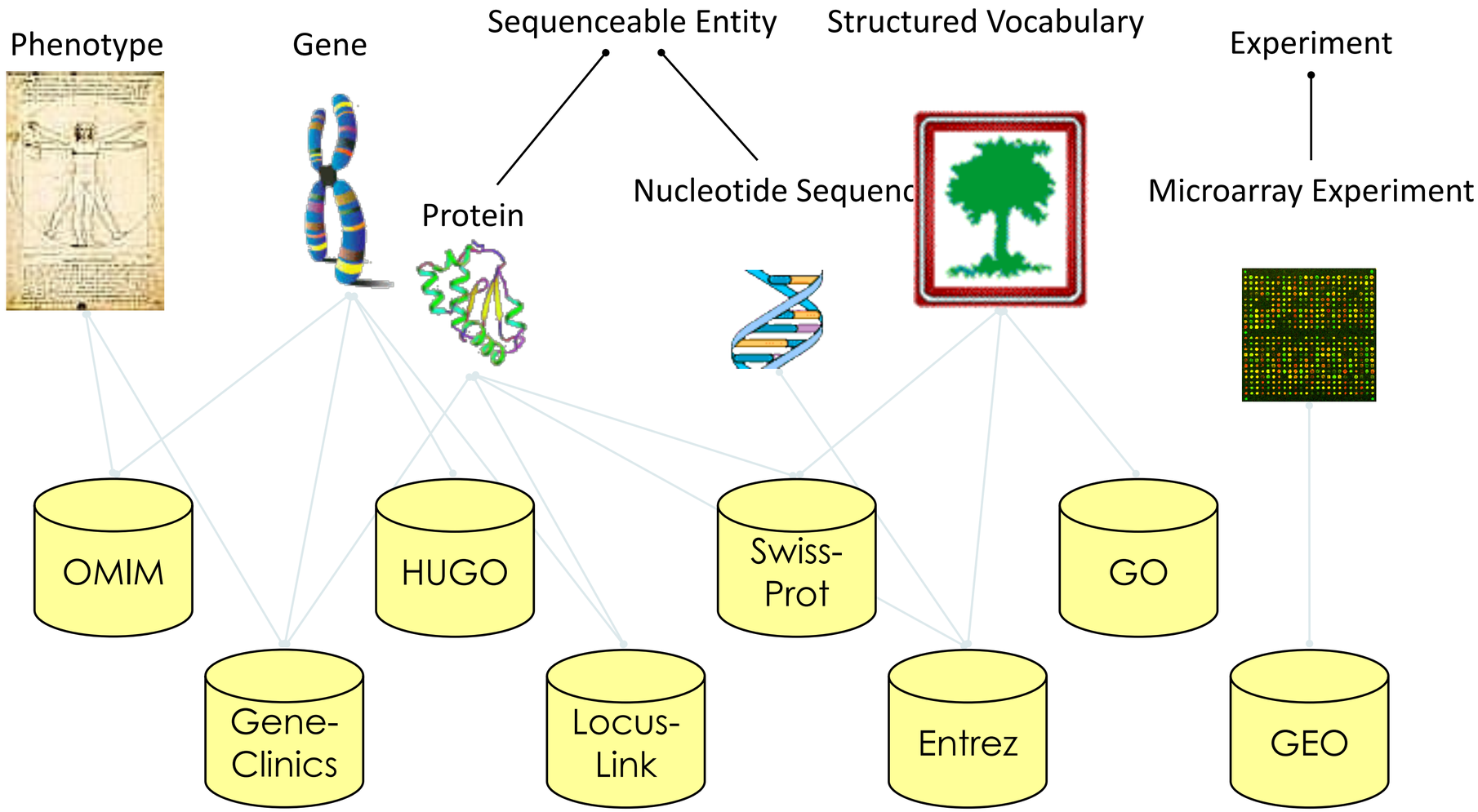
Advanced Data Management (CSCI 680/490)

