How to best manage your data to make the most of it for your research?

How to ensure that open data works for research?

Preliminary question:

Data Capture?

The main focus is on building framework and solution to store data

Daniel Jacob

INRA UMR 1332 BFP – Metabolism Group

Bordeaux Metabolomics Facility
THE FAIR DATA PRINCIPLES

The "FAIR" principles define the basis for data sharing easily to find, accessible, interoperable and reusable.

However, it is to the communities to specify the actions necessary for their implementation.

The implementation of FAIR principles is a process that must be thought of in a progressive and community-oriented way.

It must be integrated into existing practices to ensure that they evolve without interruption and in a way that is acceptable to the various actors.

The role of a data authority is to translate the principles into standards in agreement with the concerned communities.
Data flow

Concerned data

Raw Data

Processed data

Analyzed data

Published data

Specialized Data Repositories

Processed data is the raw data processed in a way so that they can highlight some features, i.e. some type of variables linked to the focus of the study. Annotation, Curation, Validation

Partly not automatically reproducible because it requires human expertise

The "data lifecycle" must therefore be integrated into the scientific research process

Data Tables

Know-how

Data capture ... as far upstream as possible

Studies with a design of experiment (DoE)

Information (often partial)

Know-how

Open access

Expected objectives
During a research project

Know-how

Data mining / Modeling
- Repetition of multiple scenarios on different subsets of data
- Selection subsets of data

Data exploration ⇒ Descriptive statistics
First glimpse of the data that can show trends. Allow the data to be well characterized, which is necessary to then choose how to analyze them.

⇒ Implying lots of data manipulation
- data capturing, data formatting, data linking, data import / data export
- Linking both metadata and data for data mining

Knowledge

Data flow

Data processing

Data / Metadata
- Data/Metadata Exploration
  - Data Visualization
- Data Mining
  - Find out “biomarkers”
- Modeling
  - Explain data

DATA
A data management plan or DMP is a formal document that outlines how data are to be handled both during a research project, and after the project is completed.

The goal of a data management plan is to consider the many aspects of data management, metadata generation, data preservation, and analysis before the project begins. This ensures that data are well-managed in the present, and prepared for preservation in the future.
Data flow

Before the project begins

During a research project

After the project is completed

Make consistent the two axes:

Motivations

⇒ Reduce data manipulation
⇒ Data sharing & data availability
⇒ Facilitate the subsequent data mining
⇒ Facilitate the data dissemination

How?

The "data life" must therefore be integrated into the scientific research process

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
Use-Case

Experiment Data Tables

**Experiment Design**

- **seeding**
- **harvesting**
- **samples preparation**
- **samples analysis**

**Sample identifiers**

**Experiment Data Tables**

- **plants.tsv**
- **harvests.tsv**
- **samples.tsv**
- **compounds.tsv**
- **enzymes.tsv**

---

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
Research question ⇒ Project ⇒ Experiment ⇒ Experimental set-up

Before the project begins

- Project building
- Then experiment design

Experiment's Mapping

During a research project

- Several operators, technics, data types, SOPs, ...

Each time we plan to share data coming from a common experimental design, the classical challenges for fast using data by every partner are data storage and data access.

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
The experimental context: recurrent needs / wishes

Experiment Design

1. data sharing & data availability
2. identifiers centrally managed
3. facilitate the subsequent publication of data
4. avoid the implementation of a RDBDMS
**Purposes**

**Objectives**
- make research **data locally or broadly accessible all along the project**
- allow any (data) scientists to be able to **explore the dataset** and then **extract a subpart or the totality of the data** according to their needs
- allow data to be selected then, **downloadable by web API**
- allow data and analysis to be **visualized online**

**Guideline keywords**
- simplicity, flexibility, efficiency
### The experimental context:

- **Needs/Wishes**
  - Seeding
  - Harvesting
  - Sample preparation
  - Sample analysis

