Extracting Dewey Decimal Classifications from Dublin Core Metadata Records With the DISTIL Project: Preliminary Findings and Observations

Michael Khoo[1]

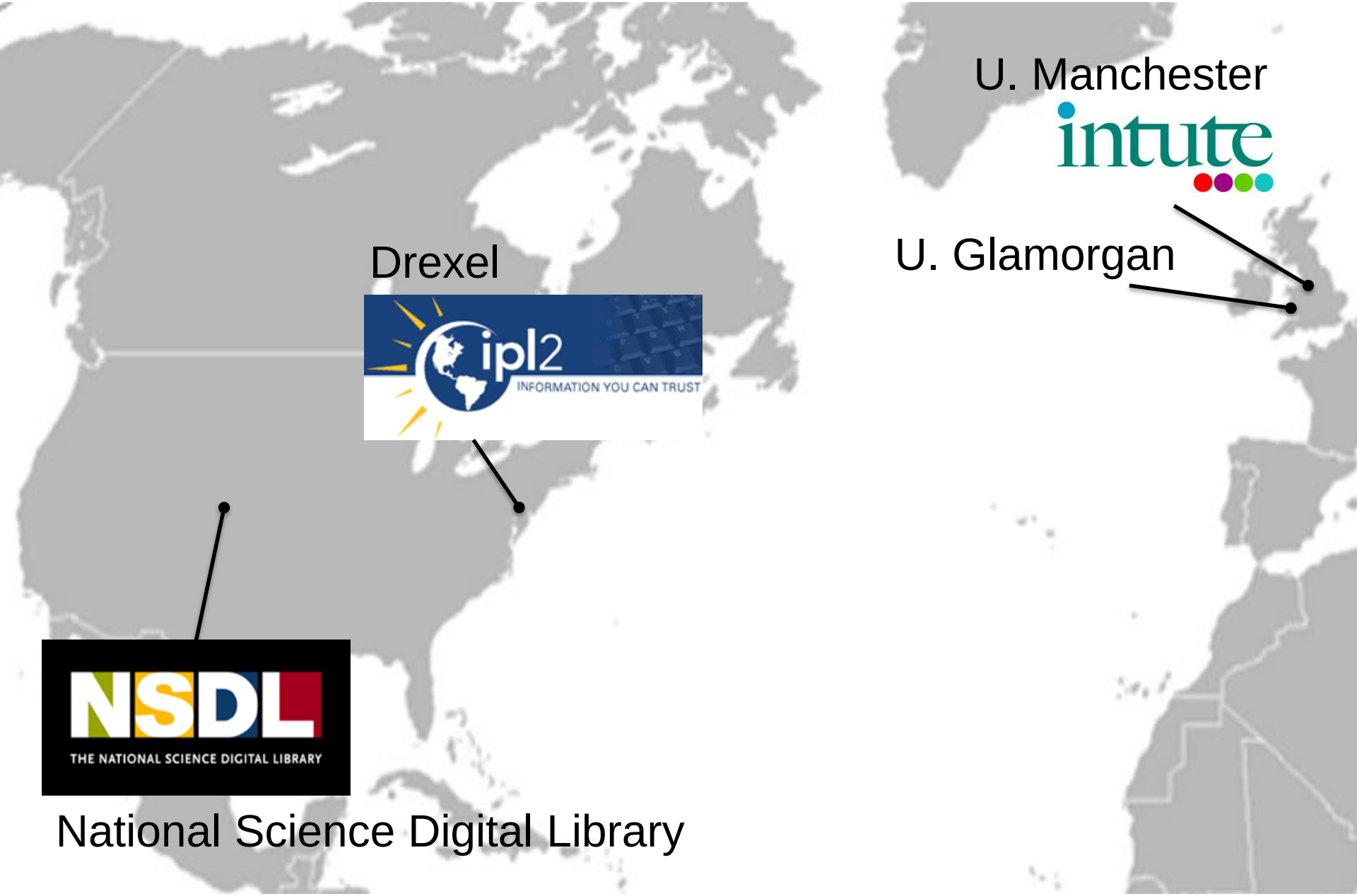
Douglas Tudhope[2], Ceri Binding[2]

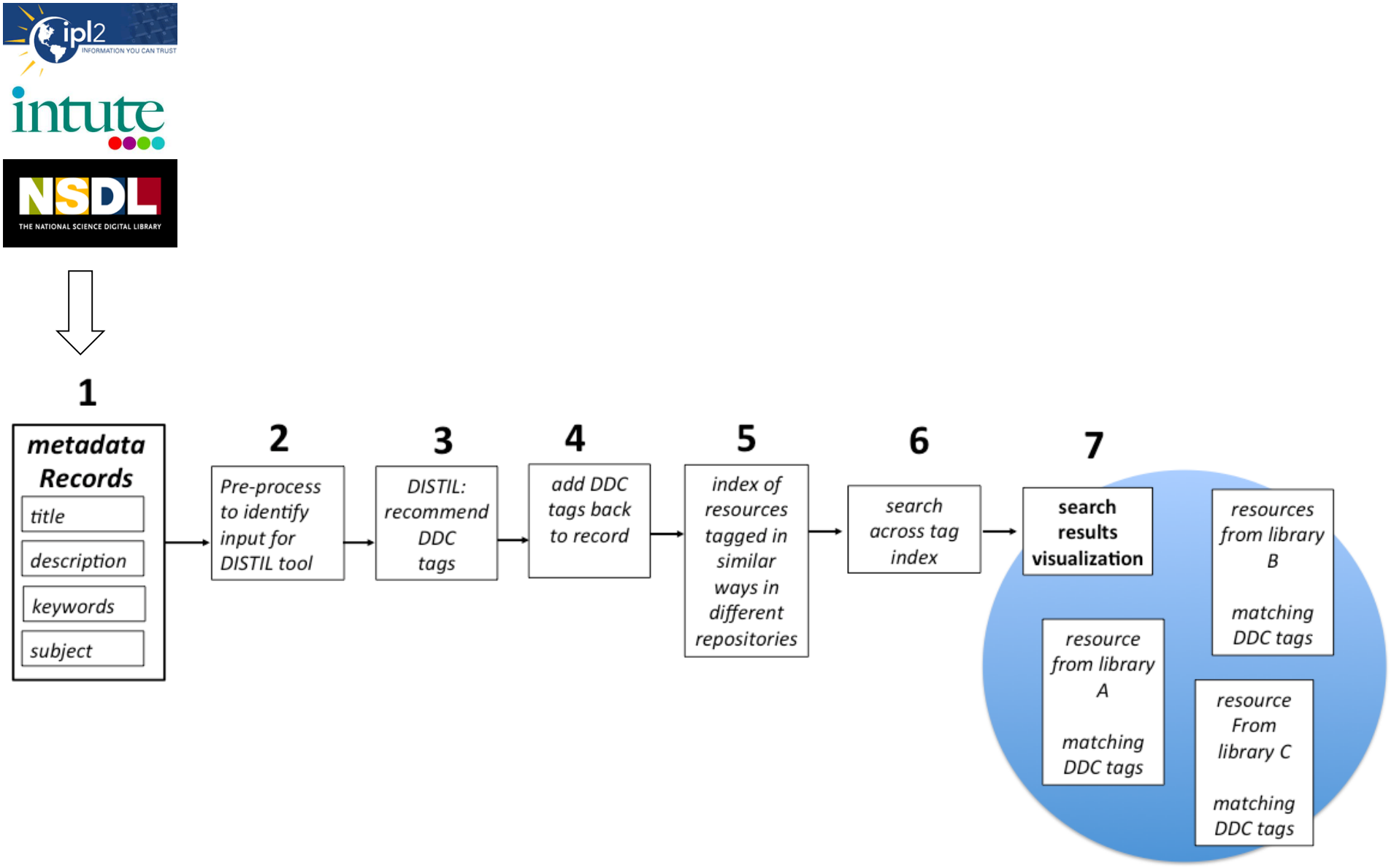
[1]Drexel University [2]University of Glamorgan

