USGS

science for a changing world

Project Open Data Publishing to data.gov: The USGS Experience

FGDC ISO Implementation Forum

14 January 2015

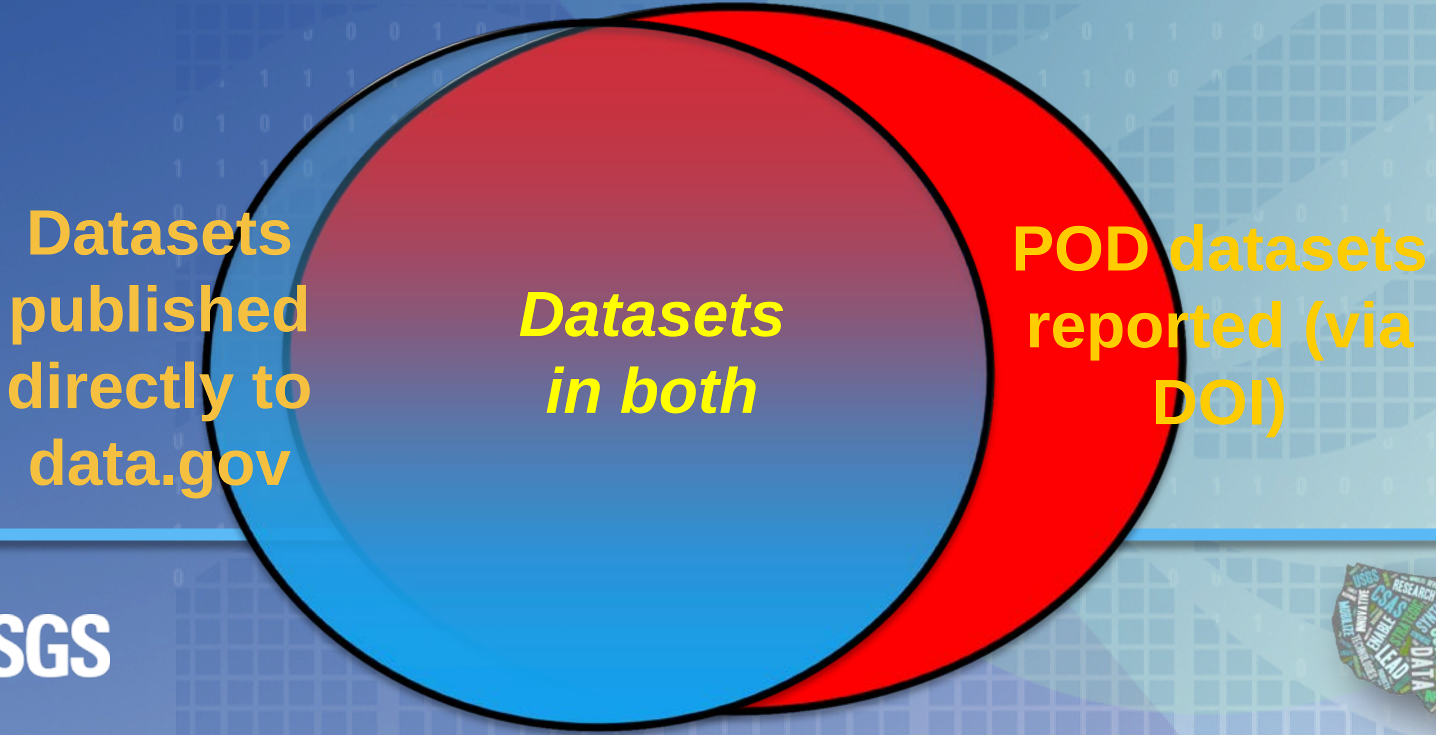
Lisa Zolly

USGS Core Science, Analytics, Synthesis & Libraries

U.S. Department of the Interior
U.S. Geological Survey

Today: separate paths and processes

- USGS reporting for Project Open Data is done *directly to Interior (DOI)*, who report for all DOI bureaus to OMB
- USGS publishing to data.gov is done *independently* by select USGS units/centers/programs


Datasets published directly to data.gov
Datasets in both
POD datasets reported (via DOI)

Today: USGS POD reporting to Interior

- Interior established CKAN Catalog, modeled on data.gov catalog, in 1Q FY2014
- Purpose 1: Public, actionable data listing for all DOI bureaus <data.doi.gov>
- Purpose 2: POD 1.0 reporting to OMB for all DOI bureaus