### Experiment Design (DoE)

- Sample identifiers

### ODAM Framework

- Open Data for Access and Mining

### EDTMS

### Use-Case

<table>
<thead>
<tr>
<th>SampleID</th>
<th>Treatment</th>
<th>DevStage</th>
<th>FruitAge</th>
<th>FruitPosition</th>
<th>FruitDiamete</th>
<th>FruitHeight</th>
<th>FruitFW</th>
<th>Rank</th>
<th>Truss</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>Control</td>
<td>FF.01</td>
<td>07DPA</td>
<td>10.89</td>
<td>41.26</td>
<td>23.67</td>
<td>0.81 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>Control</td>
<td>FF.03</td>
<td>07DPA</td>
<td>11.51</td>
<td>42.32</td>
<td>24.12</td>
<td>0.88 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>Control</td>
<td>FF.04</td>
<td>07DPA</td>
<td>12.12</td>
<td>43.56</td>
<td>25.27</td>
<td>0.92 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>353</td>
<td>Control</td>
<td>FF.05</td>
<td>07DPA</td>
<td>12.78</td>
<td>44.78</td>
<td>26.38</td>
<td>0.95 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>355</td>
<td>Control</td>
<td>FF.06</td>
<td>07DPA</td>
<td>13.33</td>
<td>45.99</td>
<td>27.49</td>
<td>0.98 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>413</td>
<td>Control</td>
<td>FF.07</td>
<td>07DPA</td>
<td>13.89</td>
<td>47.11</td>
<td>28.60</td>
<td>1.01 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>Control</td>
<td>FF.08</td>
<td>07DPA</td>
<td>14.44</td>
<td>48.22</td>
<td>29.71</td>
<td>1.04 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>117</td>
<td>Control</td>
<td>FF.09</td>
<td>07DPA</td>
<td>15.99</td>
<td>49.33</td>
<td>30.82</td>
<td>1.07 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>536</td>
<td>Control</td>
<td>FF.10</td>
<td>07DPA</td>
<td>16.55</td>
<td>50.44</td>
<td>31.93</td>
<td>1.10 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>544</td>
<td>Control</td>
<td>FF.11</td>
<td>07DPA</td>
<td>17.11</td>
<td>51.55</td>
<td>33.04</td>
<td>1.13 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>158</td>
<td>Control</td>
<td>FF.12</td>
<td>07DPA</td>
<td>17.66</td>
<td>52.66</td>
<td>34.15</td>
<td>1.16 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>109</td>
<td>Control</td>
<td>FF.13</td>
<td>07DPA</td>
<td>18.22</td>
<td>53.77</td>
<td>35.26</td>
<td>1.19 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>134</td>
<td>Control</td>
<td>FF.14</td>
<td>07DPA</td>
<td>18.77</td>
<td>54.88</td>
<td>36.37</td>
<td>1.22 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>Control</td>
<td>FF.15</td>
<td>07DPA</td>
<td>19.33</td>
<td>55.99</td>
<td>37.48</td>
<td>1.25 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>179</td>
<td>Control</td>
<td>FF.16</td>
<td>07DPA</td>
<td>19.89</td>
<td>57.11</td>
<td>38.59</td>
<td>1.28 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>383</td>
<td>Control</td>
<td>FF.17</td>
<td>07DPA</td>
<td>20.44</td>
<td>58.22</td>
<td>39.70</td>
<td>1.31 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>425</td>
<td>Control</td>
<td>FF.18</td>
<td>07DPA</td>
<td>21.00</td>
<td>59.33</td>
<td>40.81</td>
<td>1.34 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>520</td>
<td>Control</td>
<td>FF.19</td>
<td>07DPA</td>
<td>21.55</td>
<td>60.44</td>
<td>41.92</td>
<td>1.37 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>419</td>
<td>Control</td>
<td>FF.20</td>
<td>07DPA</td>
<td>22.11</td>
<td>61.55</td>
<td>43.03</td>
<td>1.40 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>138</td>
<td>Control</td>
<td>FF.21</td>
<td>07DPA</td>
<td>22.66</td>
<td>62.66</td>
<td>44.14</td>
<td>1.43 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>143</td>
<td>Control</td>
<td>FF.22</td>
<td>07DPA</td>
<td>23.22</td>
<td>63.77</td>
<td>45.25</td>
<td>1.46 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>365</td>
<td>Control</td>
<td>FF.23</td>
<td>07DPA</td>
<td>23.77</td>
<td>64.88</td>
<td>46.36</td>
<td>1.49 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>Control</td>
<td>FF.24</td>
<td>07DPA</td>
<td>24.33</td>
<td>65.99</td>
<td>47.47</td>
<td>1.52 A</td>
<td>T7</td>
<td></td>
</tr>
<tr>
<td>188</td>
<td>Control</td>
<td>FF.25</td>
<td>07DPA</td>
<td>24.89</td>
<td>67.11</td>
<td>48.58</td>
<td>1.55 A</td>
<td>T7</td>
<td></td>
</tr>
</tbody>
</table>
Use-Case

Experiment Design (DoE)

Sample identifiers

samples analysis

ODAM Framework
Open Data for Access and Mining

EDTMS

aliquots.txt

Table:

<table>
<thead>
<tr>
<th>SampleID</th>
<th>AliquotID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>31</td>
<td>101</td>
</tr>
<tr>
<td>31</td>
<td>102</td>
</tr>
<tr>
<td>31</td>
<td>103</td>
</tr>
<tr>
<td>31</td>
<td>104</td>
</tr>
<tr>
<td>32</td>
<td>111</td>
</tr>
<tr>
<td>32</td>
<td>112</td>
</tr>
<tr>
<td>32</td>
<td>113</td>
</tr>
<tr>
<td>33</td>
<td>114</td>
</tr>
<tr>
<td>34</td>
<td>115</td>
</tr>
<tr>
<td>34</td>
<td>116</td>
</tr>
<tr>
<td>35</td>
<td>117</td>
</tr>
<tr>
<td>35</td>
<td>118</td>
</tr>
<tr>
<td>36</td>
<td>119</td>
</tr>
<tr>
<td>36</td>
<td>120</td>
</tr>
<tr>
<td>97</td>
<td>301</td>
</tr>
<tr>
<td>97</td>
<td>302</td>
</tr>
</tbody>
</table>

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
The experimental context:

- seeding
- harvesting
- samples preparation
- samples analysis
- sample identifiers

Experiment Design (DoE):

- EDTMS
- ODAM Framework
  - ODAFAM Framework
  - aliquots.txt

Use-Case:

- Bibliotheques
- DATA
  - DATA (C)
  - bxflow
  - eliquots.txt
  - eliquotsattributes.tsv
  - eminocids.txt
  - nmr_Annotations.txt
  - nmrAnnot
  - nmrspec
  - compounds.txt
  - enzymes.txt
  - FRIMI_Quantities.txt
  - infs.csv
  - lipids_prot.txt
  - plante_Hexoses_P.txt
  - pools.txt
  - qnmrool.txt
  - qnmr_metao.txt
  - qpmr_pools.txt
  - s_subsets.txt
  - samples.txt

ODAM Framework:

- Open Data for Access and Mining
- EDTMS
- aliquots.txt

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
Data capture... **as far upstream as possible**

Take into account users **operating methods and work habits**

**Spreadsheet as a central tool**

Despite all their drawbacks
- e.g. multiple information in a format without internal structure
- This does not take away any of their benefits
- **Universal tool**

**But: Repetitive and tedious tasks**

**Gathering Data and Preparing Data:**
- lot of data manipulation, mainly in the form of table,
- combine data sets according to a common field (identifiers)

**Modelisation:**
- selection of subset of data then **many repetitions of complex processing operations**
- according to a very varied parametrization (**scenarios**).

Allow users to gain efficiency **where they would like to gain efficiency**

**Handling of all these tasks related to data management**

Provide services

Promote good practices
Promote good practices

samples: Sample features

Experiment Design (DoE)

Samples analysis

Sample identifiers

The experimental context:
- needs/wishes
- seeding
- harvesting
- samples preparation
- samples analysis

Data

Promote non-proprietary format like CSV or TSV

Data: well organized data
- Each variable forms a column
- Each observation forms a line
- Each type of "unit observational" forms a table

necessary and indispensable step towards « Linked Open Data ».

