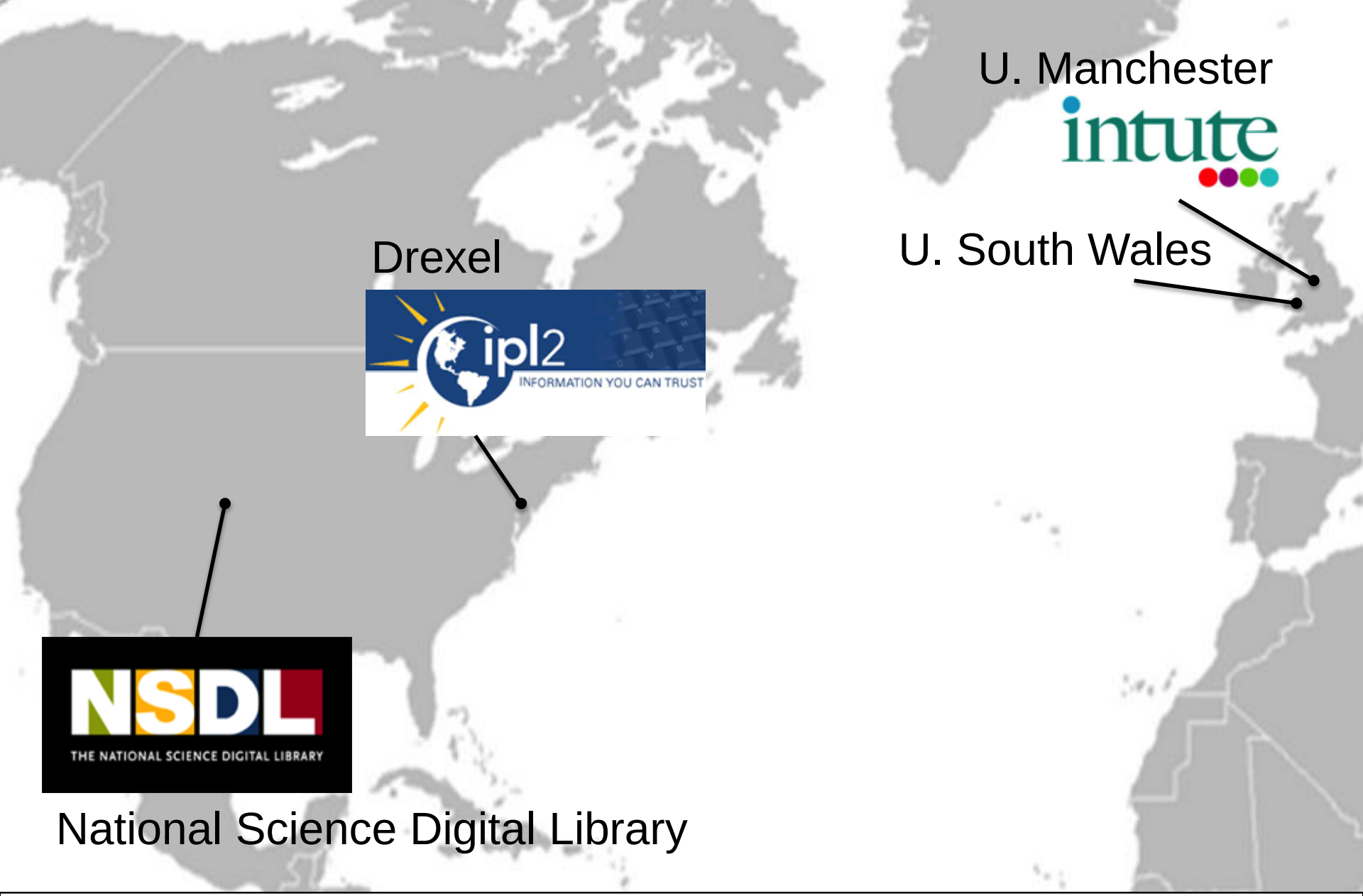
Digging into Metadata

Ceri Binding, Douglas Tudhope
(Hypermedia Research Unit, University of South Wales)
Jaewook Ahn, Mick Khoo, Xia Lin
(University of Drexel)
Diana Massam, Hilary Jones
(MIMAS, University of Manchester)

