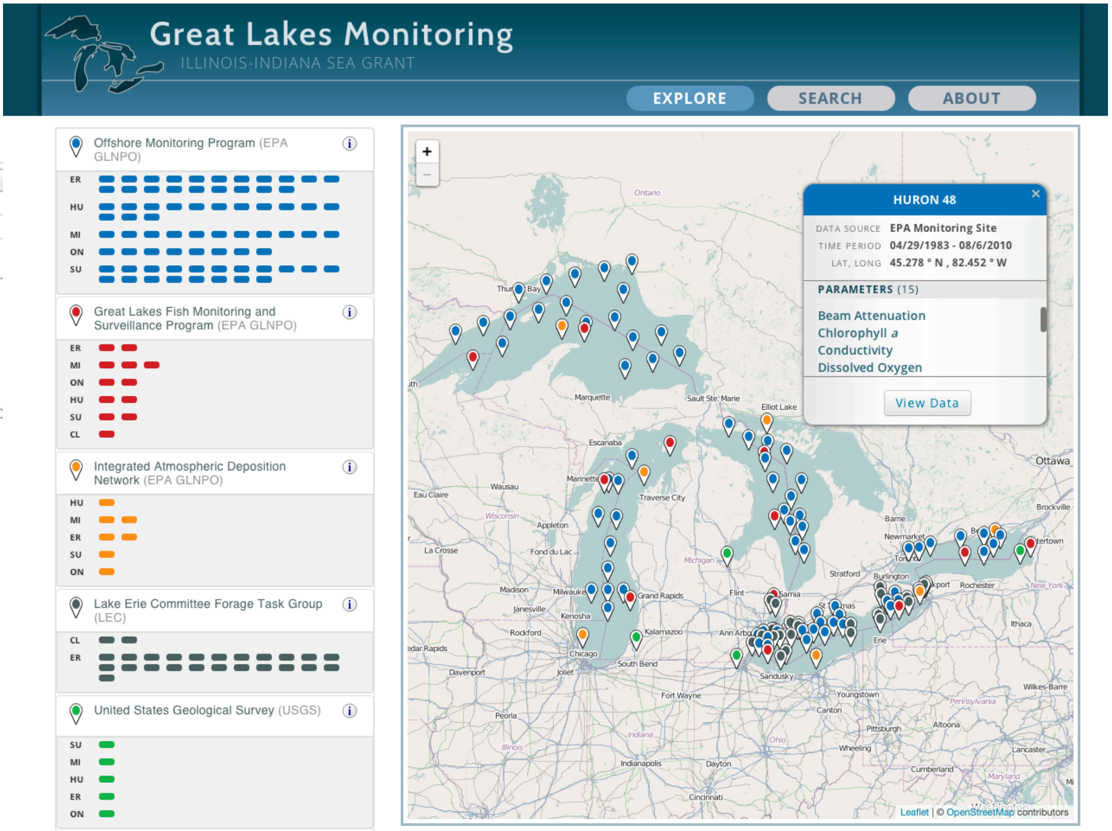
Project Spaces … organize, describe, visualize



Lake Superior

Mostafa Elag
Thursday, 2015 February 12
2 dataset(s)

dataset collections with metadata



Offshore Monitoring Program (EPA GLNPO)
ER, HU, MI, ON, SU

Great Lakes Fish Monitoring and Surveillance Program (EPA GLNPO)
ER, MI, ON, HU, SU, CL

Integrated Atmospheric Deposition Network (EPA GLNPO)
HU, MI, ER, SU, ON

Lake Erie Committee Forage Task Group (LEC)
CL, ER

United States Geological Survey (USGS)
SU, MI, HU, ER, ON

HURON 48

DATA SOURCE EPA Monitoring Site
TIME PERIOD 04/29/1983 - 08/6/2010
LAT, LONG 45.278 ° N , 82.452 ° W