- *Imminent* Purpose 3: harvest source for DOI bureaus' contributions to catalog.data.gov

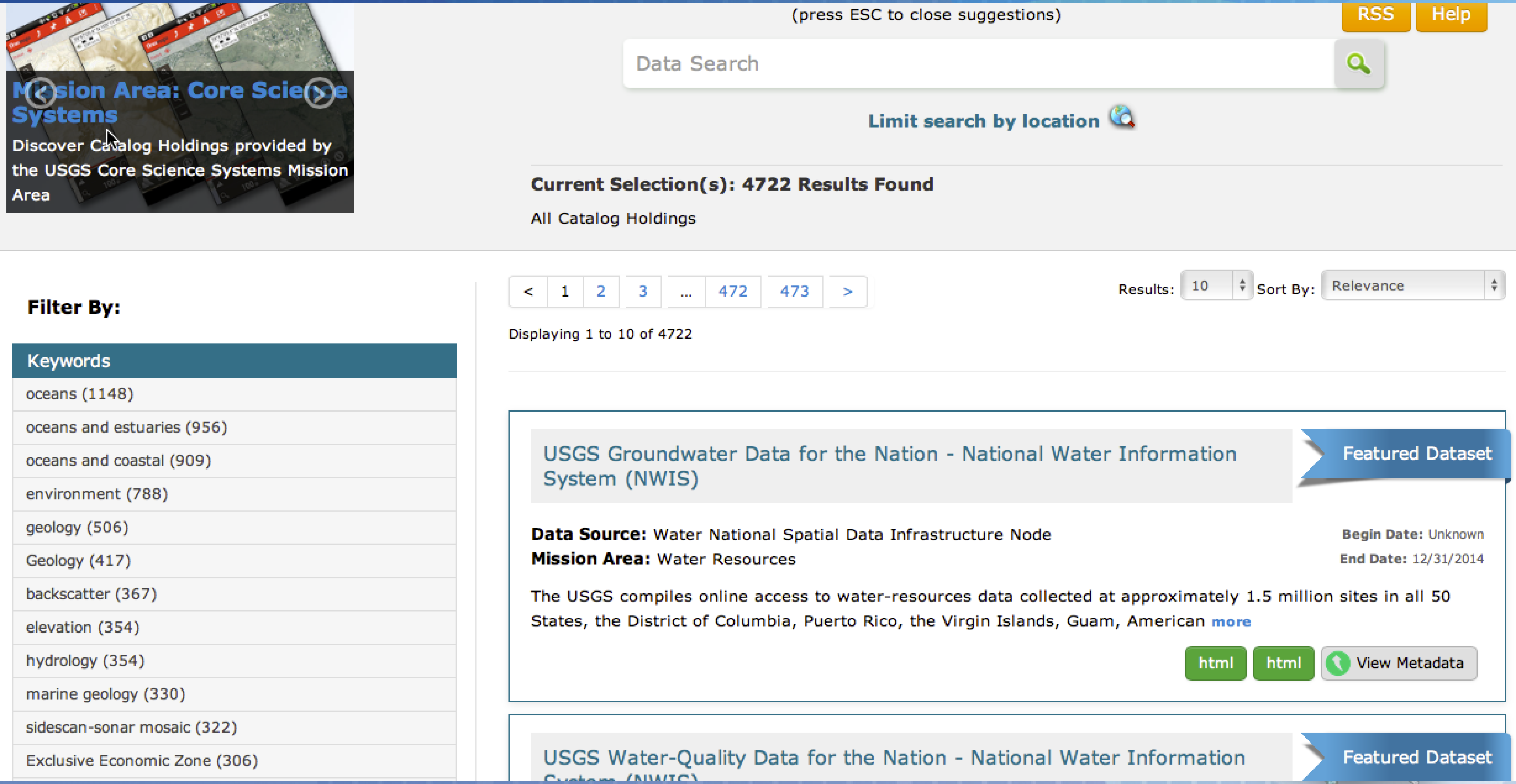
USGS

Today: USGS POD reporting to Interior

USGS Science Data Catalog established in 2Q FY2014

- Aggregation point for USGS dataset metadata
- Catalog harvests CSDGM metadata from a variety of USGS metadata WAFs and individual catalogs
- Public catalog published at data.usgs.gov
- Additional USGS metadata sources to be added in FY2015
- ISO metadata holdings to be added in FY2015

USGS Science Data Catalog



(press ESC to close suggestions)

