Co-authorship as a means of crediting data creators

Gary F. Simons
SIL International
gary_simons@sil.org
The archiving conundrum

- Given the relentless
  - entropy that degrades our field recordings, and
  - technology innovation that brings rapid obsolescence
- We know that
  - the field recordings are just as endangered as the languages they document, unless
  - they are entrusted to archives for long-term preservation
- So why do most field recordings remain unarchived?
  - “It is too much work for too little academic credit.”
  - “If I let the stuff go, someone else will publish before I do.”
An attempt to address this

- LSA Resolution Recognizing the Scholarly Merit of Language Documentation (2010)
  - “Whereas [discussion of the value of language documentation],
  - “Therefore the Linguistic Society of America supports the recognition of these materials as scholarly contributions to be given weight in the awarding of advanced degrees and in decisions on hiring, tenure, and promotion of faculty."
  - http://www.linguisticsociety.org/resource/resolution-recognizing-scholarly-merit-language-documentation
How do we make this real?

- The currency of the academic rewards system is
  - Authorship of scholarly works
  - Citation of those works by others as a measure of impact

- Thus the premise of this workshop
  - We need to ensure that archived language documentation is formally treated as scholarly work with authorship credit to the compilers and impact credit being captured through citations

- Could we go even further to give more credit by following other disciplines who credit data creators as co-authors?
A personal encounter with other rules

- My son’s PhD work in neuroscience:
  - Author contributions: D.L.S. and S.M.W. designed research; **D.L.S. performed research**; D.L.S., S.L.B., and W.W.H. contributed new reagents/analytic tools; D.L.S. and S.M.W. analyzed data; and **D.L.S. wrote the paper**.

- If this were linguistics, there would be just one author — the one who performed the research and wrote the paper.
What about the other authors?

- SMW is credited with helping to design the research and analyze the data — he was the Ph.D. advisor.
- SLB and WWH contributed a new reagent. From genetic material provided by DLS, they used the method they had previously published to grow and purify the viral vector:
  - “The plasmid transfection method using HEK293 cells as previously described (39) was used to produce and purify scAAV2/5 vectors carrying either Bbs4 or GFP.”
- They did not participate in the writing or research, but their intellectual contribution was indispensable and foundational.
The rise of co-authorship (from PubMed database)

https://thewinnower.com/papers/the-rising-trend-in-authorship
## Authors per publication by field


<table>
<thead>
<tr>
<th>Field</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical sciences</td>
<td>12</td>
</tr>
<tr>
<td>Biology</td>
<td>11</td>
</tr>
<tr>
<td>Medicine</td>
<td>10</td>
</tr>
<tr>
<td>Chemistry</td>
<td>8</td>
</tr>
<tr>
<td>Psychology</td>
<td>6</td>
</tr>
<tr>
<td>Engineering</td>
<td>4</td>
</tr>
<tr>
<td>Mathematics</td>
<td>2</td>
</tr>
<tr>
<td>Economics</td>
<td>16</td>
</tr>
<tr>
<td>History and Archaeology</td>
<td>15</td>
</tr>
<tr>
<td>Language and Culture</td>
<td>14</td>
</tr>
<tr>
<td>Philosophy and Religion</td>
<td>13</td>
</tr>
</tbody>
</table>
Another story: closer to home (at least, disciplinarily)


- Based on over 5 decades of work by the Headlands: including every birth, marriage, divorce, death, and in- and out-migration from 1950 to 2010

- The complete database is fully documented and free for anyone to download and use in their own research
A publication based on this corpus

- Cody Ross et al. In press. Bayesian analysis of Agta demography through the transition from foragers to landless peasantry (Eastern Luzon, Philippines). *Evolution and Human Behavior*
- The lead author sent the manuscript to Headland to review
- After Headland gave useful feedback, the three authors of the database were invited to be co-authors of the paper
- Headland was reluctant at first since that was not the norm he knew, but for the lead author it was the new norm and a formal “Author contributions” statement satisfied Headland
- Co-authorship does seem more appropriate than citation
  - The Agta database is not just an idea to be credited; it is the whole foundation of the work
Who is an author?

- The most widely followed criteria are those developed by the International Committee of Medical Journal Editors (ICMJE).
- Principles: intellectual contribution and responsibility for results.

Authors must meet all four conditions in order to be listed.

1. Make substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data.
2. Drafting the article or revising it critically for important intellectual content.
3. Final approval of the version to be published.
4. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.
What if we followed suit?

- The author of an analysis would be obligated to offer co-authorship to the corpus compiler.
- The corpus compiler would need not be involved in drafting the paper, but would be involved in “revising it critically for important intellectual content.”
- Corpus compilers would not only get impact credit for citations of the corpus itself but also for citations of work for which it was an indispensable foundation.
- This would provide greater incentive for timely archiving of primary language documentation data that is easy to discover and access.
Data citation & attribution in the Digital Humanities

Tanya E. Clement
tclement@ischool.utexas.edu
National Academies

• For Attribution -- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop (2012)

• http://www.nap.edu/read/13564/chapter/1

• 9- Data Citation in the Humanities: What’s the Problem?

Michael Sperberg-McQueen
Black Mesa Technologies
What counts as data in the humanities?