Data Curation

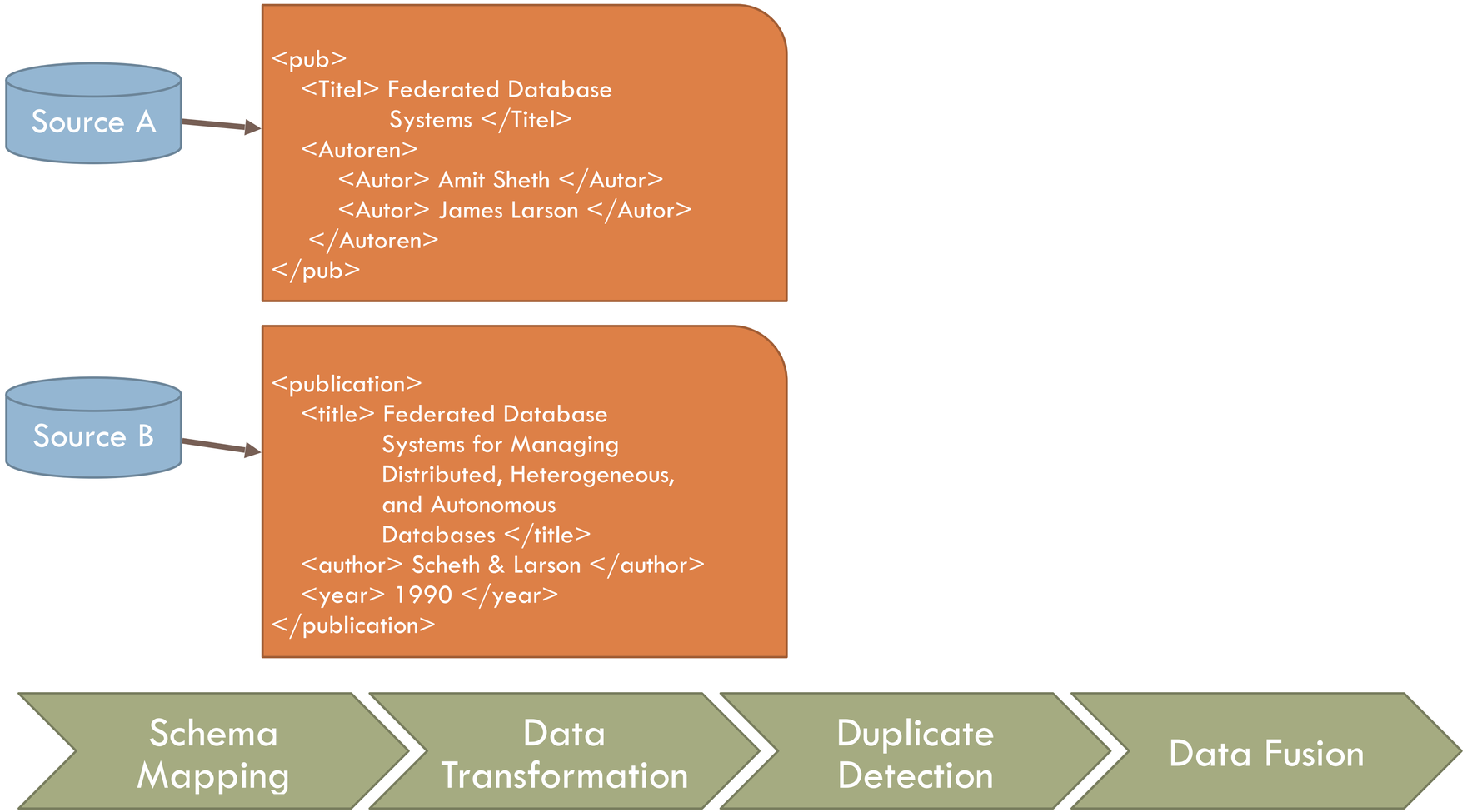
Dr. David Koop

Data Integration: Combine Datasets with Different Data



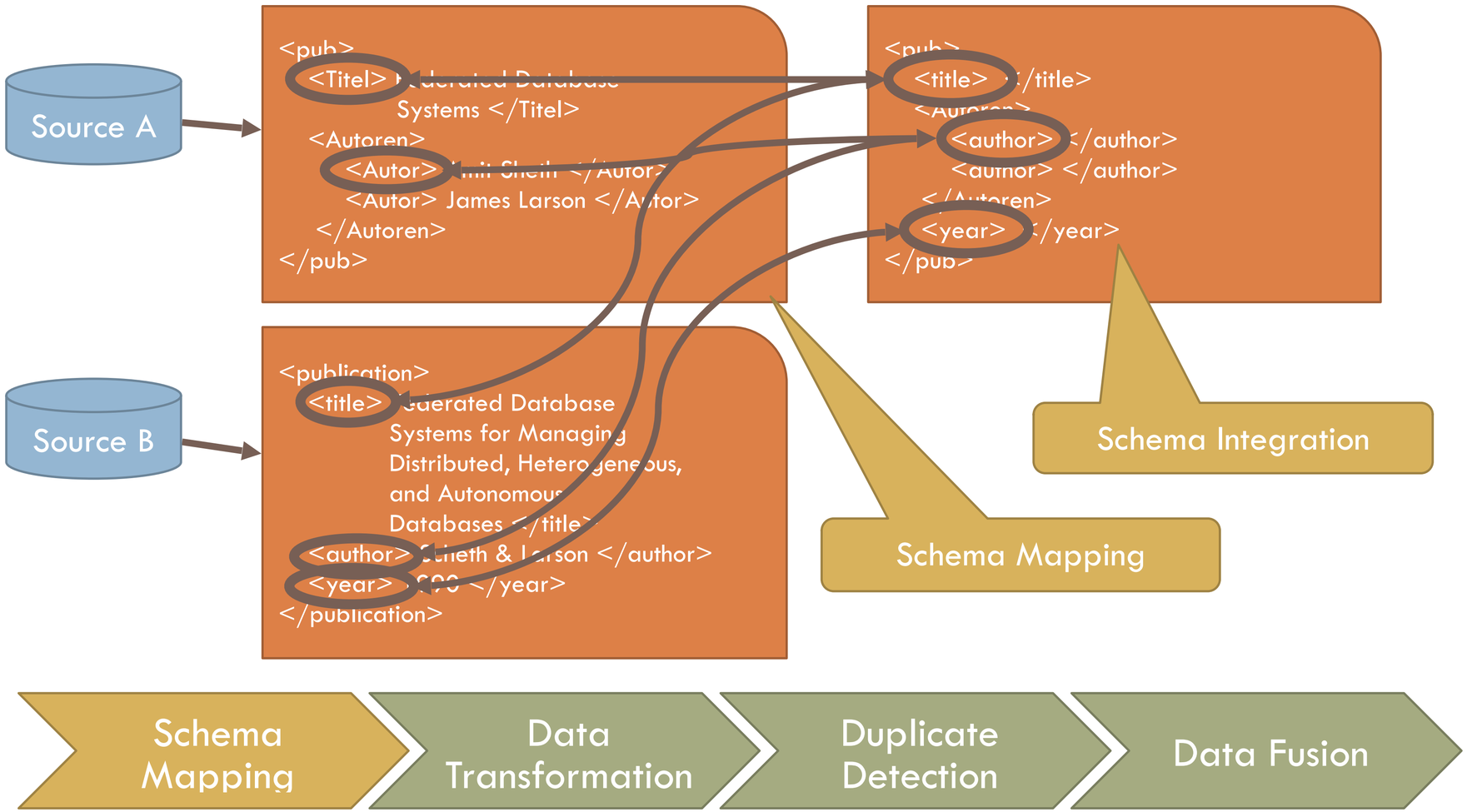
[A. Doan et al., 2012]

Information Integration



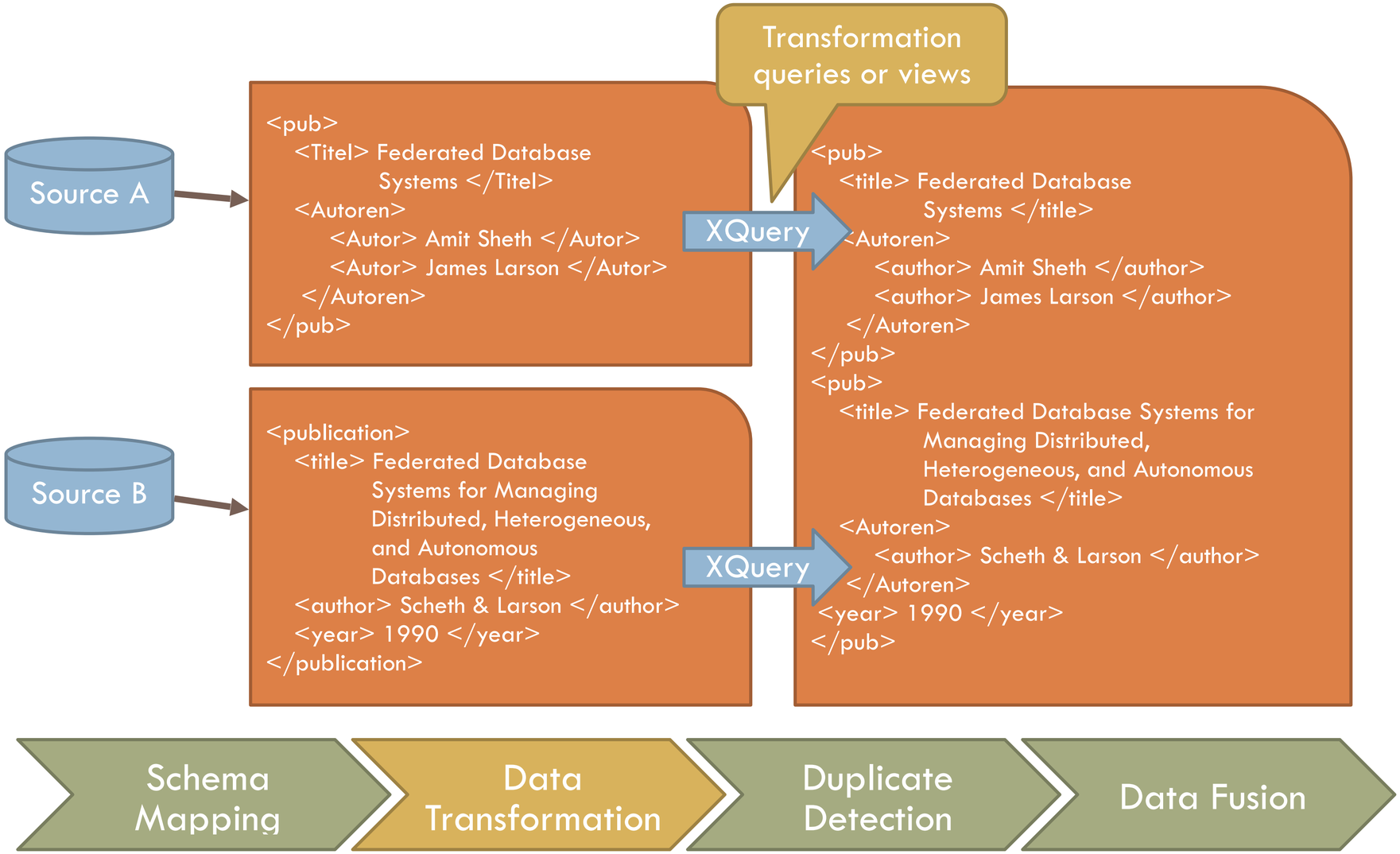
[L. Dong and F. Naumann, 2009]

Information Integration



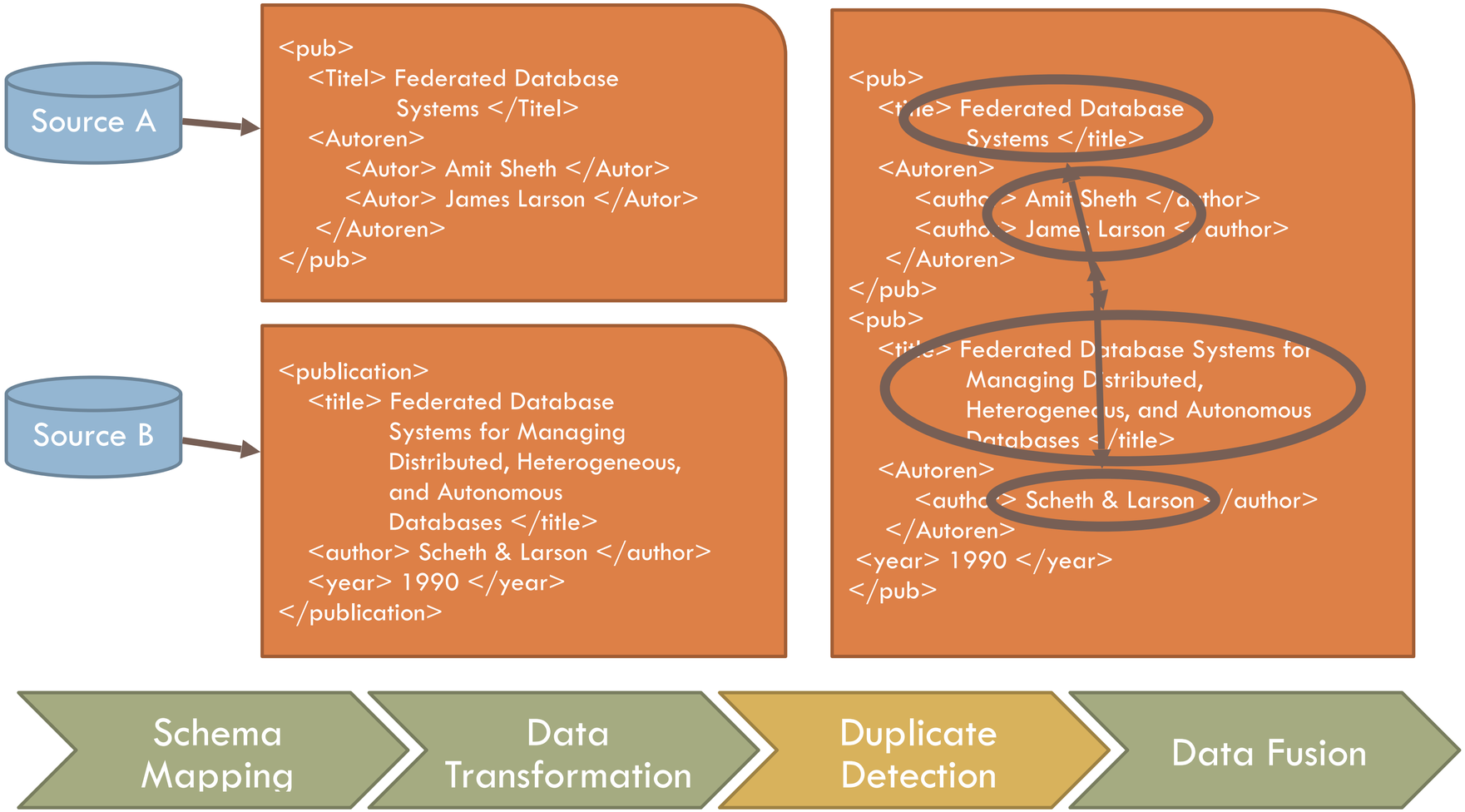
[L. Dong and F. Naumann, 2009]