RSS Help

Data Search

Limit search by location

Mission Area: Core Science Systems

Discover Catalog Holdings provided by the USGS Core Science Systems Mission Area

Current Selection(s): 4722 Results Found

All Catalog Holdings

Filter By:

Keywords

oceans (1148)
oceans and estuaries (956)
oceans and coastal (909)
environment (788)
geology (506)
Geology (417)
backscatter (367)
elevation (354)
hydrology (354)
marine geology (330)
sidescan-sonar mosaic (322)
Exclusive Economic Zone (306)

< 1 2 3 ... 472 473 >

Results: 10 Sort By: Relevance

Displaying 1 to 10 of 4722

USGS Groundwater Data for the Nation - National Water Information System (NWIS)

Featured Dataset

Data Source: Water National Spatial Data Infrastructure Node
Mission Area: Water Resources

Begin Date: Unknown
End Date: 12/31/2014

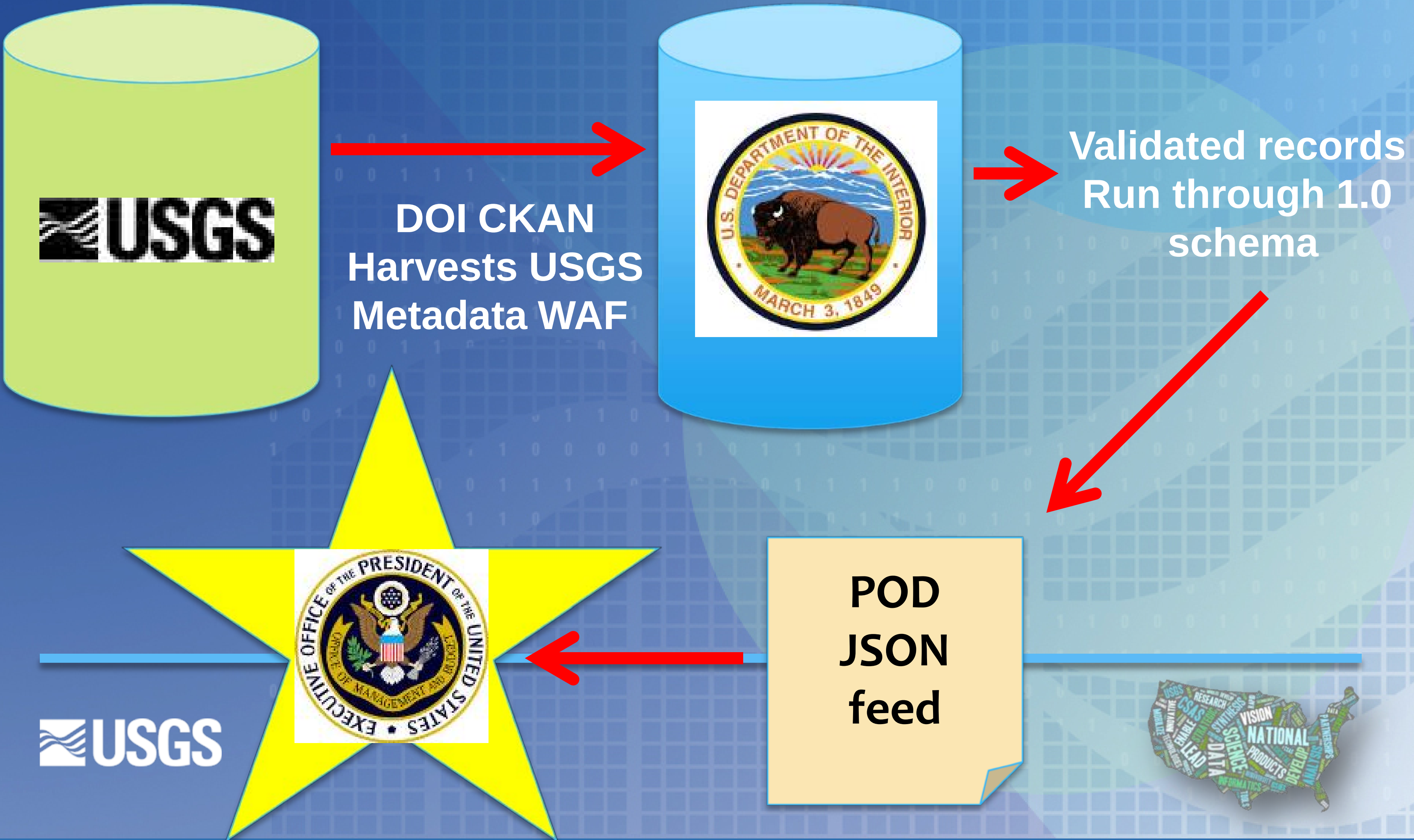
The USGS compiles online access to water-resources data collected at approximately 1.5 million sites in all 50 States, the District of Columbia, Puerto Rico, the Virgin Islands, Guam, American more

html html View Metadata

USGS Water-Quality Data for the Nation - National Water Information System (NWIS)

