The NDS Universally Accessible Data Publications Pilot

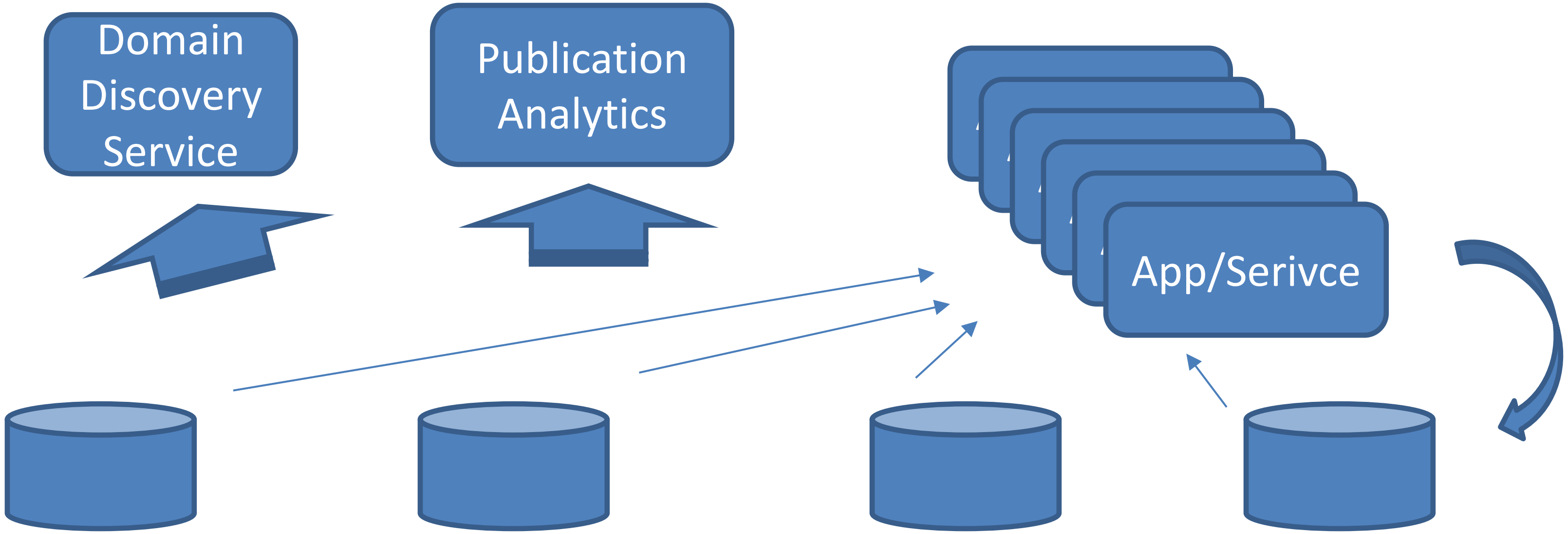
Jim Myers
Co-chairs: Sharief Youssef, Ray Plante

(Open effort, includes participation from DataOne, DFC, SEAD, and NIST)

Motivation

- The NDS/RDA/NSF/Open Data/etc. Vision
  - The ability to find, access, and reuse data, and to easily annotate and publish new work
- To take catalytic steps that support progress towards this vision
  - If it’s all about the data, let’s start by making sure it, and the information about it, can be accessed!

Use Cases



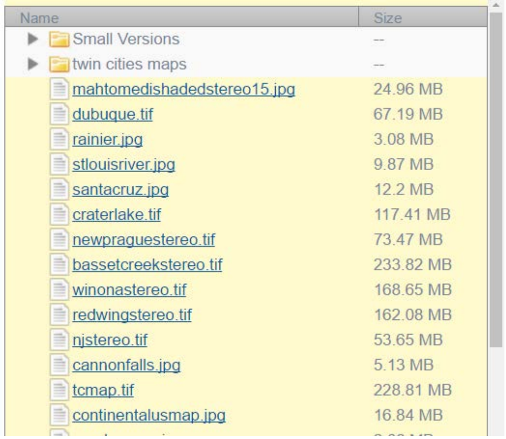
- A researcher **drops the identifier for a data publication from an arbitrary source in their analysis tool** and the **tool is able to retrieve and process a relevant data file(s)** automatically or after presenting the researcher with a brows-able display of the publications content so a selection can be made.
- The **results from a data analysis** such as the one just described **can be published in a way that they can be retrieved and then analyzed or visualized within another service**, duplicating the first use case, without any coordination between the service providers.
- A **domain-specific catalog is able to discover all data publications in targeted repositories** and **perform a deep scan** of their content to identify and index any relevant content.
- New **services can analyze the overall corpus of data publications** to discover what exists, how it is structured and annotated, what correlates with impact, ….