NKOS Workshop, TPDL 2013, Valetta, Malta, September 26, 2013

Introduction

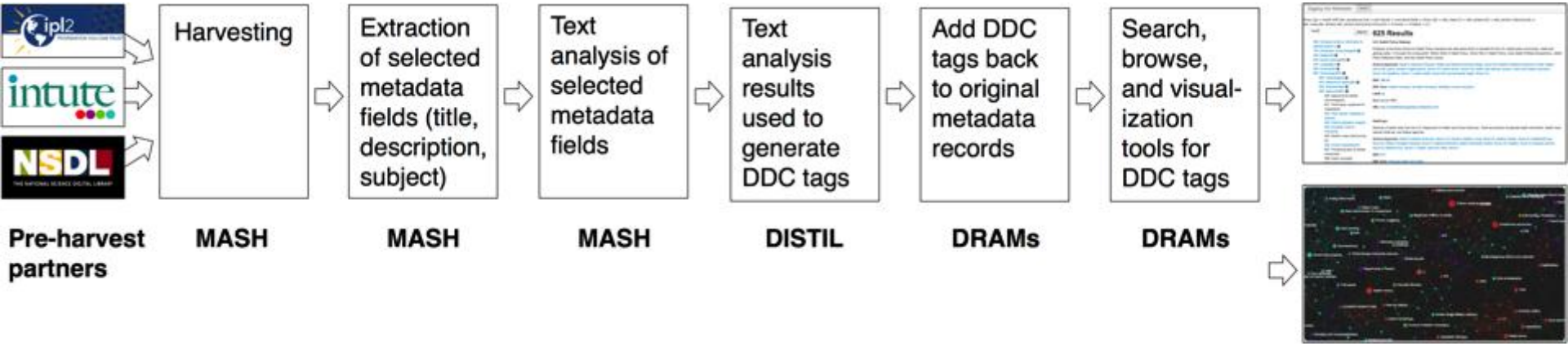
- Funding: *Digging Into Data Challenge*
- Setting
  - Small(ish)-scale, DC, educational DLs
  - Large-scale information infrastructures
- Aim: Achieve efficient federated search and discovery across heterogeneous DLs
- Method: Aggregate metadata, analyze, and generate DDC classes that can be used to support search and browse tools



National Science Digital Library

NSDL	IPL	Intute	Total
98,507	40,973	124,070	**263,550**

Project Workflow



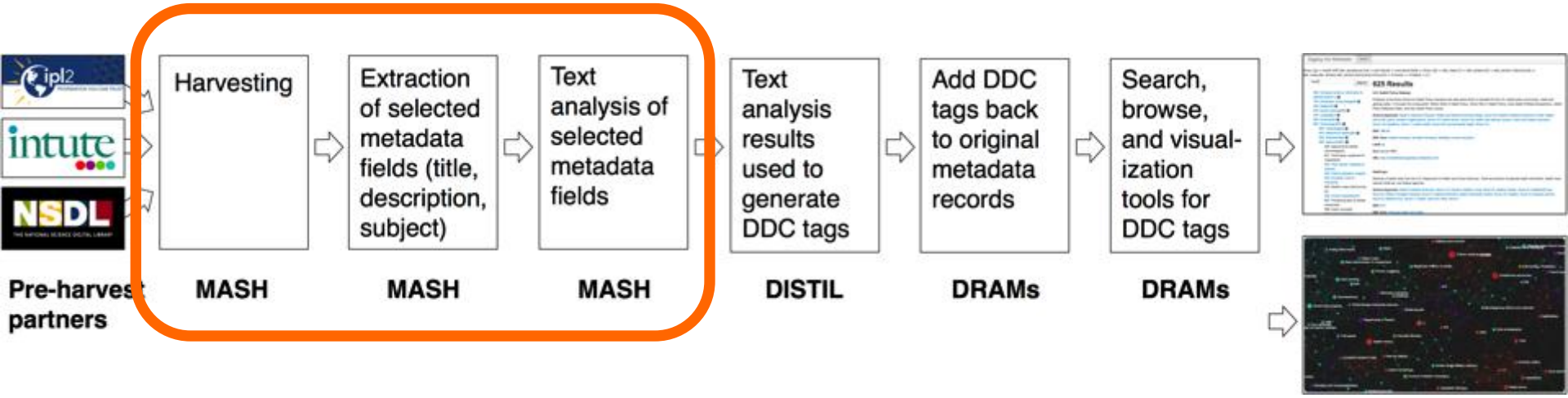
Databases:

MASH **M**etadata **A**ggregation **S**torage and **H**andling

DISTIL **D**ocument **I**ndexing & **S**emantic **T**agging **I**nterface for **L**ibraries

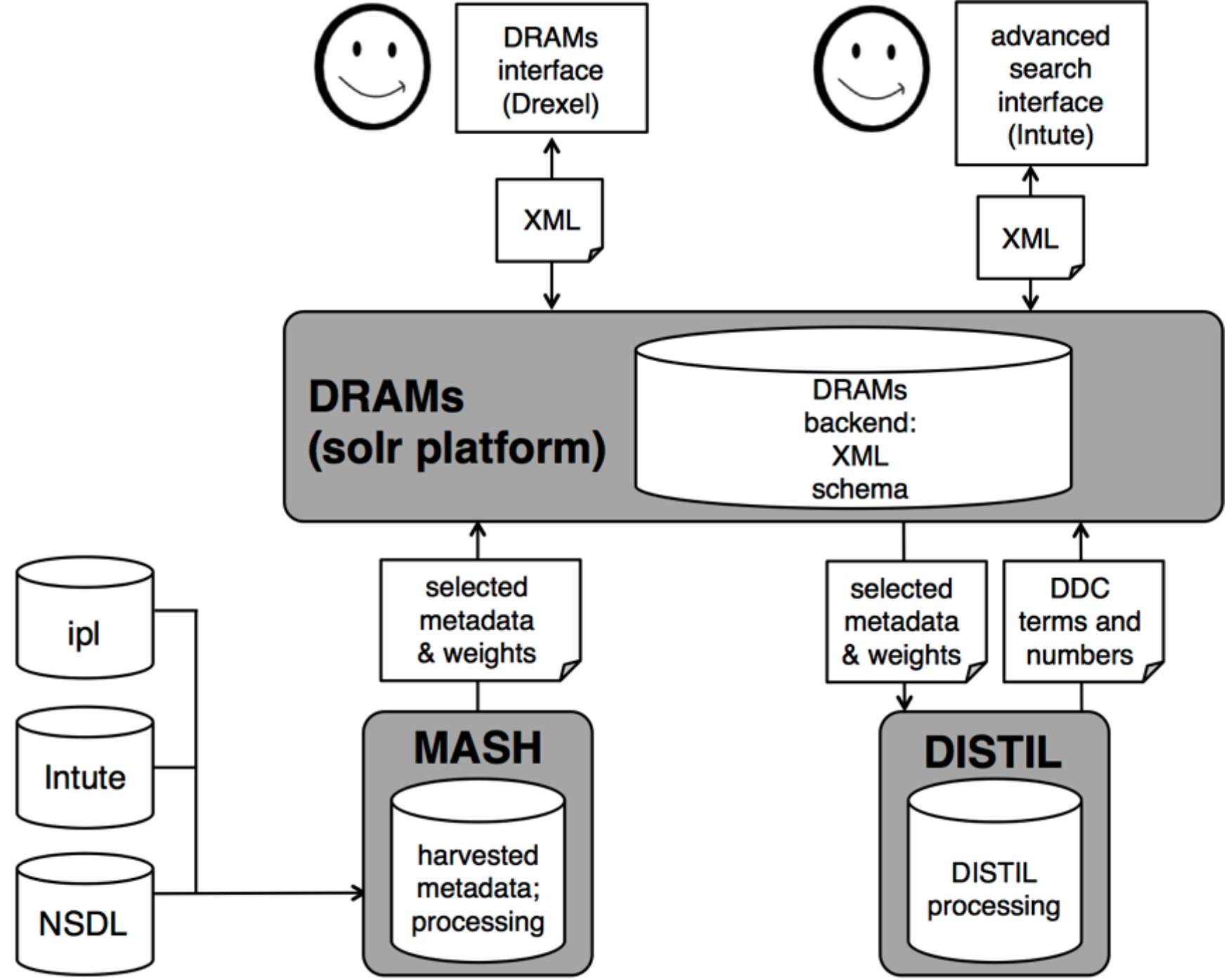
DRAMs **D**ynamic **R**epresentations of **A**nnotated **M**etadata