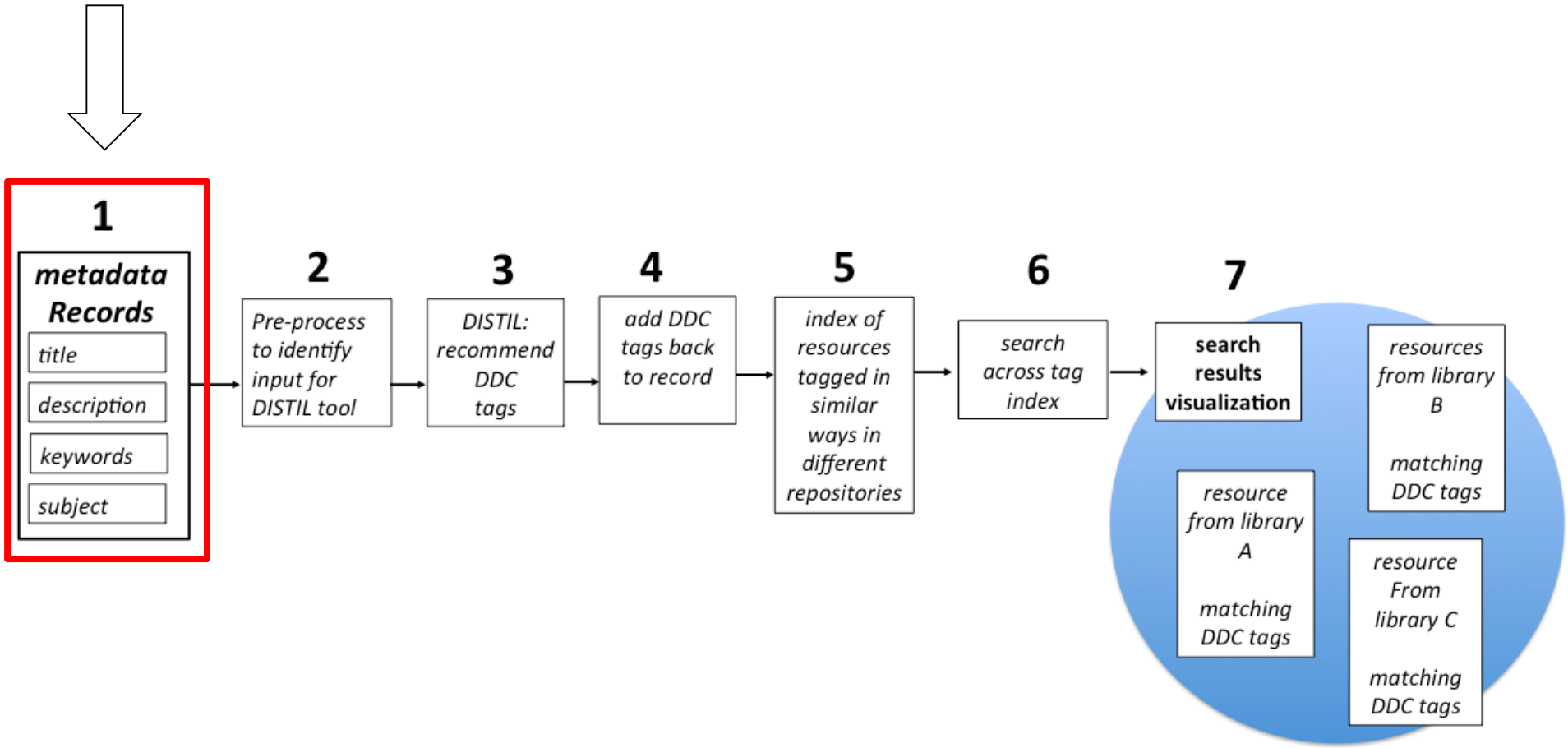
NKOS Workshop/TPDL 2012 Paphos Cyprus

DISTIL (Document Indexing & Semantic Tagging Interface for Libraries)

- Setting
  - Small(ish)-scale, DC, educational DLs
  - Large-scale information infrastructures
- Aim: Achieve efficient federated search and discovery across heterogeneous DLs
- Focus: Humanities and social sciences
- Funding: *Digging Into Data Challenge*


U. Manchester
intute
U. Glamorgan
Drexel
ipl2
INFORMATION YOU CAN TRUST
NSDL
THE NATIONAL SCIENCE DIGITAL LIBRARY
National Science Digital Library

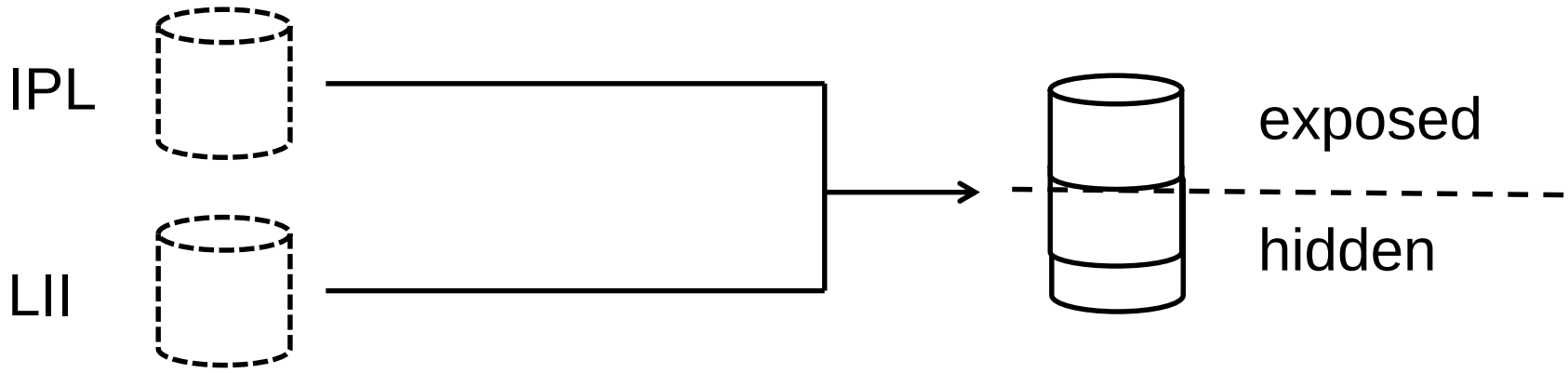

1
metadata Records
title
description
keywords
subject
2
Pre-process to identify input for DISTIL tool
3
DISTIL: recommend DDC tags
4
add DDC tags back to record
5
index of resources tagged in similar ways in different repositories
6
search across tag index
7
search results visualization
resources from library B
matching DDC tags
resource from library A
matching DDC tags
resource From library C
matching DDC tags


