2008

Curating Chemistry Data through Its Lifecycle: A Collaboration between Library and Laboratory in Scientific Data Preservation

Jeremy R. Garritano  
*Purdue University, jgarrita@umd.edu*

Jake R. Carlson  
*Purdue University, jakecarlson@purdue.edu*

Follow this and additional works at: [http://docs.lib.purdue.edu/lib_fspres](http://docs.lib.purdue.edu/lib_fspres)

Part of the [Library and Information Science Commons](http://docs.lib.purdue.edu/lib_fspres)

**Recommended Citation**

[http://docs.lib.purdue.edu/lib_fspres/23](http://docs.lib.purdue.edu/lib_fspres/23)

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
Curating chemistry data through its lifecycle: A collaboration between library and laboratory in scientific data preservation

Jeremy R. Garritano, Acting Head, Chemistry Library
& Jake R. Carlson, Data Research Scientist
Purdue University Libraries

jgarrita@purdue.edu

236th ACS National Meeting, Philadelphia, PA
August 19, 2008
Outline

• Project origins
• Needs assessment
• Creation of data archive model
• Along the way
  – Collaboration Tips
  – Instrument/Software Challenges
  – Metadata Issues
  – Preservation Concerns
The Center for Authentic Science Practice in Education

Funded by the National Science Foundation
– NSF Award #CHE-0418902

“CASPiE is a multi-institutional collaborative effort designed to address major barriers to providing research experiences to younger undergraduate science students.”

http://www.caspie.org/
Who is involved?

• **Lead Institutions**
  – Purdue University
  – Ball State University
  – University of Illinois at Chicago
  – Northeastern Illinois University

• **Partner Institutions**
  – College of DuPage
  – Harold Washington College
  – Moraine Valley Community College
  – Olive-Harvey College
(select) Goals of CASPiE

• Provide first and second year students with access to research experiences as part of the mainstream curriculum.

• Provide access to advanced instrumentation for all members of the collaborative to be used for undergraduate research experiences.

• Help faculty develop research projects so that their own research capacity is enhanced and the students at these institutions can participate in this research.
“Making Instruments Part of the Cyberinfrastructure”

- Analytical Chemistry Seminar, April 2007
- Given by Director of Instrumentation Networking
- How the instrumentation network is designed
- How authentication and scheduling is handled
- How students access the instruments
- How security is handled
After the Seminar

• Requested meetings
  – First, the Technical Side
  – Systems and instrumentation staff
  – Learned about the instrumentation network
    • Types of data generated
    • Associated metadata
    • Different modes of access
The Educational Side

• Director of CASPiE and Module Author (Assoc. Professor, Foods and Nutrition)

• Understand the workflow outside the instrumentation network

• How students generate some additional data through in-class experiments

• How students record additional information during and after lab in their notebooks

• How the final data and conclusions are forwarded to the Module Author for review and future exploration
**The Spreadsheet**