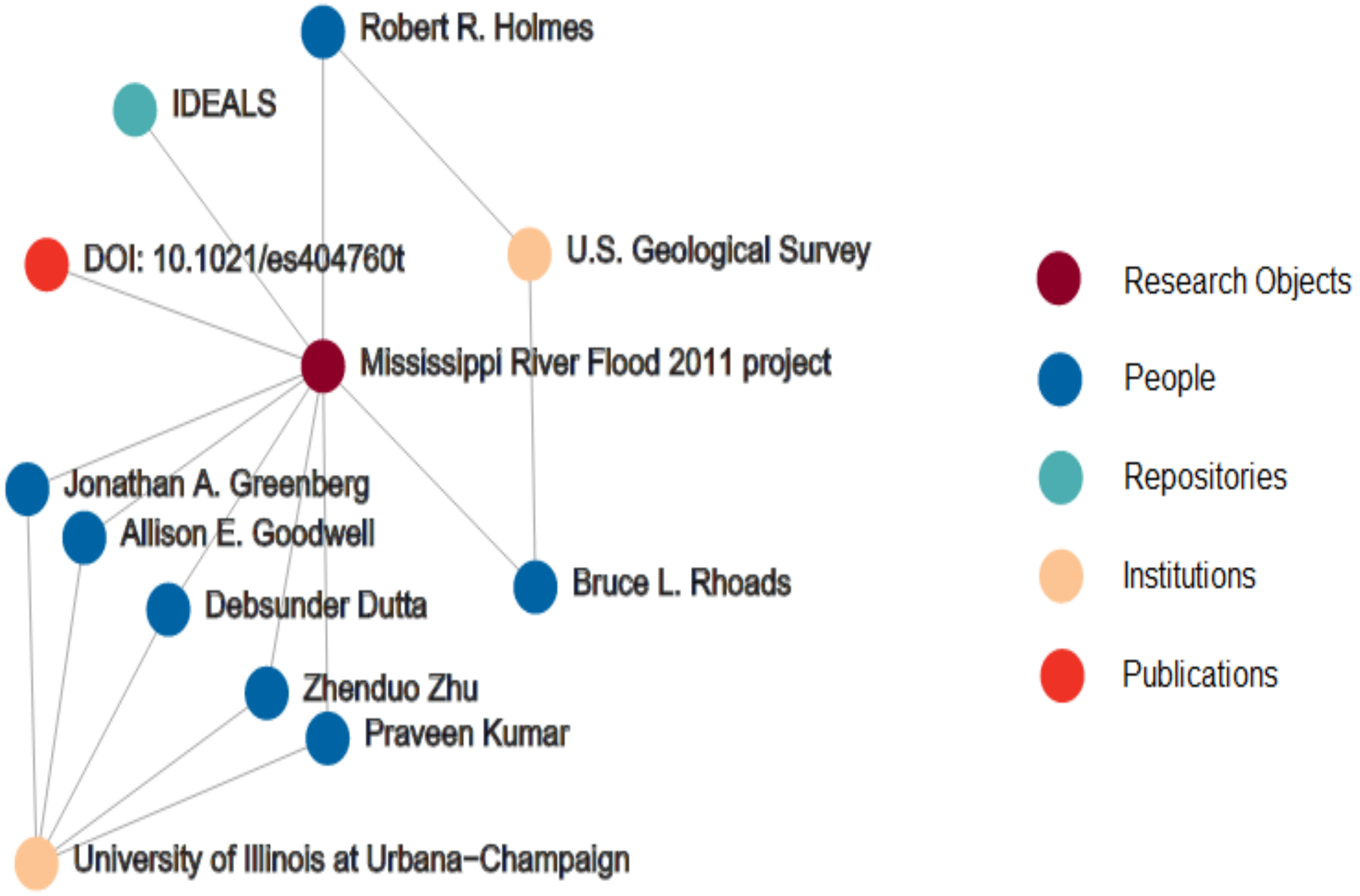
PARAMETERS (15)

