Research Data Management – An Institutional Perspective

Patricia Rankin
Associate Vice Chancellor for Research
Outline

• **Overview of University of Colorado Boulder**
  – AAU member, Research Intensive
  – Limited State Support

• **Open Questions**
  – Shifting Paradigms

• **Some Ideas**
  – Carrots work better than sticks….
Sponsored Research at CU

- $351.9 million in federally sponsored research (FY 2013)
- Annual research awards have roughly doubled over the last ten years
- Publish about 4,800 articles a year
- Lead the publics in NASA funding
- More atmospheric scientists per square mile than anywhere else in US
- Undergraduate (800+) and graduate students (1,160) participate in research

41% of sponsored research revenue goes to local salaries.
Approximately half of U.S. research output is generated by 25 universities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>469,201</td>
<td>48.5</td>
<td>AAU</td>
<td>905,522</td>
<td>56.1</td>
</tr>
<tr>
<td>25,630</td>
<td>2.65</td>
<td>Harvard University</td>
<td>68,146</td>
<td>4.22</td>
</tr>
<tr>
<td>13,071</td>
<td>1.35</td>
<td>University of Michigan System</td>
<td>33,084</td>
<td>2.05</td>
</tr>
<tr>
<td>10,567</td>
<td>1.09</td>
<td>Johns Hopkins University</td>
<td>31,503</td>
<td>1.95</td>
</tr>
<tr>
<td>16,941</td>
<td>1.75</td>
<td>University of California, Los Angeles</td>
<td>31,108</td>
<td>1.93</td>
</tr>
<tr>
<td>12,841</td>
<td>1.33</td>
<td>University of Washington System</td>
<td>30,320</td>
<td>1.88</td>
</tr>
<tr>
<td>13,366</td>
<td>1.38</td>
<td>Stanford University</td>
<td>28,318</td>
<td>1.75</td>
</tr>
<tr>
<td>10,248</td>
<td>1.06</td>
<td>University of California, San Diego</td>
<td>27,265</td>
<td>1.69</td>
</tr>
<tr>
<td>15,176</td>
<td>1.57</td>
<td>University of California, Berkeley</td>
<td>27,021</td>
<td>1.67</td>
</tr>
<tr>
<td>11,646</td>
<td>1.20</td>
<td>University of Pennsylvania</td>
<td>26,579</td>
<td>1.65</td>
</tr>
<tr>
<td>10,691</td>
<td>1.10</td>
<td>Columbia University</td>
<td>26,427</td>
<td>1.64</td>
</tr>
<tr>
<td>10,219</td>
<td>1.06</td>
<td>University of Maryland System</td>
<td>25,844</td>
<td>1.60</td>
</tr>
<tr>
<td>14,419</td>
<td>1.49</td>
<td>University of Minnesota System</td>
<td>25,497</td>
<td>1.58</td>
</tr>
<tr>
<td>13,919</td>
<td>1.44</td>
<td>University of Wisconsin, Madison</td>
<td>24,553</td>
<td>1.52</td>
</tr>
<tr>
<td>14,222</td>
<td>1.47</td>
<td>Cornell University</td>
<td>23,483</td>
<td>1.45</td>
</tr>
<tr>
<td>10,166</td>
<td>1.05</td>
<td>University of Florida</td>
<td>23,226</td>
<td>1.44</td>
</tr>
<tr>
<td>7,483</td>
<td>0.77</td>
<td>University of Pittsburgh</td>
<td>22,457</td>
<td>1.39</td>
</tr>
<tr>
<td>9,490</td>
<td>0.98</td>
<td>University of California, Davis</td>
<td>22,362</td>
