Data Repository Recommender

This page is intended to capture information related to **NDS-211** - Getting issue details... STATUS .

Overview

The goal of this project is to develop a general-purpose research data repository "recommender" service to be hosted by the NDS. The basic use case is very broad: *a research has data that they want to deposit, but they don't know where to put it*.

How is this problem currently addressed? We can find a few cases in the wild:

Service	Data Repository Recommendation
U of I Research Data Service	"Deposition of data into a web-accessible repository is generally the preferred mechanism for public data sharing because it ensures wide-spread and consistent access to the data. If your discipline already has a trusted repository, we recommend you deposit where your community knows to look. To find a repository, re3data.org is a large, vetted, and searchable catalog of data repositories. If no discipline-specific repository exists, there are several options, including Illinois' IDEALS repository (free) and other general-purpose repositories like DataDryad (fee-based)."
Elsevier	List of supported data repositories
Nature	Data availability policy "Supporting data must be made available to editors and peer-reviewers at the time of submission for the purposes of evaluating the manuscript...For information about suitable public repositories, see sections that follow."
PLOS	PLOS Data Repository Recommendation Guide "PLOS has identified a set of established repositories below, which are recognized and trusted within their respective communities. Additionally, the Registry of Research Data Repositories (Re3Data) is a full scale resource of registered repositories across subject areas. "
DCC	Where can I find a data repository? 1. Funders: Some funders stipulate that data produced in a project they fund is offered to a speific data centre or repository identified in their policy 2. Repository registries: Re3data, Biosharing.org 3. Data Journals: A data journal will not normally host data itself but recommend where it should be deposited, and then link to it. This tends to make them useful sources of advice about repositories. 4. Journal policies: Journals are increasingly requiring authors to deposit the data underlying their articles in a recognised repository, to complement or replace any in-house facility for supplementary materials 5. Learned and professional societies: A society relevant to the research domain may offer advice on data sharing that includes recommendations about where to deposit dat

Problem statement

A researcher looking for a repository has many options, all of which require manual analysis: determine funding agency requirements, identify field/domain recommendations, review publisher recommendations, or search repository registries. The NDS repository recommender will try to provide a single point where users can go to search for an appropriate repository.

There are several existing services in this space including the Registry of Research Data Repositories (RE3Data), Biosharing.org, and the SEAD C3PR service. In addition to these existing registries of research data repositories, funding agencies and publishers provide lists of recommended repositories.

To be useful, the NDS repository recommender must differentiate itself from these existing tools and services. For example

- Improved search over Re3Data through the use of priors (e.g., "trustworthiness" or some sort of impact factor)
- Accounting for user motivations (funding agency requirements, publisher requirements, data size) through guided search
- Suitable for use by publishers (via API or otherwise)

Broad vision

- Start with re3data and biosharing.org data as core
- Develop and test priors based on repository attributes
- In essence, answer the research question: what makes one repository more relevant to users than another?

  Is it really a "recommender"? Broadly speaking, a "recommender system" attempts to predict the relevance of an item to a user based on information known about the user. This could be profile information, previous ratings or related activities. It is more likely that this system will be a "search engine" in the sense that the user comes with an information need and is looking for a ranked list of candidate repositories. The information need might be a query or the dataset itself.

Work plan

1. Use an existing search engine (e.g., Solr/Lucene) to index the re3data
2. Create a test collection of datasets/queries/relevance judgements
    a. This can be done manually (find a set of researchers to give us a dataset and/or query and the repository they seleted)
    b. This can be done automatically by sampling datasets from existing repositories and assume that these are the "most relevant"
3. Develop demonstration UI

The end product will be a search engine that merges the re3data, biosharing (if available), funder and publisher lists along with models of relevance.

Search Engine

We can use either a research-oriented (Indri/Galago/Terrier) or general-purpose (Lucene) search engine platform. The goal would be to identify features /characteristics of repositories that can be used to improve rankings, aside from basic language models.

Potential features:

- Retrieval score based on name, description, subject, information crawled from associated URLs, keywords,
- language, startDate, size
- URL format (e.g. presence of non-standard ports, path depth)
- # results in Google scholar
- How much info in re3data (how complete is the record)?
- Number of policies

Test collection:

A key requirement will be to be able to evaluate the retrieval model, which requires a suitable test collection. For NDSC6, we would just pilot this.