Information Integration



[L. Dong and F. Naumann, 2009]

Information Integration

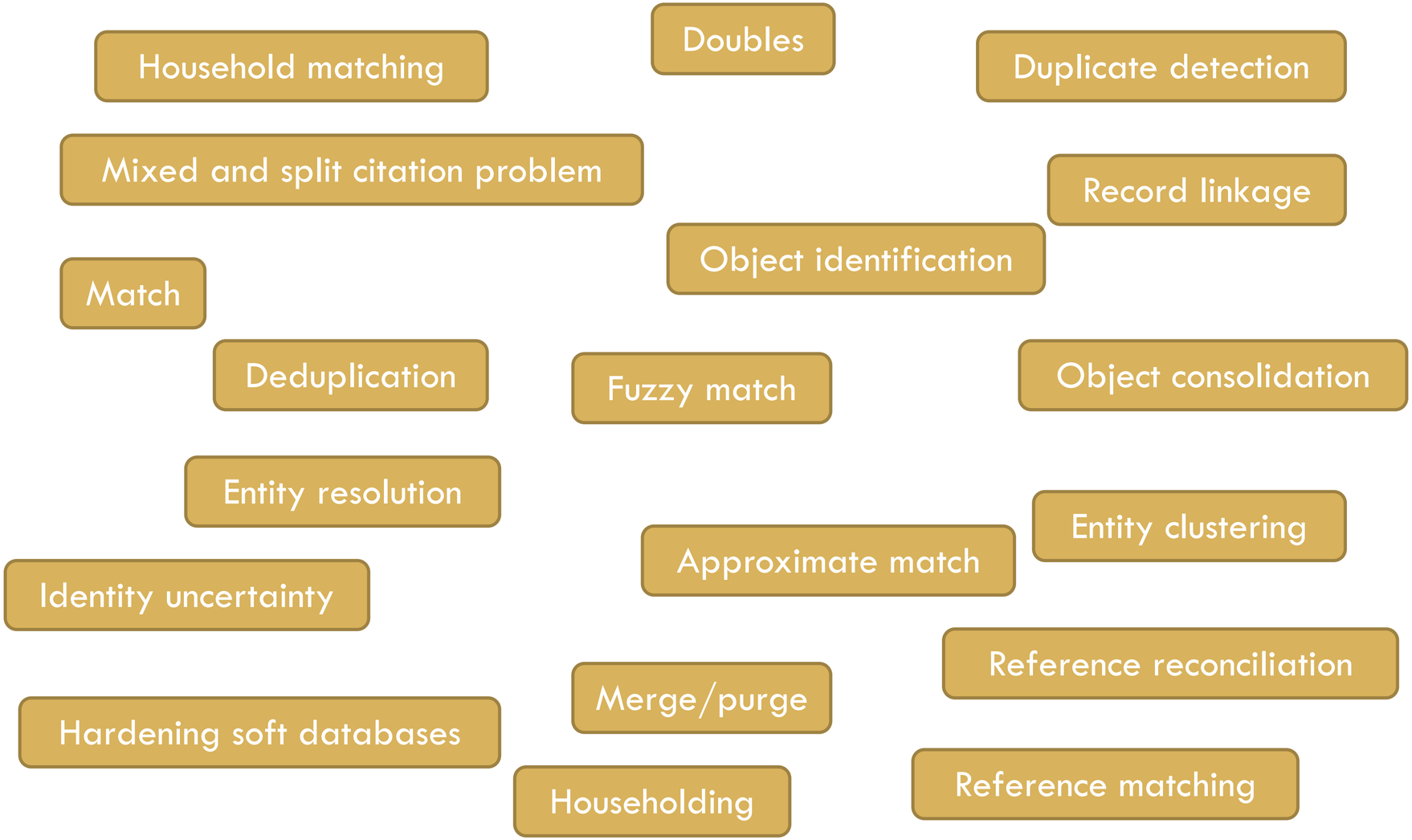


[L. Dong and F. Naumann, 2009]

"Duplicate Detection" has many Duplicates

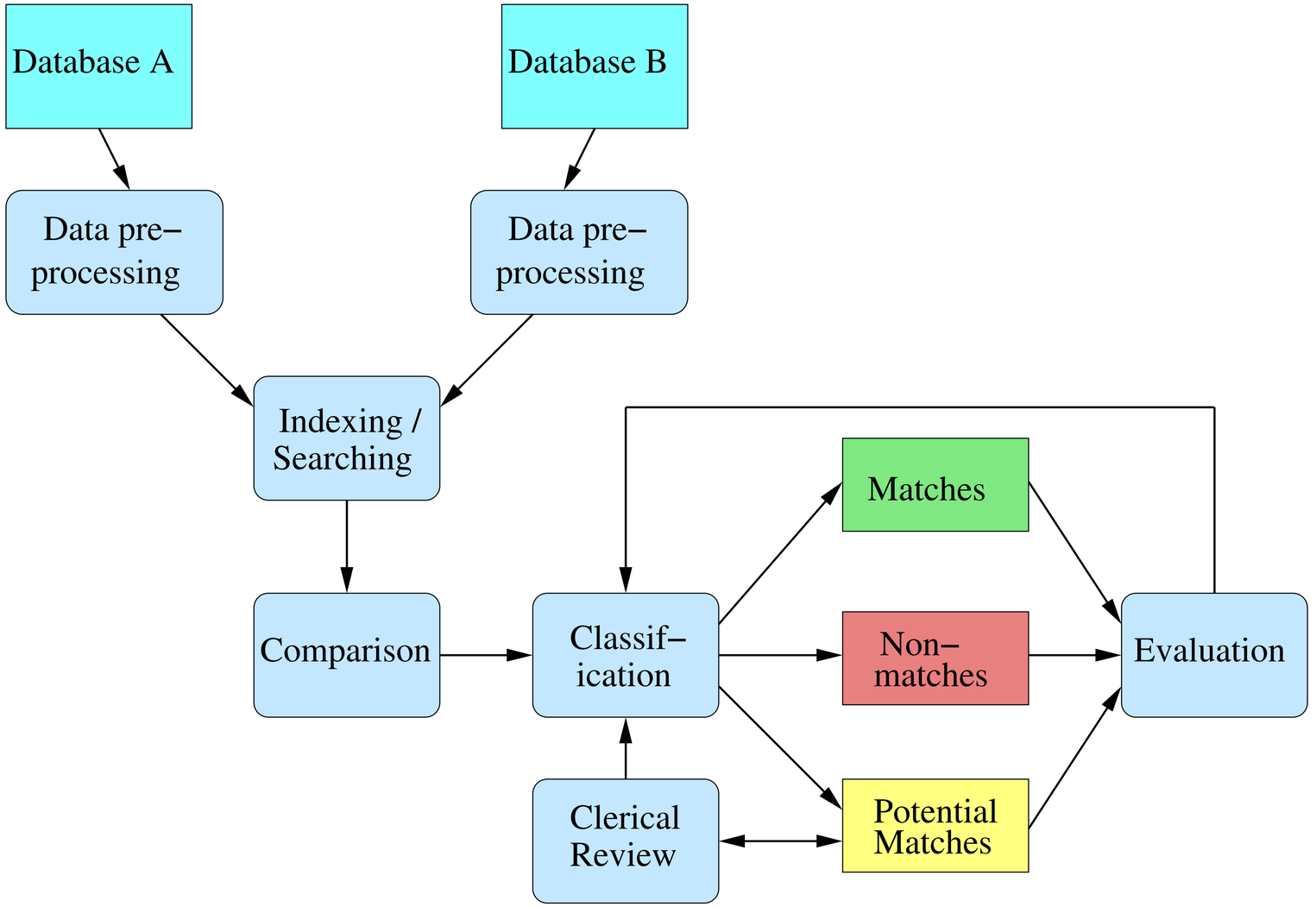
[L. Dong and F. Naumann, 2009]

"Duplicate Detection" has many Duplicates



[L. Dong and F. Naumann, 2009]

