DataONE Member Node Forum
2016-10-13

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Activity Logging

Goal

Track events associated with content in DataONE

- Create
- Read
- Update
- Delete
- Replicate
Activity Logging

Overview

1a) Get Content

1b) READ Event

2a) Get Log Records

2b) Process

2c) Store

3) Get Logs

MN

Client

CN

Log Records
Activity Logging

Challenges

Replicated content hinders complete view
Activity Logging

Solution

Coordinating Nodes collate logs, provide a complete view.
Activity Logging
Basic Log Information

- **entryId**: Record unique ID
- **identifier**: Object of log record
- **ipAddress**: IP Address of requestor (obfuscated during processing)
- **userAgent**: User Agent header identification
- **subject**: Identity of requestor ("public" or specific subject, obfuscated during processing)
- **event**: Type of event – Create, Read, Update Delete, Replicate, Synchronization_failed, Replication_failed
- **dateLogged**: When event occurred
- **nodeIdentifier**: Which node logged the event

[purl.dataone.org/architecture/apis/Types.html#Types.LogEntry](purl.dataone.org/architecture/apis/Types.html#Types.LogEntry)
Activity Logging

Processing Pipeline

Coordinating Nodes collate logs, provide a complete view.
Activity Logging
Processing Pipeline

Raw logs are processed on ingest to Coordinating Nodes

• Object properties added
• IP Address georeferenced
• Filtered for robots
• Obfuscated for privacy
• Stored for efficient retrieval and analysis
Activity Logging

What is a Individual?

- Has the user authenticated?
  - Yes
    - Use authenticated subject
  - No
    - Use IP Address

- Utility of IP Address for user ID highly variable
- Privacy concerns for both IP and Subject
- IP useful for other aspects, e.g. location, distinguishing robots
Activity Logging

What is a Individual?

Determining if events are from an individual

- Counting Online Usage NeTworked Electronic Resources
- Code of practice for recording and reporting of usage statistics
- Consistent, credible, compatible reporting
- Filter out:
  - Robots, known internal processors
  - Short term repeats
  - Unsuccessful or redirect requests

http://www.projectcounter.org/code_practice.html
## Log Entry

### After Processing

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Globally unique id for log entry</td>
</tr>
<tr>
<td>inFullRobotList</td>
<td>Request is from known robot</td>
</tr>
<tr>
<td>inPartialRobotList</td>
<td>Excludes agents such as java and perl</td>
</tr>
<tr>
<td>isRepeatVisit</td>
<td>True if this record appears to be a repeat</td>
</tr>
<tr>
<td>dateAggregated</td>
<td>Date entry was aggregated to logs</td>
</tr>
<tr>
<td>dateUpdated</td>
<td>Last time object properties updated</td>
</tr>
<tr>
<td>isPublic</td>
<td>Indicates if record is publicly available</td>
</tr>
<tr>
<td>readPermission</td>
<td>List of subjects with read permission</td>
</tr>
<tr>
<td>rightsHolder</td>
<td>Subject of object rights holder</td>
</tr>
<tr>
<td>formatId</td>
<td>ID of the object format type</td>
</tr>
<tr>
<td>formatType</td>
<td>DATA, METADATA, or RESOURCE</td>
</tr>
<tr>
<td>size</td>
<td>Size in bytes of the object</td>
</tr>
<tr>
<td>country</td>
<td>Country of event origin</td>
</tr>
<tr>
<td>region</td>
<td>State / province of origin</td>
</tr>
<tr>
<td>city</td>
<td>City of origin</td>
</tr>
<tr>
<td>geohash_1 … 9</td>
<td>Geohash of origin coordinate</td>
</tr>
<tr>
<td>location</td>
<td></td>
</tr>
</tbody>
</table>
Log Reporting
Programmatic Access

- Log access service:
  `cn.dataone.org/cn/v2/query/logsolr`
- Query using Apache Solr syntax
- All log records:
  `logsolr/?q=*:*`
- Log events for last month:
  `q=dateLogged:[NOW-1MONTH TO NOW]`
- Common parameters:
  - `rows` Number of records to return
  - `fl` Fields to show in response
  - `wt` Format of response (xml, csv, json, python)
  - `indent` Format output (useful for debugging)

https://purl.dataone.org/architecture/design/UsageStatistics.html
Log Reporting

Programmatic Access

- Events stats by formatID:
  \[ q=formatId:eml*&facet=on&facet.field=event \]