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
**Promote good practices**

**Samples : Sample features**

**Experiment Design (DoE)**

<table>
<thead>
<tr>
<th>SampleID</th>
<th>Treatment</th>
<th>DevStage</th>
<th>FruitAge</th>
<th>FruitPosition</th>
<th>FruitDiameter</th>
<th>FruitHeight</th>
<th>FruitFW</th>
<th>Rank</th>
<th>Truss</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>Control</td>
<td>FF.01</td>
<td>07DPA</td>
<td>3</td>
<td>11.95</td>
<td>10.42</td>
<td>0.81</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>121</td>
<td>Control</td>
<td>FF.03</td>
<td>22DPA</td>
<td>3</td>
<td>36.13</td>
<td>31.77</td>
<td>23.4</td>
<td>A</td>
<td>T6</td>
</tr>
<tr>
<td>164</td>
<td>Control</td>
<td>FR.03</td>
<td>42DPA</td>
<td>2</td>
<td>51.09</td>
<td>46.85</td>
<td>64.03</td>
<td>A</td>
<td>T5</td>
</tr>
<tr>
<td>790</td>
<td>Control</td>
<td>FR.02</td>
<td>55DPA</td>
<td>5</td>
<td>48.28</td>
<td>43.35</td>
<td>66.63</td>
<td>A</td>
<td>T5</td>
</tr>
<tr>
<td>564</td>
<td>Control</td>
<td>FF.02</td>
<td>47DPA</td>
<td>3</td>
<td>49.84</td>
<td>44.93</td>
<td>69.93</td>
<td>A</td>
<td>T5</td>
</tr>
<tr>
<td>512</td>
<td>Control</td>
<td>FF.03</td>
<td>21DPA</td>
<td>1</td>
<td>60.48</td>
<td>54.23</td>
<td>106.1</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>117</td>
<td>Control</td>
<td>FF.03</td>
<td>21DPA</td>
<td>NA</td>
<td>41</td>
<td>35.82</td>
<td>37.22</td>
<td>A</td>
<td>TA</td>
</tr>
<tr>
<td>536</td>
<td>Control</td>
<td>FF.03</td>
<td>21DPA</td>
<td>NA</td>
<td>13.44</td>
<td>12.39</td>
<td>1.14</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>31</td>
<td>Control</td>
<td>FF.01</td>
<td>08DPA</td>
<td>NA</td>
<td>59.4</td>
<td>49.05</td>
<td>87.28</td>
<td>A</td>
<td>TA</td>
</tr>
<tr>
<td>179</td>
<td>Control</td>
<td>FF.03</td>
<td>28DPA</td>
<td>NA</td>
<td>57.31</td>
<td>&lt;7.31</td>
<td>92.98</td>
<td>A</td>
<td>TA</td>
</tr>
<tr>
<td>383</td>
<td>Control</td>
<td>FF.04</td>
<td>34DPA</td>
<td>2</td>
<td>47.04</td>
<td>41.19</td>
<td>48.96</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>425</td>
<td>Control</td>
<td>FR.04</td>
<td>55DPA</td>
<td>5</td>
<td>62.74</td>
<td>52.07</td>
<td>115.2</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>520</td>
<td>Control</td>
<td>FF.03</td>
<td>30DPA</td>
<td>4</td>
<td>NA</td>
<td>NA</td>
<td>9.88</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>419</td>
<td>Control</td>
<td>FF.03</td>
<td>30DPA</td>
<td>3</td>
<td>53.68</td>
<td>45.43</td>
<td>60.34</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>138</td>
<td>Control</td>
<td>FF.02</td>
<td>15DPA</td>
<td>5</td>
<td>47.96</td>
<td>22.14</td>
<td>9.69</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>141</td>
<td>Control</td>
<td>FF.03</td>
<td>29DPA</td>
<td>6</td>
<td>27.96</td>
<td>22.14</td>
<td>9.69</td>
<td>A</td>
<td>T7</td>
</tr>
<tr>
<td>365</td>
<td>Control</td>
<td>FF.02</td>
<td>47DPA</td>
<td>4</td>
<td>48.45</td>
<td>42.92</td>
<td>51.33</td>
<td>A</td>
<td>T6</td>
</tr>
<tr>
<td>137</td>
<td>Control</td>
<td>FF.03</td>
<td>27DPA</td>
<td>5</td>
<td>55.11</td>
<td>49.14</td>
<td>71.82</td>
<td>A</td>
<td>T6</td>
</tr>
<tr>
<td>186</td>
<td>Control</td>
<td>FR.01</td>
<td>42DPA</td>
<td>3</td>
<td>45.71</td>
<td>43.28</td>
<td>47.9</td>
<td>A</td>
<td>T5</td>
</tr>
<tr>
<td>100</td>
<td>Control</td>
<td>FR.01</td>
<td>42DPA</td>
<td>3</td>
<td>55.38</td>
<td>47.1</td>
<td>77.5</td>
<td>A</td>
<td>T6</td>
</tr>
</tbody>
</table>

**Promote non-proprietary format like CSV or TSV**

Whatever the kind of experiment, this assumes a design of experiment (DoE) involving individuals, samples or whatever things, as the main objects of study (e.g. plants, tissues, bacteria, ...)

This also assumes the observation of dependent variables resulting of effects of some controlled experiment factors.

Moreover, the objects of study have usually an identifier for each of them, and the variables can be quantitative or qualitative.

Well organized data
- Each variable forms a column
- Each observation forms a line
- Each type of "unit observational" forms a table
Promote good practices

samples: Sample features

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Factors</th>
<th>Qualitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>SampleID</td>
<td>Treatment</td>
<td>DevStage</td>
</tr>
<tr>
<td>FruitAge</td>
<td>FruitDiameter</td>
<td>FruitHeight</td>
</tr>
<tr>
<td>FruitFW</td>
<td>Rank</td>
<td>Truss</td>
</tr>
</tbody>
</table>

Description of the different columns within data files

Metadata: not just on the "top" linked to datasets but more deeply linked to the variables.