<table>
<thead>
<tr>
<th>hypothesis</th>
<th>experimental</th>
<th>analyses</th>
<th>results</th>
<th>conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garlic would have a signi</td>
<td>Solutions of three type TEAC, Phenolics</td>
<td></td>
<td>Solution: TEAC Abs. Chg ABTS Phen.</td>
<td>Given that the values</td>
</tr>
<tr>
<td>I think that the fresh apric</td>
<td>We followed the HPLC HPLC, TEAC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparing the boiled an</td>
<td>Began the first few day HPLC, TEAC</td>
<td></td>
<td>We found that in the TEAC assay that the blueber</td>
<td></td>
</tr>
<tr>
<td>We tested the antioxidant</td>
<td>We performed the HI HPLC, Phenolics</td>
<td></td>
<td>Our results showed that ascorbate was highly pr</td>
<td></td>
</tr>
<tr>
<td>There is a greater assist</td>
<td>We chose to use HPLC HPLC, Phenolics</td>
<td></td>
<td>We used Lime juice as our standard for compar</td>
<td></td>
</tr>
<tr>
<td>we thought that the unco</td>
<td>To test for epicatechin HPLC, Phenolics</td>
<td></td>
<td>In our total polyphenols experiment, we assume</td>
<td></td>
</tr>
<tr>
<td>If you boil blueberries for</td>
<td>We did the Ascorbate HPLC</td>
<td></td>
<td>After turning in all the HPLC solutions for 1, 2, and Boiling blueberries</td>
<td></td>
</tr>
<tr>
<td>The insoluble antioxidant</td>
<td>A TEAC assay was per TEAC</td>
<td></td>
<td>Our findings, through the TEAC assay, show th</td>
<td></td>
</tr>
<tr>
<td>The antioxidant activity a</td>
<td>The individual juices w HPLC</td>
<td></td>
<td>Both types of extract were subjected to HPLC se Prior to experimen</td>
<td></td>
</tr>
<tr>
<td>Freezing apples will have</td>
<td>Antioxidant levels in th TEAC</td>
<td></td>
<td>We measured a TEAC of 0.0127 mM Trolox per L Originally the Red D</td>
<td></td>
</tr>
<tr>
<td>Find the total amount of</td>
<td>The antioxidant activi TEAC</td>
<td></td>
<td>After the TEAC value of cinnamon was found 1.51 It is important to rea</td>
<td></td>
</tr>
<tr>
<td>My original hypothesis w</td>
<td>For raw red cabbage: I TEAC</td>
<td></td>
<td>It was discovered that raw red cabbage has a sign In the future, it wou</td>
<td></td>
</tr>
<tr>
<td>Heat-treating oregano wi</td>
<td>5 gram samples of oregano HPLC, TEAC, Phen</td>
<td></td>
<td>The TEAC test results showed little variation in v After analyzing the re</td>
<td></td>
</tr>
<tr>
<td>Our initial hypothesis was</td>
<td>We tested our sample: HPLC, TEAC</td>
<td></td>
<td>Our initial hypothesis was that the heated oregano We have determined</td>
<td></td>
</tr>
<tr>
<td>Unpeeled apples will have</td>
<td>To verify our hypotheses HPLC, TEAC</td>
<td></td>
<td>After obtaining the concentration of ascorbate u At the end of 6 week</td>
<td></td>
</tr>
<tr>
<td>Boiled pineapple juice ha</td>
<td>Boiled pineapple juice HPLC, Phenolics</td>
<td></td>
<td>Our research found that boiled pineapple juice ha: Further research me</td>
<td></td>
</tr>
<tr>
<td>Based upon what I have I measured 0.044 gram</td>
<td>HPLC</td>
<td></td>
<td>We found cooked blueberries to have more vitamin By using HPLC, we l</td>
<td></td>
</tr>
<tr>
<td>As seen in the reduced a</td>
<td>We manipulated our in TEAC</td>
<td></td>
<td>From the data resulting from the TEAC assay it w Although our research</td>
<td></td>
</tr>
<tr>
<td>We believed that boiled b</td>
<td>Our group boiled blue HPLC</td>
<td></td>
<td>Our results showed that the boiled blueberries co</td>
<td></td>
</tr>
<tr>
<td>There would be a higher I</td>
<td>We used the TEAC as TEAC</td>
<td></td>
<td>The data was close, but the level of antioxidants In conclusion, this exp</td>
<td></td>
</tr>
<tr>
<td>Blueberry smoothie won</td>
<td>HPLC was used to an HPLC</td>
<td></td>
<td>Blueberry provided by the lab had a lesser amount There was more as</td>
<td></td>
</tr>
<tr>
<td>Upon boiling red cabbage</td>
<td>TEAC assay was perf TEAC</td>
<td></td>
<td>Boiled cabbag produced dark blue water and gre The color loss sh</td>
<td></td>
</tr>
<tr>
<td>The room temperature b</td>
<td>The TEAC assay and I TEAC, Phenolics</td>
<td></td>
<td>The heated blueberry solutions yielded the highest</td>
<td></td>
</tr>
</tbody>
</table>
Formal Proposal

Based on the needs identified, the Libraries proposed to offer 100 staff hours to:

• Identify a suitable module for the prototype
• Outline the scientific workflow and map it to data curation functions
• Determine needs for access/preservation
• Inventory data and determine appropriate manners of description (i.e., metadata)
• Create data repository ingest packages and archive past data
• Demonstrate prototype in Purdue e-Data service
• Document the process and challenges we faced
To do this…

We had to become familiar with:

- the particular lab module and understand the purpose of each of the analytical methods involved
- the workflow of the students and CASPiE staff as they implemented the module and generated data
- what the data generated looked like in terms of format, file size, description, etc.
- the desired outcomes for the data for all parties involved
- what metadata standards would fit these needs
Lab module

• “Phytochemical Antioxidants with Potential Health Benefits in Foods”
  – Many students have heard of antioxidants
  – Deals with “real world” items – food and drink
  – May prevent chronic diseases
  – Still has chemistry component
Lab module

• 3-4 weeks of learning analytical techniques
• 3 weeks of pursuing a research question

• Analytical techniques used:
  – Trolox equivalent antioxidant capacity (TEAC) Assay
  – Total phenolics
  – High Performance Liquid Chromatography (HPLC)
Typical student question categories

• Look at:
  – Fruits
  – Vegetables
  – Spices
  – Teas
  – Juices
  – Chocolate

• Effects of:
  – Temperature
  – Digestion
  – Storage conditions
  – Food processing
Sample student research ???’s

• Our research question was, when comparing Welch's 100% red and white grape juices, which variety has the higher antioxidant activity…

• Out of four yogurts, what will be the abundance of antioxidants within each? Which of the four will have the most antioxidants?