Beam Attenuation
Chlorophyll *a*
Conductivity
Dissolved Oxygen

View Data

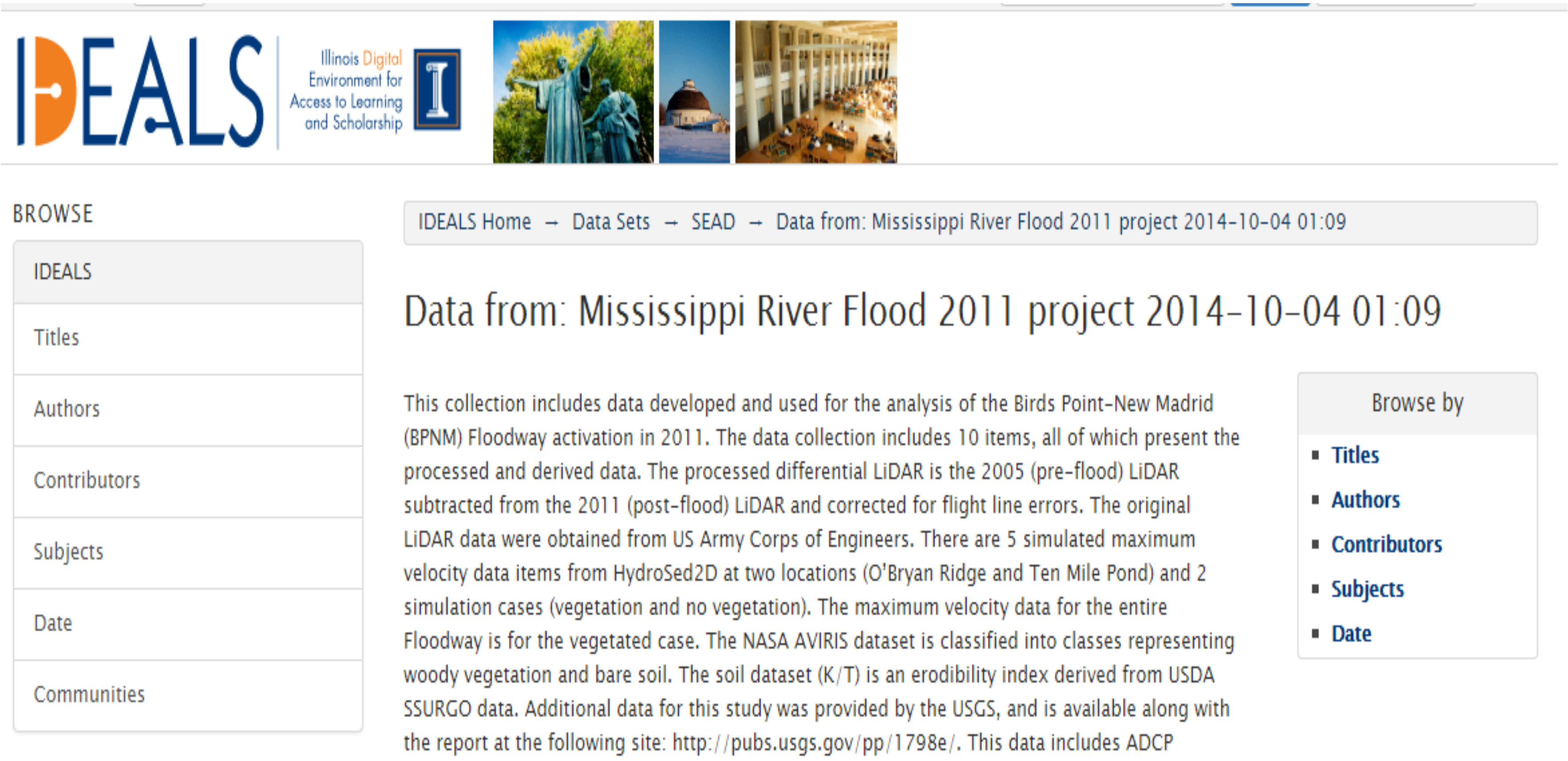
Leaflet | © OpenStreetMap contributors

geo dashboard



People, Data, Things Service: linked data of profiles for researchers, repositories, data, and Trust Threads provenance

Matchmaker : decision support service selects appropriate preservation environment based on needs of repository, data, creator



IDEALS
Illinois Digital Environment for Access to Learning and Scholarship

BROWSE

- IDEALS
- Titles
- Authors
- Contributors
- Subjects
- Date
- Communities

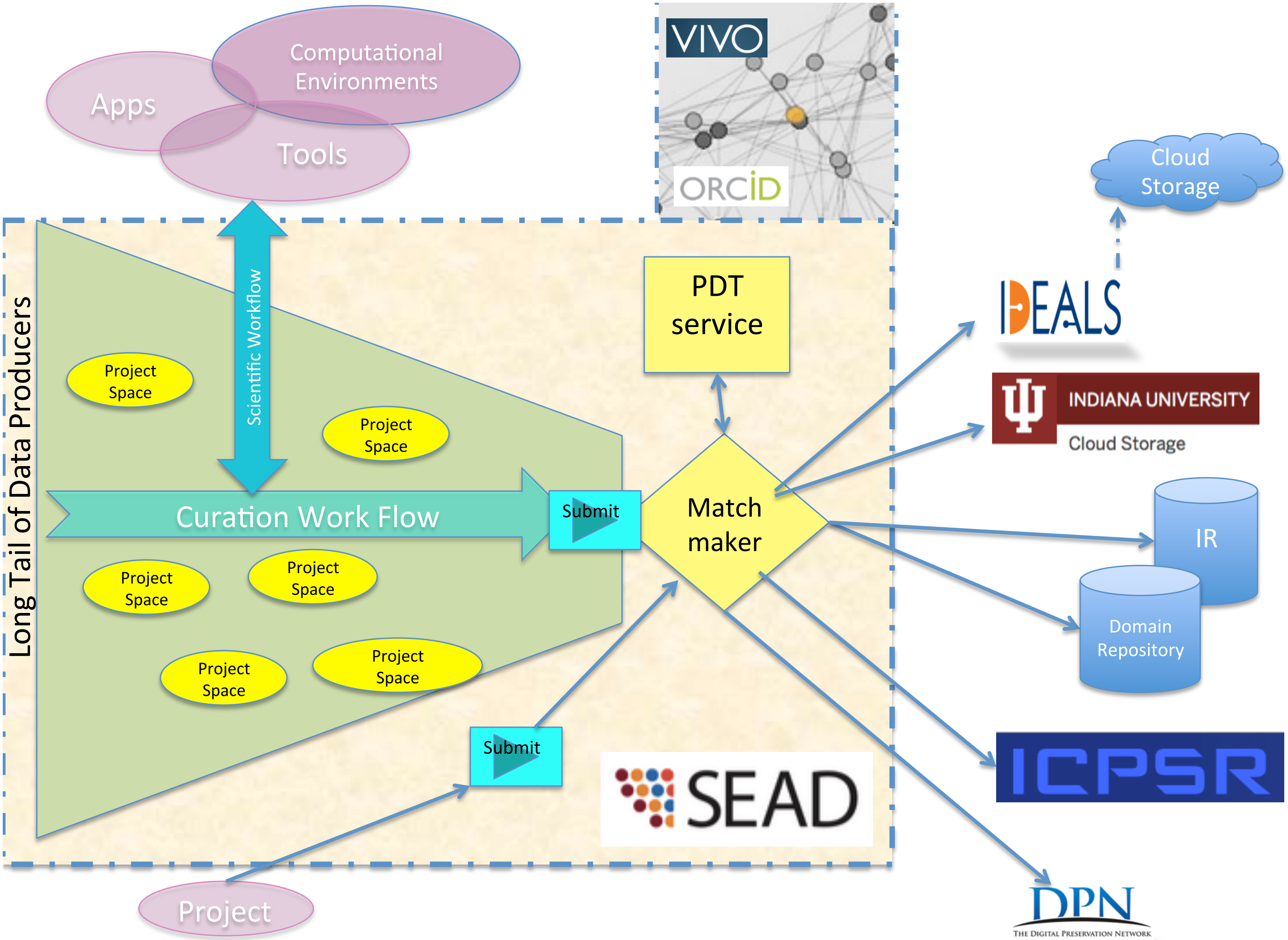
IDEALS Home → Data Sets → SEAD → Data from: Mississippi River Flood 2011 project 2014-10-04 01:09