Record Linkage Process



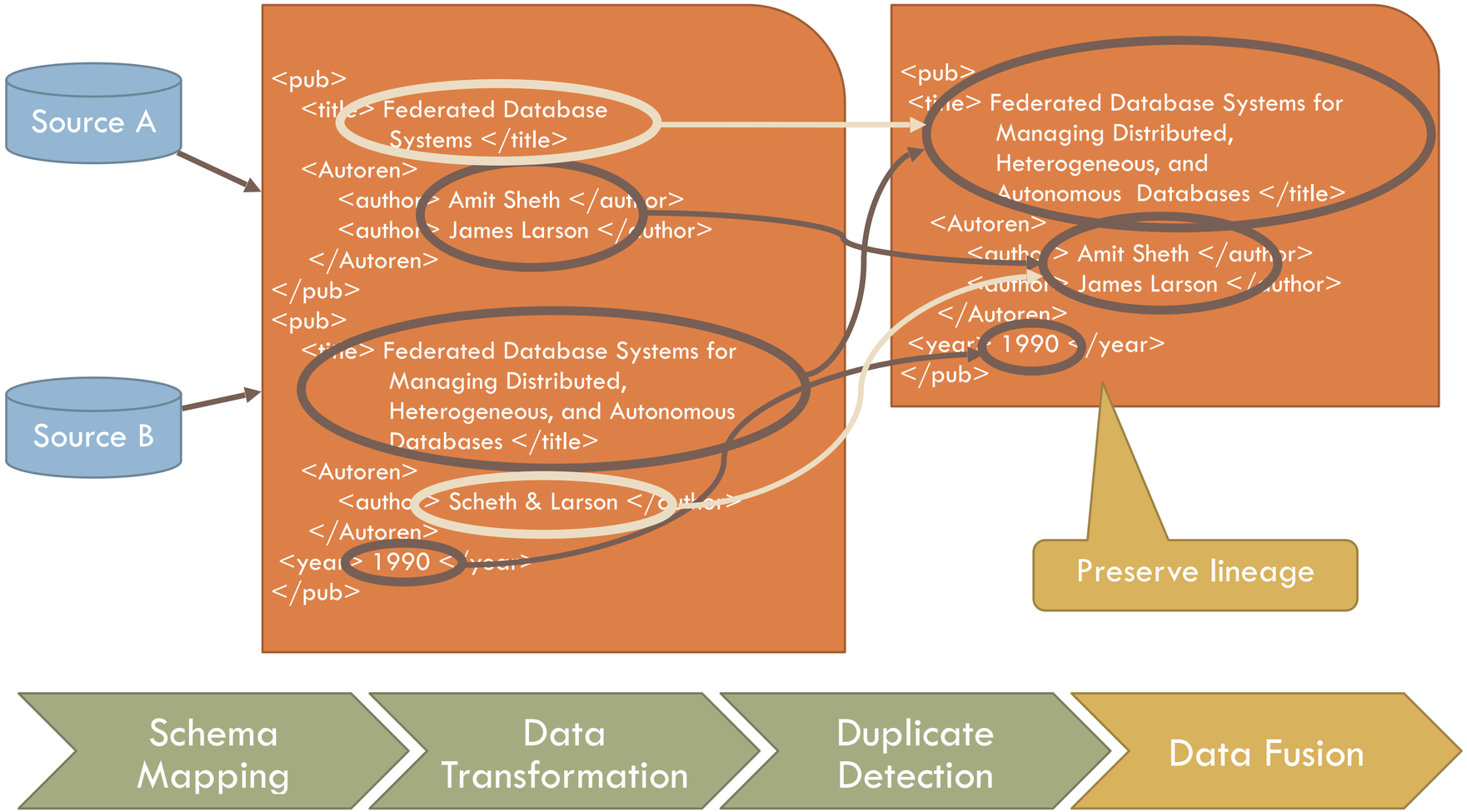
[P. Christen , 2019]

Record Linkage Techniques

- Deterministic matching
  - Rule-based matching (complex to build and maintain)
- Probabilistic record linkage [Fellegi and Sunter, 1969]
  - Use available attributes for linking (often personal information, like names, addresses, dates of birth, etc.)
  - Calculate match weights for attributes
- “Computer science” approaches
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training data (true matches)
  - Unsupervised: Clustering, collective, and graph based

[P. Christen , 2019]

Information Integration

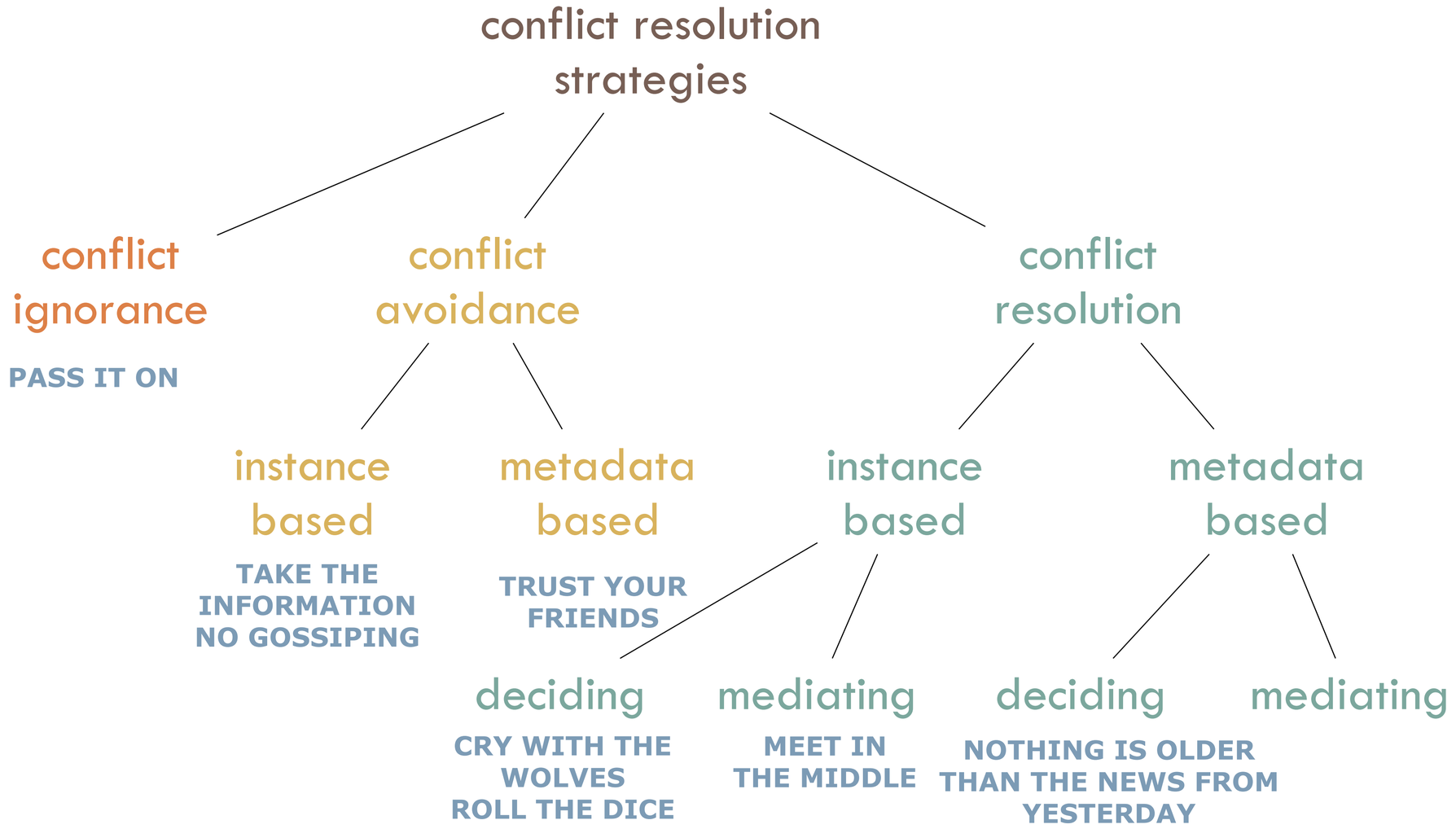


[L. Dong and F. Naumann, 2009]

Data Fusion

- Problem: Given a duplicate, create a single object representation while resolving conflicting data values.
- Difficulties:
  - Null values: Subsumption and complementation
  - Contradictions in data values
  - Uncertainty & truth: Discover the true value and model uncertainty in this process
  - Metadata: Preferences, recency, correctness
  - Lineage: Keep original values and their origin
  - Implementation in DBMS: SQL, extended SQL, UDFs, etc.

Conflict Resolution Strategies



[L. Dong and F. Naumann, 2009]

Assignment 4

- Work on Data Integration and Data Fusion
- Integrate artist datasets from different institutions (The Met, The Tate, Smithsonian, Carnegie Museum of Art)
  - Integrate information about names, places, nationality, etc.
- Record Matching:
  - Which artists are the same?
  - Which nationalities are the same? (British/English)
- Data Fusion:
  - Year of birth/death differences
  - Nationality differences

Integrating Conflicting Data: The Role of Source Dependence

X. L. Dong, L. Berti-Equille, and D. Srivastava

Integrating Conflicting Data: The Role of Source Dependence

- What is the paper's main contribution?
- Do you buy the argument? Any issues with the experiments?
- Can you think of any scenarios where the proposed technique will fail?
- Questions?

