OSDF

A cloud enabled system to store, access, and analyze scientific data

Anup Mahurkar
Institute for Genome Sciences
BOSC 2013
Motivation

• Work based on NIH Human Microbiome Project Data Analysis and Coordination Center
• Manage large collections of data sets
• Decorate data sets with rich metadata
• Develop a framework that could be reused for other collaborative or multi-center projects
• Ease of development using a language agnostic API
• Scalable and cloud-enabled
What is OSDF?

• Generic extensible framework for associating data with metadata

• Examples of data
  – Reference Databases, Sequenced Reads, Assemblies, Alignments, Annotation

• Examples of metadata
  – Sequencing platform, Library Preparation, Sequencing Strategy, Assembly method, Alignment method, Alignment reference, Subject information
What is OSDF?

- Includes a mechanism for
  - Defining data model for elements
  - Reliance on Ontologies, and controlled vocabularies
  - Defining relationships between elements
  - Generic RESTful API for accessing and placing data
  - Domain specific API with Perl/Python/Java bindings
  - Versioning and history
  - Access control
Technologies Used

• JavaScript Object Notation (JSON) objects for modeling data elements
  – Lightweight data interchange format
  – Allows sparse data in elements
  – Easy to generate and parse
  – Compact and human readable
  – JSON Schema for validation

• CouchDB for storing JSON objects
  – Document-oriented database
  – Apache project
  – RESTful JSON API out of the box
  – Real-time bi-directional replication
Technologies Used

• ElasticSearch
  – Rapid indexing on all keys and attributes of JSON
  – Query language for complex ad-hoc queries
  – Allows wild card, proximity, range, and Boolean operators

• Metadata modeled by CVs, ontologies, standards, and dictionaries
  – MIGS, MIMS, MIMARKS, GO, Relationship

• API implementation using node.js
  – Scalable server optimized for concurrency
  – Implement JSON validation using JSON Schema

• UI Implementation
  – ExtJS, GraphViz, jQuery, D3
OSDF Architecture

End users/systems

Perl, Python, and so on.

Java

Internet

HTTP

Firewall

HTTP

End users/systems

Perl, Python, and so on.

Java

Internet

HTTP

Firewall

HTTP

OSDF/Node.js

HTTP

Queries

HTTP

Retrievals

ElasticSearch

CouchDB

Runs Elasticsearch, which provides the query and indexing engine used by the OSDF HTTP/REST server. Can run on the same server as the OSDF app, or on a separate machine/VM.

Runs the Node.js code that implements the OSDF API. Exposes configurable port to the internet for HTTP access.

CouchDB document store for data storage, indexing, and versioning of data. Can run on the same server as the OSDF application, or on a separate machine/VM.

Changes API updates Elasticsearch
Schema and Node Example

```json
{
   "type": "object",
   "description": "Sample nodes describe physical samples."
   "properties": {
      "hmp_body_site": {
         "$ref": "body_sites"
      },
      "hmp_supersite": {
         "$ref": "supersites"
      },
      "mimarks_frag": {
         "$ref": "mimarks_frag"
      },
      "mims_frag": {
         "$ref": "mims_frag"
      },
      "visit_number": {
         "type": "integer",
         "minimum": 1,
         "required": true
      },
      "fma_body_site": {
         "title": "Typically a term from the FMA ontology."
         "type": "string",
         "required": true
      },
      "body_product": {
         "type": "string",
         "required": true
      }
   },
   "additionalProperties": false
}
```

```json
{
   "linkage": {
      "part_of": ["c8550ef8d3ea8c9b980650de1c6e86cf"],
      "collected_from": ["c8550ef8d3ea8c9b980650de1c87e0e1"]
   },
   "node_type": "sample",
   "meta": {
      "body_product": "",
      "hmp_supersite": "skin",
      "visit_number": 1,
      "mimarks_frag": {
         "geo_loc_name": "United States of America",
         "samp_mat_process": "",
         "lat_lon": "Unknown",
         "samp_collect_device": "Catch-All sample collection swab",
         "biome": "terrestrial biome [ENVO:00000446]",
         "samp_size": "",
         "feature": "human-associated habitat [ENVO:00009003]",
         "collection_date": "Unknown",
         "env_package": "human-associated",
         "investigation_type": "mimarks-survey",
         "material": "biological product [ENVO:02000043]",
         "rel_to_oxygen": ""
      },
      "fma_body_site": "",
      "hmp_body_site": "left_retroauricular_crease"
   },
   "id": "091f03831014afe9b2da67d698000c51",
   "ver": 1
}
```
Node Editor
**Current HMP DACC Site**

**NIH HUMAN MICROBIOME PROJECT**

**Current News**
- July 2013
  - Human Microbiome Science: Vision for the Future conference to be held in Bethesda, MD July 29-30
- May 2013
  - Human Microbiome Consortium Virtual Meeting: Approaches in Microbiome Assembly
- May 2013
  - Booth at ASM 2013 (#839)

**Publications**
- Colitis-induced Bone Loss is Gender Dependent and Associated with Incr...
- Topographic diversity of fungal and bacterial communities in human skin...
- Comparative metagenomic and rRNA microbial diversity characterization...