- Find researchers with real datasets and have them identify the top repositories from re3data?
- For some subset of repositories, go find a dataset.

Background/Analysis

What tools already exist in this space?

Registries of Research Data Repositories

Registry	Description	Notes
Re3Data	Registry of research data repositories	Started from Databib, crowd-sourced. Metadata is too general for search; user feedback "precision is horrible"; not based on natural language

Biosharing.org	Registry of databases and policies for life /environmental/bio sciences	Schema based on BioDBCore: http://biocuration.org/community/standards-biodbcore/ Data is not available, but will be. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences
Cinergi	Community Inventory of EarthCube Resources for Geosciences Interoperability	Curated database of geoscience information resources
OpenAIRE	OpenAIRE data provider search	Publishes guidelines for data archives
LA Referencia		
bioCADDIE	Data discovery index	Index of data "do for data what pubmed did for literature"
OpenDOAR	Directory of open-access repositories	
SHARE		Index of research activities/outputs including data management plans, grant proposals, preprints, presentations, and data repository deposits

Approved and Recommended Repositories

Publishers, funding agencies, research/domain organizations(e.g., AGU, ACM), and libraries often provide lists of recommended or supported repositories for depositing research data. The motivations and requirements are often different, but the lists themselves might serve as the basis for our analysis. We can review these (and other) lists to determine the factors in recommending data repositories to researchers.

(This list is not exhaustive – it's likely that many publishers, agencies, and organizations will provide similar lists):

NIH	https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html	Note that the Biosharing database already includes information about whether a repository is recommended by a funding agency: • https://biosharing.org/policy/1947 • https://biosharing.org/policy/1931
Elsevier	Supported Data Repositories Public repositories to store and find data (Data in Brief)	• List of databases with bi-directional linking
Nature	Data policy Recommended Data Repositories Data Policies: Nature Scientific Data	• Includes mandates • Drawn from re3data and biosharing
PLOS	http://blogs.plos.org/everyone/2015/07/02/plos-recommended-data-repositories/ http://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories	
Libraries	https://library.uoregon.edu/datamanagement/sharingdata.html http://www.library.cmu.edu/datapub/dms/respositories	
Other	http://www.ijdc.net/index.php/ijdc/article/viewFile/9.1.152/349 http://www.rdc-drc.ca/wp-content/uploads/Review-of-Research-Data-Repositories-2015.pdf AMS: https://www.ametsoc.org/ams/index.cfm/publications/authors/journal-and-bams-authors/journal-and-bams-authors-guide/data-archiving-and-citation/ AGU: http://publications.agu.org/files/2014/06/Data-Repositories.pdf http://openarchaeologydata.metajnl.com/about/#repo https://www.datacite.org/services/find-repository.html	

SEAD Publication API

See also and the actual

A primary function of the SEAD Publication API (C3PR) is to match or recommend a repository given a research data object based on a set of technical requirements implemented as rules:

- Maximum dataset size
- Purpose
- Organization match
- Acceptable data types (based on mimetypes)
- Minimal required metadata
- Maximum total size
- Maximum collection depth
- Rights-holder requirements

SEAD 2.0 introduces the "publish" workflow. The user selects "publish" and the "live object" is copied to a staging area into a "curation object". The user is able to modify the curation object – adding removing files, metadata, etc. There can be many curation objects for a live object. During the publish workflow, SEAD/Clowder represents the curation object as an ORE MAP and sends a request to the C3PR service. The C3PR service matches the ORE-MAP to available repositories based on a set of rules/criteria. The user is presented with a ranked list of repositories based on a best-match against the ORE. The user can opt to publish to any listed repository.

See also:

- SEAD 2.0 Publication API Walkthrough
- C3PR API server
- Workflow of Operating Your Repository with SEAD
- Plale et al (2013). SEAD Virtual Archive: Building a Federation of Institutional Repositories for Long-Term Data Preservation in Sustainability Science

- Git repository https://github.com/Data-to-Insight-Center/sead2

Other sources of information:

What other sources of information might we include in a recommender service?