1
metadata Records
title
description
keywords
subject
2
Pre-process to identify input for DISTIL tool
3
DISTIL: recommend DDC tags
4
add DDC tags back to record
5
index of resources tagged in similar ways in different repositories
6
search across tag index
7
search results visualization
resources from library B
matching DDC tags
resource from library A
matching DDC tags
resource From library C
matching DDC tags

Stage 1: Harvesting

- Some metadata is exposed – other metadata is hidden
- Building the harvest is requiring some communication and negotiation with the original metadata curators

Stage 1: Harvesting - IPL

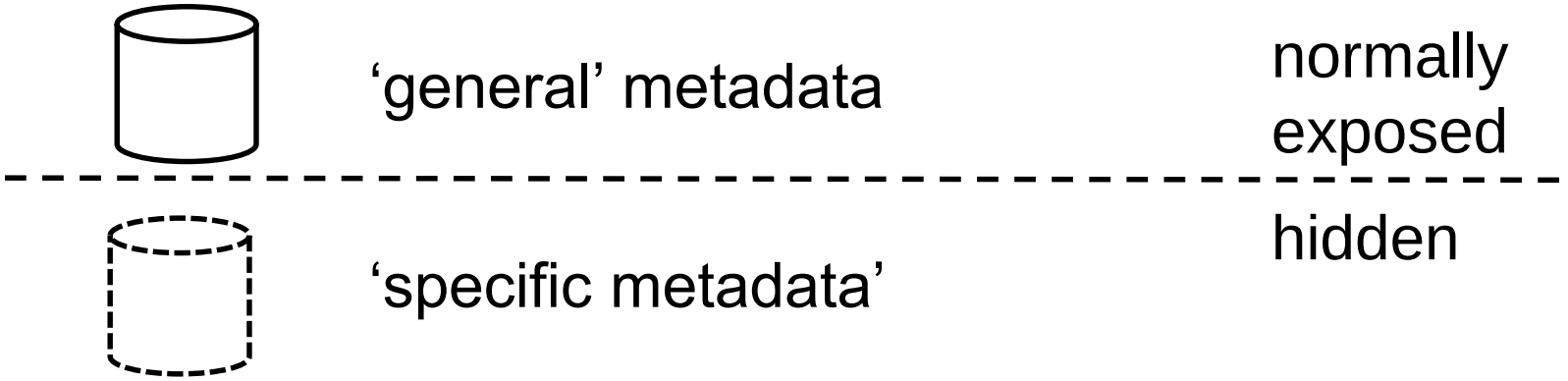


1990s
Separate organizations
Homebrewed metadata
& SQL databases

2008
Merge
> DC

2012
Dublin Core
Fedora database
with multiple datastreams

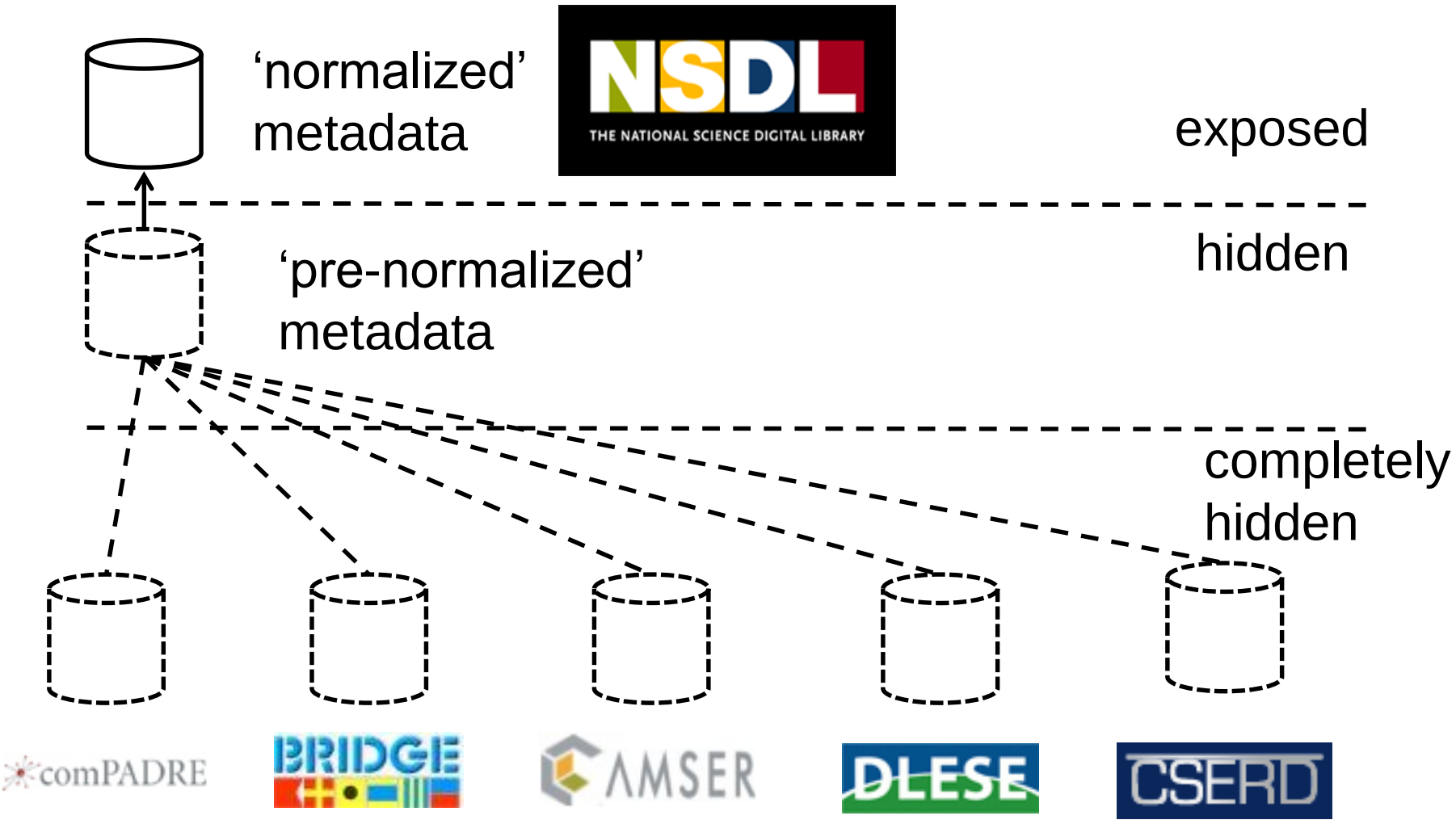
Stage 1: Harvesting - Intute



Intute stores metadata for each resource in unrelated tables

- One database contains the main record
- Additional tables contain discipline-specific metadata that supports different focused search and browsing views on the collections (e.g. some collections indexed with specific controlled vocabularies)