Example Problem

[X L Dong et al., 2009]

Example Problem

	S1	S2	S3
Stonebraker	MIT	Berkeley	MIT
Dewitt	MSR	MSR	UWisc
Bernstein	MSR	MSR	MSR
Carey	UCI	AT&T	BEA
Halevy	Google	Google	UW

[X L Dong et al., 2009]

Naive Voting Works

	S1	S2	S3
Stonebraker	MIT	Berkeley	MIT
Dewitt	MSR	MSR	UWisc
Bernstein	MSR	MSR	MSR
Carey	UCI	AT&T	BEA
Halevy	Google	Google	UW

[X L Dong et al., 2009]

Naive Voting Only Works if Data Sources are Independent

[X L Dong et al., 2009]

Naive Voting Only Works if Data Sources are Independent

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

S4 and S5 copy from S3

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

S4 and S5 copy from S3

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

2. With only a snapshot it is hard to decide which source is a copier.

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

[X L Dong et al., 2009]

Challenges in Dependence Discovery

1. Sharing common data does not in itself imply copying.

2. With only a snapshot it is hard to decide which source is a copier.

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

3. A copier can also provide or verify some data by itself, so it is inappropriate to ignore all of its data.

[X L Dong et al., 2009]

Source Dependence

- Source dependence: two sources S and T deriving the same part of data directly or transitively from a common source (can be one of S or T).
  - Independent source
  - Copier
    - copying part (or all) of data from other sources
    - may verify or revise some of the copied values
    - may add additional values
- Assumptions
  - Independent values
  - Independent copying
  - No loop copying

[X L Dong et al., 2009]

Core Case

- Conditions
  - Same source accuracy
  - Uniform false-value distribution
  - Categorical value
- Proposition: W. independent “good” sources, Naïve voting selects values with highest probability to be true.

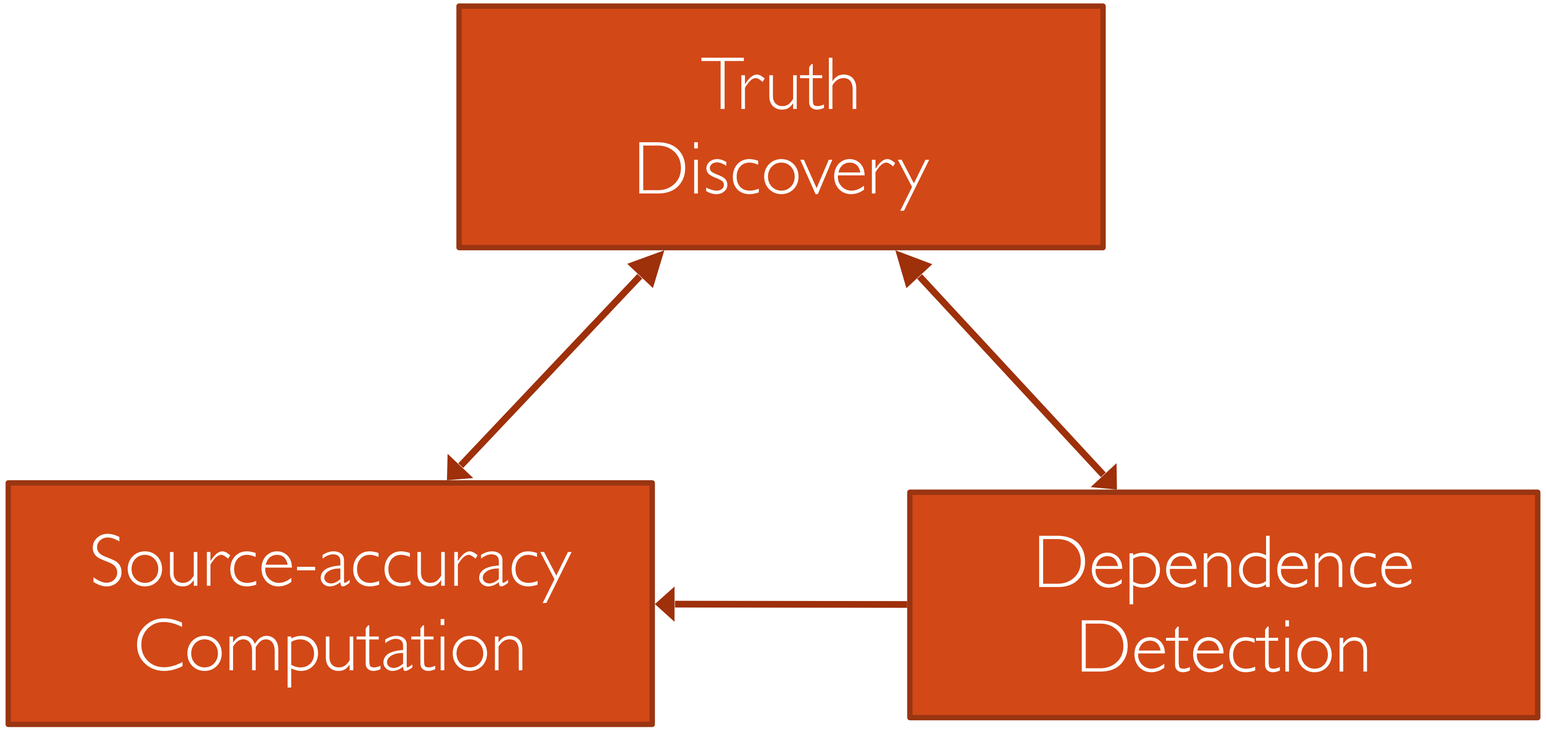
[X L Dong et al., 2009]

Ideas

- If two sources share a lot of false values, they are more likely to be dependent.
- S1 is more likely to copy from S2, if the accuracy of the common data is highly different from the accuracy of S1.

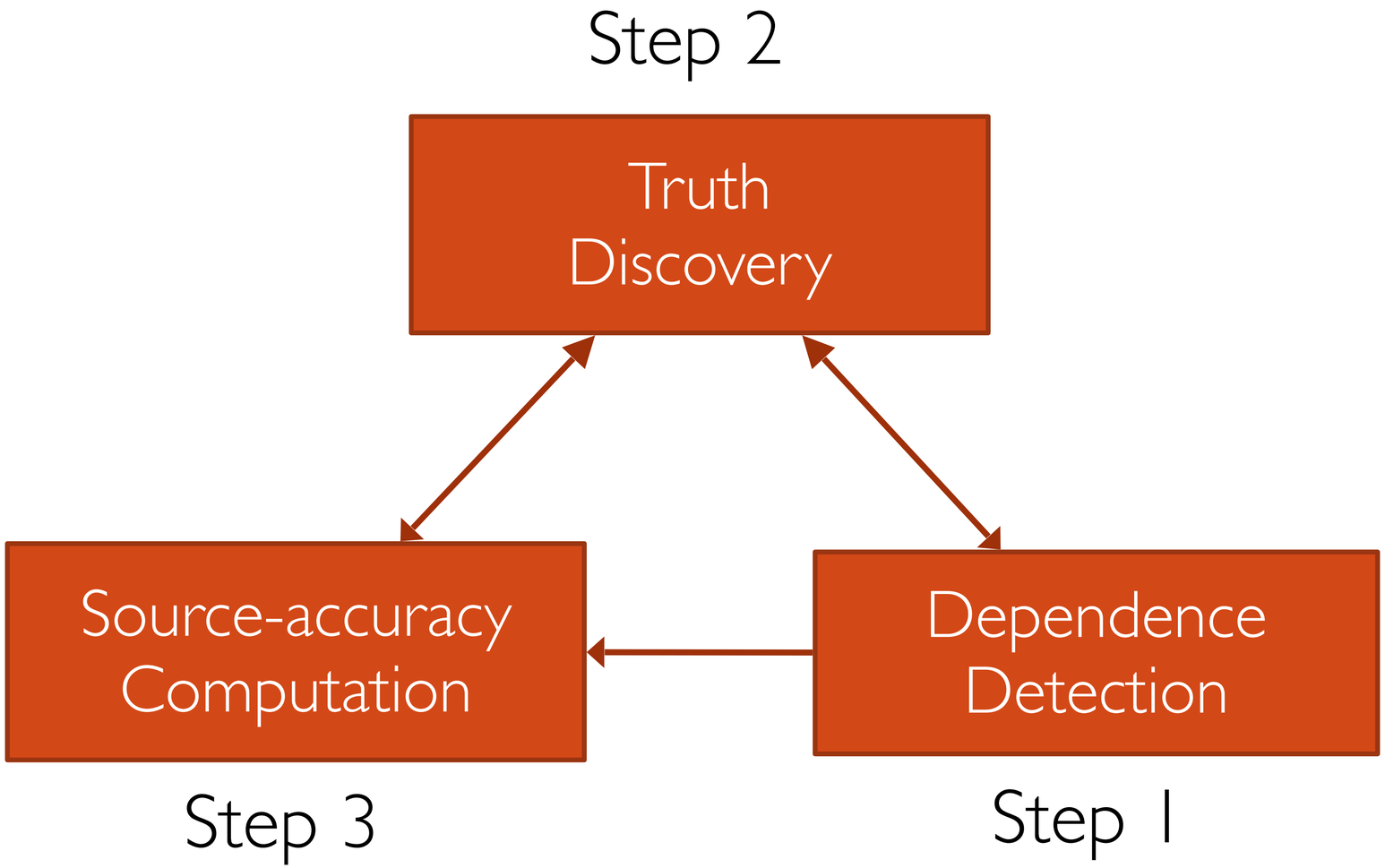
[X L Dong et al., 2009]