Featured Dataset

Today: USGS POD reporting



DOI CKAN Harvests USGS Metadata WAF

Validated records Run through 1.0 schema

POD JSON feed

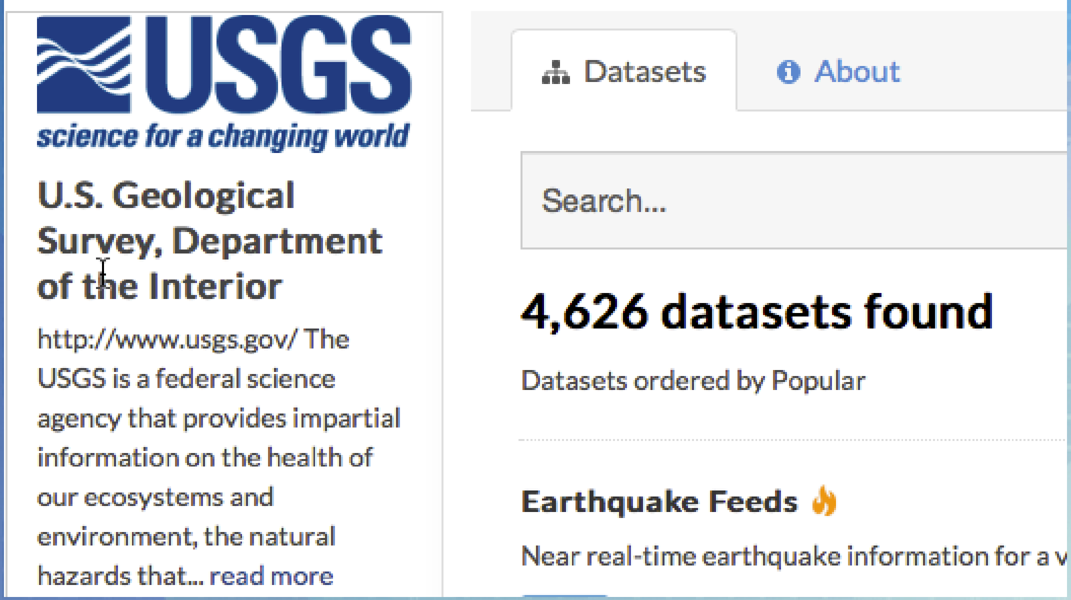
Current challenges

- Critical content not being retained between harvest from USGS and translation into 1.0 and 1.1
    - Loss of important content such as taxonomy (no mappings)
    - Loss of link back to original metadata source
    - Loss of originating programs, contacts
- Issues will be introduced into USGS holdings in data.gov if POD 1.1 is used as data.gov source for DOI bureaus

- Working with DOI contact to try to address problems
- Is POD 1.1 the best way to deliver geospatial metadata to data.gov?

USGS in data.gov: today's snapshot

- Number of records fluctuates by 10s and even 100s each day
  - Additions & deletions at sources
  - Problems with records passing and then failing ISO Transform



U.S. Geological Survey, Department of the Interior

http://www.usgs.gov/ The USGS is a federal science agency that provides impartial information on the health of our ecosystems and environment, the natural hazards that... read more