Stage 1: Harvesting - NSDL



NSDL Pathway metadata

Stage 1: Harvesting - NSDL

comPADRE BRIDGE AMSER DLESE CSERD

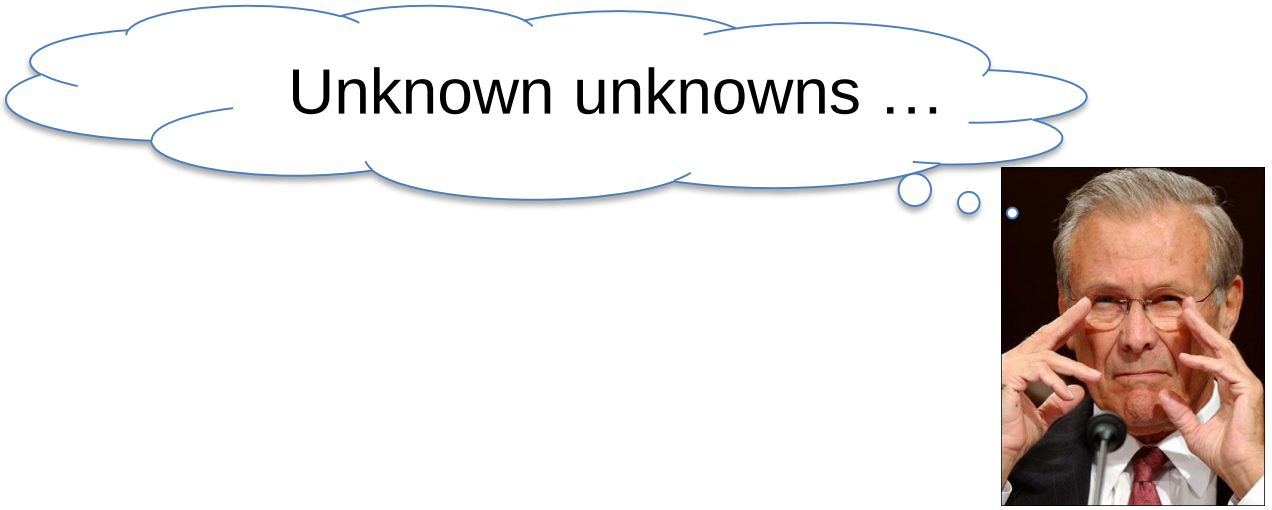
Environmental science
teacher resource
professional development
teaching awards
Professional organization
Ecology, Forestry and Agriculture
Geoscience
Social Sciences
Education
Chemistry
Physics
Space Science

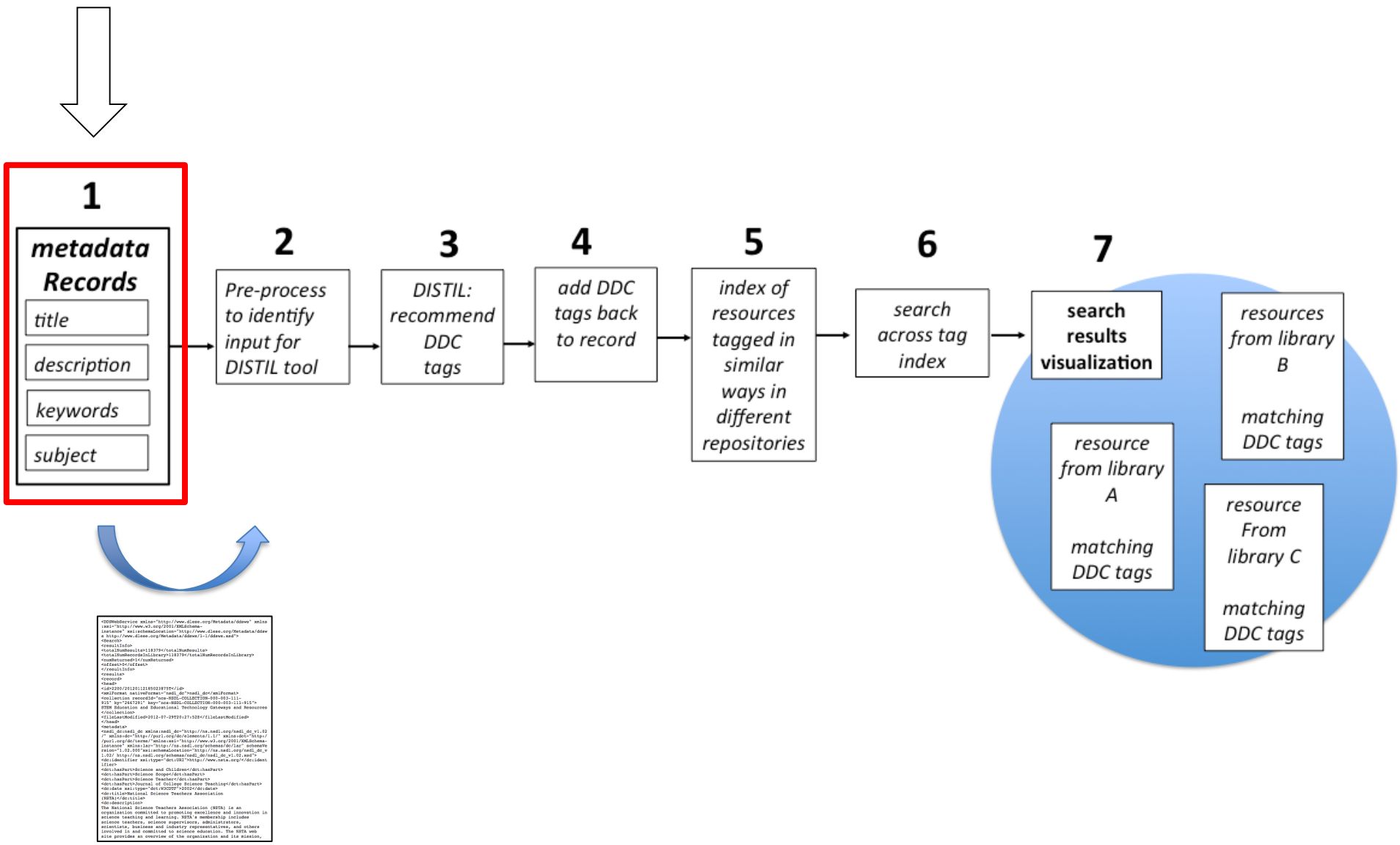
Educational theory and practice
Environmental science
Policy issues
Space science
Science
Earth science
Physical sciences
Chemistry
Biology
Education (General)
Physics
Astronomy
Space sciences
Education
Ecology, Forestry and Agriculture
Geoscience
Social Sciences
History/Policy/Law
Space Science
Chemistry
Physics
Life Science
Technology

