DDI Core Technology Development Team
Outline

- Introduction
- Dataset Indexing Infrastructure
  - Data indexing pipeline
  - ElasticSearch Endpoint
  - Data transformation
- Terminology service
- Search interface
  - Workflow
  - Functionalities
  - Evaluation
- System Demo
Introduction

- Help users find accessible data
  - By disease, data modality, demographics, etc.
  - Engage these users and the data community

- Assist data producers on how to publish data for maximal discoverability
  - Tools to index, submit, visualize

- Build a prototype that docks targeted pilot products
  - RFAs for others to help build DataMed

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
BioCADDIE prototype architecture

Repositories

Metadata

Ingestion

ElasticSearch

Terminology server

User Interface

Supported by the NIH grant #1-U24-AI117966 to UCSD
DDI Core Technology Development Team

Dataset Indexing Infrastructure
Data Indexing Pipeline

1. Configuration file developed by curator
2. Extraction of metadata/data from data resource or dataset via ingestion module
   - Cache information for further processing
3. Process metadata/data via a set of processing modules
   - e.g. ID conversion, keyword extraction, data normalization
4. Mapping of metadata/data to metadata model(s)
5. Export to target endpoint(s) via export modules
6. Search via ElasticSearch APIs

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Data Indexing Pipeline

Current Technologies

1. JSON based documents and services

2. **mongoDB** (Apache 2 license) being used to manage cached dataset description documents

3. Processing pipeline components can take advantage of cloud deployment for scalability

4. Document processing coordinated via messaging queue (Apache ActiveMQ (Apache 2 license))

5. **ElasticSearch** (Apache 2 license) being used as index endpoint
   - Simple cloud deployment and management
   - Sophisticated RESTful API
   - Advanced index customization
   - Full power of lucene and plug-ins

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
MongoDB used for cache of dataset descriptions
- Store “original” information from remote repository
- Allows us to check if information is updated

All cached documents converted to JSON
- Small description stored directly within document
- Larger descriptions stored using GridFS
Search Endpoint

- Distributed, scalable, and highly available search and analytics platform
  - Will detect new or failed nodes, and reorganize and rebalance data automatically, to ensure that your data is safe and accessible.

- A RESTful API, based on JSON, where almost any action can be performed

- Builds distributed capabilities on top of Apache Lucene

- Store complex real world entities as structured JSON documents
  - End-to-end ingestion pipeline (after initial conversion) is JSON

- Sophisticated analytics capabilities through aggregations and aggregation pipelines

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Scalability and Sustainability

Three potential mechanisms for population of the DDI:

1. Standard submission format for ingestion
   • Similar to how journals currently provide data to PubMed.

2. Submission formats for ingestion based on community standards
   • Would mirror how researchers provide data to certain repositories.
   • In this case data that conforms to an existing standard (e.g. MINiML format from GEO)
   • Once a single adaptor for the standard has been written that data can be easily submitted.

3. Repository specific submission format
   • Would need to develop specific adaptors for specific resources.

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Data Transformation

### Transformations in JSON Documents

- **Xpath-like syntax for JSON documents**
- **Ability to perform transformations/operations on values**

```javascript
transform column "$.PDBx:datablock'.PDBx:chem_compCategory'.PDBx:chem_comp['$.'].PDBx:name'.$." to "materialEntity[]\.name";
```

```javascript
apply {{
  arr = re.split("\s*,\s*", value1,)
  arr.extend(re.split("\s*,\s*", value))
  result = arr
}};
```
Data Mapping

map biocaddie-0003["\$.Small Molecule Name"]\$.InChi to "materialEntity[0].InChi";

- Ability to incorporate data from external documents into description
- Support for multiple descriptor files describing a single dataset
- Supports enhancement of dataset description by external data (e.g. information from PubMed article describing the experimental dataset and its results)
ElasticSearch – Structure

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
ElasticSearch – via CURL

curl -XGET 'https://data.biocaddie.org/biocaddie/pdb/_search' -d '{ "query" : { "match" : { "dataItem.keywords" : "GPCR" } },"size":100 }' > ES.out

JSON data visualized via JSON Editor for the Mac
https://search.itunes.apple.com/WebObjects/MZContentLink.woa/wa/link?path=mac%2fJSONEditor

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Current Work with Data Ingest

- Addition of more data repositories
- Working with WG3 - ensure consistency with metadata model
  - Revisions of metadata model based on data currently being ingested
- Addition of enhancement modules
  - Semantic Expansion
  - Metadata Enhancement

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
DDI Core Technology Development Team

Terminology Services
Roles for terminology service

- Indexing
  - Process metadata

- Searching
  - Search expansion
    - Brain tumor
      - Glioblastoma, astrocytoma, ...
      - Frontal lobe tumor, ...
  - Suggestions (autocompletion)
  - Fix typo (did you mean?)
  - Facet management
    - Search entity type: gene, disease, ...