4,626 datasets found

Datasets ordered by Popular

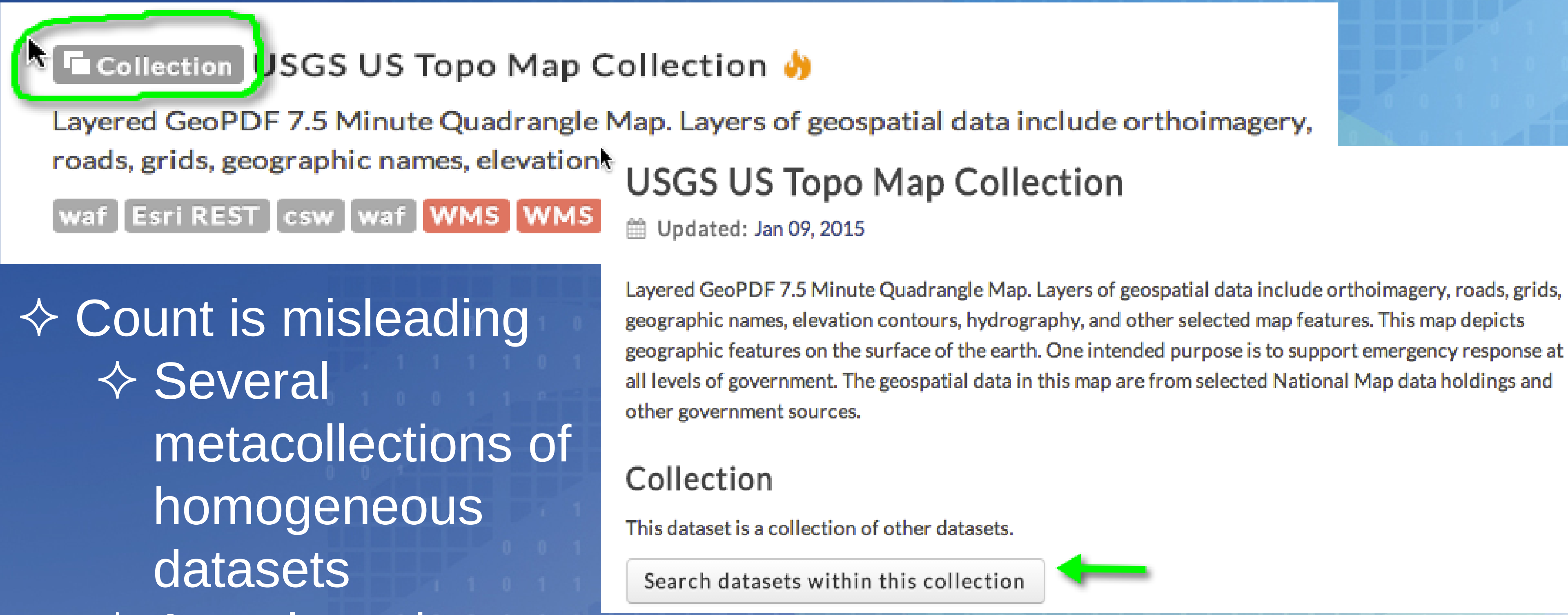
Earthquake Feeds

Near real-time earthquake information for a v

USGS in data.gov: today’s snapshot

- Number does *not* reflect the number of USGS datasets available
- Number reflects holdings from the segment of USGS units/programs/centers
  - grandfathered in from GOS
  - recruited by data.gov communities to provide specific datasets (e.g. NGDA datasets)
  - Currently willing and (somewhat) able to attempt to understand process and maintain data.gov harvest points

USGS in data.gov: today's snapshot



USGS US Topo Map Collection

Layered GeoPDF 7.5 Minute Quadrangle Map. Layers of geospatial data include orthoimagery, roads, grids, geographic names, elevation

USGS US Topo Map Collection

Updated: Jan 09, 2015

Layered GeoPDF 7.5 Minute Quadrangle Map. Layers of geospatial data include orthoimagery, roads, grids, geographic names, elevation contours, hydrography, and other selected map features. This map depicts geographic features on the surface of the earth. One intended purpose is to support emergency response at all levels of government. The geospatial data in this map are from selected National Map data holdings and other government sources.