• Does sugar affect the antioxidant levels in green tea?
Sample student conclusions

- Our data supports our hypothesis. We believed that the strawberry yogurt would have more antioxidants than the other yogurts. However, we found that it was not the yogurt that has the antioxidants but rather the fruit put into the yogurt.

- Our results show that red grape juice has a higher antioxidant concentration, by both TEAC and total polyphenolic standards, in comparison to white grape juice. This verifies that our hypothesis was correct.

- Inconclusive.
Sample ID: SMP Green tea and lemon juice 1/25'
Data File: Z:\Data\UIC18648B-12-24Apr-6
Method: \Method\AscorbicAcid.met
Vial: A:B6
Inj. Vol.: 10 uL

<table>
<thead>
<tr>
<th>Name</th>
<th>Retention Time</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascorbic Acid</td>
<td>2.277</td>
<td>481748</td>
</tr>
</tbody>
</table>
Sample “Raw” HPLC Data

<table>
<thead>
<tr>
<th>Version:</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxchannels:</td>
<td>1</td>
</tr>
<tr>
<td>Sample ID:</td>
<td>SMP Green tea and lemon juice 1/25'</td>
</tr>
<tr>
<td>Vial Number:</td>
<td>A;B6</td>
</tr>
<tr>
<td>Data File:</td>
<td>Z:\Data\Week4\UIC18648B-12-24Apr-6</td>
</tr>
<tr>
<td>Method:</td>
<td>K:\Method\AscorbicAcid.met</td>
</tr>
<tr>
<td>Volume:</td>
<td>10</td>
</tr>
<tr>
<td>Pretreat Name:</td>
<td>(None)</td>
</tr>
<tr>
<td>User Name:</td>
<td>central.purdue.lcl\1393steffen</td>
</tr>
<tr>
<td>Acquisition Date and Time:</td>
<td>4/25/2008 12:43:13 PM</td>
</tr>
<tr>
<td>Sampling Rate:</td>
<td>10.000000 Hz</td>
</tr>
<tr>
<td>Total Data Points:</td>
<td>1801 Pts.</td>
</tr>
<tr>
<td>X Axis Title:</td>
<td>Minutes</td>
</tr>
<tr>
<td>Y Axis Title:</td>
<td>mAU</td>
</tr>
<tr>
<td>X Axis Multiplier:</td>
<td>0.016667</td>
</tr>
<tr>
<td>Y Axis Multiplier:</td>
<td>0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>124833</th>
<th>6241</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>71</td>
<td>126503</td>
<td>5542</td>
</tr>
<tr>
<td></td>
<td>146</td>
<td>127544</td>
<td>4915</td>
</tr>
<tr>
<td></td>
<td>232</td>
<td>127959</td>
<td>4354</td>
</tr>
<tr>
<td></td>
<td>334</td>
<td>127759</td>
<td>3853</td>
</tr>
<tr>
<td></td>
<td>455</td>
<td>126963</td>
<td>3406</td>
</tr>
<tr>
<td></td>
<td>598</td>
<td>125600</td>
<td>3009</td>
</tr>
<tr>
<td></td>
<td>768</td>
<td>123705</td>
<td>2657</td>
</tr>
<tr>
<td></td>
<td>971</td>
<td>121317</td>
<td>2344</td>
</tr>
<tr>
<td></td>
<td>1214</td>
<td>118481</td>
<td>2068</td>
</tr>
<tr>
<td></td>
<td>1505</td>
<td>115245</td>
<td>1824</td>
</tr>
<tr>
<td></td>
<td>1856</td>
<td>111659</td>
<td>1610</td>
</tr>
<tr>
<td></td>
<td>2278</td>
<td>107774</td>
<td>1420</td>
</tr>
<tr>
<td></td>
<td>2786</td>
<td>103642</td>
<td>1254</td>
</tr>
<tr>
<td></td>
<td>3396</td>
<td>993312</td>
<td>1108</td>
</tr>
<tr>
<td></td>
<td>4128</td>
<td>94833</td>
<td>979</td>
</tr>
<tr>
<td></td>
<td>5001</td>
<td>90251</td>
<td>866</td>
</tr>
<tr>
<td></td>
<td>6041</td>
<td>85611</td>
<td>767</td>
</tr>
<tr>
<td></td>
<td>7271</td>
<td>80952</td>
<td>680</td>
</tr>
<tr>
<td></td>
<td>8718</td>
<td>76312</td>
<td>604</td>
</tr>
<tr>
<td></td>
<td>10410</td>
<td>71723</td>
<td>537</td>
</tr>
<tr>
<td></td>
<td>12372</td>
<td>67216</td>
<td>477</td>
</tr>
<tr>
<td></td>
<td>14633</td>
<td>62816</td>
<td>425</td>
</tr>
<tr>
<td></td>
<td>17215</td>
<td>58544</td>
<td>379</td>
</tr>
<tr>
<td></td>
<td>20139</td>
<td>54418</td>
<td>337</td>
</tr>
<tr>
<td></td>
<td>23421</td>
<td>50451</td>
<td>301</td>
</tr>
<tr>
<td></td>
<td>27071</td>
<td>46656</td>
<td>268</td>
</tr>
<tr>
<td></td>
<td>31092</td>
<td>43040</td>
<td>238</td>
</tr>
<tr>
<td></td>
<td>35479</td>
<td>39609</td>
<td>211</td>
</tr>
<tr>
<td></td>
<td>40217</td>
<td>36366</td>
<td>187</td>
</tr>
<tr>
<td></td>
<td>45284</td>
<td>33312</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>50646</td>
<td>30446</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>56261</td>
<td>27766</td>
<td>127</td>
</tr>
<tr>
<td></td>
<td>62077</td>
<td>25268</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>68035</td>
<td>22947</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>74069</td>
<td>20796</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>80105</td>
<td>18809</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>86069</td>
<td>16979</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>91884</td>
<td>15298</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>97473</td>
<td>13757</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>102764</td>
<td>12349</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>107588</td>
<td>11066</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>112179</td>
<td>9899</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>116182</td>
<td>8841</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>119649</td>
<td>7883</td>
<td></td>
</tr>
</tbody>
</table>
“Paper” Data