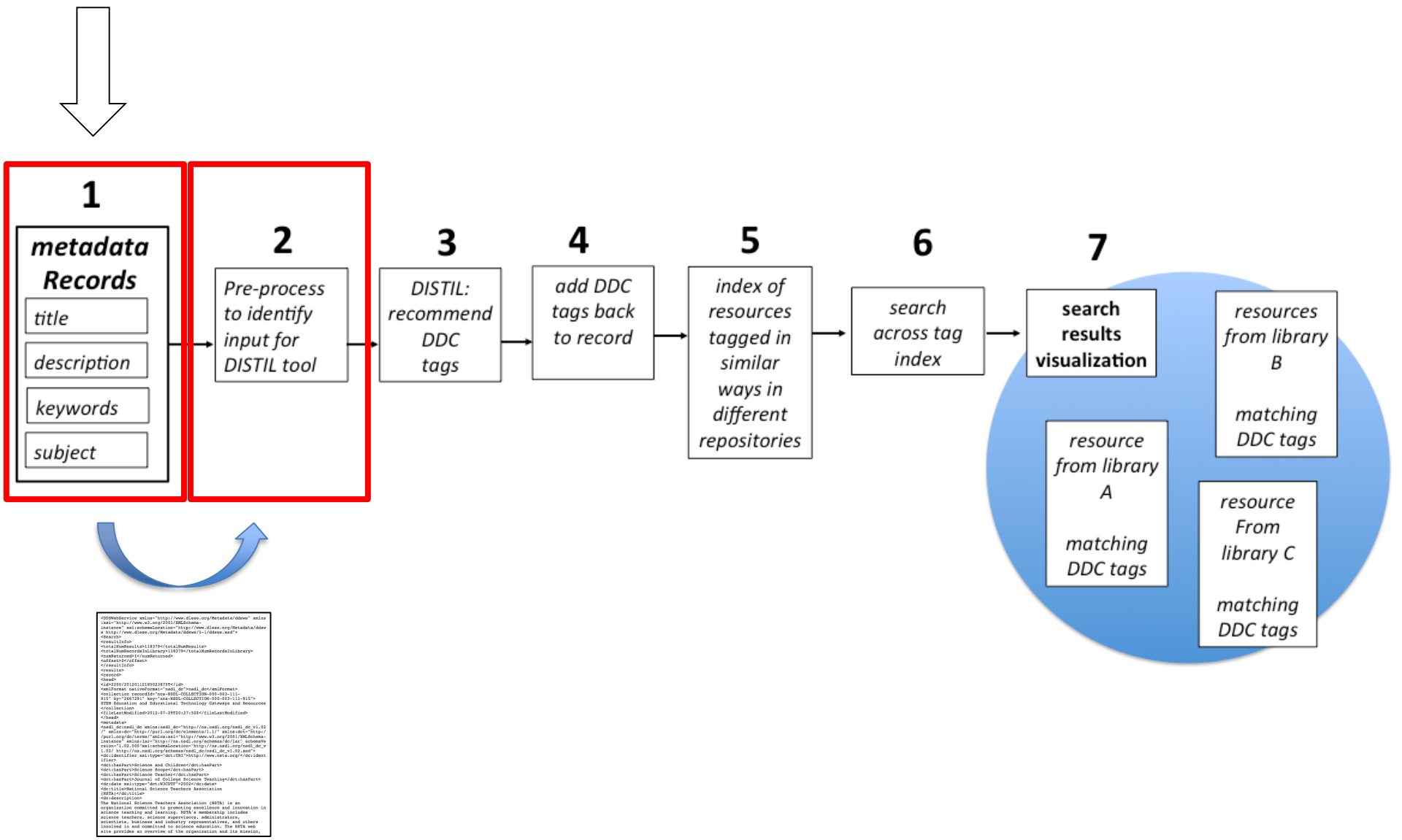
Biology
Physics
Education
Life Science
Chemistry