What’s required to enable…

http://doi.org/10.5967/M0NP22DR

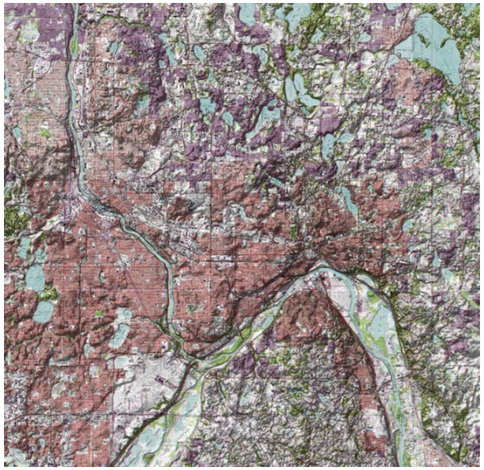
Drop a PID

Retrieve and display components and metadata



Name	Size
Small Versions	--
twin cities maps	--
mahtomedishadedstereo15.jpg	24.96 MB
dubuque.tif	67.19 MB
rainier.jpg	3.08 MB
stlouisriver.jpg	9.87 MB
santacruz.jpg	12.2 MB
craterlake.tif	117.41 MB
newpraguestereo.tif	73.47 MB
bassetcreekstereo.tif	233.82 MB
winonastereo.tif	168.65 MB
redwingstereo.tif	162.08 MB
njstereo.tif	53.65 MB
cannonfalls.jpg	5.13 MB
tcmap.tif	228.81 MB
continentalusmap.jpg	16.84 MB

Retrieve specified file and process it



“Title”: “njstereo.jpg”

“rdf:Label”: “njstereo.jpg”

“Name”: “njstereo.jpg”

Use Cases

- Analysis tools can use a standard library/have one mechanism to access data regardless of publisher
  - 1-1 collaborations/agreements not needed
- Catalogs can retrieve any/all metadata
  - Geospatial, domain, provenance, any other advanced searches can be supported over any sources that can provide the metadata
- Researchers can use source-agnostic browsing interfaces that show metadata terms/values and data items
- Everyone is pressured to standardize/bridge existing metadata, expand required sets because that is now the primary issue in automating integration…
- A data inventory and analysis of holdings across systems becomes possible…

The Challenge

- Given an arbitrary data publication identifier today, how much of the process to access it can be automated?
- Not much!
  - Different IDs may or may not be URLs, and have different resolvers
  - ID schemes may or may not provide some metadata, through schema-specific mechanisms, but they don’t provide all
  - IDs may only resolve to a human-readable page
  - The mechanism for retrieving data once finding a landing page is repository/publisher specific
  - The mechanism for retrieving metadata once finding a landing page is repository/publisher specific
  - There is no standard syntax for the metadata if/when you find it

All of which is independent of any standards for data format or required metadata fields

UADP Pilot

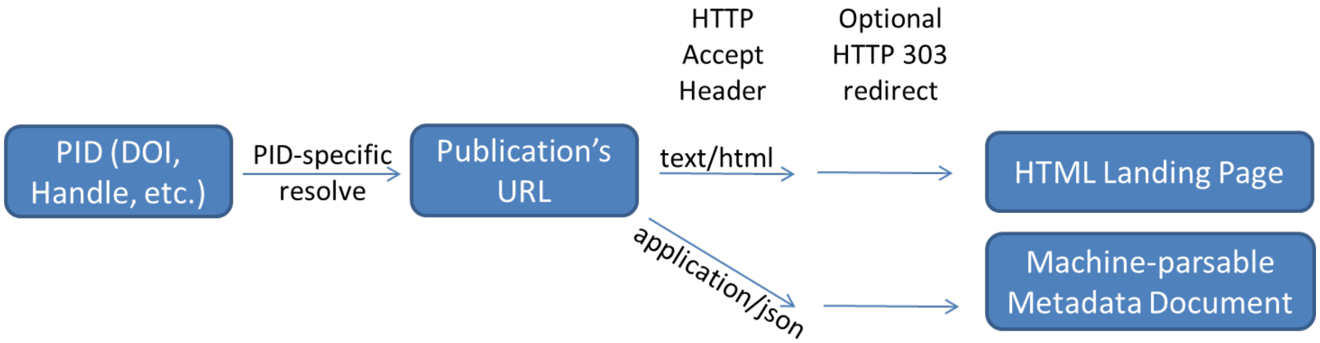
- Task 1: Implement a standard way to get data/metadata given an identifier (API, gateway service, export format, …)
- Task 2: Implement a standard harvesting mechanism (that provides, or ties to Task 1 to provide, full data/metadata access)
- Task 3: Document the metadata provided by repositories and needed by services
- Task 4: Standardize the means to identify and retrieve individual files from within a publication