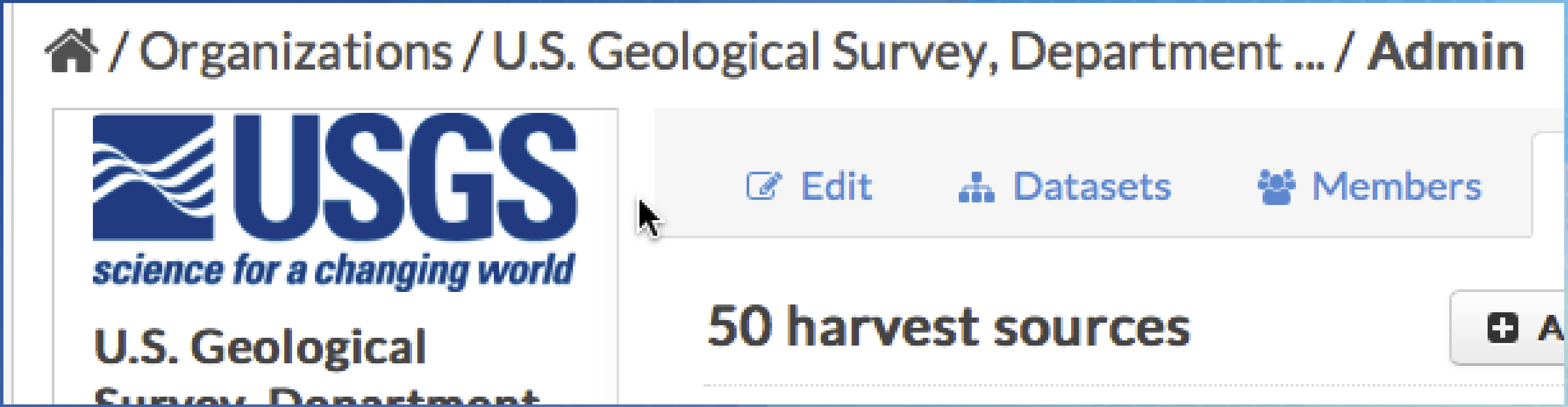
Collection

This dataset is a collection of other datasets.

- Count is misleading
  - Several metacollections of homogeneous datasets
  - Actual number exceeds 1M

USGS

USGS in data.gov: today's snapshot



/ Organizations / U.S. Geological Survey, Department ... / **Admin**

U.S. Geological Survey, Department

50 harvest sources

Why on earth does USGS have *50* harvest sources?!!?

USGS metadata universe

- Metadata historically a highly distributed activity within USGS
- With a few notable exceptions, USGS does not have 'data centers'
- Units, science centers, regional offices, field stations, programs generally responsible for
  - Creating metadata
  - Publishing metadata
  - Distributing metadata

USGS metadata universe

- ✧ Some units have established metadata assistance and process
- ✧ In other units, research teams ‘on their own’ to produce metadata
- ✧ Metadata validation and quality control varies
- ✧ Most USGS metadata still produced as CSDGM
- ✧ ~20-25% is CSDGM+Biological Data Profile
- ✧ Select programs are transitioning to ISO
  - ✧ Mostly basic 19115-2 records, not robust

USGS presence in Data.gov

- 50 USGS ‘collections’ of metadata from various programs, centers, units
  - Many are heterogeneous, several are homogeneous
- ~25% were grandfathered over from GOS
- Size of collections varies widely
- Most of the NGDA datasets are included
- Current holdings in data.gov do *not* represent the total USGS metadata holdings across the Bureau

USGS presence in Data.gov

Why aren't all USGS data holding now in data.gov?