**Partner Resources**
- NIH Common Fund
- NCSI HMP Data Repository

---

**HMSCP - Shotgun community profiling**
Reads generated by Illumina wgs sequencing were mapped on to a database of reference genomes in order to calculate organism abundance. For each sample, we provide three files:
- A tab delimited abundance table, indicating depth and breadth of mapping to each reference
- A metrics file, summarizing the number of reads mapped versus the number that aligned to a reference
- Mapping alignment files in bam format

- Data Table
- Protocols and Tools
- Related Pages

---

**HMSCP**

<table>
<thead>
<tr>
<th>Description</th>
<th>Download</th>
<th>Size</th>
<th>MD5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundance Table</td>
<td><img src="#" alt="Download" /></td>
<td>6.4 KB</td>
<td>21f28e1f2b5e2359b025eeb5bd37e4dfe</td>
</tr>
<tr>
<td>BAM File</td>
<td><img src="#" alt="Download" /></td>
<td>7.5 GB</td>
<td>30b05e68b52bfed41bd316e7e3c7f9</td>
</tr>
<tr>
<td>Metrics File</td>
<td><img src="#" alt="Download" /></td>
<td>64.0 bytes</td>
<td>5633a925bfed142fd31fde215fe5924</td>
</tr>
</tbody>
</table>

**HMSCP**

<table>
<thead>
<tr>
<th>Description</th>
<th>Download</th>
<th>Size</th>
<th>MD5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abundance Table</td>
<td><img src="#" alt="Download" /></td>
<td>3.1 KB</td>
<td>428f2c2d9a1354a574bd93501a1d217c</td>
</tr>
<tr>
<td>BAM File</td>
<td><img src="#" alt="Download" /></td>
<td>145.7 MB</td>
<td>7ee19443bd98e5e759e6388df5d45a1</td>
</tr>
<tr>
<td>Metrics File</td>
<td><img src="#" alt="Download" /></td>
<td>182.0 bytes</td>
<td>1f4eb8ecab453e7deaa7290e034d690a</td>
</tr>
</tbody>
</table>
Programmatic Access

• Generic RESTful API
  – Get
  – Put
  – Update
  – Delete

• Domain Specific API
  – Metagenomics Perl API
    • Sequence data R operations
    • Assembly CR operations
    • Compress files
    • Upload files
    • Link nodes
Performance Testing

• Server specifications
  – VM running on VMWare ESX server
  – Ubuntu 12.04
  – 4 vCPUs, 16GB RAM
  – Node.js v0.10
  – Elastic Search v0.90
  – CouchDB v1.2

• Client specifications
  – 8 cores, 6GB RAM

• Benchmark
  – Use ‘ab’ Apache Server Benchmarking Tool
  – Run 5000 operations with increasing concurrency from 20 to 500 operations
  – { "query": { "term": { "node_type": "sample" } } }
Performance Numbers – Node Retrieval

Node retrieval
5000 req/run, 1-0 runs/step
20 to 500 req/sec, step 20

Average

Requests per second

Response time across 1-9 runs (sec)
Future Development

- Deployment in the cloud
- Replication
- Distributed deployment
- Performance enhancements
- Richer microbiome domain API
- Language bindings in Perl/Python/Java
- Host 1000 Genomes Data
Acknowledgements

- Principal Investigator
  - Owen White

- Institute for Genome Sciences
  - Sonia Agrawal
  - Cesar Arze
  - Jonathan Crabtree
  - Noam Davidovics
  - Victor Felix
  - Aaron Miller
  - Mike Schor

- University of Colorado Boulder
  - Rob Knight

- University of Chicago/Argonne National Lab
  - Folker Meyer
  - Narayan Desai
  - Jared Wilkening

Funding provided by:
NHGRI
U01 HG006537
Resources

• OSDF Website
  http://osdf.igs.umaryland.edu

• OSDF Code
  http://sourceforge.net/projects/osdf/

• Metagenome API site
  http://sourceforge.net/projects/metagenosdf/

• DACC Website
  http://hmpdacc.org

We invite you to take a look and participate
Performance Numbers – Node Creation

- Node Creation
  - 5000 rq/run, 1-8 runs/step
  - 20 to 500 rq/sec, step 20

The graph shows the performance numbers for node creation over different request rates per second. The x-axis represents requests per second, while the y-axis represents the number of runs across 1-8 runs/step.
Performance Numbers – Node Query

Simple query (all samples)
1000 req/run, 1 runs/step
20 to 500 req/sec, step 20

Average

Requests per second

Average response across 1 run(s) (ms)