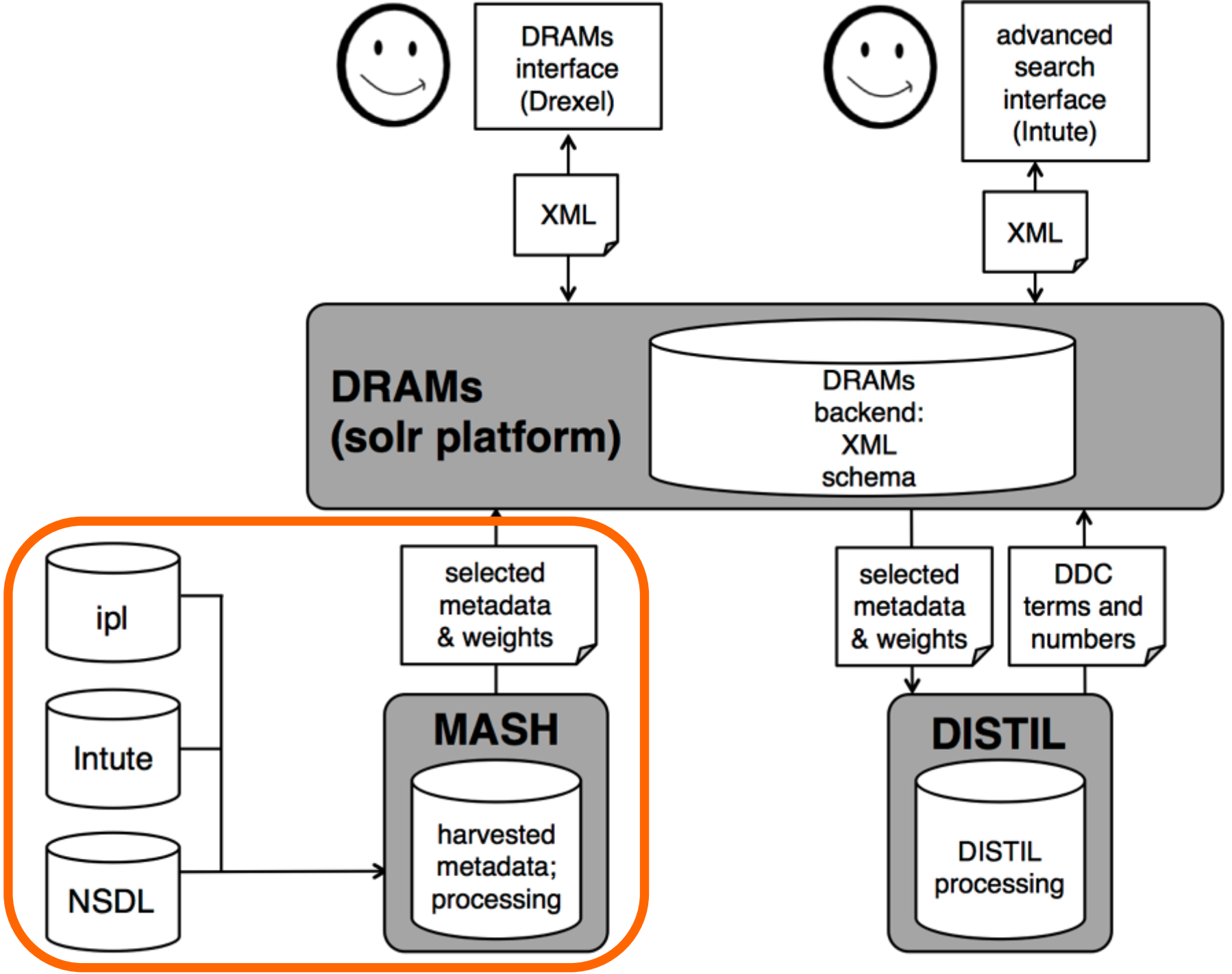
Project Workflow



IPL	Intute	NSDL	Total
40,973	124,070	98,507	**263,550**

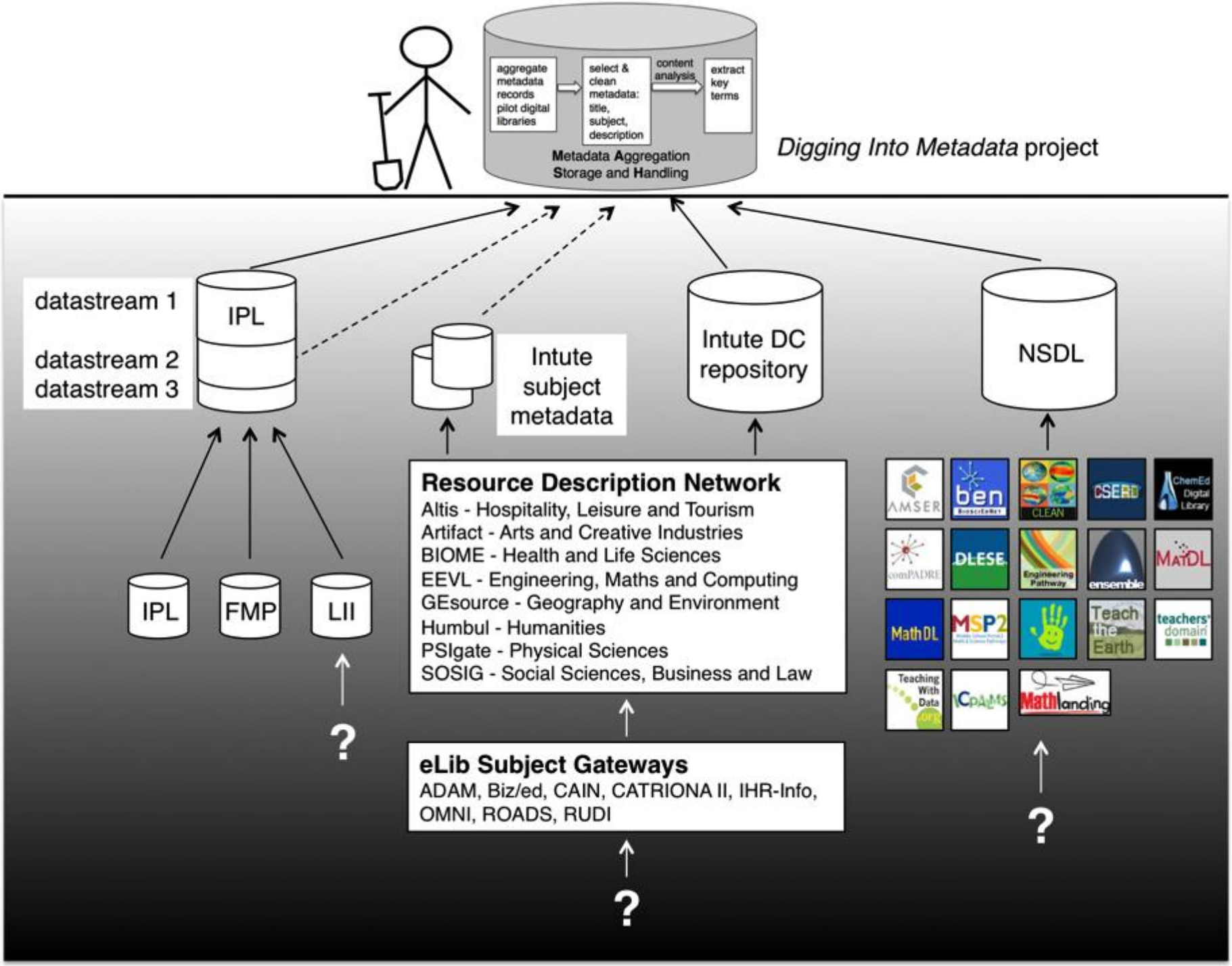
MASH	Metadata Aggregation Storage and Handling
DISTIL	Document Indexing & Semantic Tagging Interface for Libraries
DRAMs	Dynamic Representations of Annotated Metadata


DRAMs interface (Drexel)
advanced search interface (Intute)
XML
XML
DRAMs (solr platform)
DRAMs backend: XML schema
selected metadata & weights
selected metadata & weights
DDC terms and numbers
ipl
Intute
NSDL
MASH
harvested metadata; processing
DISTIL
DISTIL processing


DRAMs interface (Drexel)
advanced search interface (Intute)
XML
XML
DRAMs (solr platform)
DRAMs backend: XML schema
ipl
Intute
NSDL
selected metadata & weights
selected metadata & weights
DDC terms and numbers
MASH
harvested metadata; processing
DISTIL
DISTIL processing