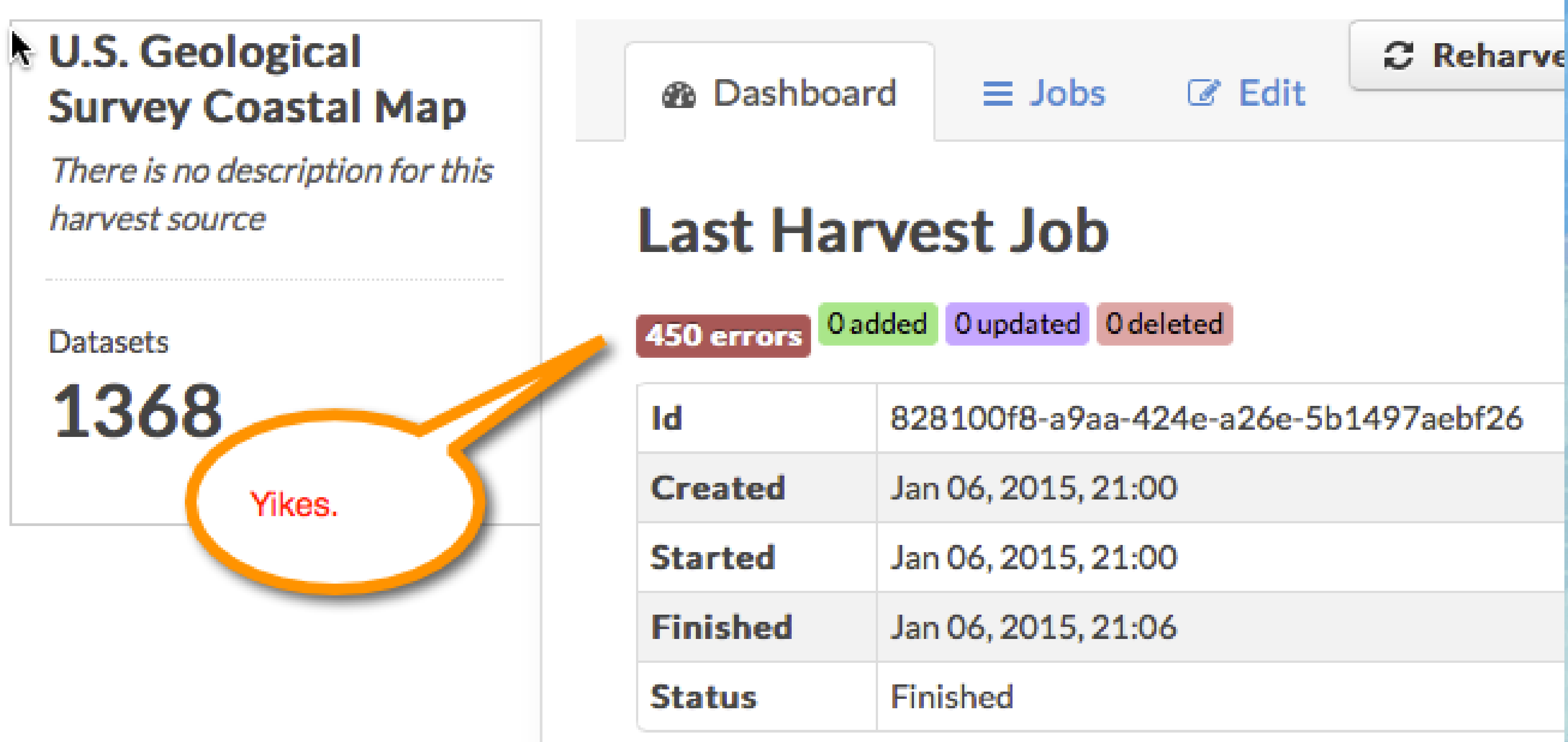
- Many units/programs/centers lack:
  - Operational or technical understanding of how to publish metadata outside local holdings
  - Technical infrastructure to aggregate metadata and data
  - Personnel to perform these tasks routinely
- Lack of comprehensive metadata policies and processes
  - Varying quality and compliance
  - Varying levels of access/publishing

USGS

Harvest challenges in Data.gov

- All ‘managed’ by different people
- Different harvest locations, frequencies, levels of engagement
- Struggles include
  - Opaqueness of ingest process/workflow
  - Records failing ISO Transform
  - In some collections, 10-50% of submitted records are not getting published in data.gov