- Researcher identifiers, such as ORCID Persistent digital identifier for researchers: these might be helpful in collecting researcher profile information that can be used for recommendation.
- Journal/publication information: We can relate specific journals to data repositories. If the user is publishing in a specific journal, we can recommend where to put the data.
- Abstract: Use text matching techniques to match an abstract to a repository.
- https://www.datacite.org/
- BrownDog: Can we use information from extractors to identify criteria for recommendation?

Harvesting information

- Many of the data repositories are crawl-able or implement standard APIs (OAI-PMH) for harvesting metadata. It might be interesting to consider whether we can harvest descriptive metadata – particularly citation information – and use journal or other publication metadata as part of the recommendation process.

What would make the existing tools better?

- Natural language search
- Ranking basic on different characteristics
    - Does it support my (identifier, metadata, etc)
    - Is it trusted (sustainability/certification). How long is the commitment?
    - Repository "impact factor"
    - Additional value adds (curatorial, linked)
    - Specialized vs geneal

Use cases

Who are the users?

Researchers with data and they don't know where to put it, for various reasons.

User	Situation
No community repository	The researcher is in a community without a repository

Doesn't fit neatly	A researcher is becoming interdisciplinary, moving to a new discpline, or has data they think might be useful for other disciplines
Novice/lazy	New research not aware of existing resources (note, most advice would come from social media, conferences, training)

What are their motivations?

- Responding to request from funding agency. Might need different characteristics (needs DOI, linking etc)
- Has very large data (university can't handle it, domain repos can't handle it)
- Has specific availability requirements (5 years, 10 years)
- Is really complicated (has a lot of contextual information, does the service support it)
- Sharing – not responding to regulatory requirement – just wants to make things available for reuse

Reviewing the above publisher lists and registries, we can identify factors in the recommendation of repositories to researchers:

Factor	Description
Funding agency approval	Funding agencies (e.g. NIH) have lists of approved repositories
Researcher communities	Some repositories restrict to researchers in certain communities
Publisher integration	Publishers (e.g., Elsevier) have arrangements with repositories (e.g., bi-directional linking)
Domain/Field	Repositories are often restricted by domain, with some generalist services
Technical restrictions	Repositories have technical restrictions (e.g., maximum file size, supported formats)
Community mandates	Some research communities have mandated repositories (see Nature list)
Data type	*Does the repository take the data you want to deposit?* Some repositories are restricted to specific types of data. These criteria vary, for example: ○ Protein structures ○ Human or non-human derived ○ Phenotypes Data types are often directly related to domain/field of study.
Metadata format	Some repositories are restricted to specific types of metadata (e.g., MIAME)
Licensing	Free and unrestricted use or public domain (PLOS)
Best practices	Repository adhere's to best practices pertaining to responsible data sharing, digital preservation, citation, and openness (PLOS)

Additional factors from the DCC:

- is a reputable repository available?
- will it take the data you want to deposit?
- will it be safe in legal terms?
- will the repository sustain the data value?
- will it support analysis and track usage?

Publishers, funding agencies, and libraries construct these lists of approved repositories to meet the needs of researchers, Many of these sites now link to centralized services, such as re3data.org. However, re3data.org does not capture all of the information needed to make a recommendation (e.g., C3PR technical restrictions).

Draft Questions for service providers?

1. Do researchers come to you looking for places to put their data?
    a. Of those that come to you, do you have some estimate of the percentage of those that eventually do find a place to put their data?
2. Thinking about the researchers that come to you, what is the typical consultation like? What types of questions or concerns do they have?
3. Do you notice any common challenges or themes across the campus for researchers looking for places to deposit data?
4. What are some of the tools you recommend and how well do they meet the needs of the researcher?
5. Do you have any ideas of tools or services that could help you/them better?
6. We're thinking of this service (describe current vision of recommender), what do you think? Would it be useful?

7. Are there any departments/researchers/labs that you think are representative of this problem that we could talk to? (Looking for most common cases)
8. Is there anyone else working in this space that you think we should talk to?

References

Elsevier. Supported Data Repositories.

Myers, Jim. (2016). SEAD 2.0 Publication API Walkthrough:.

Nature. Availability of data and materials.

PLOS ONE. Data availability.

UI RDS. Saving and Sharing your Data.

Whyte, A. (2015). 'Where to keep research data: DCC checklist for evaluating data repositories' v.1.1 Edinburgh: Digital Curation Centre.