A data sharing model for decentralized research data management

April 13, 2018

Nassib Nassar
Index Data ApS
nassib@indexdata.com
"Wilhelm Ostwald divided scientists into the classical and the romantic . . . . John R. Platt calls them Apollonian and Dionysian . . . .

Support mostly takes the form of grants, and the present methods of distributing grants unduly favor the Apollonian . . . .

A discovery must be, by definition, at variance with existing knowledge."

—Albert Szent-Györgyi (Science, June 2, 1972)
Science ≠ Workflows

- Science is methodical and orderly, but also instinctive and chaotic.

- Workflows suggest process. Science is not only about process; it is also about innovating, which can involve an unexpected departure from process.

- In building systems, too much focus on workflows will lead to overly rigid models, a bias in favor of centralization, and monolithic systems.

- A better approach is to insist on simple, independent "software tools", which scientists can either use together in expected ways or arrange in new, unforeseen ways.
Managing data: research and libraries

Researchers . . .

would like to focus on doing research

Libraries . . .

would like to offer research data storage and preservation services
Managing data: research and libraries

Researchers . . .

would like to focus on doing research

use files, databases, spreadsheets, etc.

Libraries . . .

would like to offer research data storage and preservation services

use repositories
Managing data: research and libraries

<table>
<thead>
<tr>
<th>Researchers . . .</th>
<th>?</th>
<th>Libraries . . .</th>
</tr>
</thead>
<tbody>
<tr>
<td>would like to focus on doing research</td>
<td></td>
<td>would like to offer research data storage and preservation services</td>
</tr>
<tr>
<td>use files, databases, spreadsheets, etc.</td>
<td></td>
<td>use repositories</td>
</tr>
</tbody>
</table>
Research data management concepts

- Data sharing
- Research workflow
- Data curation
- Data reuse
- Data lifecycle
Focus in on data sharing

Data sharing

Research workflow

Data curation

Data reuse

Data lifecycle
Data sharing: a ubiquitous activity

Levels of data sharing

Data collection
Automated pipeline
Research team
Scholarly communications
Data sharing: a ubiquitous activity

Points of engagement with research activity

- Data collection
- Automated pipeline
- Research team
- Scholarly communications
Data sharing: a ubiquitous activity

- Data collection
- Automated pipeline
- Research team
- Scholarly communications
- Data curation
Data sharing: a ubiquitous activity

- Data collection
- Automated pipeline
- Research team
- Scholarly communications

Data curation
Sharing a data set
Sharing a data set

The receiver needs more information about the data
Glint is software that adds a thin layer of services to data.

Glint

"cell membrane"

Communicate

Curate

Integrate
Glint is software that adds a thin layer of services to data.
Repositories tend to internalize and accumulate features

Glint strives to do one thing well, to be easy to install, and to integrate easily with other software

Integrate with database, analysis, and visualization software, fit into diverse research workflows, and curate to the extent possible when data are created.
The data can reside in a repository, on a lab server, etc.
Using Glint

Web-based user interface: for general users (work in progress)

Command line interface: for technical users & software integrators
Posting data on a Glint server

$ glint post ocean.csv
https://glintcore.net/izzy/ocean
Sharing data & retrieving it in a web browser