Data from: Mississippi River Flood 2011 project 2014-10-04 01:09

This collection includes data developed and used for the analysis of the Birds Point-New Madrid (BPNM) Floodway activation in 2011. The data collection includes 10 items, all of which present the processed and derived data. The processed differential LiDAR is the 2005 (pre-flood) LiDAR subtracted from the 2011 (post-flood) LiDAR and corrected for flight line errors. The original LiDAR data were obtained from US Army Corps of Engineers. There are 5 simulated maximum velocity data items from HydroSed2D at two locations (O'Bryan Ridge and Ten Mile Pond) and 2 simulation cases (vegetation and no vegetation). The maximum velocity data for the entire Floodway is for the vegetated case. The NASA AVIRIS dataset is classified into classes representing woody vegetation and bare soil. The soil dataset (K/T) is an erodibility index derived from USDA SSURGO data. Additional data for this study was provided by the USGS, and is available along with the report at the following site: http://pubs.usgs.gov/pp/1798e/. This data includes ADCP

Browse by

- Titles
- Authors
- Contributors
- Subjects
- Date


Computational Environments
Apps
Tools
VIVO
ORCID
Cloud Storage
Long Tail of Data Producers
Scientific Workflow
PDT service
Project Space
Project Space
Curation Work Flow
Submit
Match maker
IDEALS
INDIANA UNIVERSITY
Cloud Storage
IR
Project Space
Project Space
Domain Repository
Project Space
Project Space
Submit
SEAD
ICPSR
Project
DPN
THE DIGITAL PRESERVATION NETWORK