Harvest challenges in Data.gov



U.S. Geological Survey Coastal Map

There is no description for this harvest source

Datasets

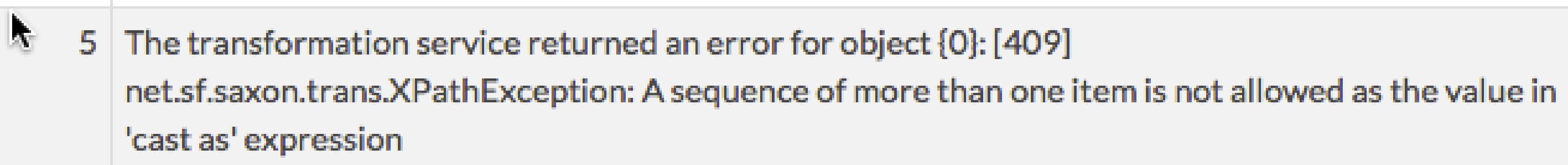
1368

Last Harvest Job

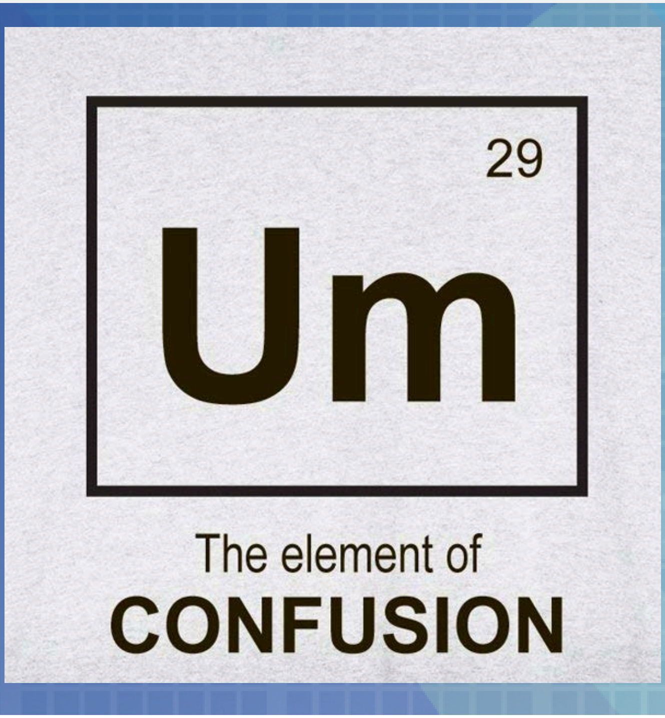
450 errors 0 added 0 updated 0 deleted

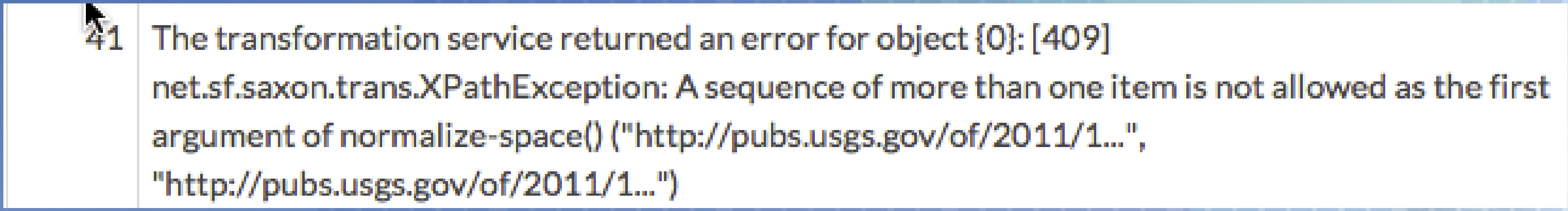
Id	828100f8-a9aa-424e-a26e-5b1497aebf26
Created	Jan 06, 2015, 21:00
Started	Jan 06, 2015, 21:00
Finished	Jan 06, 2015, 21:06
Status	Finished

Harvest challenges in Data.gov



```
5  The transformation service returned an error for object {0}: [409]
   net.sf.saxon.trans.XPathException: A sequence of more than one item is not allowed as the value in
   'cast as' expression
```





```
41  The transformation service returned an error for object {0}: [409]
    net.sf.saxon.trans.XPathException: A sequence of more than one item is not allowed as the first
    argument of normalize-space() ("http://pubs.usgs.gov/of/2011/1...",
    "http://pubs.usgs.gov/of/2011/1...")
```

Harvest challenges in Data.gov

- Heterogeneous collections contain mix of records using CSDGM as well as CSDGM+BDP
  - Problem: can specify only one validation schema per collection
- Variability in the robustness of CSDGM records within heterogeneous collections
  - Problem: some CSDGM validation schema choices expect robust records

We recommend “FGDC Minimal Validation” for all CSDGM harvest sources

Harvest challenges in Data.gov

- Current data.gov validation goes beyond what data.gov is actually using in the catalog.data.gov index
  - Qualitative AND quatitative
  - Why validate on what’s not being used?
- Agencies should be responsible for enforcing metadata quality
- Data.gov should validate only on what it needs to support its index and POD

MAJOR thanks to FGDC for facilitating conversation that has led to ad hoc focus group on this issue!

Harvest challenges in Data.gov

- **New mystery:** records that were passing validation in October began failing in November
- **Discovery:** someone edited github Transforms and introduced errors
- **Recommendation:** Need some governance on access and edits to Transforms….*these impact the entire data.gov universe*

HUGE thank you to Anna Milan, Jaci Mize, and Kathy Martinolich at NOAA for helping us to troubleshoot confusing harvest report errors and for recognizing recent errors introduced to ISO Transform!

USGS

Change is underway at USGS

- Increasing emphases on data management at all levels
- Increasing awareness of open data policies at all levels
- Imminent release of new, Bureau-wide policies on data management, metadata, data release
- 2014 release of USGS Science Data Catalog
  - Aggregation point for USGS metadata
  - Public window to USGS data

Migration to ISO

- Reluctance to move to ISO related mostly to
  - Comfort-levels with CSDGM
  - Lack of form-based tools to do ISO
  - Concerns about loss of *details* in 19115/19115-2 related to entity & attribute, methodology (i.e. 19110 and 19157)
- Movement to ISO *will* happen, albeit gradually

USGS

Thanks!

Lisa Zolly
lisa_zolly@usgs.gov