Timeline

- Draft Proposal started after NDS 6 meeting.
- Initial wiki materials and refinement of proposal through Nov-Dec 2016.
- Kick-off meeting Jan. 13, 2017
- Input from pilot participants on current practices and technical approaches
- Decision, initial implementations, test data, write-up -

A start on task 1

- Sub-group met, discussed, and proposes:



Next steps

- If you’re interested, as a source or user of data publications or other interested party, **join the pilot!**
- Discussion tomorrow morning
- Mail list
- Group wiki
- Semi-regular meetings
- Access to NDS Labs for dev/testing

Thanks!

- https://nationaldataservice.atlassian.net/wiki/display/NDSC/Universally+Accessible+Data+Publications+Pilot

Community
Participants
Committees and Working Groups
TAC Interoperability Task Force
DataNet/DIBBs Vision Paper
Universally Accessible Data Publications Pilot

Title? Abstract? License? Topic? Coordinates? Quality? …

- All very important – but when we can't even retrieve what's there, being able to interpret it is an academic question
- The work in this pilot can be seen as a pre-requisite & catalyst to efforts to standardize vocabularies/required kernels, etc.
- It also has value as a stand-alone effort (applications can display metadata to users and retrieve data files they select without further standardization)

DOI
Handle
ARK
RDA PIT
DataCite
EzID

Current examples: For published data, can you find the Data and Metadata …?

- https://dx.doi.org/10.6084/m9.figshare.4515722.v2
  - https://figshare.com/articles/Data-Driven_Decision-Management_A_Values-focused_Approach_to_Enable_Traceable_Decision_Analytics_for_Adaptive_Climate_Resilience/4515722
    - https://ndownloader.figshare.com/files/7341470
      - ESIPandNCSEPosterv1.5.pdf
  - Or Figshare API (or DataCite: https://api.datacite.org/works/10.6084/m9.figshare.4515722.v2)
- doi:10.13012/J8CC0XM
  - http://dx.doi.org/10.13012/J8CC0XM
    - http://www.isws.illinois.edu/warm/datatype.asp
      - http://www.isws.illinois.edu/warm/datalist.asp
        - » http://www.isws.illinois.edu/warm/stationlist.asp?y=2016&m=6&site=&stn=&from=
          - http://www.isws.illinois.edu/warm/icndata.asp?y=2016&m=6&stn=Freeport
            - http://www.isws.illinois.edu/warm/data/2016/June/Freeport.txt
    - Or http://www.isws.illinois.edu/warm/data/cdfs/alldata.zip which contains two zips which contain data files
    - Which will include Jan 2017 data next month…
  - Metadata subset at http://ezid.cdlib.org/id/doi:10.13012/J8CC0XMK - html or xml link or https://api.datacite.org/works/10.13012/J8CC0XMK
- http://hdl.handle.net/102.100.100/15
  - https://store.synchrotron.org.au/experiment/view/879/
    - https://store.synchrotron.org.au/download/experiment/879/tar/ or
    - login from SFTP button or
    - Download button that doesn't have URL, requires selection of dataset parts elsewhere on page
- http://doi.org/10.5967/M0NP22DR
  - https://nced.ncsa.illinois.edu/refrepository/landing.html#tag:sead-data.net,2015:RO_FSrI6AEmuKutlOBEuDif8g
    - Metadata in ORE-JSON-LD syntax: https://nced.ncsa.illinois.edu/refrepository/api/researchobjects/tag:sead-data.net,2015:RO_FSrI6AEmuKutlOBEuDif8g/meta/oremap.jsonld.txt
    - Data in BagIT zip file with ORE-JSON-LD metadata: https://nced.ncsa.illinois.edu/refrepository/api/researchobjects/tag:sead-data.net,2015:RO_FSrI6AEmuKutlOBEuDif8g/bag