Promote non-proprietary format like CSV or TSV
Promote good practices       Minimal but relevant Metadata

Metadata Files

• Our approach, by adjoining some minimal but relevant metadata, gives access to the data themselves with the possibility to explore and mine them.

For that, 2 definition files for associate metadata are required

• **s_subsets.csv**: a definition file allowing each data subset file to be associated with a concept

• **a_attributes.csv**: a definition file allowing each attribute (concept/variable) to be annotated with some minimal but relevant metadata
**metadata Files**

Metadata file allowing to associate a key concept to each **data subset file**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>rank</td>
<td>obtainedFrom</td>
<td>subset</td>
<td>identifier</td>
<td>file</td>
<td>description</td>
<td>CV_term_id</td>
<td>CV_term_name</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>plants</td>
<td>PlantID</td>
<td>plants.tsv</td>
<td>Plant features</td>
<td><a href="http://purl.obolibrary.org/obo/PD:0000003">http://purl.obolibrary.org/obo/PD:0000003</a></td>
<td>whole plant</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>harvests</td>
<td>Lot</td>
<td>harvests.tsv</td>
<td>Harvest features</td>
<td><a href="http://purl.obolibrary.org/obo/PD:0000161">http://purl.obolibrary.org/obo/PD:0000161</a></td>
<td>tissue harvesting</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>samples</td>
<td>SampleID</td>
<td>samples.tsv</td>
<td>Samples features</td>
<td><a href="http://purl.obolibrary.org/obo/PD:0000162">http://purl.obolibrary.org/obo/PD:0000162</a></td>
<td>fruit</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>compounds</td>
<td>SampleID</td>
<td>compounds.tsv</td>
<td>Compound quantifications</td>
<td><a href="http://purl.obolibrary.org/obo/PD:0000163">http://purl.obolibrary.org/obo/PD:0000163</a></td>
<td>chemical entity</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>enzymes</td>
<td>SampleID</td>
<td>enzymes.tsv</td>
<td>Enzyme Features</td>
<td><a href="http://purl.obolibrary.org/obo/PD:0000164">http://purl.obolibrary.org/obo/PD:0000164</a></td>
<td>enzyme</td>
</tr>
</tbody>
</table>

---

"Is Obtained From"
Promote good practices

Minimal but relevant Metadata

Metadata Files

s_subsets.tsv

a_attributes.tsv

Entity ⇔ unit observational (e.g. samples, compounds, …)

Attribute ⇔ Variable, Feature, ... (e.g. Plants, Fruits, Glucose, Rank, ….)

Additional data subsets can be added step by step, as soon as data are produced.
Data Emancipation

Data

Tools

ODAM framework allows experimental data tables to be widely accessible and fully reusable including through a scripting language such as R, and this with minimal effort on the part of the data provider.

The approach consists in building a web-based data network, based on appropriate technologies (Web API), and using standard data formats (TSV, JSON).

Web applications, each with a clearly defined objective, then operate this network.

A data can therefore be used for several applications and vice versa. The data management system becomes completely independent of its operation.

The data is thus “decompartmentalized”, a sine qua non condition for the Web of Data.

With the help of the two metadata files

⇒ Data emancipation regarding Tools

Data ⇔ API ⇔ Tools

Multiscale deployment
Local / Intranet / Internet
Merely dropping *data files* in a *data repository* (e.g. a local disk, NAS or distant storage space) should allow users to **access them by web API**

No database schema, no programming code are required
Additional data subsets can be added step by step, as soon as data are produced.
Infrastructure: PC Desktop, Laptop
Host OS: MS Windows 10, Mac OS
Hypervisor: Virtual Box - https://www.virtualbox.org/

ODAM
Open Data for Access and Mining

Multiscale deployment
Local / Intranet / Internet

See the ODAM installation Guide
http://pmb-bordeaux.fr/odam/

Data can be downloaded, explored and mined

At the first stage, set up your data management within your own workspace
Make both metadata and data available for data mining

ODAM
Open Data for Access and Mining
http://pmb-bordeaux.fr/odamsw/

REST Services: hierarchical tree of resource naming (URL)

With the help of the two metadata files

Data can be explored and mined based on some minimal but relevant metadata
Example based on FRIM - Fruit Integrative Modelling

- Identifiers
- Merging & selection of data subsets
- Avoids lots of data manipulation
- Facilitates linking both metadata and data for data mining

http://pmb-bordeaux.fr/getdata/.tsv/frim1/(activome,qNMR_metabo)/sample/3652?limit=10