Combining Accuracy and Dependence



[X L Dong et al., 2009]

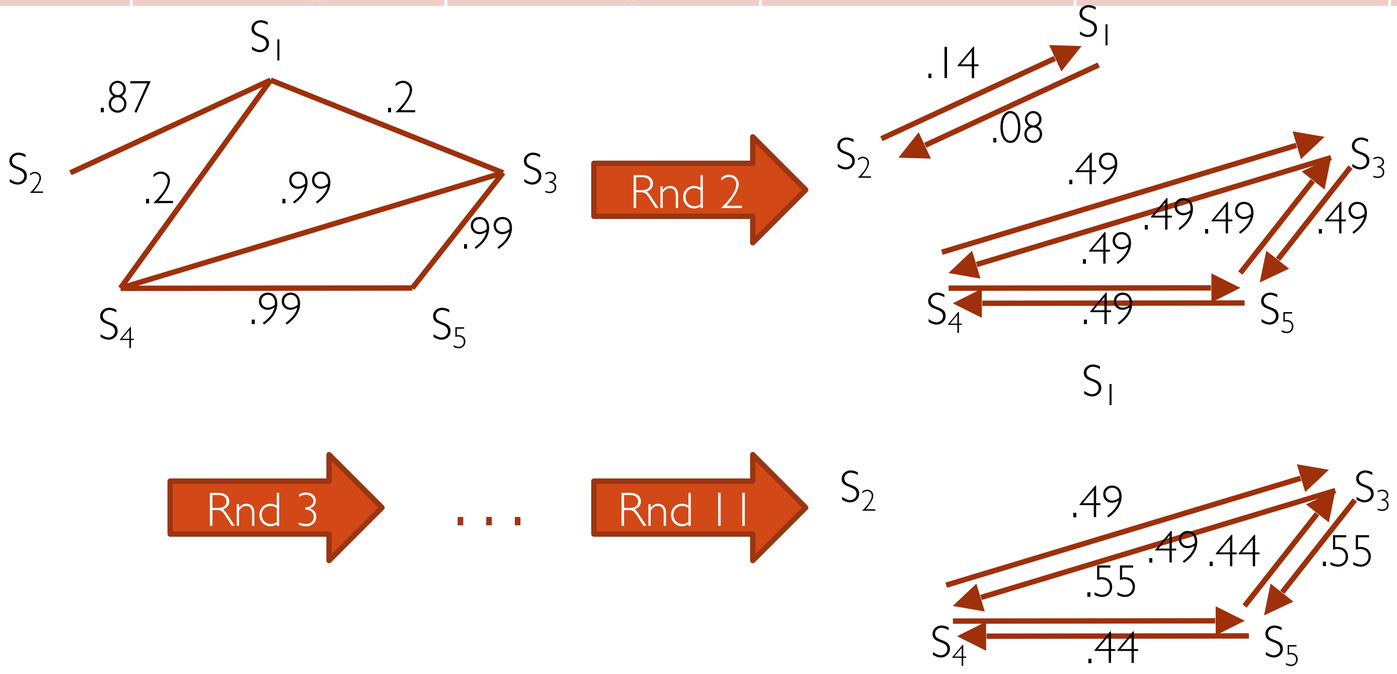
Combining Accuracy and Dependence



[X L Dong et al., 2009]

The Motivating Example

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW



[X L Dong et al., 2009]

The Motivating Example

Accuracy	S1	S2	S3	S4	S5
Round 1	.52	.42	.53	.53	.53
Round 2	.63	.46	.55	.55	.55
Round 3	.71	.52	.53	.53	.37
Round 4	.79	.57	.48	.48	.31
…	…	…	…	…	…
Round 11	.97	.61	.40	.40	.21

Value Confidence	Carey			Halevy	
	UCI	AT&T	BEA	**Google**	UW
Round 1	1.61	1.61	2.0	2.1	2.0
Round 2	1.68	1.3	2.12	2.74	2.12
Round 3	2.12	1.47	2.24	3.59	2.24
Round 4	2.51	1.68	2.14	4.01	2.14
…	…	…	…	…	…
Round 11	4.73	2.08	1.47	6.67	1.47

[X L Dong et al., 2009]

Data Curation

Why?

Big Data, Little Data, or No Data?

C. L. Borgman

What is data and why share it?

- "Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship."
  [C. L. Borgman]
- Data can be digital but can also be physical (e.g. sculptures)
- Semantics are important (e.g. temperature to engineer and biologist)
- Grey Data: surveys, student records—think about **privacy**
- Sharing Data
  - Required/encouraged by universities, funding agencies, publishers
  - "Publications are arguments made by authors, and **data are the evidence** used to support the arguments." [C. L. Borgman]

Data attribution and citation

- Publications are counted, authorship is negotiated
- For data:
  - Often compound
  - Ownership is rarely clear
  - Attribution?
  - What about derived data?
- Bibliometrics and Altmetrics

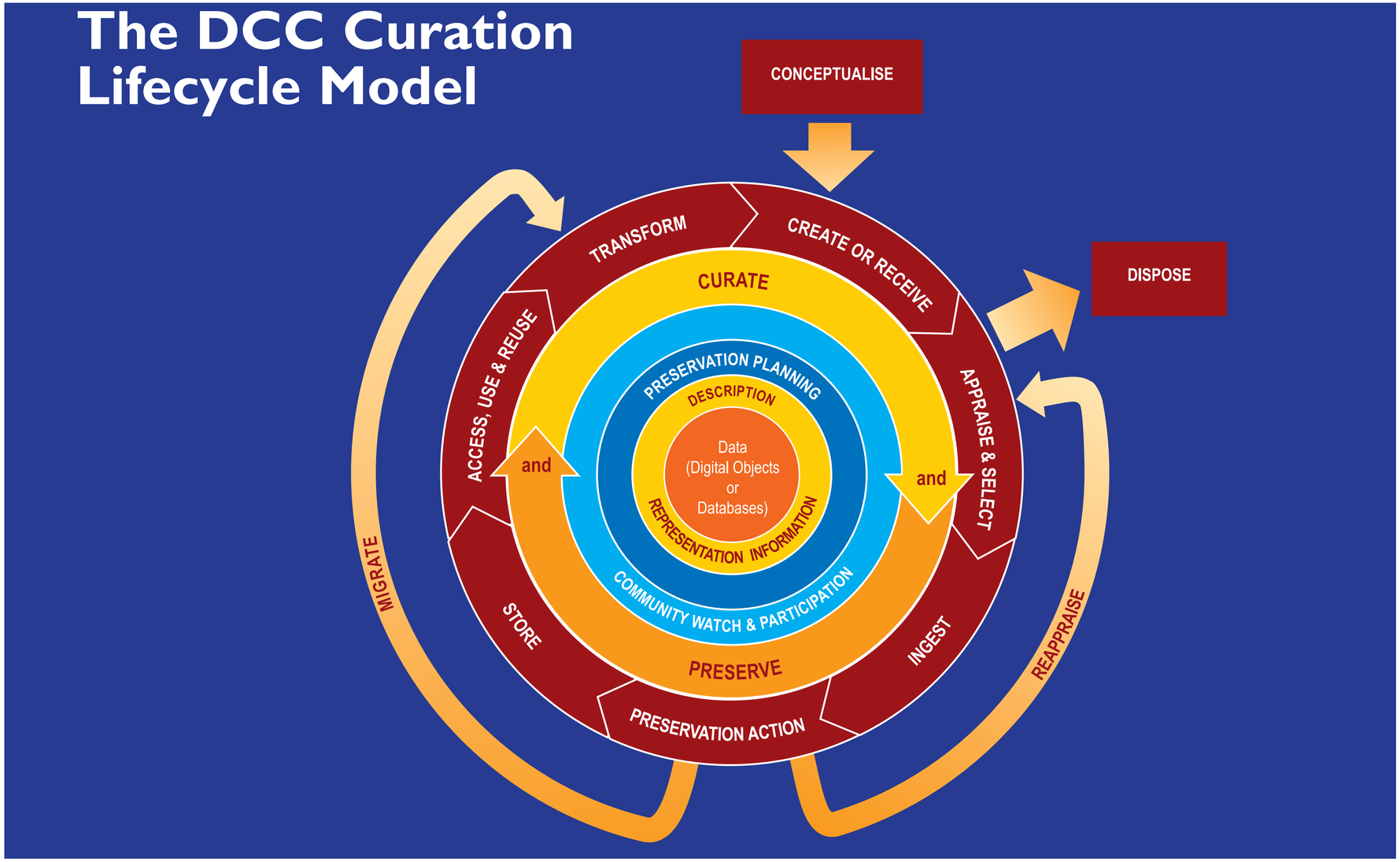
Data Identity

- Identifiers: DOIs, URIs
- Naming and namespaces: ORCID, KEGG Identifier
- Description: Metadata, Self-describing

Data Persistence

- How long should this data be kept?
  - Perishable
  - Long-lived
  - Permanent
- Who is responsible for keeping the data?
  - Scientists/investigators?
  - Publishers?
  - Librarians?
- Privacy should be considered from the beginning

Data Curation Lifecycle



[DCC]

Data (Digital Objects or Databases)

- Data, any information in binary digital form, is at the centre of the Curation Lifecycle. This includes:
  - **Digital Objects**
    - Simple Digital Objects are discrete digital items; such as textual files, images or sound files, along with their related identifiers and metadata.
    - Complex Digital Objects are discrete digital objects, made by combining a number of other digital objects, such as websites.
  - **Databases**: Structured collections of records or data stored in a computer system.

