Data Sharing

(a curation perspective)

LIS 598 Data Curation II

Agenda

- Introduce our weekly exercises
- Introduce “curation intervention” and “affordances”
- Describe curation perspective on data sharing:
    - Data Storage
    - Data Access

Protocol Exercises

- Each week (starting this week) there will be an exercise that builds off of or relates to our topic.

- Your group should follow the directions to complete the exercise - we will use this exercise to guide our weekly check-in.

- You will not be graded on these exercises - however, the more effort and time you put into these the easier your final deliverable (protocol) will be.

“All data curation work is a form of intervention - it mediates a relationship between collection and use…In many instances curators are responsible for the transformation of information objects into forms of evidence…data curators hold a great deal of power to afford certain ways of seeing, interacting with, and drawing inferences from data. These are not simple acts- they should be seen and thought of through the lens of affordance theory.”

Allen Renear (2017)

Data Storage

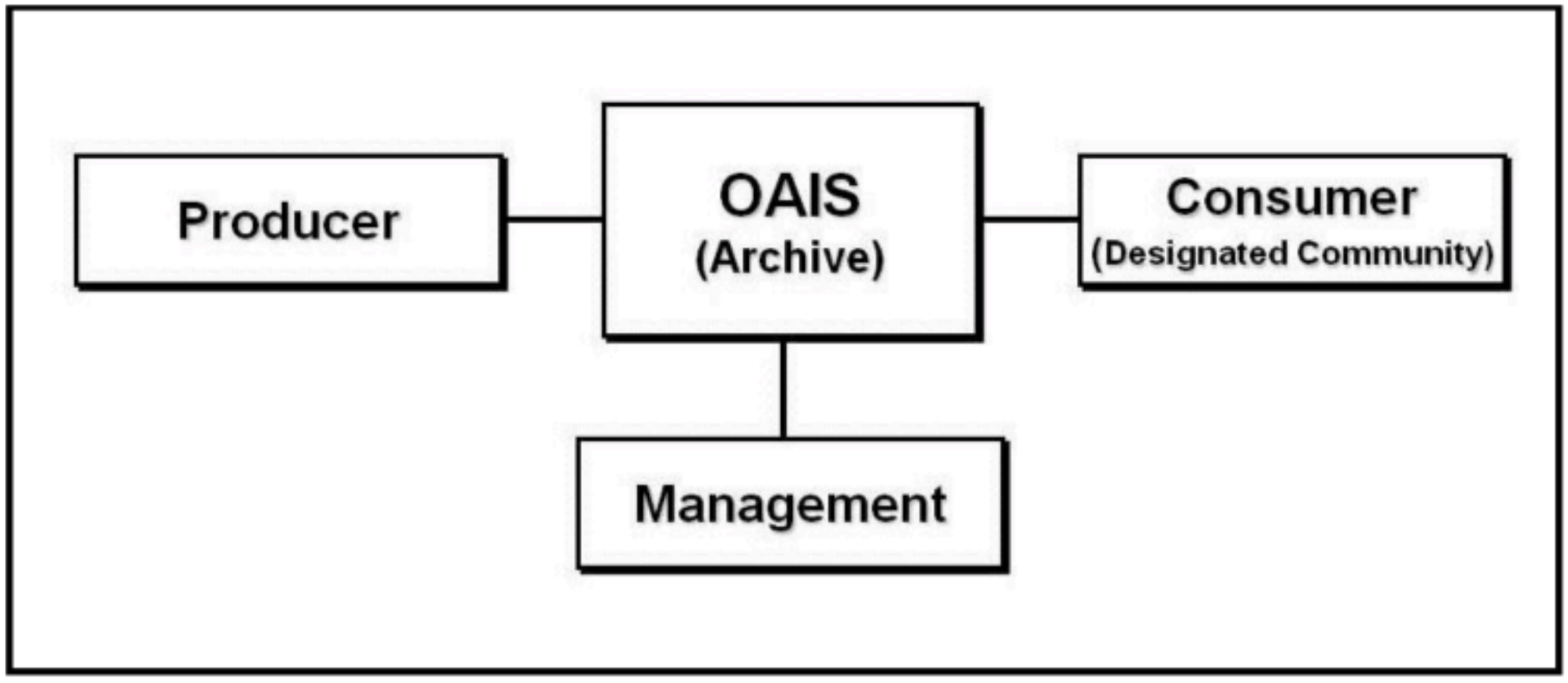
10 basic rules…

- **Anticipate use [1] ; Know use case(s) [2]**
    - Today
- Keep raw data raw [3] ; Store data in open formats [4] ; Structure data for analysis [5]
    - We talked about most of these Week 1-2
- Identifiers should be unique [6] ; Metadata should be linked [7]
    - Data curation I
- Adopt proper privacy controls [8];
    - Weeks 7-8
- **Have a systematic backup scheme [9]; Location and method storage depend on scale [10]**
    - Today

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005097

Anticipate Use

- OAIS - Notion of a designated community is the guiding principle for understanding and anticipating use. The service of our DC is what drives our development of curation affordances .



Use Cases / User Stories

- User stories are short, simple ways of describing what a user wants to achieve in taking some action. These can take the following generic form:
    - As **USER** I want to take an **ACTION** to achieve a **GOAL**

- Use cases are focused on the system or the feature - they define ways of interacting with a feature so that a complicated task can be achieved.