Some Harvesting Issues and Questions …

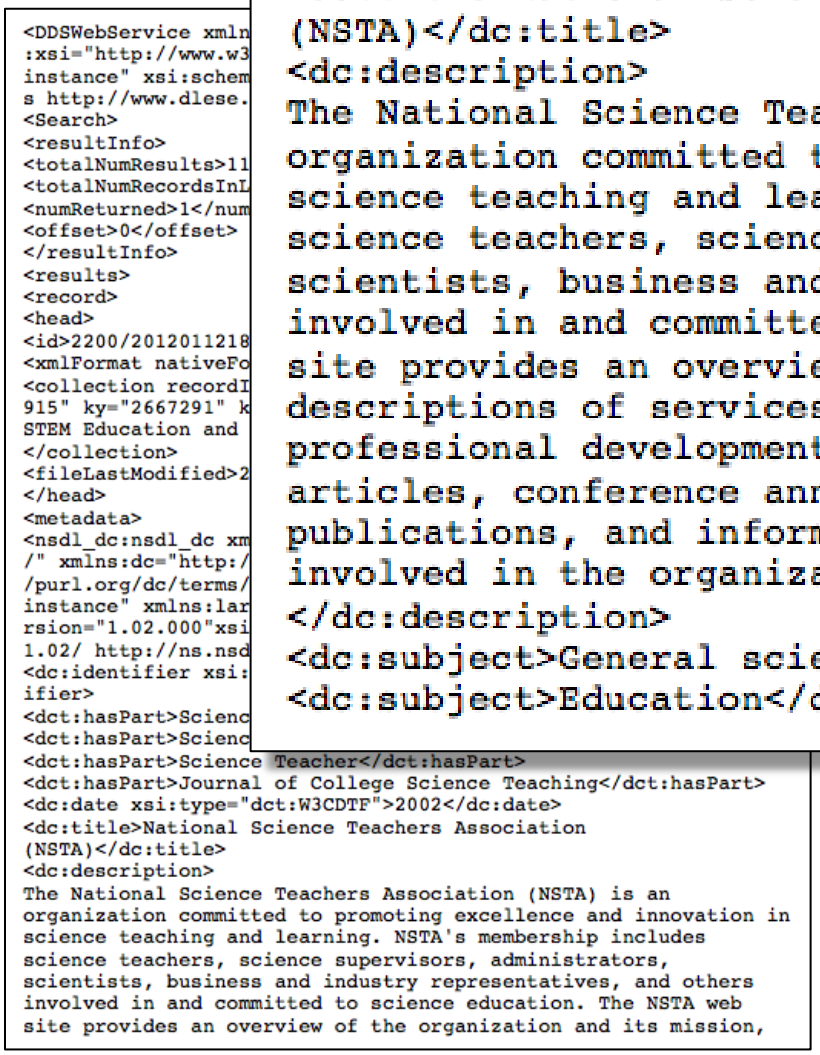
- Some stubborn legacy issues
- These required some communication and manual rectification – for databases as much as for individual fields
- This is expected – but is it normal?
- If it is normal, how can this be understood and modeled in order to develop better harvesting workflows?


aggregate metadata records pilot digital libraries
select & clean metadata: title, subject, description
content analysis
extract key terms
Metadata Aggregation Storage and Handling
Digging Into Metadata project
datastream 1
datastream 2
datastream 3
IPL
Intute subject metadata
Intute DC repository
NSDL
IPL
FMP
LII
?
Resource Description Network
Altis - Hospitality, Leisure and Tourism
Artifact - Arts and Creative Industries
BIOME - Health and Life Sciences
EEVL - Engineering, Maths and Computing
GEsource - Geography and Environment
Humbul - Humanities
PSIgate - Physical Sciences
SOSIG - Social Sciences, Business and Law
eLib Subject Gateways
ADAM, Biz/ed, CAIN, CATRIONA II, IHR-Info, OMNI, ROADS, RUDI
?
?