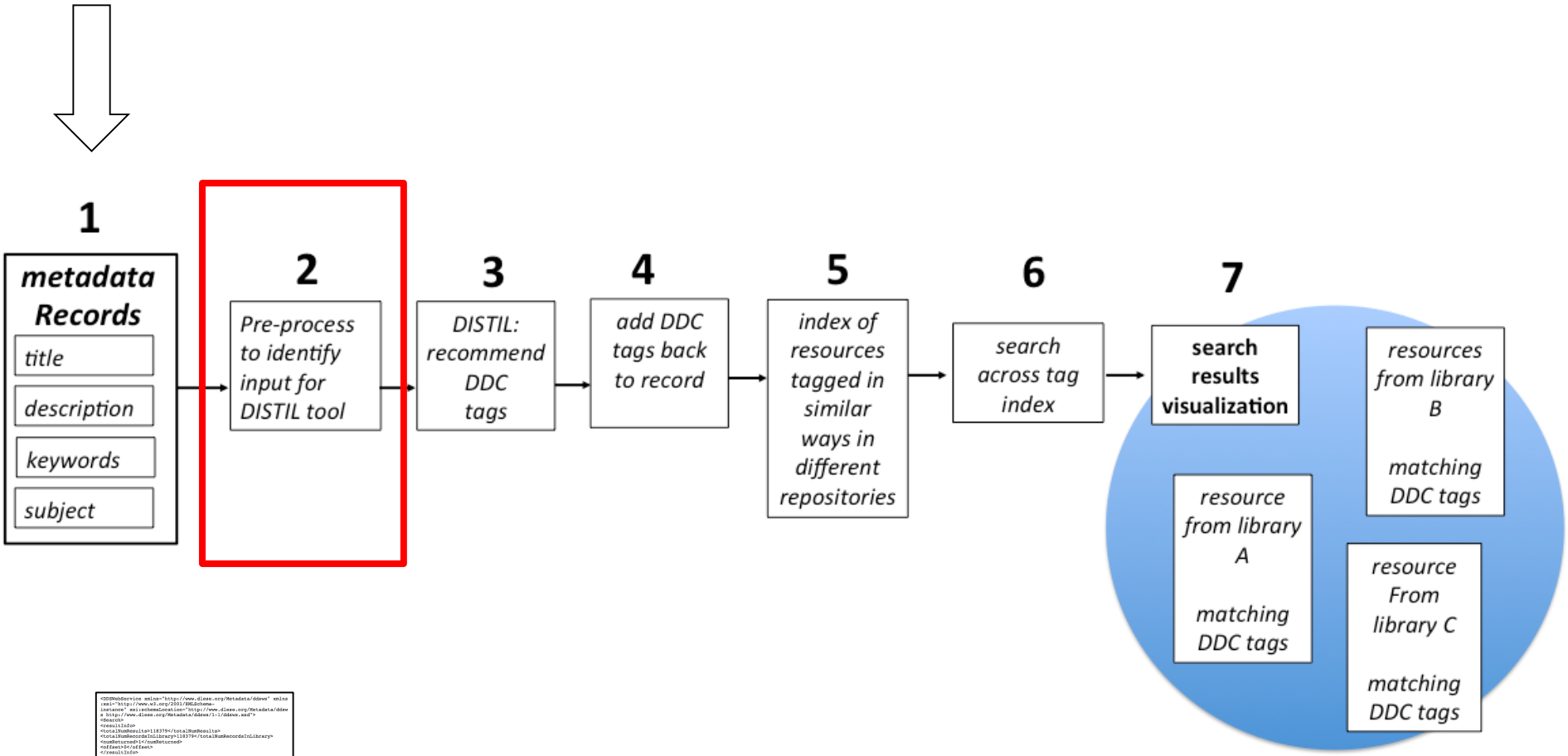
Observation

- Easy in theory
- In practice, organizational histories and legacy factors complicate the process
- Each DL's metadata is requiring:
  - Custom approaches in order to harvest and process
  - Access to specific people with specific knowledge



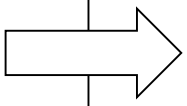

1
metadata Records
title
description
keywords
subject
2
Pre-process to identify input for DISTIL tool
3
DISTIL: recommend DDC tags
4
add DDC tags back to record
5
index of resources tagged in similar ways in different repositories
6
search across tag index
7
search results visualization
resources from library B
matching DDC tags
resource from library A
matching DDC tags
resource From library C
matching DDC tags


1
metadata Records
title
description
keywords
subject
2
Pre-process to identify input for DISTIL tool
3
DISTIL: recommend DDC tags
4
add DDC tags back to record
5
index of resources tagged in similar ways in different repositories
6
search across tag index
7
search results visualization
resources from library B
matching DDC tags
resource from library A
matching DDC tags
resource From library C
matching DDC tags


1
metadata Records
title
description
keywords
subject
2
Pre-process to identify input for DISTIL tool
3
DISTIL: recommend DDC tags
4
add DDC tags back to record
5
index of resources tagged in similar ways in different repositories
6
search across tag index
7
search results visualization
resources from library B
matching DDC tags
resource from library A
matching DDC tags
resource From library C
matching DDC tags

Stage 2: Pre-processing

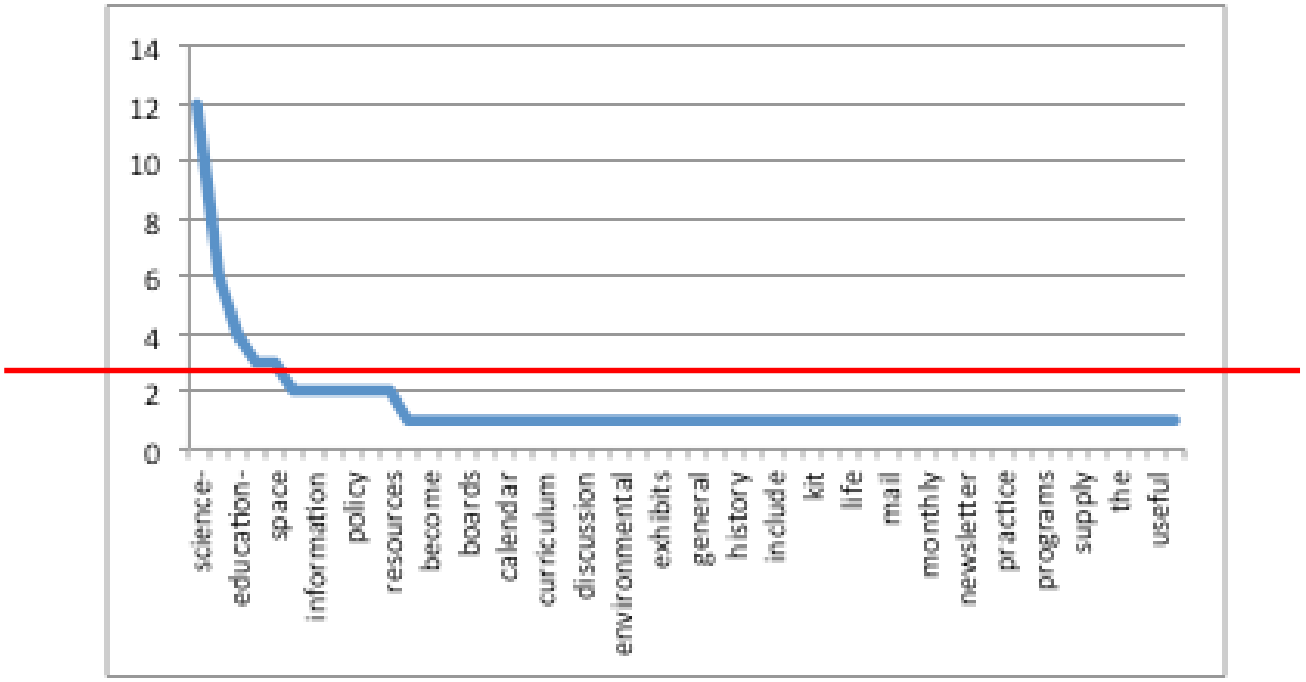
```
<DDSWebService xmlns="http://www.dlese.org/Metadata/ddsws" xmlns
:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:schemaLocation="http://www.dlese.org/Metadata/ddsw
s http://www.dlese.org/Metadata/ddsws/1-1/ddsws.xsd">
<Search>
<resultInfo>
<totalNumResults>118379</totalNumResults>
<totalNumRecordsInLibrary>118379</totalNumRecordsInLibrary>
<numReturned>1</numReturned>
<offset>0</offset>
</resultInfo>
<results>
<record>
<head>
<id>2200/20120112185023875T</id>
<xmlFormat nativeFormat="nsdl_dc">nsdl_dc</xmlFormat>
<collection recordId="ncs-NSDL-COLLECTION-000-003-111-
915" ky="2667291" key="ncs-NSDL-COLLECTION-000-003-111-915">
STEM Education and Educational Technology Gateways and Resources
</collection>
<fileLastModified>2012-07-29T20:27:52Z</fileLastModified>
</head>
<metadata>
<nsdl_dc:nsdl_dc xmlns:nsdl_dc="http://ns.nsdl.org/nsdl_dc_v1.02
/" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dct="http:/
/purl.org/dc/terms/"xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xmlns:lar="http://ns.nsdl.org/schemas/dc/lar" schemaVe
rsion="1.02.000"xsi:schemaLocation="http://ns.nsdl.org/nsdl_dc_v
1.02/ http://ns.nsdl.org/schemas/nsdl_dc/nsdl_dc_v1.02.xsd">
<dc:identifier xsi:type="dct:URI">http://www.nsta.org/</dc:ident
ifier>
<dct:hasPart>Science and Children</dct:hasPart>
<dct:hasPart>Science Scope</dct:hasPart>
<dct:hasPart>Science Teacher</dct:hasPart>
<dct:hasPart>Journal of College Science Teaching</dct:hasPart>
<dc:date xsi:type="dct:W3CDTF">2002</dc:date>
<dc:title>National Science Teachers Association
(NSTA)</dc:title>
<dc:description>
The National Science Teachers Association (NSTA) is an
organization committed to promoting excellence and innovation in
science teaching and learning. NSTA's membership includes
science teachers, science supervisors, administrators,
scientists, business and industry representatives, and others
involved in and committed to science education. The NSTA web
site provides an overview of the organization and its mission,
```



```
<dc:title>National Science Teachers Association
(NSTA)</dc:title>
<dc:description>
The National Science Teachers Association (NSTA) is an
organization committed to promoting excellence and innovation in
science teaching and learning. NSTA's membership includes
science teachers, science supervisors, administrators,
scientists, business and industry representatives, and others
involved in and committed to science education. The NSTA web
site provides an overview of the organization and its mission,
descriptions of services for members, and information on
professional development opportunities. There are also news
articles, conference announcements, information on NSTA
publications, and information for those who wish to become
involved in the organizaton's activities.
</dc:description>
<dc:subject>General science</dc:subject>
<dc:subject>Education</dc:subject>
```