<td>1.38</td>
</tr>
<tr>
<td>7,880</td>
<td>0.81</td>
<td>Duke University</td>
<td>21,954</td>
<td>1.36</td>
</tr>
<tr>
<td>8,715</td>
<td>0.90</td>
<td>Penn State University System</td>
<td>21,689</td>
<td>1.34</td>
</tr>
<tr>
<td>11,150</td>
<td>1.15</td>
<td>Yale University</td>
<td>21,676</td>
<td>1.34</td>
</tr>
<tr>
<td>8,792</td>
<td>0.91</td>
<td>Ohio State University</td>
<td>21,380</td>
<td>1.32</td>
</tr>
<tr>
<td>8,889</td>
<td>0.92</td>
<td>University of Colorado System</td>
<td>21,066</td>
<td>1.30</td>
</tr>
<tr>
<td>10,027</td>
<td>1.04</td>
<td>University of California, San Francisco</td>
<td>20,691</td>
<td>1.28</td>
</tr>
<tr>
<td>11,651</td>
<td>1.20</td>
<td>MIT</td>
<td>20,609</td>
<td>1.28</td>
</tr>
<tr>
<td>6,975</td>
<td>0.72</td>
<td>Texas A&amp;M University System</td>
<td>19,432</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Approximately half of U.S. research citations generated by 19 universities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT</td>
<td>2.14</td>
<td>2.16</td>
<td>2.28</td>
</tr>
<tr>
<td>Caltech</td>
<td>2.13</td>
<td>2.02</td>
<td>2.18</td>
</tr>
<tr>
<td>Princeton University</td>
<td>2.19</td>
<td>2.07</td>
<td>2.11</td>
</tr>
<tr>
<td>University of California, Santa Barbara</td>
<td>1.75</td>
<td>2.28</td>
<td>2.04</td>
</tr>
<tr>
<td>Stanford University</td>
<td>2.05</td>
<td>2.08</td>
<td>1.96</td>
</tr>
<tr>
<td>Harvard University</td>
<td>1.98</td>
<td>2.14</td>
<td>1.94</td>
</tr>
<tr>
<td>University of California, Berkeley</td>
<td>1.79</td>
<td>1.77</td>
<td>1.92</td>
</tr>
<tr>
<td>University of Colorado, Boulder</td>
<td>1.67</td>
<td>1.65</td>
<td>1.86</td>
</tr>
<tr>
<td>University of Chicago</td>
<td>1.98</td>
<td>1.92</td>
<td>1.85</td>
</tr>
<tr>
<td>University of Washington System</td>
<td>1.78</td>
<td>1.76</td>
<td>1.82</td>
</tr>
<tr>
<td>University of Pennsylvania</td>
<td>1.62</td>
<td>1.73</td>
<td>1.77</td>
</tr>
<tr>
<td>University of California, San Francisco</td>
<td>1.86</td>
<td>1.89</td>
<td>1.76</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>1.69</td>
<td>1.85</td>
<td>1.74</td>
</tr>
<tr>
<td>Columbia University</td>
<td>1.70</td>
<td>1.83</td>
<td>1.74</td>
</tr>
<tr>
<td>University of California, Los Angeles</td>
<td>1.62</td>
<td>1.61</td>
<td>1.74</td>
</tr>
<tr>
<td>Northwestern University</td>
<td>1.62</td>
<td>1.69</td>
<td>1.73</td>
</tr>
<tr>
<td>Boston University</td>
<td>1.35</td>
<td>1.59</td>
<td>1.71</td>
</tr>
<tr>
<td>Yale University</td>
<td>1.91</td>
<td>1.89</td>
<td>1.71</td>
</tr>
<tr>
<td>University of Rochester</td>
<td>1.46</td>
<td>1.60</td>
<td>1.71</td>
</tr>
<tr>
<td>U.S. UNIVERSITY average</td>
<td>1.37</td>
<td>1.40</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Research Initiatives

Key Areas

- Aerospace Sciences and Engineering
- Biotechnology and Biosciences
- Renewable and Sustainable Energy
- Geosciences/Environmental Sciences
- Computational Sciences
- STEM Education (Science, Technology, Engineering, Mathematics)

World-class interdisciplinary research at CU-Boulder advances society and the economy.
Computational Sciences

A Broad Spectrum of Faculty Partner with Universities, Government and Industry in:

- High-performance scientific computing
- Artificial intelligence
- Nanotechnology
- Next-generation internet
- Biotechnology
- Genomics
- Fluid dynamics
- Climate modeling
- Laser sciences

A great deal of research in the science and engineering disciplines is driven by simulations, requiring significant advances in computational technologies.
Data sets include

- Artifacts from Indian tribes in arctic regions
- Bee population studies
- Sounds from endangered languages
- NMR scans
- Ice Cores
- Collision data from LCH Higg’s search and reconstructed events
- Musical Performances
- Genomic studies
- Simulations of likely material behaviors
General guidelines

• Major collaborations and networks tend to have discipline specific archives
• Some agencies require data to be stored in specific repositories
• In many cases computing/data management is delegated to a postdoc or graduate student (aka “technically savvy native”)
• Many assumed technically savvy natives are not (and often information does not cross the barrier when a postdoc or graduate student moves on)
Big Data

• Universities are becoming major consumers of analytics
  – Research Productivity/Rankings
  – Student Retention “Smart” systems
• What questions can we answer because we have
  – Access to larger data sets?
  – Better ways to connect data sets?
  – More compute power?
• Who gets to use the data?
• Who sets the standards for allowed use?
Changing Times

• Federal funding of basic research is increasingly becoming a political issue
  – Economic Driver/Translational Research
  – Value of Social Science
  – Distrust of “expert” opinion
    • *Data produced in research funded by the public should be available to public*
    • *Results of research should be broadly disseminated/easily available*
      – Data Management Plans, Open Access
Details Matter

• NSF has indicated that people can budget for data management plans in their proposal requests… *but*
  – Budgets not growing to accommodate extra demands
  – Not clear that quality of data management plans matters to many reviewers yet
  – Communities still working to define data management plan standards
Questions I have

• How long do researchers get to keep data private?
  – IP issues, publication rights
• How do we determine a sensible amount of time to preserve data for?
  – Some simulations that took a few months some years back can be redone in a fraction of the time
  – The raw data may require analysis code that has evolved over time
• What happens if a researcher’s data management plan requires campus level resources that they don’t ask for in advance?
• Who pays once grant has ended?
More questions

• And how do we deal with publications based on data that are not high quality/do not meet discipline standards?
  – Statistics
  – Data selection
  – Equal time or proportional representation?
    • **BBC in UK has changed policy on allowing all sides in a debate to speak...**
Open Access

• How does this impact tenure/promotion?
  – How do we figure out merit factors for open access journals?
    • Peer Review
    • Quality of other papers published
    • Long term reliability, reputation, accumulated social capital
  – How does providing a data set weigh towards tenure/promotion?
    • Reward what we value

• How do we sustain?
  – $2K publishing fee multiplied by 4,800 articles…not going to work
Peerage of Science

- [www.Peerageofscience.org](http://www.Peerageofscience.org)
- Interesting model
  - Authors submit manuscripts and deadlines for four stages
    - Reviews
    - *Peer Review of Peer Review (reviews get a quality index)*
    - *Manuscript Revision*
    - *Final Evaluation* – breadth, impact, originality, data, methods, inference, literature coverage – leads to a quality index
Next Steps

• Subscribing journals can offer to publish or authors can choose to submit to another journal (that journal can have access to existing reviews)
• Quality indexes include article quality, number of reviews, quality of reviews
• Issues
  – Seems mostly bio related right now
  – Early days – will be interesting to see adoption rate
What we are doing at CU Boulder?

• Research Computing – reports to Office of Information Technology and Office of Vice Chancellor for Research
  – Regular meetings between Head of Research Computing and Associate Vice Chancellor for Research
  – Regular meetings of both with Library leadership
• Research Data Management Task Force
• Data Management Audit
Research Data Advisory Committee

• Mix of disciplines and roles
  – Co-chairs from English, Evolutionary Biology
  – Research Staff, Library Staff
  – Looking to add post-doc, graduate student
• Goal – to develop definitions (what is “data”), policies, best practices, campus outreach
Data Management Plans

• Now required for campus competitions (competitions run to select CU nominee if have a restriction on allowed number of proposals)

• Seed grant competition
  – About 80 proposals from across campus
  – RDAC Committee analyzed data managements plans
    • Not a selection criteria this year – will be next
    • Lots of information on current state of data management practices – lots of room to improve
So -

• Running a competition to search for the best data management plans
  – 5 broad areas – including arts and humanities, social sciences
  – Open to graduate students, post docs, and faculty
  – Encouraging use of tools available to develop data plans, review of best practices documents developed from seed grant study
Closing Words

• Data Management is an emerging field
• Interesting mix of technical, social issues
  – How do we store
  – What do we store
  – Why do we store
  – How do we use
  – When do we delete
• Important to form broad alliances