https://glintcore.net/izzy/ocean

<table>
<thead>
<tr>
<th>id</th>
<th>t</th>
<th>record</th>
<th>site_id</th>
<th>air_temp_avg</th>
<th>baro_press_avg</th>
<th>rel_hum_avg</th>
<th>dew_pt_avg</th>
<th>vpr_</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2016-12-19 17:04:00</td>
<td>8109</td>
<td>1</td>
<td>792.5</td>
<td>171.399993896484</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2016-12-19 17:34:00</td>
<td>8110</td>
<td>1</td>
<td>789</td>
<td>163.699996948242</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2016-12-19 18:04:00</td>
<td>8111</td>
<td>1</td>
<td>790.400024414062</td>
<td>169.699996948242</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2016-12-19 18:34:00</td>
<td>8112</td>
<td>1</td>
<td>12.6400003433228</td>
<td>1012</td>
<td>92.699996948242</td>
<td>11.5</td>
<td>1.3555</td>
</tr>
<tr>
<td>5</td>
<td>2016-12-19 19:04:00</td>
<td>8113</td>
<td>1</td>
<td>13.2600002288818</td>
<td>1011</td>
<td>92.5</td>
<td>12.079999237061</td>
<td>1.4075</td>
</tr>
</tbody>
</table>
Retrieving data in R

```r
> ocean <- read.csv("https://glintcore.net/izzy/ocean")

> ocean

<table>
<thead>
<tr>
<th>id</th>
<th>t</th>
<th>record</th>
<th>site_id</th>
<th>air_temp_avg</th>
<th>baro_press_avg</th>
<th>rel_hum_avg</th>
<th>dew_pt_avg</th>
<th>vpr_press_avg</th>
<th>wind_speed</th>
<th>wind_dir</th>
<th>stdev</th>
<th>wind_gust</th>
<th>wtr_lvl_avgreal</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2016-12-19</td>
<td>17:04:00</td>
<td>8109</td>
<td>NA</td>
<td>792.5</td>
<td>171.4</td>
<td>NA</td>
<td>NA</td>
<td>0.443</td>
<td>26.72</td>
<td>0.048</td>
<td>0.443</td>
<td>1.238093</td>
</tr>
<tr>
<td>2</td>
<td>2016-12-19</td>
<td>17:34:00</td>
<td>8110</td>
<td>1</td>
<td>789.0</td>
<td>163.7</td>
<td>NA</td>
<td>NA</td>
<td>0.443</td>
<td>26.72</td>
<td>0.048</td>
<td>0.443</td>
<td>1.237691</td>
</tr>
<tr>
<td>3</td>
<td>2016-12-19</td>
<td>18:04:00</td>
<td>8111</td>
<td>1</td>
<td>790.4</td>
<td>169.7</td>
<td>NA</td>
<td>NA</td>
<td>0.000</td>
<td>0.00</td>
<td>0.000</td>
<td>0.000</td>
<td>1.238556</td>
</tr>
<tr>
<td>4</td>
<td>2016-12-19</td>
<td>18:34:00</td>
<td>8112</td>
<td>1</td>
<td>12.64</td>
<td>1012.0</td>
<td>92.7</td>
<td>11.50</td>
<td>1.355</td>
<td>0.00</td>
<td>0.000</td>
<td>0.000</td>
<td>1.237252</td>
</tr>
<tr>
<td>5</td>
<td>2016-12-19</td>
<td>19:04:00</td>
<td>8113</td>
<td>1</td>
<td>13.26</td>
<td>1011.0</td>
<td>92.5</td>
<td>12.08</td>
<td>1.408</td>
<td>0.00</td>
<td>0.000</td>
<td>0.000</td>
<td>1.236872</td>
</tr>
</tbody>
</table>
```
Changing how data are retrieved

https://glintcore.net/izzy/ocean?show(t,air_temp_avg) as (tsv)

<table>
<thead>
<tr>
<th>id</th>
<th>t</th>
<th>record</th>
<th>site_id</th>
<th>air_temp_avg</th>
<th>baro_press_avg</th>
<th>rel_hum_avg</th>
<th>dew_pt_avg</th>
<th>vpr_</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2016-12-19 17:04:00</td>
<td>8109</td>
<td>1</td>
<td></td>
<td>792.5</td>
<td>171.399993896484</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2016-12-19 17:34:00</td>
<td>8110</td>
<td>1</td>
<td></td>
<td>789</td>
<td>163.699996948242</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2016-12-19 18:04:00</td>
<td>8111</td>
<td>1</td>
<td></td>
<td>790.400024414062</td>
<td>169.69996948242</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2016-12-19 18:34:00</td>
<td>8112</td>
<td>1</td>
<td>12.6400003433228</td>
<td>1012</td>
<td>92.6999969482422</td>
<td>11.5</td>
<td>1.3550</td>
</tr>
<tr>
<td>5</td>
<td>2016-12-19 19:04:00</td>
<td>8113</td>
<td>1</td>
<td>13.2600002288818</td>
<td>1011</td>
<td>92.5</td>
<td>12.0799999237061</td>
<td>1.4075</td>
</tr>
</tbody>
</table>
Adding metadata

$ glint md ocean.t dc:date
$ glint md ocean.wind_speed yamz:h3846
Integrating data with services (1)

This plotting service can plot any time series data containing a dc:date or yamz:h1317 column.

Data set URL: 

Submit
Integrating data with services (2)

https://glintcore.net/izzy/ocean?show(t,air_temp_avg,wind_speed)

https://glintcore.net/izzy/ocean?show(t,air_temp_avg,wind_speed)md()

t{dc:date},air_temp_avg,wind_speed{yamz:h3846}
Integrating data with services (3)

t{dc:date},air_temp_avg,wind_speed

![Graph showing air temperature and wind speed over time]
"The generation of most biomedical data is highly distributed and is accomplished mainly by individual scientists or relatively small groups of researchers. Moreover, data also exist in a wide variety of formats, which complicates the ability of researchers to find and use biomedical research data generated by others and creates the need for extensive data 'cleaning.' According to a 2016 survey, data scientists across a wide array of fields said they spend most of their work time (about 80 percent) doing what they least like to do: collecting existing data sets and organizing data. That leaves less than 20 percent of their time for creative tasks like mining data for patterns that lead to new research discoveries."

—Draft NIH Strategic Plan for Data Science (2018)
"The value of research data arises from its use, and the more it is used the greater the social benefits and the higher net welfare."

—Business models for sustainable research data repositories (OECD report, Dec. 6, 2017)
Effective data sharing can accelerate cooperation around data.

Suppose that we could share and cooperate around data—forming communities to discuss and understand data better—as easily as we share and discuss interesting articles today.

https://glintcore.net