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
Data Emancipation

With the help of the two metadata files

⇒ Data emancipation regarding Tools
Data ↔ API ↔ Tools

ODAM

Data Explorer

Multiscale deployment
Local / Intranet / Internet

Develop if needed, lightweight tools
- R scripts (Galaxy), lightweight GUI (R shiny)

https://shiny.rstudio.com/
https://plot.ly/dash/

https://biostatflow.org/
Visual data exploration
a first key step for deeper analyses

Example online
https://pmb-bordeaux.fr/dataexplorer/?ds=frim

ODAM - Data Explorer

FRIM

Data Information

Tomato

Culture location
- INVENIO, Sia Livrade, France

ERASysBio++ FRIM Project

Fruit Integrative Modelling
The project aimed to build a virtual tomato fruit that enables the prediction of metabolite levels given genetic and environmental inputs, by an iterative process between laboratories which combine expertise in fruit biology, ecophysiology, theoretical and experimental biochemistry, and biotechnology.

INRA UMR 1332 BFP - Metabolism Team - Yves Gibon - 2017

R shiny
With the help of the two metadata files

ODAM - Data Explorer

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Description</th>
<th>Identifier</th>
<th>WSEntry</th>
<th>CV_Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>samples</td>
<td>Sample features</td>
<td>SampleID</td>
<td>sample</td>
<td>[DB_1110046] organ harvesting</td>
</tr>
<tr>
<td>cellwall_metabo</td>
<td>Cell wall Compound quantifications</td>
<td>aliquotsID</td>
<td>aliquots</td>
<td>[CHEBI_24431] chemical entity</td>
</tr>
<tr>
<td>cellwall_metabFW</td>
<td>Cell Wall Compound quantifications (FW)</td>
<td>aliquotsID</td>
<td>aliquots</td>
<td>[CHEBI_24431] chemical entity</td>
</tr>
<tr>
<td>activome</td>
<td>Actvome Features</td>
<td>aliquoteID</td>
<td>aliquote</td>
<td>[CHEBI_24431] chemical entity</td>
</tr>
<tr>
<td>qMS_metabo</td>
<td>MS Compounds quantification</td>
<td>poolID</td>
<td>pool</td>
<td>[DB_1110046] organ harvesting</td>
</tr>
<tr>
<td>qNMR_metabo</td>
<td>NMR Compounds quantifiation</td>
<td>poolID</td>
<td>pool</td>
<td>[DB_1110046] organ harvesting</td>
</tr>
<tr>
<td>plato_hexosesP</td>
<td>Hexoses Phosphate</td>
<td>aliquoteID</td>
<td>aliquote</td>
<td>[CHEBI_24431] chemical entity</td>
</tr>
<tr>
<td>lipids_A3</td>
<td>Lipids A3</td>
<td>aliquoteID</td>
<td>aliquote</td>
<td>[CHEBI_24431] chemical entity</td>
</tr>
<tr>
<td>AminoAcid</td>
<td>Amino Acids</td>
<td>aliquoteID</td>
<td>aliquote</td>
<td>[CHEBI_24431] chemical entity</td>
</tr>
</tbody>
</table>
With the help of the two metadata files
Explore several possibilities by interacting with the graph.

With the help of the two metadata files.
Explore your data in several ways according to your concerns and always by interacting with the graphs.

With the help of the two metadata files.
As far as possible, keep the old way of using the scientist's worksheets...
The R package Rodam

The Comprehensive R Archive Network
https://cran.r-project.org

Using Data

... while allowing a way to be more efficient ...
The R package Rodam

Example based on FRIM - Fruit Integrative Modelling

GET http://pmb-bordeaux.fr/getdata/tsv/frim1/(samples)/sample/365
Retrieving Data within R

Data and metadata are all available and accessible by scripting languages (R, API).

**Merge** the [activome data](#) AND the metabolome data acquire on the same samples

```r
### Load the R ODM package
library(Rodam)

### Get the merged data of two subsets based on their common identifiers
dh <- nav('odams/', weURL='https://rmb.bordeaux.fr/getdata/', dname='friml')
setNamesList <- c('activome', "qNR_H_meta")
dMerged <- dh$getSubsetByName(setNamesList)
```

- **Example based on FRIM - Fruit Integrative Modelling**

---

The Comprehensive R Archive Network

[https://cran.r-project.org](https://cran.r-project.org)