[DCC]

Full Lifecycle Actions

- Description and Representation Information: Assign metadata, using appropriate standards, to ensure adequate description and control
- Preservation Planning: Plan for preservation throughout the curation lifecycle of digital material
- Community Watch and Participation: Watch standards, tools, software.
- Curate and Preserve: Promote curation and preservation throughout the curation lifecycle

[DCC]

Sequential Actions

- Conceptualize: Plan creation of data—capture method and storage options.
- Create or Receive: Create/receive data and make sure metadata exists
- Appraise and Select: Evaluate data and select for long-term curation and preservation
- Ingest: Transfer data to an archive, repository, data centre or other custodian
- Preservation Action: Data cleaning, validation (ensure that data remains authentic, reliable and usable)
- Store: Store the data in a secure manner adhering to relevant standards Access, Use and Reuse: Make sure is accessible to users and reusers
- Transform: Create new data from the original (migrate formats, subsets, etc.)

[DCC]

Occasional Actions

- Dispose: Transfer to another archive or perhaps destroy data
- Reappraise: Return data which fails validation procedures for further appraisal and reelection
- Migrate: Migrate data to a different format—ensure the data's immunity from hardware or software obsolescence

[DCC]

The FAIR Guiding Principles for Scientific Data Management and Stewardship

M. D. Wilkinson et al.

Who and Why?

- Who: People from academia, industry, funding agencies, & scholarly publishers
- Why?
  - Data management leads to knowledge discovery, innovation, and reuse
  - Existing digital ecosystem **prevents** maximum benefit
  - Need to specify what "good" data management/curation/stewardship is
  - Enhance the ability of machines to automatically find and use the data
  - Principles should also apply to **tools**

[M. D. Wilkinson et al., 2016]

FAIR Principles

- Findable: Metadata and data should be easy to find for both humans and computers
- Accessible: Users need to know how data can be accessed, possibly including authentication and authorization
- Interoperable: Can be integrated with other data, and can interoperate with applications or workflows for analysis, storage, and processing
- Reusable: Optimize the reuse of data. Metadata and data should be well-described so they can be replicated and/or combined in different settings

[GO FAIR]

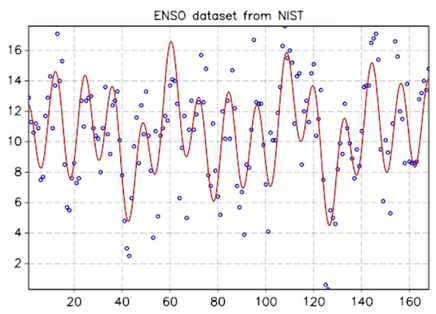
To be Findable

- F1. (Meta)data are assigned a **globally unique and persistent identifier**
- F2. Data are described with **rich metadata** (defined by R1)
- F3. Metadata clearly and explicitly include the **identifier** of the data it describes
- F4. (Meta)data are **registered or indexed** in a searchable resource

[M. D. Wilkinson et al., 2016]

DataCite Workflow

1. Take a dataset



2. Describe it

- Title
- Authors
- Year
- Description
- And others…

3. Assign a DOI

10.1234/exampledata

4. Reuse and reference!

ATLAS Collaboration, "Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \to \gamma\gamma$,"
http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

☑ Unique ☑ Persistent

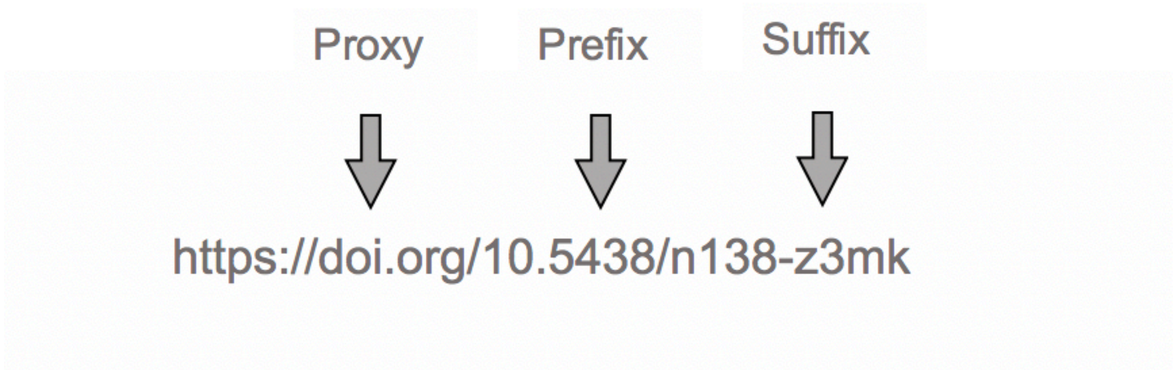
5. Enjoy the benefits

- Findability
- Track citations
- Reusability
- Measure impact

[DataCite]

Digital Object Identifier

- Name: Proxy + Prefix + Suffix



- Metadata: description of the object
- URL: resolves to a digital location, which contains object's details

[DataCite]

DataCite Metadata

Mandatory Properties	Details
Identifier	with mandatory type sub-property
Creator	with optional name identifier and affiliation sub-properties
Title	with optional type sub-properties
Publisher	
PublicationYear	
ResourceType	with mandatory general type description sub-property

Recommended Properties	Details
Subject	with scheme sub-property
Contributor	with type, name identifier, and affiliation sub-properties
Date	with type sub-property
RelatedIdentifier	with type and relation type sub-properties
Description	with type sub-property
GeoLocation	with point, box, and polygon sub-properties

Optional Properties
Language
AlternateIdentifier
Size
Format
Version
Rights
FundingReference

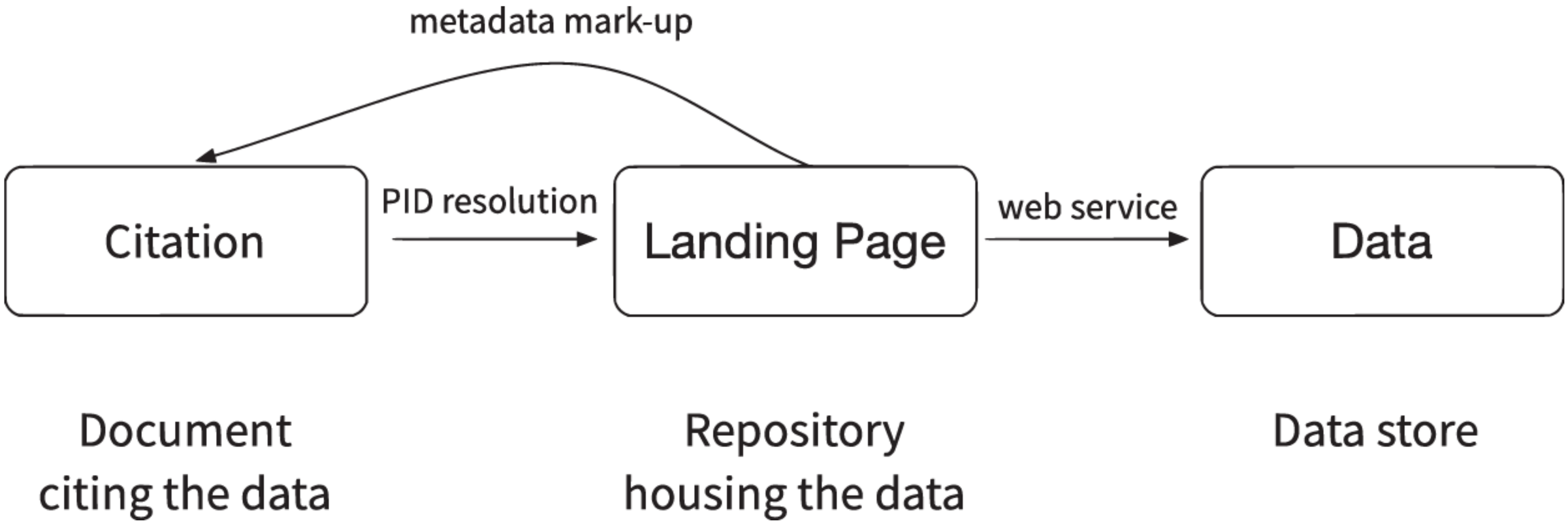
[DataCite]

