Identifier Services: Tracking Objects and Metadata across Time and Distributed Storage Systems

Maria Esteva, Ramona Walls
DCMI 2016, Copenhagen, Denmark
Panorama

- Bio data collections evolve in a redundant, unstable, distributed research environment
- Role of multiple types of identifiers
- Data is “poorly bound” (Wynholds, 2000)
- Data is big, has multiple components at different stages of completion and is stored across different platforms
- Repositories can be silos
- Pre and post data publication events are difficult to record

Challenges to assign and maintain identifiers
A single identifier type and number does not scale
Identifying everything for study reproducibility
Realistic mandates for data persistence
Each domain may need its own identifier types
Evolving data

Jungle of identifiers
Quality disparity in identifiers
Persistence is not binary
Identifiers metadata is not useful for reuse
Gaps between identification of samples and outputs of research
ID transitions in data lifecycle

Cyberinfrastructure can support identifier tasks
Team members

Biologists, bioinformaticians, data curators, archivists, software developers, and data scientists

Ramona Walls
Weijia Xu
Matt Hanlon
Andrew Magill
Jawon Song
James Carson
Ruizhu Huang
Ming Chen
Research: Identifier Services (IDS)

- Lifecycle identifiers management
- Use-case driven
- Focus on genomics data
- IDS services to:
  - Bind dispersed data objects
  - Track and represent provenance
  - Validate data integrity over time
  - Aid data identity
- Cyberinfrastructure
  - Deal with large data/metadata
  - Deal with evolving data over time
  - Deal with data and metadata in distributed environment
- Evaluation
Service Framework

- Infrastructure based on Agave and national HPC resources
- Web interface to the services
- Data model as anchor to represent relations between data and processes (provenance)
- Data is stored elsewhere
- Registers data from other repositories
- Users select and schedule services
  - Verify data location, integrity, and identity
  - Assign and manage multiple identifiers
- Gathers and integrates evolving events metadata
- Dataset representation
User registers repository(s) data, providing access mechanism(s). User selects files from a registered storage system.

1. User
   - Public Cloud
     - IDS Web
       - identifierservices.org

2. IDS queues corresponding Agave apps.
   - Agave Tenant
     - Agave API
     - Apps API
     - Jobs API
     - Systems API
     - Files API
     - Metadata API

3. Data app pulls data from repository, computes analysis on high performance computing system.

4. IDS updates metadata with data analysis results.
   - Data Repository
   - HPC System
   - User
     - User
     - User
     - User
     - User
Data Model

Generic genomic model

Abstract data model
Use Case: 5 Maize Genotypes

5 maize genotypes: B73, Mo17, CML322, Oh43, Tx303

- Fastq files (count: 5)
- Bam files (count: 5)
- 100 bp tiles (CpG, CHG, CHH) (count: 5)
Interface to IDS

- User creates a project
- Inputs access mechanisms to repositories
Creating a project

- User selects the repository/ies
- Agave acts on its behalf accessing the repository
- User selects the investigation type
  - Generic data model is mapped to the investigation type
  - Yaml file generates a form with the descriptive names for corresponding entities
Adding data to processes

- User selects files
  - Can bulk select
- Add metadata
  - Can be bulk uploaded
- Relations between processes and data are established
- More data and metadata can be added over time
- Existing identifiers are part of the metadata
- Users create a dataset for publication
Dataset representation

Project: Maize 5 genotypes bisulfite sequencing

Name: Maize 5 genotypes bisulfite sequencing
Creator: Jawon Song
Title: Maize 5 genotypes bisulfite sequencing
Investigation Type: Genomic
Description: Whole genome bisulfite sequencing data of 5 Maize Genotypes (B73, Mo17, CML322, Oh43, Tx303)

Project's Related Objects

- Specimen: Oh43 Zea Mays
  - Process: Sequencing
    - Data: SRX731433
    - Data: Oh43_R2.allreads_val_2.fq
    - Data: Oh43_R1.allreads_val_1.fq
    - Data: SRX731434
    - Data: B73_all3_R1_val_1.fq

- Specimen: Tx303 Zea Mays
  - Process: Sequencing
    - Data: Tx303_R2.allreads_val_2.fq
    - Data: Tx303_R1.allreads_val_1.fq
    - Data: SRX731434

- Specimen: B73 Zea Mays
  - Process: Analysis
Integrity checking

- Checks and records data location
- Calculates checksums
- Brings back the information to the user
- Actions can be repeated/scheduled
- Event metadata is recorded over time
Assigning global unique identifiers

- [Diagram showing processes and data flow](https://identifierservices.org/projects)
Data identity

- Content-based comparison algorithm
- Determine the composition of the collections to be compared
- Decide record readers: gff,fasta,fastq
- List of record pairs for comparison
- 3 to 40 Gb file sizes 
  ~19,000,000 to 165,993,413 records to compare
- Compare all the records from the lists using Spark framework
- Report results
Content-based analysis

- Data in different repositories
- Different metadata records
- Content-based analysis
  - Not identical data
- Researcher and curator evaluate results
  - Infer causes for difference
  - Identification decision making
  - Update metadata

- Reports have to be summarized for understandability

<table>
<thead>
<tr>
<th></th>
<th>Records matched in both collections</th>
<th>Identical records in both collections</th>
<th>Different records in both collections</th>
<th>Records found only in A</th>
<th>Records found only in B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>990,278</td>
<td>0</td>
<td>990,278</td>
<td>9,722</td>
<td>0</td>
</tr>
</tbody>
</table>
Evaluation

- System evaluation: performance and functionalities
- Follows use case steps
  - User understands the data modeling process
  - Can register repositories
  - Registers files efficiently
  - Data representation
    - Dataset coming together
IDS cont.

- Manages key aspects of the identified data
- Scales processes
- Repositories or individuals can use the service
- Generalize access to data and metadata
- Big data interfaces
- Results presentation
- Generalize to other domains
- Could be integrated to other services
  - Tracking other publications