Systematic Backup Scheme

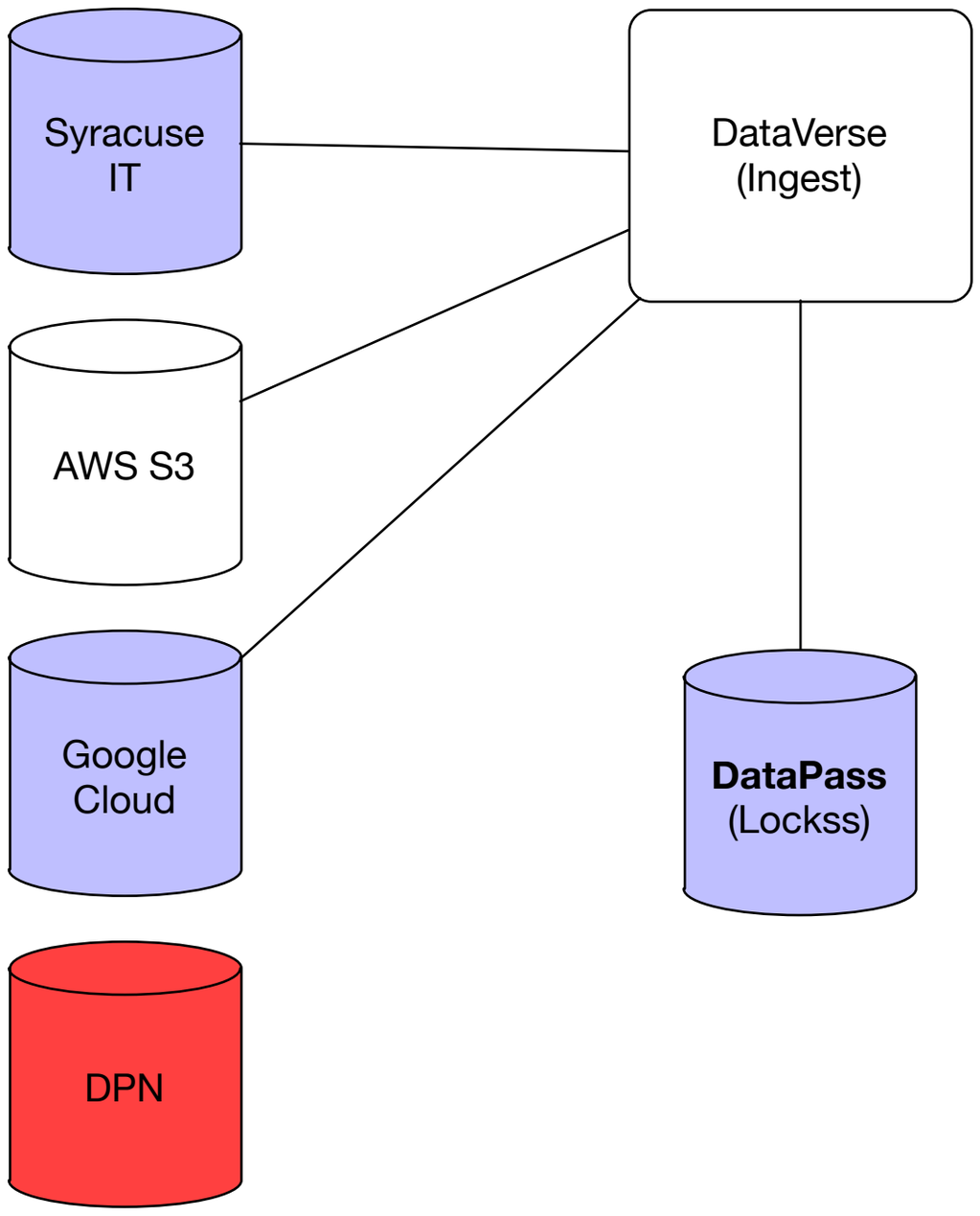
- Regular methods of creating duplicate copies, snapshots, or offsite backups of data packages.

- Monitoring those backups for integrity (or fixity)

- The NDSA Levels of Digital Preservation describes a best practice of at least three copies in different geographic locations with different disaster threats.

    - Many repositories can be configured to distribute data through institutional partnerships or cloud services.

Systematic Backup at Scale



Data Access

Sensitive Data

- “Information that, if exposed, can put a person, group, or institution at risk of harm.”

- For personally identifiable information (PII) this often includes demographic or biometric data such as:
    - Racial or ethnic origin
    - Political opinions
    - Religious or philosophical beliefs
    - Trade union membership
    - Genetic data
    - Biometric data for the purpose of uniquely identifying a natural person
    - Data concerning health or a natural person’s sex life and/or sexual orientation

Credentialing

- Depends upon creating a vetting system for secure access. The approach is akin to a digital passport stating who can go where and for what purposes.

    - Credentialed access in data repositories:
        - IRB clearance to access data
        - Data producer’s consent
        - Application of intent and Data Sharing Agreement that declares penalties for misuse
        - Data Enclaves

Mediated Analysis

- Synthetic Data: Creation of statistically viable copies of data that impact identity, but do not impact significance

- Multi-party Computation: Network protocol that enables the connection to and running of software against a dataset such that answers are obtained, proofs are stored, but no data is exchanged

- Differential Privacy: Some combination of the two above.