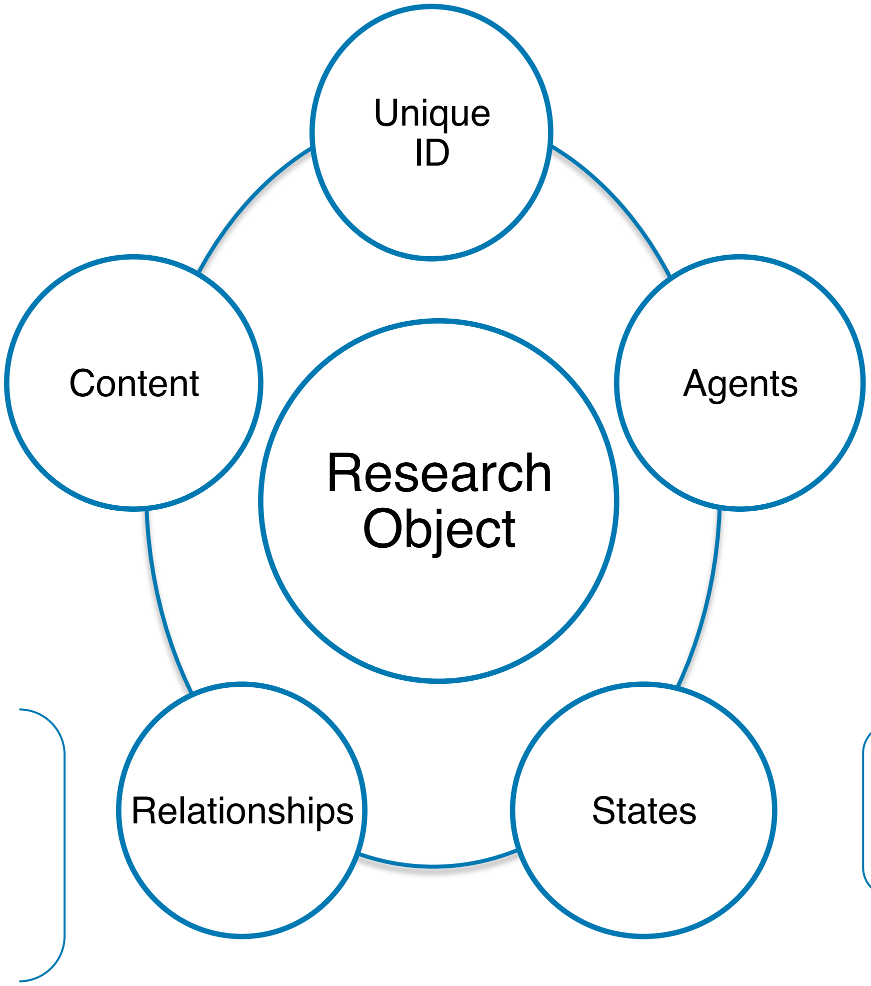
Intent to publish

- At point in time researcher ready to publish data, she makes decision about what data – from amongst all data sources used, consulted, and created, should be include in publishable result.
- Data Curator: At this point (at latest) is when data curator should get engaged
- Curator works in-situ and alongside researcher to assemble publishable object
    - Occurs in SEAD Project Spaces

Tracking life of research object

Research Object Framework

Bundles of resources that use common standards and services to transfer and consume them



Research Object

Unique ID

Content

- Files
- Bitstreams
- Pointers
- Annotative information

Agents

- Data creator
- Curator
- Data re-use scientist

Relationships

- Aggregates
- Related to
- Describes
- Derived from
- Versioned from

States

- Live
- Curated
- Published

Research Object Publish reuse lifecycle

- Research Object begins as *Live Object (LO)*, this is when all future content is still in state of high flux.
- Live Object state is collaborative, multi-person project team working together in single shared project space in SEAD environment