**The R package**

Rodam
Retrieving Data within R

Data and metadata are all available and accessible by scripting languages (R, API).

```
library(Rodam)

# Get the metadata of the 'frilm' dataset
dh <- new('rodamw', 'http://pmb-bordeaux.fr/getdata/', 'frilm')

# Get 'activome' data subset
ds <- dh$getDataByName('activome')

d$samples   # Show the identifier defined in the data subset
[1] "AliquotID"

d$factors   # Show all factors defined in the data subset
[1] "Treatment" "DevStage" "FruitAge"

d$vnames    # Show all quantitative variables defined in the data subset
[1] "FAC1"   "FAC2"   "FAC3"   "FAC4"   "FAC5"

# Boxplot of all variables defined in ds$vnames
par(mar=c(7,8,1,2))
boxplot(log10(ds$data[, ds$vnames]), horizontal=T, outline=F, las=2, cex.axis=0.7)
```

Data mining / Modeling

⇒ Selection subsets of data
⇒ Repetition of multiple scenarios on different subsets of data
Reproducible Research ... with R and RStudio

Chap II **Data Gathering and Storage** (70 pages out of 300)

II. 6 - **Gathering Data with R**

“How you gather your data directly impacts how reproducible your research will be. If all of your data gathering steps are tied together by your source code, then independent researchers (and you) can more easily regather the data”

II. 7 - **Preparing Data for Analysis**

“Once we have gathered the raw data that we want to include in our statistical analyses we generally need to clean it up so that it can be merged into a single data file.”

This is exactly what the ODAM framework aims to answer in a normalized way the easier and faster as possible


Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
Reproducible Research ... with R and RStudio

The R package Rodam

ODAM Framework

https://rmarkdown.rstudio.com/authoring_quick_tour.html
Data Dissemination

The Dataverse Project

Schematic Diagram of a Dataset in Dataverse 4.0

Container for your data, documentation, and code.

https://data.inra.fr/

FRIM - Fruit Integrative Modelling

Bénard, Camille; Bernillon, Stéphane; Blais, Benoit; Maucourt, Mickael; Ballias, Patricia; Deborde, Catherine; Colombié, Sophie; Cabasson, Cécile; Jacob, Daniel; Gibon, Yves; Moing, Annick, 2018, "FRIM - Fruit Integrative Modelling"; https://doi.org/10.15454/95JUTK; Portail Data Inra, V1

Description

The project aimed to build a virtual tomato fruit that enables the prediction of metabolite levels given genetic and environmental inputs, by an iterative process between laboratories which combine expertise in fruit biology, omics, plant health and pathology.

Subject

Computer science; Information management; Omics; Plant Health and Pathology

Related Publication


Link to data

https://pmb-bordeaux.fr/dataexplorer/?ds=frim1

R scripts (Rmd)

Jupyter Notebook (ipynb)

If applicable

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
**ODAM Framework Overview**

Promote good practices

- Experiment Data Tables + Metadata

Provide services

- Storage and sharing with all partners
- Multi-scale deployment
  - Local / Intranet / Internet

ODAM
Open Data for Access and Mining

ODAM

- PUT: Data can be downloaded, explored and mined
- GET: Data can be downloaded

Multiscale deployment

- Local / Intranet / Internet

F A I R

INTEROPERABLE

Allow users to gain efficiency where they would like to gain efficiency

Additional data subsets can be added step by step, as soon as data are produced.

Merging & selection of data subsets

- Repetition of complex treatments according to very varied parameters

ODAM

- Link
- Development of lightweight tools
  - R/Python scripts (Galaxy), lightweight GUI (R shiny)
  - Python (Plotly Dash)

- API

Data is accessible:
Additional data subsets can be added step by step, as soon as data are produced.

The "data lifecycle" is thus integrated into the scientific research process.

Data is “keep alive” in the cycle.

Data can be downloaded, explored and mined.

Data is "keep alive" in the cycle.

The "data lifecycle" is thus integrated into the scientific research process.

Data is “keep alive” in the cycle.

ODAM Framework Overview

ODAM
Open Data for Access and Mining

ODAM

Open Data for Access and Mining
ODAM Framework Overview

Advantages of this approach

data sharing & data availability
- The array of the "plants" may be created even before planting the seeds.
- Similarly, the array of the "harvests" can be created as soon as the harvests are done, and this before any analysis.
- Thus, these arrays are generated only once in the project and we can set up the sharing soon the seed planting. Then each analysis comes to complement the set of data as soon as they produce their own sub-dataset.
- data are accessible to everyone as soon as they are produced,

identifiers centrally managed
- data are archived and compiled, so that it becomes useless to proceed a laborious investigation to find out who possesses the right identifiers, etc.

facilitate the subsequent publication of data
- data are already readily available online by web API,
- But nothing prevents to take this data to fill in existing a data repository, by adjoining more elaborate annotations such as Dataverse.
ODAM Framework Overview

Advantages of this approach

minimal effort, maximum efficiency

Format the data
- Based on TSV: choice to **keep the good old way of scientist to use worksheets**, thus *i*) using the same tool for both data files and metadata definition files, *ii*) no programmatic skill are required

Give an access through a web services layer
- based on current standards (REST)

Use existing tools
- Spreadsheets, R studio, Spyder, Jupyterlab, BioStatFlow, ...