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Terminology services

- Term/Phrase analyzer/normalizer
  - Search and indexing
  - Synonyms/hypo-hypernyms (term mapper)
    - cancer $\rightarrow$ \{carcinoma, sarcoma, ...\}
    - But not carcinoma $\rightarrow$ cancer
  - Build metadata structure

- Term mapper
  - Semantic mappings
  - Other Terminologies

- Term processor
  - Suggestions (autocompletion)
  - Fix typo (did you mean?)

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Terminology

- **Primary**
  - Disease (Condition)
  - Drug (Chemical, Substance, Compound)
  - Omic: Gene - Protein

- **Auxillary**
  - Body Structure
  - Procedure
    - Laboratory
  - Organism

- **Maybe**
  - Physical object

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Sources

- MeSH
- SNOMED CT
- NCBI
- GO
- HGNC
- FMA
- RXNORM
- UBERON

Condition
Chemical
Omic
Procedure
Body Loc/Sys
Organism
...

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Relationships

- Is a, part of
- Synonym
- Sibling

- MeSH $\rightarrow 10^6$
- SNOMED CT $\rightarrow 10^6$
- GO $\rightarrow 2 \times 10^6$
- NCI $\rightarrow 10^5$
- HGNC $\rightarrow \sim 10^5$ (synonyms)

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Infrastructure for terminology service

- SciGraph
  - Neo4j Graph database
- ElasticSearch
  - Term search

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
DDI Core Technology Development Team

Prototype Search Interface Development
MVC Structure

Model

- Responsible for constructing ES query and retrieving search results.

Controller

- Responsible for responding to user input; Instructing the Model to respond to the user input.

View

- User interaction; Responsible for rendering of model.

Generate corresponding ES index, type, search keyword, search fields, facets fields, filter fields. Integrate with terminology server.

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
User Interface workflow

Query Entry → Entity Identification → Expansion
- Synonym
- Hypernym
→ Query Execution
→ ElasticSearch
→ Organize results
→ Facets
→ BioCADDIE backend
→ Presentation
→ Visualization

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
Search Functionalities

Y2 – Q1
Sep. – Nov., 2015

Interface design
- Global statistics
- New interface for prototype v0.2
- Improving interface based on feedback

Searching algorithms
- Finder similar datasets
- Add Boolean search
- Add advanced search
- Add data repositories search

Ranking algorithms
- Refine search results based on user’s selection
- Report from WG 8

Dataset result display
- Sort datasets
- Group metadata
- Accessibility of dataset
- Summarize all returned results (iSEE-DELVE)
- Allow users to select multiple repositories
- Improve faceted browsing

Personalized search
- Search history
- Share search results
- User account
- Save search results

Link dataset to external resources
- PubMed
- Grants (via PubMed?)

Y2 – Q2

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
## Search Functionalities

### Integration of pilot projects
- Integrate PP 1.1 GWAS Finder
- Integrate PP 2.2 iSEE-DELVE for pdb
- Integrate PP 2.1 DataRank for GEO
- Explore gene expression data using PP 2.2
- Integrate PP 3.2 for PDB
- Integrate PP 2.1 for other dataset