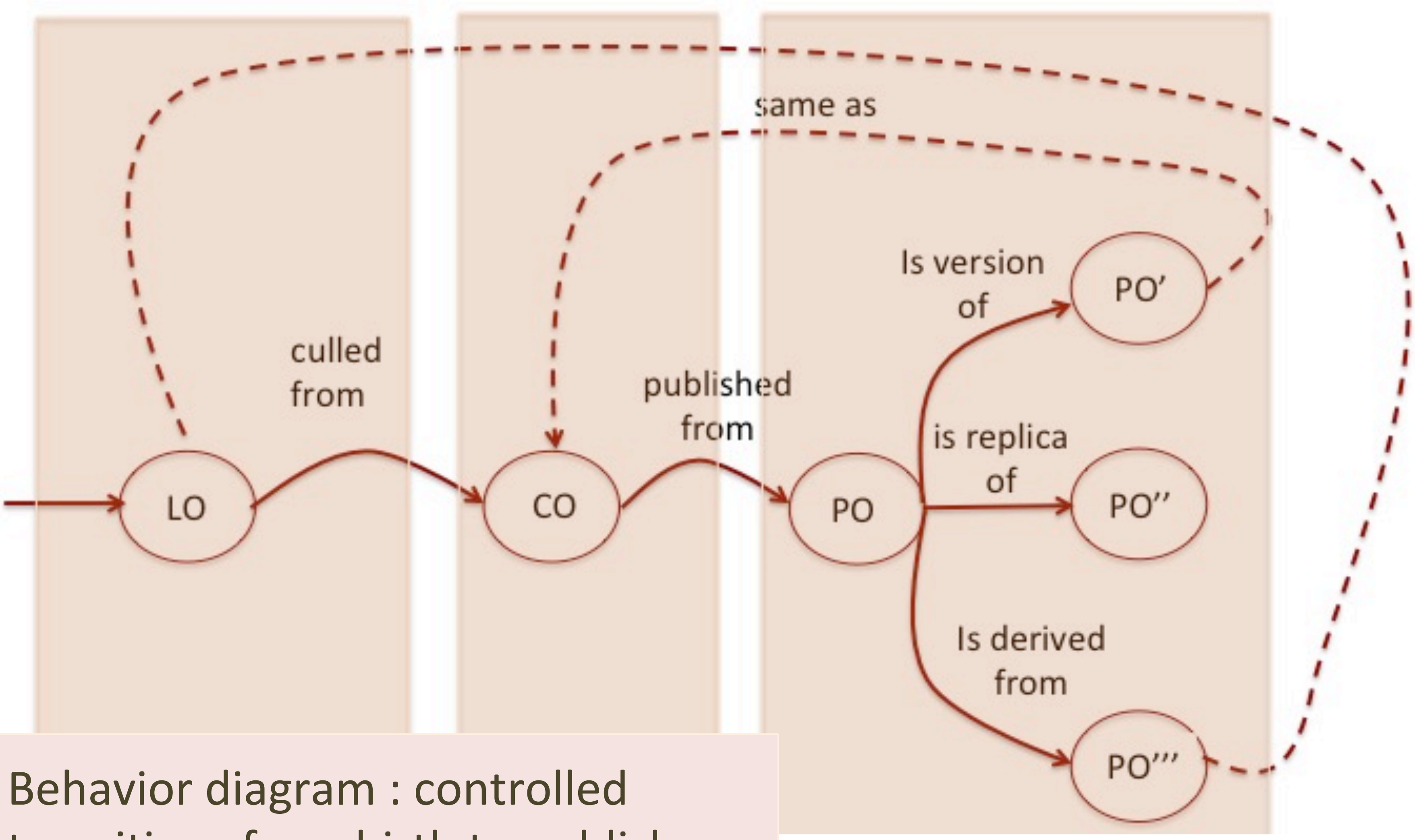
Research Object Publish reuse lifecycle

- At point in time when researcher begins assembly for publishing, she will prune and organize material to publish into new directory or set of directories, or tag specific files.
- This pruned content becomes the *Curation Object (CO)*

Research Object Publish reuse lifecycle

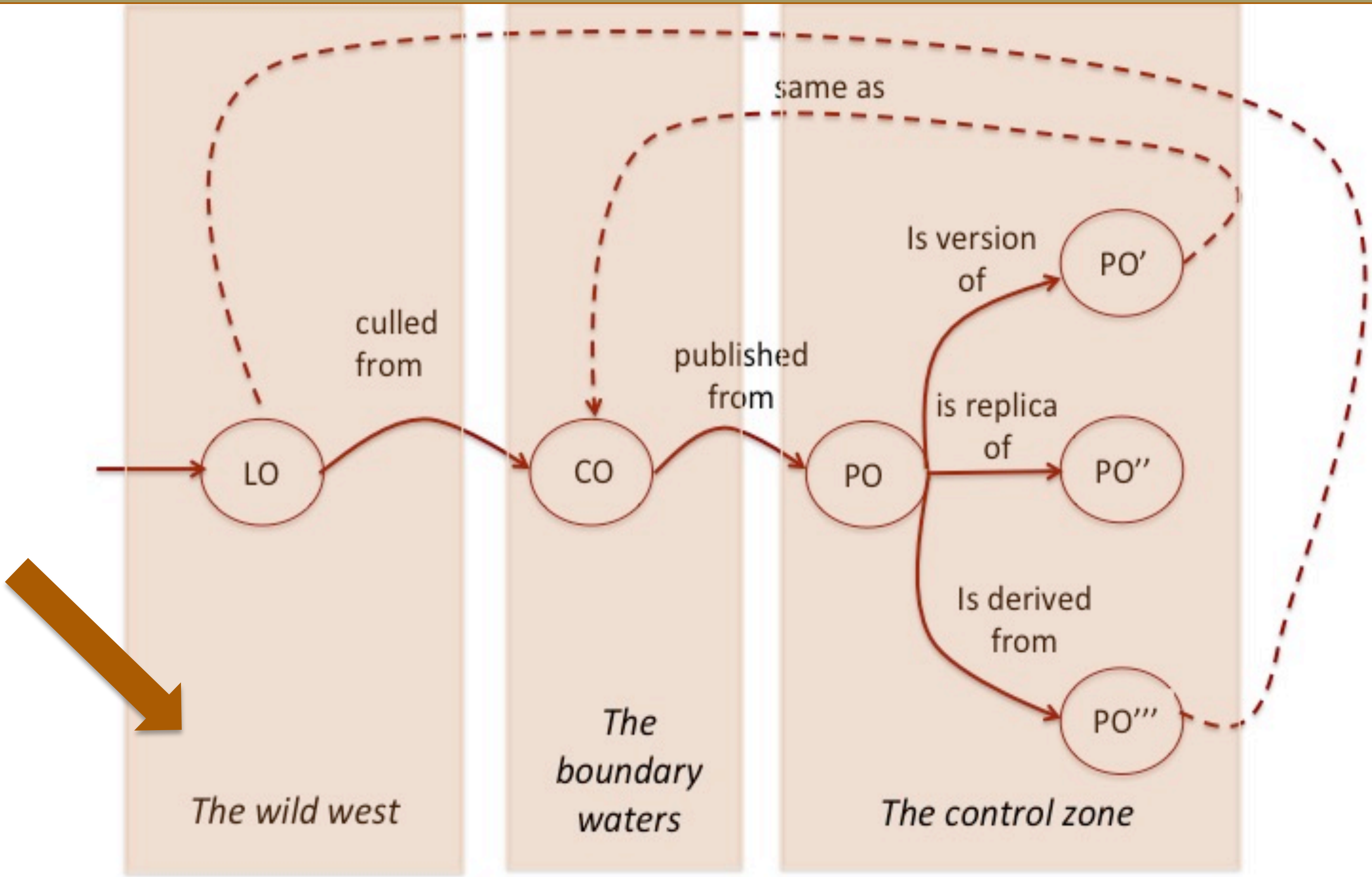
- Once researcher and digital curator agree that content and descriptions of research product are ready, researcher signals *intent to publish* whereupon research object moves from its state as a Curation Object to a new state as a *Published Object (PO)*.
- PO exists in “control zone” where changes are tracked in a rigorous manner.