Develop if needed, lightweight tools
- R/Python scripts, lightweight GUI (R shiny, Dash), Galaxy, ...
Tools for data science

Reproducible Research
Tools for data science

Anaconda Distribution

https://www.anaconda.com/distribution/

R packages for data science

https://www.tidyverse.org/

In particular dplyr provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges

http://intelligency.org/ai_r.php
Tools for data science
Reproducible Research

ODAM Framework

+ Jupyter Notebook

EDTMS

+ Binder

R

python

GitHub

Jupyter

nbviewer
**Jupyter Notebook**

**Get the data from ODAM**

```r
In [3]: # API call to retrieve data
    dataset = 'Frias'
    subset = 'qMS_metadata'
    # retrieve odam from dataset, subset
    ODAM["factor"]
```

**ODAM**

**EDTMS**

**Initialize the 'ODAM' object**

```r
In [9]: dh <- read('odam.csv', url='https://geb-bordeaux.fr/getdata/', dataframe='Frias')
```

**Plot the data tree of the whole dataset**

```r
In [11]: dn <- dnGetDatasetTree()
    data.tree <- setNodeStyle(color = "grey35", penwidth = 2)
    data.tree <- setNodeStyle(stroke = "filled", rounded = "box", fillcolor = "GreenYellow", fontname = "helvetica", tooltip = "
    data.tree <- setNodeStyle(disabled = "false", fillcolor = "Thistle", penwidth = "2px")
    plotTree()
```

**Compute then plot PCA**

```r
In [15]: # Compute PCA
    res_pca = pca_2dplot(x, y, n, scale=TRUE)
    fac1 = factor("Frias", levels = c("Frias", "Frias2", "Frias3", "Frias4"))
    # plot PCA scores
    plotPCA(res_pca, 1, 3, Fac1)
```
Reproducible Research  Jupyter Notebook

GitHub

djacob65 / binder_odam

Manage topics

1 commit  1 branch  0 releases  1 contributor

Branch: master  New pull request

<table>
<thead>
<tr>
<th>File</th>
<th>Latest commit</th>
<th>Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>PyODAM_api_PCA.ipynb</td>
<td>update</td>
<td>5 minutes ago</td>
</tr>
<tr>
<td>PyODAM_api_demo.ipynb</td>
<td>update</td>
<td>2 minutes ago</td>
</tr>
<tr>
<td>README.md</td>
<td>update</td>
<td>6 days ago</td>
</tr>
<tr>
<td>Rodam_api_demo.ipynb</td>
<td>update</td>
<td>2 days ago</td>
</tr>
<tr>
<td>Rodam_api_graphics.ipynb</td>
<td>update</td>
<td>6 days ago</td>
</tr>
<tr>
<td>environment.yml</td>
<td>update</td>
<td>6 days ago</td>
</tr>
</tbody>
</table>

environment.yml

```yaml
name: ipython-environment
channels:
  - conda-forge
dependencies:
  - requests
  - numpy
  - matplotlib
  - pandas
  - scikit-learn
```

https://github.com/djacob65/binder_odam

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019
A simple way to share Jupyter Notebooks

Enter the location of a Jupyter Notebook to have it rendered here:

https://nbviewer.jupyter.org/github/djacob65/binder_odam/
The Binder Project is an open community that makes it possible to create sharable, interactive, reproducible environments.

https://mybinder.org/

The Binder Project is an open community that makes it possible to create sharable, interactive, reproducible environments.
The Binder Project is an open community that makes it possible to create sharable, interactive, reproducible environments.

Push the docker container into the several nodes.
Jupyter Interactive Notebook

[Image of Jupyter Notebook interface]

Shareable link: https://mybinder.org/v2/gh/djacob65/binder_odam/master

See more complete examples:
https://github.com/nuest/reproducible-research-and-giscience
https://ajstewartlang.github.io/SIPS_2019/SIPS_presentation.html
ODAM Framework
Open Data for Access and Mining

https://fr.slideshare.net
“Make your data great now” towards Open Data and Reproducible Research
“Make your data great again” towards Linked Open Data

https://bio.tools/ODAM

Thank you for your attention

Daniel Jacob – INRA UMR 1332 BFP – Oct 2019