Data Science Training for the Future: Building a Carpentries Consortium

Julie Goldman
Harvard Medical School
New England Software Carpentry Library Consortium

Follow this and additional works at: https://escholarship.umassmed.edu/ner

Part of the Library and Information Science Commons, and the Public Health Commons
Creative Commons Attribution 4.0 License
This work is licensed under a Creative Commons Attribution 4.0 License.

Repository Citation
https://escholarship.umassmed.edu/ner/68

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in National Network of Libraries of Medicine New England Region (NNLM NER) Repository by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Data Science Training for the Future

Building a Carpentries Consortium

Julie Goldman on behalf of the New England Software Carpentry Library Consortium
NNLM NER eScience Forum | March 29, 2019
Community of Practice

- Desire for more advanced/technical training for librarians
  - For teaching students
  - For librarians to gain skills too
- Wanted training to teach coding and data skills
  - Carpentries offered this opportunity
  - Long waiting list for training
  - Shorter with a membership
- Expensive to go alone
  - Less risk as a group
- Conversation at RDAP 2017
  - Could we bring together institutional contributions?
  - What would be the benefits to our institutions?
The Carpentries

Non-profit organization and international community of volunteers that develop lessons, train instructors, and organize workshops in data management and software development best practices for and by researchers and librarians across disciplines.

https://carpentries.org
Workshops

- 2-days, active/hands-on learning
- Trained/certified instructors
- Feedback to learners throughout the workshop (Surveys, post-its, +/-)
- Friendly learning environment (Code of Conduct)
Who takes workshops?

66% of Data Carpentry workshop attendees are early career

What’s the impact?

85% report they gained more confidence after a workshop

Analysis of Software and Data Carpentry’s Pre- and Post-Workshop Surveys
https://doi.org/10.5281/zenodo.1325463

Analysis of the Carpentries Long-Term Feedback Survey: Perception of Workshop Impact
nuary_long_term_report.html
Forming a Consortium

New England Software Carpentry Library Consortium (NESCLiC)

- Gathered a group of like minded individuals
- Library and IT staff from academic institutions

Administrative Steps

- Written proposals for each school
- Signing/Billing Institution & Letters of Intent
- Gold Tier Membership
  - $15,000 per year
  - $1000 per instructor trained

MIT Workshop

NEASIST & MIT host a Library Carpentry workshop

NESCLiC Formed

First NESCLiC meeting held; first-year schools chosen

Joined Carpentries

NESCLiC officially joins Software Carpentry as a member organization

1st Workshop

NESCLiC members attend SWC workshop at Tufts

Instructor Training

NESCLiC members attend Instructor Training at Brown & Amherst

Year Two

NESCLiC began planning year-two; reached out to new schools for interest
NESCLiC Members

**Brown University**
Andrew Creamer
Bruce Boucek
Patrick Rashleigh
Kelsey Sawyer
Ashley Champagne

**Tufts University**
Kristin Lee
Ari Gofman
Kyle Monahan
Susan Remondi

**University of Connecticut**
Jennifer Chaput

**Dartmouth**
Lora Leiligdon
Christian Darabos
James Adams
Pamela Bagley
Stephen Gaughan

**Yale University**
Joshua Dull
Catherine DeRose
Kate Nyhan
David Cirella
Barbara Esty
Sawyer Newman

**University of Massachusetts**
Thea Atwood
Ann Kardos
Erin Jerome

**Harvard Library**
Julie Goldman
Ceilyn Boyd
Meghan Kerr

**Mount Holyoke**
Sarah Oelker
James Burke

**University of Massachusetts Medical School**
Sally Gore
Tess Grynoch
Instructor Training

- Virtual instruction in two locations
  - Brown & UMass Amherst
- Introduce you to evidence-based best-practices of teaching
  - A different way to look at pedagogy!
- Create a positive environment for learners at your workshops
- Provide opportunities for you to practice and build your teaching skills
- Help you become integrated into the Carpentries community
Example Lesson: OpenRefine

Library Carpentry OpenRefine

Introduction to OpenRefine

What is OpenRefine?

OpenRefine is described as "a power tool for working with messy data" by David Huynh - but what does this mean? It is probably easiest to describe the kinds of data OpenRefine is good at working with and the sorts of problems it can help you solve.

OpenRefine is most useful where you have data in a simple tabular format such as a spreadsheet, a comma separated values file (csv) or a tab delimited file (tsv) but with internal inconsistencies either in data formats, or where data appears, or in terminology used. OpenRefine can be used to standardize and clean data across your file. It can help you:

- Get an overview of a data set
- Resolve inconsistencies in a data set, for example standardizing date formatting
- Help you split data up into more granular parts, for example splitting up cells with multiple authors into separate cells
- Match local data up to other data sets, for example in matching local subjects against the Library of Congress Subject Headings
- Enhance a data set with data from other sources

Some common scenarios might be:

- Where you want to know how many times a particular value (name, publisher, subject) appears in a column in your data
- Where you want to know how values are distributed across your whole data set
- Where you have a lot of dates which are formatted in different ways, and want to change all the dates in the list to a single common data format. For example:

<table>
<thead>
<tr>
<th>Data you have</th>
<th>Desired data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st January 2014</td>
<td>2014-01-01</td>
</tr>
<tr>
<td>01/01/2014</td>
<td>2014-01-01</td>
</tr>
<tr>
<td>Jan 1 2014</td>
<td>2014-01-01</td>
</tr>
<tr>
<td>2014-01-01</td>
<td>2014-01-01</td>
</tr>
</tbody>
</table>

- Where you have a lot of names or terms that differ from each other but refer to the same people, places or concepts. For example:

<table>
<thead>
<tr>
<th>Data you have</th>
<th>Desired data</th>
</tr>
</thead>
<tbody>
<tr>
<td>London</td>
<td>London</td>
</tr>
<tr>
<td>London</td>
<td>London</td>
</tr>
</tbody>
</table>

https://librarycarpentry.org/lc-open-refine
Unexpected Benefits

- Not just instructor training
  - Keep data/coding skills sharp
  - A group of people who help each other
- Expanded Local Network
  - Communicate and collaborate on other projects
- Collaboration with other regional organizations
  - NEASIST
  - NNLM New England Region

Moving Forward

- The Carpentries
  - Evolving as an organization
- Changing roles within libraries
  - Evolution of Library Carpentry
  - More ‘non-traditional’ library training
- Train-the-Trainer
  - Train others in Carpentries pedagogy
- Further participation in lesson development & curriculum areas
  - Becoming lesson maintainers
  - Digital Humanities at Yale
  - Biodiversity at Harvard
The New England Software Carpentry Library Consortium or NESCLiC is an association of academic libraries, joined as a community of practice focused on building data science skills in research computing, and extracting, wrangling, storing, analyzing, and visualizing data. NESCLiC offers member institutions to share in a Gold-level membership with the Carpentries.

Our Model

NESCLiC member institutions share a Gold-level Carpentries membership, which affords NESCLiC B free-coordinated workshops, a 20% discount on the cost for additional coordinated workshops, no charge for self-organized workshops, and fifteen instructors trained, per year. NESCLiC provides the opportunity for members to:

- Develop instructional materials
- Build a network of data-scientist researchers and instructors
- Refine skills by participating as instructors at other member institutions

Financial commitment: Cost of shared Carpentry membership ($1000/person) plus training expenses (travel or hosting)

Time commitment: 2-2 day training workshops plus committee work. NESCLiC has established bylaws and members contribute to the consortium by participating in one of several roles:

- Carpentry liaison
- Membership/funding
- Assessment
- Outreach
- Workshop coordination

Become a Member

Interested in joining NESCLiC?

Email: neslic@googlegroups.com

Workshops

Check back for future workshops offered by NESCLiC members.

This list includes official Carpentries' workshops and workshops adopted from the Carpentries' materials where NESCLiC members served as hosts, instructors, or helpers.

<table>
<thead>
<tr>
<th>Venue</th>
<th>Dates</th>
<th>Lessons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tufts University</td>
<td>April 12-13, 2019</td>
<td>Unix, Git, Programming with Python</td>
</tr>
<tr>
<td>Tufts University</td>
<td>February 15-16, 2019</td>
<td>Data Organization in Spreadsheets, OpenRefine for Data Cleaning, R for Social Scientists</td>
</tr>
<tr>
<td>Brown University</td>
<td>January 16-17, 2019</td>
<td>Data Organization in Spreadsheets, OpenRefine for Data Cleaning, Introduction to R, Data Analysis and Visualization in R, Data Management with SQL</td>
</tr>
<tr>
<td>Harvard University</td>
<td>January 15-16, 2019</td>
<td>Data Intro for Librarians, Shell Lessons for Librarians, Tidy Data for Librarians, OpenRefine for Librarians</td>
</tr>
<tr>
<td>UMass Amherst</td>
<td>January 8-9, 2019</td>
<td>Unix, Python, Git</td>
</tr>
<tr>
<td>Dartmouth College</td>
<td>November 29-30, 2018</td>
<td>Unix, Git, Python</td>
</tr>
<tr>
<td>Tufts University</td>
<td>November 9-10, 2018</td>
<td>Unix, Git, Programming with Python</td>
</tr>
<tr>
<td>Brown University Library</td>
<td>October 22-23, 2018</td>
<td>Data Intro for Librarians, Shell Lessons for Librarians, Git Intro for Librarians, OpenRefine for Librarians</td>
</tr>
<tr>
<td>Yale University Library</td>
<td>August 15-16, 2018</td>
<td>Tidy Data for Librarians, Shell Lessons for Librarians, OpenRefine for Librarians, Python Intro for Librarians</td>
</tr>
</tbody>
</table>