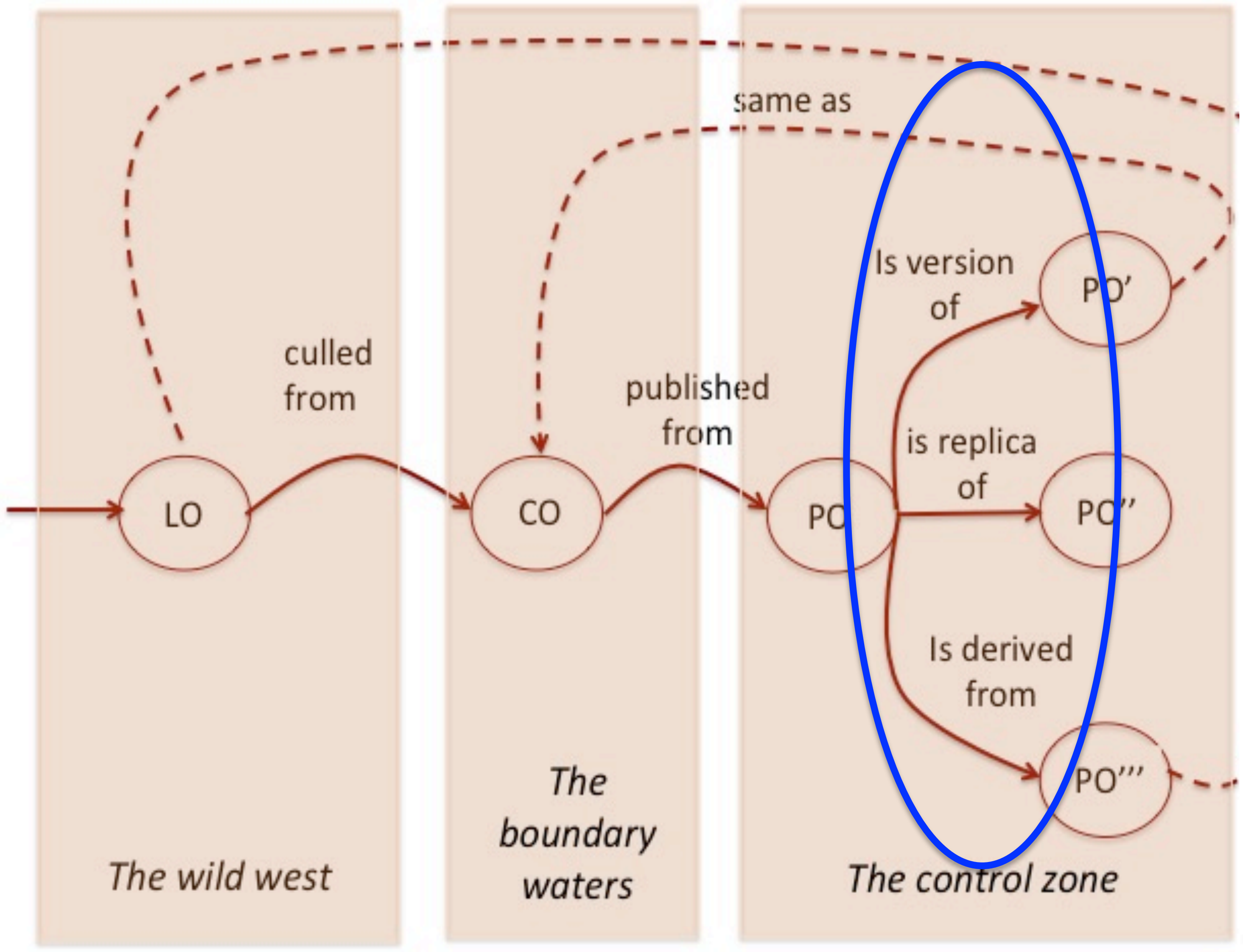
Publish reuse lifecycle



Behavior diagram : controlled transitions from birth to publish

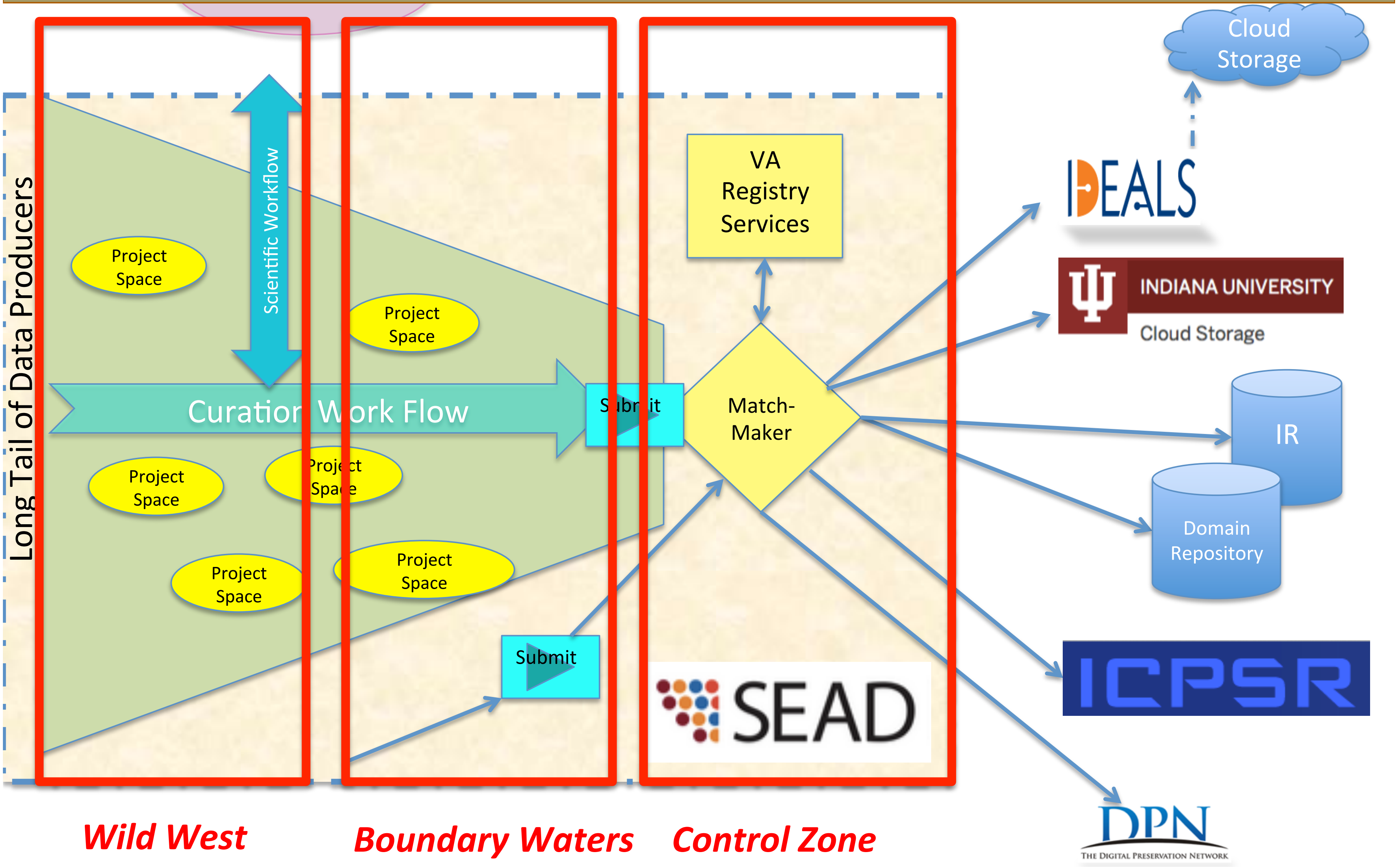
Objects transition from little control in change tracking to full tracking

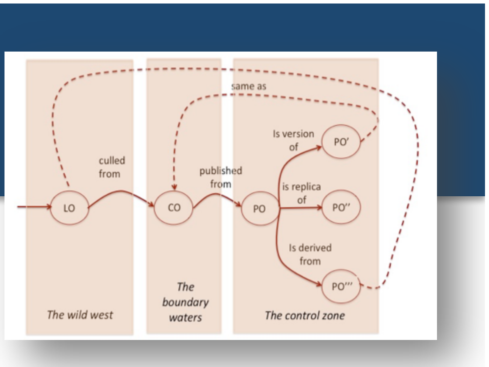




"is version of" relationship means something substantially different from *"is derived from"*. Critical to be able to tell them apart in future.

Stages implemented in Active Curation and Publishing Services





From point in time where data is first conceived as ready for publication through to its first reuse is critical stage in data's life.

If we are unable to bring technical rigor to handling object curation, relationships, and derivations in this early stage, we lose critical opportunity for enhancing trustworthiness of data object:

an opportunity that is irrecoverable.

With special thanks

- Hydrologist postdoc Allison Goodwell, Professor Praveen Kumar, UIUC
- Dr. Inna Kouper, IU
- IU SEAD developer team: Scott McCaulay, Isuru Suriarachchi, Aravindh Varadharaju, Charitha Arachchige, Yuan Luo
- National Science Foundation

RDA RESEARCH DATA ALLIANCE

SEAD Sustainable Environment Actionable Data

NSF