Select fields and remove tags …

Stage 2: Pre-processing



Frequency counts

Sum (total occurrences) = 81

Mean = 1.6

Std Dev = 1.7

Cut off (Mean + Std Dev) = 3.3

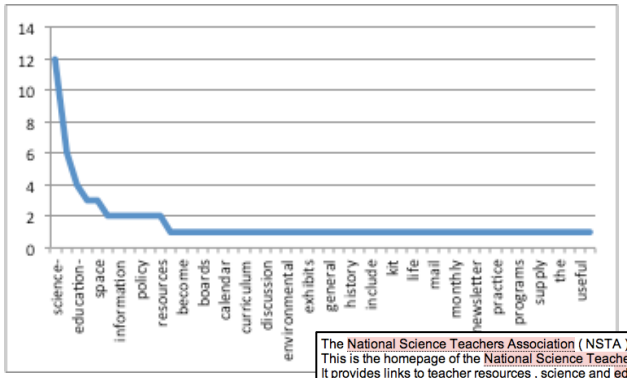
Stage 2: Pre-processing

The National Science Teachers Association (NSTA).
This is the homepage of the National Science Teachers Association (NSTA).
It provides links to teacher resources , science and education news , a calendar of exhibits , discussion boards , a monthly e-mail newsletter , information on teacher programs for professional development , and an opportunity to become an NSTA member.
Teacher resources include a curriculum kit about science and the food supply , information on books for teaching evolution , and useful websites.
Educational theory and practice.
Environmental science.
Policy issues.
Space science.
Science.
Earth science.
Physical sciences.
Biology.
Education (General).
Astronomy.
Space sciences.
Education.
Geoscience.
History/Policy/Law.
Chemistry.
Life Science.
Physics.
Space Science.
Technology.

Noun phrases

Frantzi, K., Ananiadou, S. and Mima, H. (2000) Automatic recognition of multi-word terms. International Journal of Digital Libraries 3(2), pp.117-132.
http://www.nactem.ac.uk/software/termine/