- Read events for last year grouped by month:
  \[ q=event:read \text{ AND formatType=METADATADATA} \]
  &facet=true&facet.range=dateLogged
  &facet.range.start=NOW-1YEAR
  &facet.range.end=NOW
  &facet.range.gap=+1MONTH

- Bytes downloaded:
  \[ q=event:read\text{ AND stats=true} \text{ AND stats.field=size} \]
  &rows=0&stats.facet=formatId

https://purl.dataone.org/architecture/design/UsageStatistics.html
Log Reporting

Search Interface - Package Stats
Log Reporting
Search Interface - Data Set Stats

David Bloom. 2015. VertNet Portal - Class Mammalia Records. KNB Data Repository. doi:10.5063/F1GQ6VPM.

<table>
<thead>
<tr>
<th>Name</th>
<th>File type</th>
<th>Size</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata: VertNet Portal - Class Mammalia Records</td>
<td>EML v2.1.1</td>
<td>7 KB</td>
<td>441 views</td>
</tr>
<tr>
<td>um:uuid:1d09e64b-d25a-46e7-bdc1-0a91f7b7f8bb</td>
<td>GZIP Format</td>
<td>403 MB</td>
<td>107 downloads</td>
</tr>
</tbody>
</table>

General

<table>
<thead>
<tr>
<th>Title</th>
<th>Identifier</th>
</tr>
</thead>
</table>
Log Reporting
Search Interface - User Stats

David Bloom
Username: ORCID
http://orcid.org/0000-0003-1273-1807

1 year, 6 months Contributor since April 3, 2015
12 contributions 2,201 downloads

Datasets 1 to 5 of 5

1 2 Next

doi:10.5063/F1GQ6VM.

doi:10.5063/F1MG7MDB.

doi:10.5063/F1A49NQB.
Log Reporting

Search Interface - Member Node Stats

Datasets 1 to 5 of 19,758

1 2 3 ... 3,952 Next

PISCO

doi:10.6085/AA/publication_data.40.3.


Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Bruce Menge, and Francis Chan. 2016. PISCO: Intertidal: site temperature data: Manipulation Bay, Oregon, USA (MBYX00). PISCO MN. doi:10.6085/AA/MBYX00_XXXITBPXLCR03_20130708.50.2.

Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Bruce Menge, and Francis Chan. 2016. PISCO: Intertidal: site temperature data: Manipulation Bay, Oregon, USA (MBYX00). PISCO MN. doi:10.6085/AA/MBYX00_XXXITBPXLCR03_20130525.50.2.

Partnership for Interdisciplinary Studies of Coastal Oceans (PISCO), Bruce Menge, and Francis Chan.

4 years, 3 months DataONE Member Node since 2012
39,169 contributions
2,386,582 downloads
Log Reporting

Search Interface - Member Node Stats

Downloads

- 2,227,437 metadata
- 159,145 data

The number of individual metadata and data files downloaded over time. These download counts are partially COUNTER compliant, meaning that downloads from some internet robots and repeat downloads within a certain time window are excluded.

Uploads

- 19,758 metadata
- 19,411 data

The number of individual metadata and data files uploaded over time. Only the first version of each file is counted.
Aggregated Logs
Member Node Reporting

- Event logging benefits all participants
- Member Nodes should implement `getLogRecords()`

| Tier | 1.0, 2.0 | GET /log?[fromDate={fromDate}][&toDate={toDate}][&event={event}][&idFilter={idFilter}][&start={start}][&count={count}] | MNCore.getLogRecords() | (session, [fromDate], [toDate], [event], [idFilter], [start=0], [count=1000]) -> Types.Log |

purl.dataone.org/architecture/apis/MN_APIs.html#MNCore.getLogRecords
Technical Resources

Education
- purl.dataone.org/architecture
- ...architecture/design/UseCases/06_uc.html
- dataone.org/previous-webinars

Communications
- mnforum@dataone.org
- developers@dataone.org
- DataONEorg.slack.com (visit slack.dataone.org)
- irc.ecoinformatics.org #dataone

Sources
- repository.dataone.org/software/cicore
- github.com/DataONEorg

Tools
- search.dataone.org... search-dev... search-sandbox... search-stage...
- cn.dataone.org/cn/v2/diag/subject /sysmeta /object
- Browser, curl, Python, Java libraries
- examples.dataone.org