• Student lab notebooks
  – Pre-labs
  – Notes and data collected during lab
  – Calculations
  – Post-lab reports
• Hard to read
• Hard to extract relevant information
Instrument/Software Challenges

• Make it easy
• Proprietary instruments mean…
• Security and access
• File name generation
• Actual instrument data generation
Revised Proposal

Based on the needs identified, the Libraries proposed to:

- Identify a suitable module for the prototype
- Outline the scientific workflow and map it to data curation functions
- Determine needs for access/preservation
- Inventory data and determine appropriate manners of description (i.e., metadata)
- Create data repository ingest packages and archive past data
- Demonstrate prototype in Purdue e-Data service
- Document the process and challenges we faced
Technical Metadata

• Consulted with Indigo Biosystems

• Chose to go with MIAPE for HPLC
  – Minimum Information About a Proteomics Experiment
    • MIAPE Column Chromatography subset

• Others considered
  – mzData, netCDF, AnIML, FuGE, and GAML
Sample Fields in MIAPE Standard

- **Date/Time Stamp**
- **Product details about the column**
  - Make
  - Model
- **Physical characteristics of the column**
  - Length
  - Diameter
  - Description of the stationary phase
- **Mobile Phase**
  - Name of mobile phase
  - Description of constituents
- **Properties of the column run**
  - Time
  - Gradient
  - Flow rate
  - Temperature
  - Separation purpose
- **Column outputs**
  - Detection
  - Equipment used for detection
  - Type
  - Equipment settings
  - Timescale over which data was collected
  - Trace
Additional Fields Needed

• Surveyor autosampler settings:
  – Flush/Wash
  – Injection Mode
  – Tray set temperature

• Peak table:
  – Name
  – Expected Retention Time
  – Expected Retention Window

• Integration events: event type: width:
  – Start
  – Stop
  – Value

• Software
Additional Metadata Needed

- Unique identifier for the lab group
- Number of lab partners in the group
- First and last name of each student in the lab group
- Institution (ex. Purdue University)
- Course – abbreviation and number (ex. CHEM 116)
- Section number
- Name of the professor (ex. Jay Burgess)
- Name of the teaching assistant
- Semester / Year (ex. Spring 2007)
Additional Metadata Needed

- Hypothesis
- Experiment
- Food category
- Food type
- Food description
- Methods of analysis
- Analysis of results
- Conclusion
  - etc.
Metadata Issues

• Make it easy
• Different modules mean…
• What standards to use
• How to “tag” the data – print and electronic
Proposed CASPiE Data Archive Model

Initial Data Interview
- Extract: Student Information, Lab Group ID
- Extract: Experiment Design, Hypothesis
- Extract: Description of Sample, Steps Taken to Prepare Sample
- Extract: Instrument Settings, Methods File
- Extract: Observations, Results
- Extract: Changes Made to the Data
- Extract: Results of the Experiment
- Extract: Student Research Papers, Lab Notebooks
- Researcher Reviews Student Data

Class Begins
- Data set #1
  - Bad Data
- Data set #2
  - Original
- Data set #3
  - Original
- Data set #4
  - Original

Sample Preparation
- Data set #1
- Data set #2
- Data set #3
- Data set #4

HPLC Instrument
- Data set #1
- Data set #2
- Data set #3
- Data set #4

Analysis
- Data set #1
- Data set #2
- Data set #3
- Data set #4

Data Reduction, Smoothing
- Data set #1
- Data set #2
- Data set #3
- Data set #4

Conclusion of the Experiment
- Data set #1
- Data set #2
- Data set #3
- Data set #4

Conclusion of the Course
- Data set #1
- Data set #2
- Data set #3
- Data set #4

Metadata

Active “Education” Database
- Data set #2a & b
  - Inconclusive
  - Little/No Value
  - High Value
Proposed CASPiE Data Archive Model

1. Review of the Data by the Researcher
   - Metadata
     - Data set #2a & b
       - Inconclusive
     - Data set #3a & b
       - Little/No Value
     - Data set #4a & b
       - High Value
2. Technical Metadata are Reviewed
   - Extract Relevant Metadata
     - Extract Module/Lab Manual Information
   - Technical Metadata are Mapped and Validated
3. High Level Metadata are Reviewed
   - High Level Metadata is Mapped and Validated
   - Full Review and Approval
     - Create Submission Information Package
4. Create/Finalize Policies For Data Sets

Document Process and Workflows
Preservation Concerns

• Make it easy
• Student data means…
• How long to keep
• How to identify most important
Funding

CASPiE looking to acquire additional funding.
- To continue operations
- To improve their instrumentation network
- To enable additional capabilities (incl. a data archive)
Collaboration Tips

• Make it easy
• Undergraduate research means…
• Chemical Education component
• Need funding to survive
• Increase interdisciplinary nature
• Show what we bring to the table
Suggestions

• Explore options for having students do their work electronically

• Have students input more granular and/or specific information

• Provide students with controlled vocabulary when submitting information

• Use consistent ID and naming conventions

• Extract data and metadata into open formats

• Develop a system of review and quality control for the data and metadata
First attempt

- Qualtrics survey software
  - Site license
  - Can allow others read only of the data
  - Exports data in CSV, XML, HTML, and SPSS
  - Allows branching
Sample Data Collection Form

Sample information

What type of food are you testing?

Describe your food samples in detail. (For example: If you selected "Fruit or Fruit Juice" in the previous question, you might answer: "Welch's purple grape juice and Welch's white grape juice". Be sure to enter the specific type and brand of each sample.)

What is the treatment or process that your food is subjected to?

- Digestion
- Temperature changes
- Storage conditions
- Other

Describe your treatment or process in detail. (For example: If you selected "Temperature changes" in the previous question, you might answer: "30 mL of each sample were boiled in 250 mL of water for 20 minutes". Be sure to enter the relevant information about time, temperature etc.)
Which tests did you carry out? How many trials did you complete for each test? How many samples did you have for each trial?

<table>
<thead>
<tr>
<th></th>
<th>Number of trials</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>didn't run this test</td>
<td>1 trial</td>
</tr>
<tr>
<td>TEAC</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Ascorbate by HPLC OR titration</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Total polyphenolics</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>
Future exploration: e-Lab Notebooks

- Make it easy
- Exporting features
- Flexibility
- Data pre-population and/or matching
- Importing features
Future directions

• Explore additional funding
• Demonstrate in e-Data Service
• Map additional modules
Acknowledgements

• Jake Carlson
• Andrea and Rob
• Fred Lytle, Debbie Steffen, Phil Wyss
• Gabriela Weaver, Jay Burgess