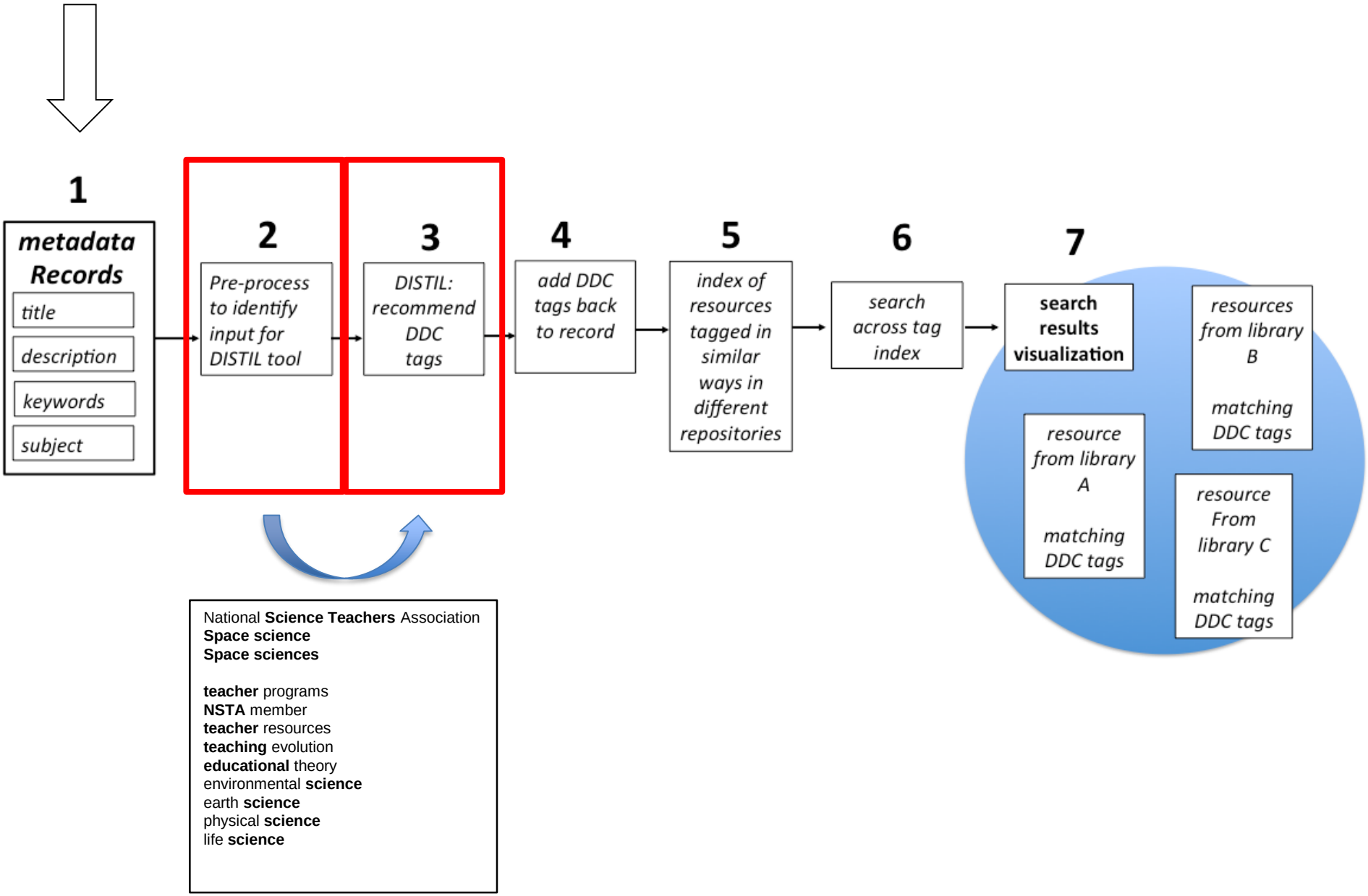
Stage 2: Pre-processing

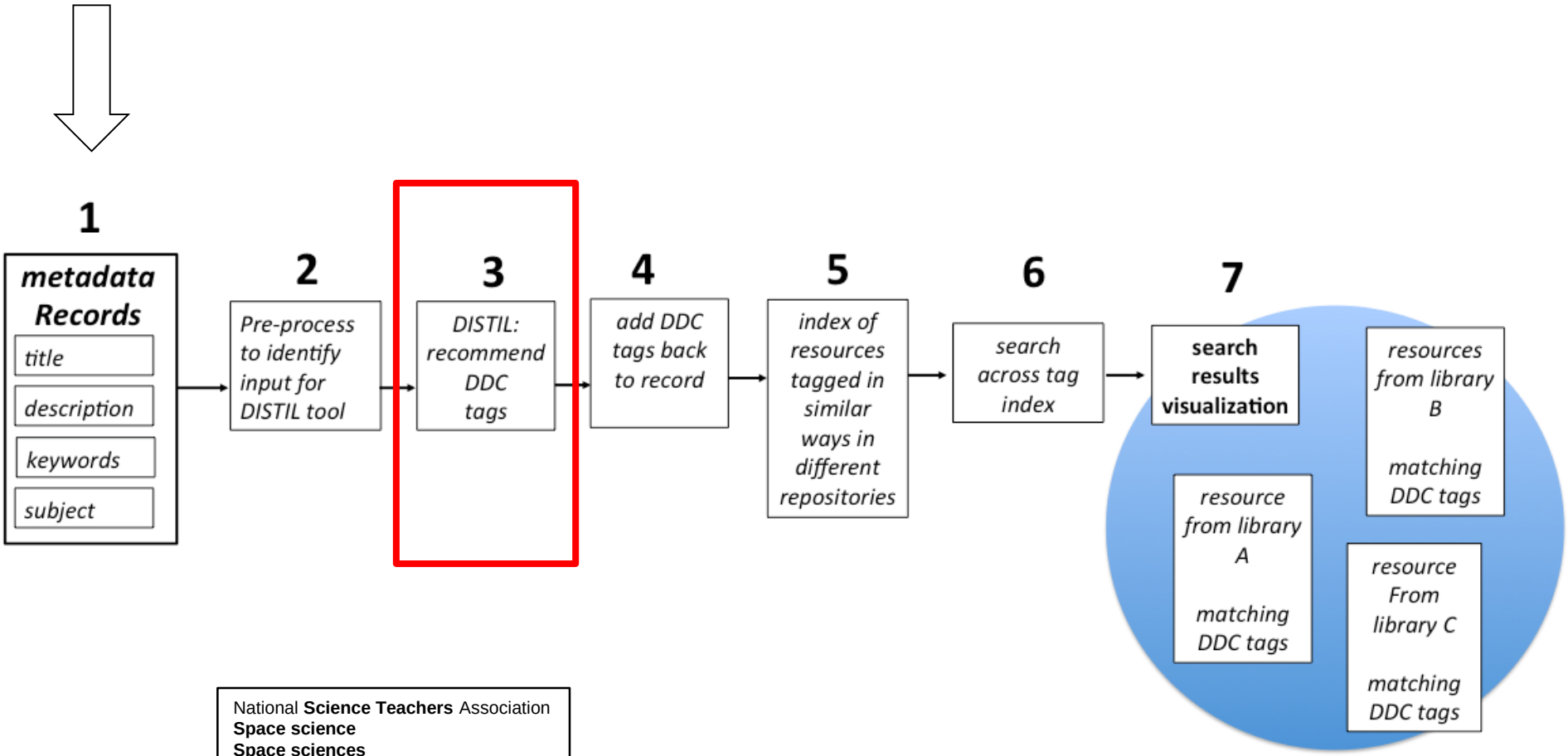


The National Science Teachers Association (NSTA).
This is the homepage of the National Science Teachers Association (NSTA).
It provides links to teacher resources , science and education news , a calendar of exhibits , discussion boards
newsletter , information on teacher programs for professional development , and an opportunity to become an
Teacher resources include a curriculum kit about science and the food supply , information on books for teach
websites.
Educational theory and practice.
Environmental science.
Policy issues.
Space science.
Science.
Earth science.
Physical sciences.
Biology.
Education (General).
Astronomy.
Space sciences.
Education.
Geoscience.
History/Policy/Law.
Chemistry.
Life Science.
Physics.
Space Science.
Technology.

National **Science Teachers** Association
Space science
Space sciences

teacher programs
NSTA member
teacher resources
teaching evolution
educational theory
environmental **science**
earth **science**
physical **science**
life **science**


1
metadata Records
title
description
keywords
subject
2
Pre-process to identify input for DISTIL tool
3
DISTIL: recommend DDC tags
4
add DDC tags back to record
5
index of resources tagged in similar ways in different repositories
6
search across tag index
7
search results visualization
resources from library B
matching DDC tags
resource from library A
matching DDC tags
resource From library C
matching DDC tags
National Science Teachers Association
Space science
Space sciences
teacher programs
NSTA member
teacher resources
teaching evolution
educational theory
environmental science
earth science
physical science
life science



National **Science Teachers** Association
Space science
Space sciences

teacher programs
NSTA member
teacher resources
teaching evolution
educational theory
environmental **science**
earth **science**
physical **science**
life **science**

Summary

- Work is complex but do-able (so far)
- Many subsidiary steps
- Harvesting work has a significant organizational knowledge dimension, and requires organizational communication*
  - Suggests a need for organizational models, processes, and best practices to account for and address the general nature of these phenomena

Khoo, M., Hall, C. (2012). Rethinking organizational distance: Networks of practice, legacy issues, and metadata work in a digital library project. Accepted, *Information and Organization*.

Lagoze, C., Krafft, D. B., Cornwell, T., Dushay, N., Eckstrom, D., & Saylor, J. (2006). Metadata aggregation and 'automated digital libraries': a retrospective on the NSDL experience. 6th ACM-IEEE Joint Conference on Digital Libraries (JCDL), June 11–15, 2006, Chapel Hill, North Carolina, USA, pp. 230-239.

Lagoze, C., & Patzke, K. (2011). A research agenda for data curation in cyberinfrastructure. Paper presented at the 11th ACM-IEEE Joint Conference on Digital Libraries (JCDL), June 13-17, 2011, Ottawa, Canada.

Thank you – and …

Questions?