### Feedback collection
- Github
- Feedback form

### Documentation
- Source codes on github
- Tutorials

### Usability Study
- UI Analysis
- User Study
- Track user’s actions

### Data duplication problem
- Metadata management
- Architecture/Scalability

### Back up

---

**Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego**

---

### Y2 – Q1
**Sep. – Nov., 2015**

### Y2 – Q2
**Dec. 2015 – Feb., 2016**
# Pilot project integration

<table>
<thead>
<tr>
<th>PP</th>
<th>Title</th>
<th>Integrated As</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Linking publications and underlying datasets using natural language processing</td>
<td>Specialized advanced search for GWAS datasets</td>
</tr>
<tr>
<td>2.1</td>
<td>Data Recommendation Using Machine Learning and Crowdsourcing</td>
<td>Ranking function based on citation metrics for GEO series data</td>
</tr>
</tbody>
</table>
| 2.2 | Intelligent Search Expansion and Visualization of Datasets (iSee-DELVE) | a) iSee similarity metric in ElasticSearch  
b) DELVE implementation as exploratory search and visualization option.  
   (i) for PDB  
   (ii) for gene expression data |
| 3.2 | Development of Citation and Data Access Metrics applied to RCSB Protein Data Bank and related Resources | Ranking function based on citation metrics (dataset mentions) for PDB data |
Evaluation

❖ System-centric evaluation (WG4):
  • Evaluated in the context of pre-defined use cases.
  • Queries and benchmark datasets being developed
  • Evaluation using standard relevance metrics, such as precision, recall, mean average precision (MAP) and the F-measure.

❖ User-centric evaluation:
  • Collect user behavioral data to improve user interface design
It’s all about users!

- User need analysis
  - 13 interviews (face-to-face or by phone)
- User feedback during development
  - Issue reporting at GitHub - https://github.com/biocaddie/prototype_issues
  - Brief survey at the BD2K meeting
- User-centric evaluation in addition to system-centric evaluation
- Formal Usability Study:
  - User Analysis
  - Function Analysis
  - Representation Analysis
  - Task Analysis
DDI Core Technology Development Team

Prototype Demo
Prototype availability

- **Web URL:**
  - datamed.biocaddie.org

- **Login information**
  - User name: biocaddie
  - Password: biocaddie

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
Core Development Roadmap

DDI architecture
- Setup website for searching for datasets
- Set up infrastructure for web portal

Data identifier
- Implement Data identifier into the DDI

Data indexing
- Set up indexing using metadata from WG 3.0

Dataset result display
- Sort datasets
- Group metadata

Terminology server
- Import ontology
- Integrate to Scigraph API
- Integrate autocomplete feature to prototype

Interface design
- New interface for prototype v 0.2
- Global statistics

Usability study
- UI Analysis
- Ranking algorithm
- Results from PP 2.1
- Search function
- Expand the function to 7 repositories
- Find similar datasets
- Search history

Architecture
- Code refactoring

Personalized search
- Share/save search results
- User account

Link dataset to external resources
- PubMed
- Grants

Search algorithm
- Boolean/advanced search
- Data repository search function

Data duplication problem

Metadata management

Wrap up of pilot projects
- PP 1.1 literature/dataset link: Advanced search for domain specific repository
- PP 2.1 Recommender System: Ranking results
- PP 2.2 iSEE/DELVE: Innovative visualization
- PP 3.2 PDB citation pipeline

Usability study
- User study
- Track user’s action

Ranking algorithm
- Refine search results based on user’s selection
- Report from WG 8

Data ingestion
- Determine datasets
- Decide on scalable data/metadata input routes
- Metadata mapping

Feedback collection
- Github

RFA for pilot on Harvester for DDI schema
- RFA announced
- Review, selection and award

Version 0.1
September 2015

Version 0.2
December 2015

Version 0.5
February 2016

Version 1.0
June 2016
Acknowledgement

UCSD
Claudiu Farcas
Jeffrey Grethe
Yueling Li
Larry Lui
Burak Ozyurt
Xiaoqian Jiang
Hyeoneui Kim

Oxford
Alejandra Gonzalez-Beltran
Philippe Rocca-Serra
Susanna-Assunta Sansone

NIH
Ian Fore
Ron Margolis
Dawei Lin
Alison Yao

UTHealth
Xiaoling Chen
Ruiling Liu
Ergin Soysal
Min Jiang
Saeid Pournejati
Mandana Salimi
Vidya Narayana
Anupama Gururaj
Todd Johnson
Trevor Cohen
Cui Tao
Hua Xu

Pilot projects team members
1.1 Hua Xu, Xiao Dong, Yaoyun Zhang
2.1 Xiaoqian Jiang
2.2 Todd Johnson, Hyeoneui Kim, Deevakar Rogith,
3.1 Peter Rose, Chun-nan Hsu

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego