Anatomy of the chemistry research enterprise in the academic sector: Serving the underserved in a large research institution

Leah McEwen
Chemistry Librarian
Cornell University
ACS San Diego
March 2016
Abstract

The Research Life Cycle (RLC) at any research institution involves a myriad of scientific and technical support roles, including instrumentation, data management, information access, environment health and safety. Researchers engage with many of these services and these providers in turn liaise across numerous disciplines and departments. All of these functions involve the use of technical information for analysis, interpretation and documentation. In supporting these other research support groups, libraries contribute more fully to the RLC and engage more broadly across the research community. This talk will outline outreach services developed for a variety of service groups on an academic university campus, including chemical analysis labs, chemistry IT services, Environmental Health & Safety and Occupational Medicine.
“The Division of Chemistry recognizes that in most cases, principal investigators will publish data (and relevant supplementary information) in peer-reviewed journal articles within a reasonable time, and that the chemistry research community maintains a significant number of databases that provide for access to data. **Such disclosure of data meets the majority of needs for robust and open scientific discourse.** The purpose of the Data Management Plan is to provide a means for highlighting the existing practices of the principal investigator’s laboratory and larger research community, and to **encourage innovations that, where appropriate and practical, take advantage of emerging information technologies and cyber-infrastructure.”
Goal: Digital Data Culture

enabling ‘conversation’
machine-to-human-to-machine

• Machine readable and interpretable
  – Corollary:
    digitally managed cradle to cradle
• Sustainably accessible for scientific record
  – Corollary:
    systematically deposited at publication

[conversari | Latin: ‘keep company (with)’]
Analog Management -> Digital Management

Current practice in chemistry:
Paper notebooks, Excel spreadsheets, local analysis systems, individual hardware
Current practice in chemistry:

Paper notebooks, Excel spreadsheets, local analysis systems, individual hardware

Shift in Data Practices

Source: Cerys Whilloughby, Univ. Southampton
Current practice in chemistry:
Paper notebooks, Excel spreadsheets, local analysis systems, individual hardware.

Shift in Data Practices

Source: Cerys Whilloughby, Univ. Southampton
Publication -> Deposit

Current practice in chemistry:

Included in journal article text or supplemental files
Most often processed data in figures/tables in .pdf and images
**Scheme S6. Synthesis of S4**

**Synthesis of S4:** Anhydrous PhMe (31 mL) and freshly distilled iPr₂NH (10 mL) were added to a 100 mL flask and sparged with N₂ for 20 min. 1,4-dibromo-2,5-diiodobenzene (3.000 g, 6.15 mmol), 1-(tert-butyl)-4-ethynylbenzene (2.433 g, 15.38 mmol), Pd(PPh₃)₂Cl₂ (216 mg, 0.308 mmol) and CuI (117 mg, 0.615 mmol) were added to the solution, in sequence. The mixture was stirred at rt for 14 h. The crude reaction mixture was filtered through a pad of SiO₂ gel and washed with additional CH₂Cl₂ (200 mL). Evaporation of the solvent provided the crude product as a solid, which was purified by chromatography (SiO₂, hexanes) to provide S4 (2.310 g, 68% yield) as a white powder. ¹H and ¹³C NMR spectroscopy were consistent with the report of Hseuh et al. **S4:** ¹H NMR (300 MHz, CDCl₃) δ 7.77 (s, 2H), 7.58 – 7.47 (m, 4H), 7.46 – 7.35 (m, 4H), 1.34 (s, 18H). ¹³C NMR (75 MHz, CDCl₃) δ 152.68, 136.05, 131.70, 126.58, 125.63, 123.81, 119.46, 97.03, 86.49, 77.58, 77.16, 76.74, 35.06, 31.30.
The Challenge

Making Sense for Researchers...

• An early question: if we build it (on best intentions), will they come?
• A later question: if we mandate it (on best principles), will they participate?
• A recent question: if it makes sense to them, will they adapt their practice?

> How can research support groups help with this transition?
Data Management in Context

Premise: moving towards the future is a trajectory from current practice, built on historic motivations

• What does chemistry data management (CDM) for re-use look like today, and how did it function in the past?
• What is the interplay between in-lab data workflows and post-publication data handling as two critical areas of data management?
• What roles can service groups play to support researchers in digitally managing lab generated data and preparing them for sustainable publication?
CDM Origins: Post-Publication

- Compiled data collections from primary literature
- Critically evaluated data by agencies and unions
- Post-publication, downstream of lab work
- Highly curated and indexed
- Value-added commercial or subscription
CDM Current Blend: Post/At/Pre

- **Post**-publication: highly curated compilation continues
- **At** publication: journal requirements for editorial and peer review
  - Primarily supplemental files
  - Some specific sub-discipline repositories
- **Pre**-publication: LIMS and ELN systems (primarily in industry, enterprise systems), data not generally published
Service Group Data Functions

- **Instrument labs**
  - Collection instruments, specialized software licenses, data storage, analysis expertise
- **Departmental IT support**
  - Local data storage, software license support, implementation, troubleshooting
- **Health & Safety**
  - Chemical inventories, chemical safety information, green chemistry protocols
- **Libraries**
  - Published resources, DMP consultation (w/ OSP), dialogue w/ publishers, Institutional Repositories

-**Economy of scale, many researchers, many contact hours**
Case Type: Materials Synthesis

- Notebook page
- MS Word
- Lab group report
- ChemDraw
- Excel
- iPhone
- MNova

**Table:**

<table>
<thead>
<tr>
<th>Date</th>
<th>S/N-Tag</th>
<th>Type</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>4/16/2014</td>
<td>SJI-02-036</td>
<td>0.40</td>
<td>Set up reaction same way that Fernando described but Crude mixture shows mainly starting material. Tony ran a DART of my crude which also shows SM as the main mass in the mixture and no product forming.</td>
</tr>
<tr>
<td>4/20/2014</td>
<td>SJI-02-051</td>
<td>0.40</td>
<td>Set up exp, flame dried, FPT, with BuLi. DART shows dehydrobrominated product. Added boronate ester but when I went to take the DART the machine was down. Told Ivan about this and he hasn’t gotten back to me about it yet.</td>
</tr>
</tbody>
</table>

*When complete, the reaction was washed with sat. NaHCO₃ aq and extracted three times with DCM. The organic fractions were collected and dried with MgSO₄, filtered and solvent was removed under vacuum. The product was purified on SiO₂ column chromatography (3:7 DCM:Henvane). The product was isolated as a light yellow solid.*
A. Materials.

B. Synthetic Procedures

Scheme SX. Synthesis of SX.

Synthesis of Product: Experimental Procedure to provide Product (X XXX g, XX% yield) as a [form]. SX: H NMR (400 MHz, CDCl₃) δ X XX (s, YH), ¹³C NMR (100 MHz, CDCl₃) δ XXXXX, IR (solid, ATR) XXXX cm⁻¹, HRMS (EI, m/z): calcd for [CₓHₓNₓ]⁺ XXXXXXX, found XXX XXXX.

C. NMR Spectra

Figure SX.
[Images of NMR fraction spectra, with peak assignments and image of product structure]

D. [other spectral analyses and calculations]
[Images of other spectra]
[Images of other analyses]

E. References to Supporting Information

Scheme SX. Synthesis of SX.
[Image of reaction schema]

Synthesis of Product: Experimental Procedure to provide Product (X XXX g, XX% yield) as a [form]. SX: H NMR (400 MHz, CDCl₃) δ X XX (s, YH), ¹³C NMR (100 MHz, CDCl₃) δ XXXXX, IR (solid, ATR) XXXX cm⁻¹, HRMS (EI, m/z): calcd for [CₓHₓNₓ]⁺ XXXXXXX, found XXX XXXX.

HRMS (EI, m/z): calcd for [CₓHₓNₓ]⁺ XXXXXXX, found XXX XXXX.
Data Flow

Collect
- Quantities
- Observation
- TLC
- Instrument files

Process
- Manual entry
- Photo images
- Spectra

Analyze
- Drawing
- Calculation
- Rendering
- Modeling

Collect
- Images
- Text

Document

Notebook pages
(per experimental run)

Work Reports
(per transformation)

Working SI
(per schema)

ESI

-> final published supporting data
Practice Commonalities

• Intra-experiment core data+documentation packages (e.g., protocol & measurement)
• Inter-experiment contextual variability and dependency
• Variable data types, formats, sources and cycle points
• Process/protocol focus
• Use of templates

-> Existing ‘Minimum Requirement’ management system features to manage variability
Practice Drivers

- Intra-lab and intra-domain re-use
- Community norms
- Publication practices
- Provenance, integrity

-> Existing ‘Control Points’
   to manage perceived expectations
At Cornell: can we implement an ELN that supports these activities across different labs?

Needed support:

- Upstream instrument documentation
- Downstream publisher guidelines
- and a cooperative link from publishers to a repository for characterization data a la CSD
At Cornell: can we implement an ELN that supports these activities across different labs?

Needed support:
• Upstream instrument documentation
• Downstream publisher guidelines
• and a cooperative link from publishers to a repository for characterization data a la CSD

Towards Ideal Practice
Source: Cerys Whilloughby, Univ. Southampton
### Reaction

- **Reaction Run:** SJH-01-227 dated 2/12/2014;  
  - FailedReaction: false;  
  - Experiment Stage: Planned

<table>
<thead>
<tr>
<th>Label</th>
<th>Reaction Component</th>
<th>Substance</th>
<th>Amounts</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>SJH-01-223</td>
<td>Role: Limiting Reactant Compound</td>
<td>State: Solid</td>
<td>Equivalence: 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molecular Mass: 756.95</td>
<td></td>
<td>Moles: 0.132 mMol</td>
<td></td>
</tr>
<tr>
<td></td>
<td>State: Solid</td>
<td></td>
<td>Mass: 0.1 g</td>
<td></td>
</tr>
<tr>
<td>benzaldehyde</td>
<td>Role: Reactant Compound</td>
<td>State: Solid</td>
<td>Equivalence: 6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molecular Mass: 206.24</td>
<td></td>
<td>Moles: 0.293 mMol</td>
<td></td>
</tr>
<tr>
<td></td>
<td>State: Solid</td>
<td></td>
<td>Mass: 0.163 g</td>
<td></td>
</tr>
<tr>
<td>Cu(OTf)$_2$</td>
<td>Role: Reactant Compound</td>
<td>State: Liquid</td>
<td>Equivalence: 0.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molecular Mass: 361.67</td>
<td>Purity: 98%</td>
<td>Moles: 0.013 mMoles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Source: 283673-5G, Sigma Aldrich</td>
<td></td>
<td>Mass: 0.005 g</td>
<td></td>
</tr>
<tr>
<td>DCE</td>
<td>Role: Solvent Compound</td>
<td>State: Liquid</td>
<td>Volume: 1.321 mL</td>
<td>Concentration in line 1: 0.1 M</td>
</tr>
<tr>
<td></td>
<td>Purity: 99-100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Source: 283673-5G, Sigma Aldrich</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TFA</td>
<td>Role: Solvent Compound</td>
<td>State: Liquid</td>
<td>Equivalence: 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molecular Mass: 114.02 Density: 1.49 g/ml</td>
<td>Purity: 99%</td>
<td>Moles: 0.396 mMol</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Source: T6508-500mL, Sigma Aldrich</td>
<td></td>
<td>Mass: 0.045 g</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Volume: 0.030 mL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SJH_01_227</td>
<td>Role: Product Compound</td>
<td>State: Solid</td>
<td>Equivalence: 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Molecular Mass: 1063.36</td>
<td></td>
<td>Moles: 0.132 mMol</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Source: Roya l Society of Chemistry</td>
<td></td>
<td>Mass: 0.140 g</td>
<td></td>
</tr>
</tbody>
</table>
**Reaction**

- **Reaction Run:** Reaction SJH-01-227 dated 2/12/2014;  
  **FailedReaction:** false;  
  **Experiment Stage:** Planned

<table>
<thead>
<tr>
<th>Ordinal</th>
<th>Parent</th>
<th>Title</th>
<th>Description</th>
<th>Parameter/Substances</th>
<th>Parameter</th>
<th>Equipment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Reaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Reaction</td>
<td>Add</td>
<td>Add SJH-01-223 (0.1 g, 0.132 mmol) to a 5 mL round bottom flask with a reflux condenser with a schlenk adaptor</td>
<td>SJH-01-223 stoichiometry table row</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Reaction</td>
<td>Add</td>
<td>Add Cu(OTf)$_2$ (0.005 g, 0.013 mmol) and put under a N$_2$ environment</td>
<td>Cu(OTf)$_2$ stoichiometry table row</td>
<td></td>
<td>round bottom flask</td>
</tr>
<tr>
<td>4</td>
<td>Reaction</td>
<td>Dissolve</td>
<td>Dissolve the benzaldehyde (0.163 g, 0.791 mmol) in DCE (1.3 mL) in a vial</td>
<td>Benzaldehyde stoichiometry table row</td>
<td></td>
<td>vial</td>
</tr>
<tr>
<td>5</td>
<td>Reaction</td>
<td>Transfer</td>
<td>Transfer this solution via syringe to the reaction round bottom flask</td>
<td></td>
<td></td>
<td>syringe</td>
</tr>
</tbody>
</table>

Source: Royal Society of Chemistry
"An Idealised Scientific Research Activity Lifecycle Model"

Source: Simon Coles, Univ. Southampton
Honoring Digital Data Needs

- DMP
- Documentation
- Storage
- Backup
- Visualization
- Archive
- QC
- Discovery
- Annotate

Transport
Escort
Feeding
Care
Oversight

- Metadata
- File Formats
- Validation Software
- Identifiers
- Repositories
- Semantic Frameworks
- Review Tracking

http://www.twistypuzzles.com/
Honoring Researcher Data Needs

It's hard to decide what a typical day is, so I think I'll just describe my day today. This morning I arrived at work at around 7:15 am. I checked my email, put the finishing touches on my lecture slides (which I finished tweaking around 11pm last night), gathered the handouts I printed yesterday, and made my way down to the lecture hall for class. I taught two back-to-back, one-hour sections of General Chemistry, then returned to my office to respond to some emails and modify my lecture slides to post on the course website. At 11am, I had a department meeting where we discussed the progress of our current faculty search. At noon I met with my research group (which consists of 3 senior thesis students) to check in on their lab progress over the last couple of weeks... During the meeting, I stole glances at my email to see if I had heard back from the funding agency that was supposed to announce their decision on my grant application in "mid November"... Then I headed over to the dining hall at 12:30 to catch the last half of the monthly Junior Faculty lunch meeting, where I learned about the college's current development campaign. At 1pm I huddled back to my office to meet with a student who had done poorly on the first quiz in Gen Chem. Then I helped my research student set up a nitrogen tank in the lab before meeting with more Gen Chem students during my office hours from 2-4. At 4pm I had a phone conference with a representative from an online learning site who wants me to promote his product to students in my organic chemistry class next semester. At 5:00, I logged into the college's Academic Advisory site to send notices to students at risk of failing General Chemistry. For the last hour, I've been working through my inbox to reply to assorted emails (like this one) before going home for dinner.

Making the Case for Researchers

• Improved accuracy of data collection
• Streamlined and verifiable analyses
• Easier processing of outputs (reports)
• Improved lab group communication
• More consistent record keeping

-> already happening across large chemical companies

• In the context of responsive service
• Chance favors only the prepared mind*
• Pearl growing and avalanches

*[Pasteur]*
Does CDM Make Sense to Chemists?

- Does it look like chemistry research?
- Is it faster and better for their research methods than current dm and notebook practices?
- Will it improve intra-lab communication and instill better habits in students?
- Will it reinforce scientific integrity?
- Will it increase visibility and chance of funding?
- Is it robust and accepted by greater chemistry enterprise (e.g., patents, industry standards)?
Thanks!

- Sam Hein, Cornell Chemistry & Chemical Biology
- Cerys Whilloughby and Simon Coles, University of Southampton Chemistry
- Ye Li, University of Michigan Science Library

- Cornell University Library
- Cornell University chemistry labs
- Royal Society of Chemistry
- University of Southampton
- ACS CINF data discussions
- RDA Chemistry Research Data Interest Group
Questions?