Software in Scholarship

Daniel S. Katz
Assistant Director for Scientific Software & Applications, NCSA
Research Associate Professor, CS
Research Associate Professor, ECE
Research Associate Professor, iSchool
dskatz@illinois.edu, d.katz@ieee.org, @danielskatz

FORCE11 Scholarly Communications Institute
WT02: Software Citation: Principles, Usage, Benefits, and Challenges
2–3 August 2017
Data Science vs Computational Science

- Oversimplified definitions and examples:
  - Data science - trying to use data to produce an understanding of something
    - Does drug X or drug Y better cure disease A? Give some people with disease A drug X, and some drug Y, use data to see what happens over time
  - Computational science - trying to use models and simulations to understand something
    - Build models of the molecular structure of drugs X and Y. Build a model for how the body acts with disease A, and without it. Combine the models to see how drugs X and Y interact with the body with disease A. Which has more effect in moving the body model towards the model without disease A.
Computational science research

Create Hypothesis

Acquire Resources (e.g., Funding, Software, Data)

Perform Research (Build Software & Data)

Gain Recognition

Publish Results (e.g., Paper, Book, Software, Data)

Knowledge
Data science research

1. Create Hypothesis
2. Acquire Resources (e.g., Funding, Software, Data)
3. Perform Research (Build Software & Data)
4. Publish Results (e.g., Paper, Book, Software, Data)
5. Gain Recognition
6. Acquire Resources (Data)

Software
Software vs. data

- Software is data, but it is not just data
- Data (in computing and information science): anything that can be processed by a computer
- Software: special kind of data that can be a creative, executable tool that operates on data
- Software & data are similar in with regard to credit and metrics, and both traditionally have not been cited in publications

Software in research

- **Claim:** software (including services) essential for the bulk of research
- **Evidence from surveys**
  - UK academics at Russell Group Universities (2014)
  - Members of (US) National Postdoctoral Research Association (2017)
  - My research would not be possible without software: 67% / 63% (UK/US)
  - My research would be possible but harder: 21% / 31%
  - It would make no difference: 10% / 6%

S. Hettrick, “It's impossible to conduct research without software, say 7 out of 10 UK researchers,” Software Sustainability Institute, 2014. Available at: https://www.software.ac.uk/blog/2016-09-12-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers


Software in scholarship

• Claim: software (including services) essential for the bulk of research
• Evidence from journals:
  • About half the papers in recent issues of Science were software-intensive projects
  • In Nature Jan–Mar 2017, software mentioned in 32 of 40 research articles
    • Average of 6.5 software packages mentioned per article

Why is capturing software in research useful?

• Scientific research is becoming:
  • More open – scientists want to collaborate; want/need to share
  • More digital – outputs such as software and data; easier to share
• Significant time spent developing software & data
• Efforts not recognized or rewarded
• Citations for papers systematically collected, metrics built
  • But not for software (& data)

• Want to appropriately reward software developers
• Want to better understand research by including software
How to better measure software usage

• Citation system was created for papers/books
• We need to either/both
  1. Jam software into current citation system
  2. Rework citation system
     • Most people focus on 1; 2 is very hard.
• Challenge: not just how to identify software in a paper
  • How to identify software used within research process
Software citation today

- Software and other digital resources currently appear in publications in very inconsistent ways
- Howison: random sample of 90 articles in the biology literature -> 7 different ways that software was mentioned

<table>
<thead>
<tr>
<th>Mention Type</th>
<th>Count (n=286)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cite to publication</td>
<td>105</td>
<td>37%</td>
</tr>
<tr>
<td>Cite to users manual</td>
<td>6</td>
<td>2%</td>
</tr>
<tr>
<td>Cite to name or website</td>
<td>15</td>
<td>5%</td>
</tr>
<tr>
<td>Instrument-like</td>
<td>53</td>
<td>19%</td>
</tr>
<tr>
<td>URL in text</td>
<td>13</td>
<td>5%</td>
</tr>
<tr>
<td>In-text name only</td>
<td>90</td>
<td>31%</td>
</tr>
<tr>
<td>Not even name</td>
<td>4</td>
<td>1%</td>
</tr>
</tbody>
</table>

- Studies on data and facility citation -> similar results