To be Accessible

- A1. (Meta)data are **retrievable** by their identifier using a standardized communications protocol
  - A1.1. The protocol is **open**, free, and universally implementable
  - A1.2. The protocol allows for an **authentication** and authorization procedure, where necessary
- A2. Metadata are accessible, even when the data are **no longer available**

[M. D. Wilkinson et al., 2016]

How data accessibility might work within publications



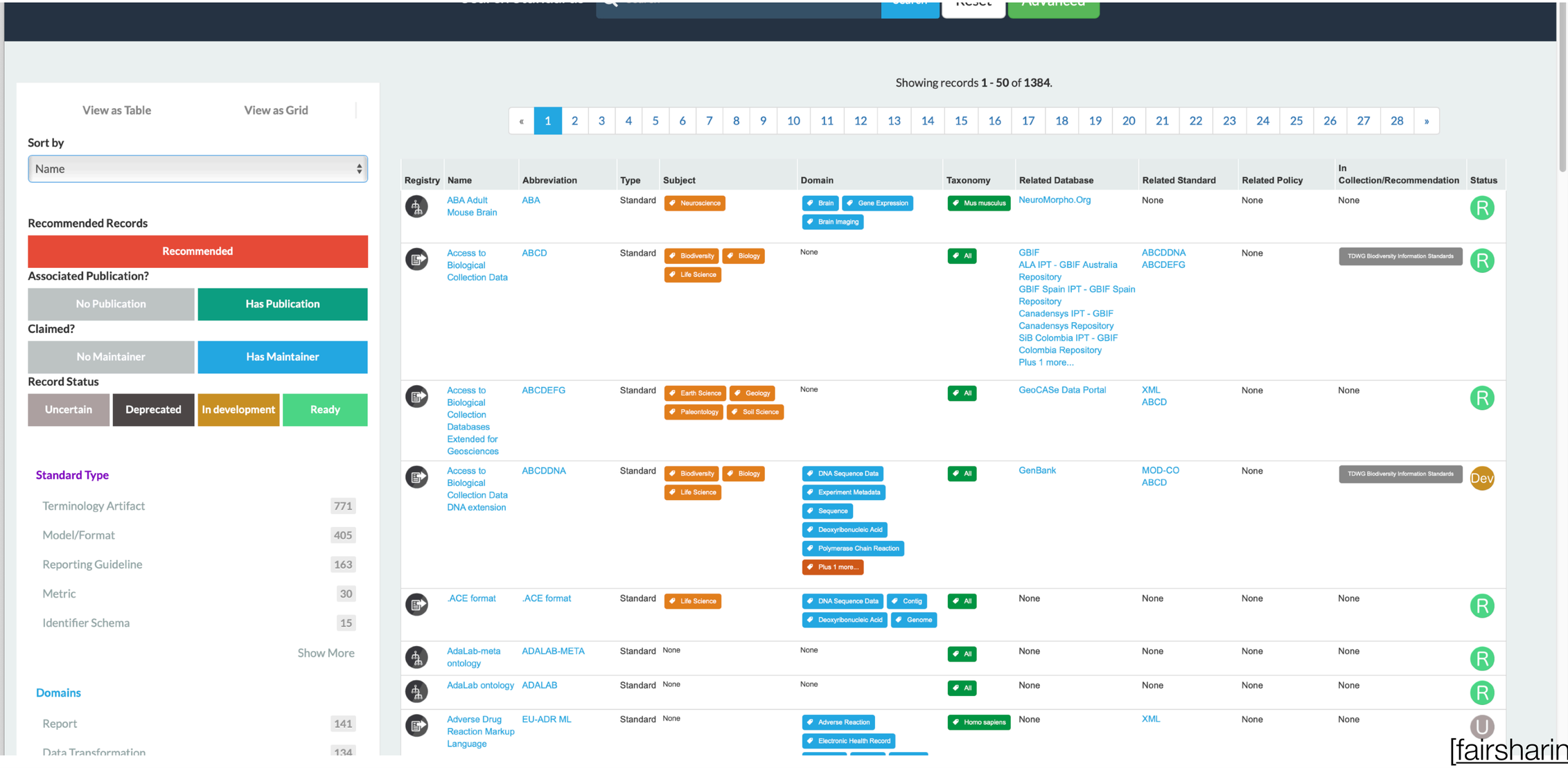
[M. Fenner et al., 2019]

To be Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable **language** for knowledge representation.
- I2. (Meta)data use **vocabularies** that follow FAIR principles
- I3. (Meta)data include **qualified references** to other (meta)data

[M. D. Wilkinson et al., 2016]

Standard vocabularies



[fairsharing.org]

To be Reusable

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage **license**
  - R1.2. (Meta)data are associated with detailed **provenance**
  - R1.3. (Meta)data meet domain-relevant **community standards**

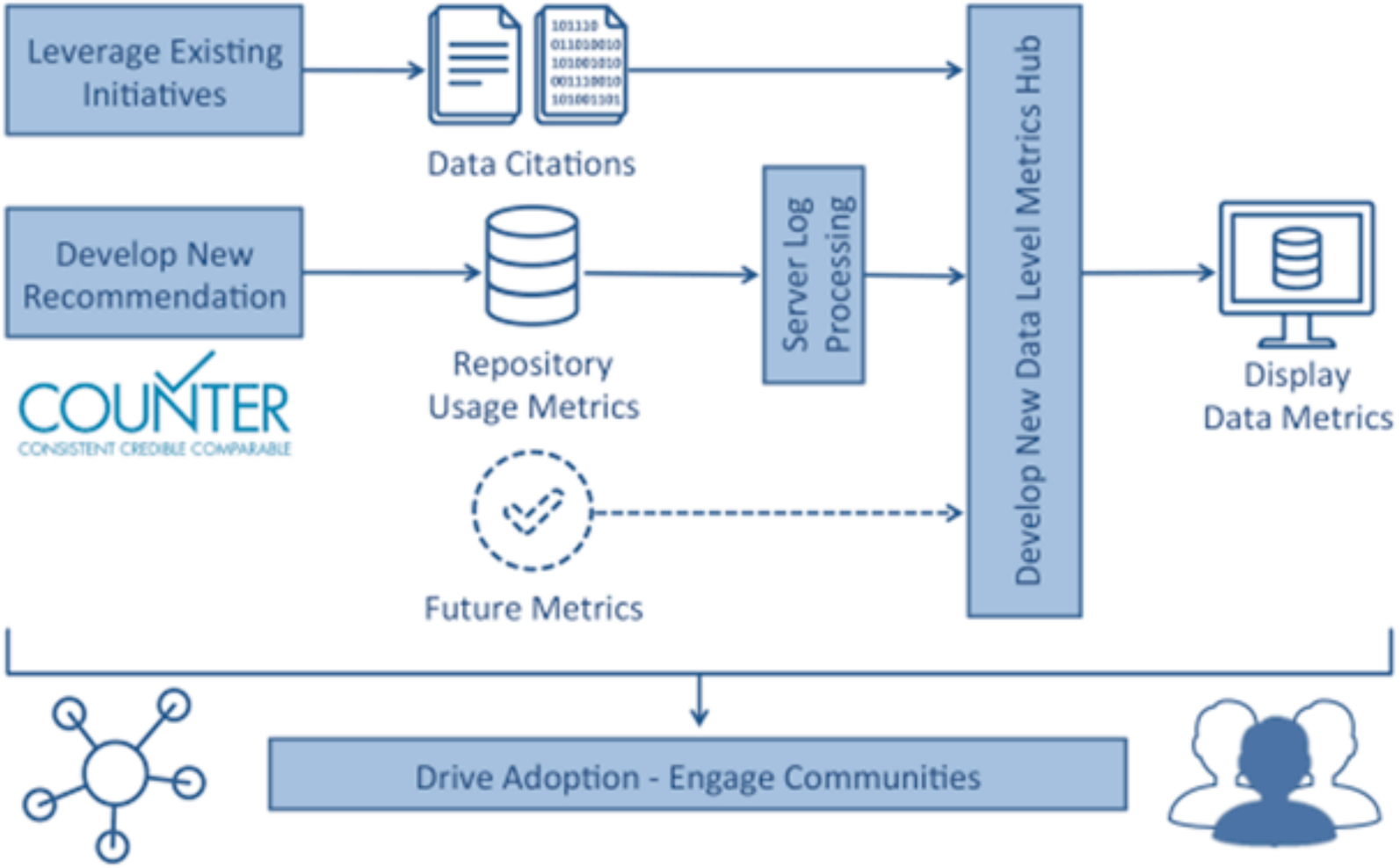
[M. D. Wilkinson et al., 2016]

Licensing

- Citation of a dataset is expected as a scholarly norm, not by law
- CC0:
  - "I hereby waive all copyright and related or neighboring rights together with all associated claims and causes of action with respect to this work to the extent possible under the law"
- CC BY: license, not a waiver as CC0
  - "You must give appropriate credit, provide a link to the license, and indicate if changes were made."
- Data Use Agreements (DUA): Used when data are restricted due to proprietary or privacy concerns.

[M. Crosas]

Make Data Count



[H. Cousijn et al., 2019]