Pre-processing/cleaning

```
<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,
```

Pre-processing/cleaning




```
<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
- Remove tags
- Tokenization
- Stop word removal
- Porter stemming

Term Extraction

Terms selected over a specific threshold:

$$Threshold_{term} = mean(TF) + standarddeviation(TF)$$

where TF = Term Frequency

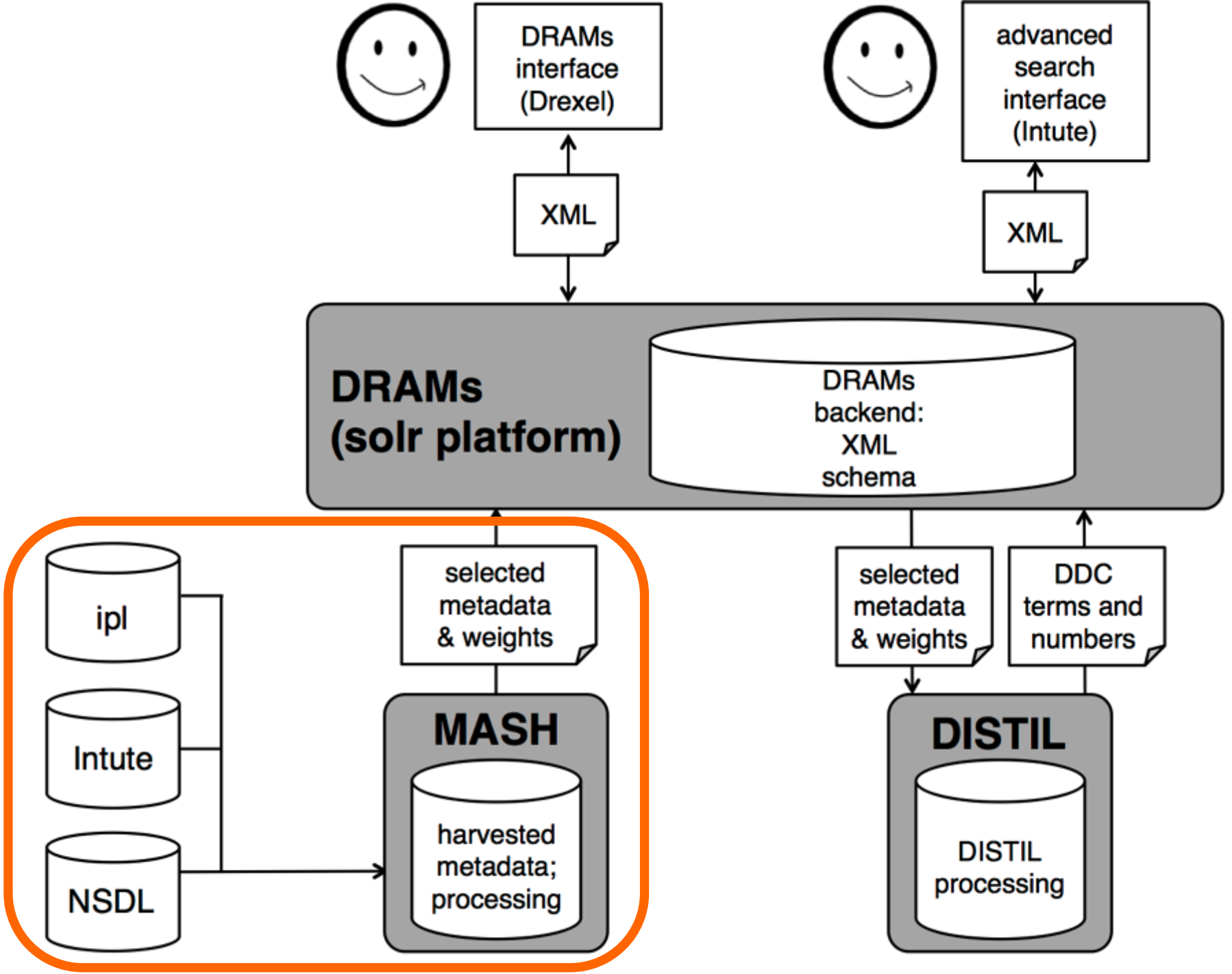
Phrase Extraction

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/

From MASH to DISTIL

- Aggregated metadata from the original digital library in original format (e.g., title, description, and subjects)
- Terms and phrases extracted from these records – these act as inputs to DISTIL to generate the DDC class for each record
- Administrative metadata


DRAMs interface (Drexel)
advanced search interface (Intute)
XML
XML
DRAMs (solr platform)
DRAMs backend: XML schema
ipl
Intute
NSDL
selected metadata & weights
selected metadata & weights
DDC terms and numbers
MASH
harvested metadata; processing
DISTIL
DISTIL processing