- digitized editions of major works;
- transcriptions of manuscripts;
- thematic collections (e.g., author, period, genre);
- language corpora (balanced or opportunistic; monolingual or multilingual [parallel structure or parallel-text translation equivalents]);
- images of artworks (e.g., Rossetti, Blake, DeYoung Museum ImageBase); and
- maps.
Some problems in humanities data citation

- Citation standards (How?)
- Reliable metadata to cite (Creators aren’t noted.)
- A desire for turn-key systems (not easy to make citations)
- Fear of copyright issues (What if the thing I’m citing is copyrwrong?)
- Anti-scientism (We don’t cite data sets.)
- Lack of citation chains (No one has cited it before.)
- Versioning (which one am I citing?)
- Quiddity (which thing or part am I citing?)
- Longevity (what if it disappears?)
But, before we think about citation and attribution . . .
let’s ask, what’s the data? [What’s the level of granularity . . . ? Issues of foundational, secondary, tertiary data . . . ]
HathiTrust Library

- Founded in 2008

- Grew out of large-scale digitization initiative at academic research libraries
  - Google Books project

- Over 100 member institutions (nationally and globally) continue to contribute
HathiTrust

HathiTrust Digital Library

Access → Formulate

Preserve → Gather

Collect → Analyze

Disseminate → Produce

HathiTrust Research Center
HathiTrust

HathiTrust Digital Library

Access

Preserve

Collect

Disseminate

Formulate

Gather

Analyze

Produce

HathiTrust Research Center
HathiTrust Digital Library

13+ million volumes | 5+ million book titles | 29k serial titles | 3+ billion pages

Languages

- English 50%
- German 9%
- French 7%
- Spanish 5%
- Other 29%

Around 50% of volumes are in English. Many other languages included as well.
HathiTrust Digital Library

Dates span the 15th - 21st centuries

70% in copyright or undetermined | 30% out of copyright
HathiTrust
“Reading” with computers

Digitized text

Computational methods

Analysis

HathiTrust Research Center

at scale
HTRC Algorithms ...make Data?

1. Extracted Features
   Rsync Script Generator
2. MARC Downloader
3. Meandre Classification
   Naive Bayes
4. Meandre Dunning Log-Likelihood to Tag Cloud
5. Meandre OpenNLP
   Date Entities to Simile
6. Meandre OpenNLP
   Entities List
7. Meandre OpenNLP
   Report per Volume
8. Meandre Tag Cloud
9. Meandre Tag Cloud
   with Cleaning
10. Meandre Topic
    Modeling
11. Simple Deployable
    Word Count
Welcome to HiPSTAS

The HiPSTAS application is now available!
DEADLINE EXTENDED to February 1, 2013
Please apply.

Welcome to HiPSTAS (High Performance Sound Technologies for Access and Scholarship).

http://www.hipstas.org
HiPSTAS team

1. Tanya Clement, [PI] Assistant Professor, University of Texas at Austin

2. Loretta Auvil [Co-PI] Senior Project Coordinator at the Illinois Informatics Institute (I3) at the University of Illinois at Urbana-Champaign

3. David Tcheng [Co-PI] Research Scientist at I3; ARLO developer

4. Tony Borries, Research Programmer working as a consultant with I3; ARLO programmer

5. David Enstrom, Biologist, University of Illinois at Urbana-Champaign; consultant
HiPSTAS Institute, 2013-2014

- 9 librarians and archivists
- 8 humanities scholars
- 3 advanced graduate students in humanities and information science
Participating collections

- poetry from PennSound at the University of Pennsylvania 30,000 audio files

- folklore at the Dolph the Briscoe Center for American History at UT Austin, 57 feet of tapes (reels and audiocassettes)

- storytelling traditions at the Native American Projects (NAP) at the American Philosophical Society in Philadelphia, 50 tribes, 3,000 hours
Other Collections of interest to HiPSTAS Participants

- Field recordings (200,000 recordings) American Folklife Center, Library of Congress
- 30,000 hours, Oral histories, Storycorps
- Speeches in the Southern Christian Leadership Conference recordings, Emory University
- 700 recordings in the Elliston Poetry Collection at the University of Cincinnati
HiPSTAS: primary goals

To develop a virtual research environment in which users can better access and analyze spoken word collections of interest to humanists through:

1. an assessment of scholarly requirements for analyzing sound
2. an assessment of technological infrastructures needed to support discovery
3. preliminary tests that demonstrate the efficacy of using such tools in humanities scholarship
4. A freely available, open-source, API-driven version for general use
ARLO (Adaptive Recognition with Layered Optimization)

HZ, a unit of frequency

Energy represented by a heat based color scheme.
- **White** – hottest, most intense
- **Yellow**
- **Red**
- **Green**
- **Blue** – coolest, least intense

![Image showing color scheme with HZ units and time axis]
Choose parameters
Label examples
Divide corpus into frames

.5 second time slices in ARLO
Create the model

```
<table>
<thead>
<tr>
<th>Band1</th>
<th>Band2</th>
<th>Band3</th>
<th>Band4</th>
<th>FileID</th>
<th>FrameID</th>
<th>Instr</th>
<th>Speak</th>
<th>Sing</th>
</tr>
</thead>
<tbody>
<tr>
<td>5269</td>
<td>30272</td>
<td>697668</td>
<td>61495</td>
<td>180674</td>
<td>8107</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>7945</td>
<td>46834</td>
<td>750671</td>
<td>59708</td>
<td>180674</td>
<td>8108</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>4856</td>
<td>42833</td>
<td>566044</td>
<td>50430</td>
<td>180674</td>
<td>8109</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>6091</td>
<td>36418</td>
<td>461895</td>
<td>45732</td>
<td>180674</td>
<td>8110</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>4082</td>
<td>34732</td>
<td>661432</td>
<td>43447</td>
<td>180674</td>
<td>8111</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>6281</td>
<td>39368</td>
<td>752475</td>
<td>55325</td>
<td>180674</td>
<td>8112</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>
```

**Figure 2** Basic dataset statistics

- Total data entries/lines: 257830
- Min band value: 1
- Max band value: 1000000-200000000
A typical desktop computer has 2 or 4 processing cores, and may also include an accelerator such as a graphics card. With so many more processing cores available, Stampede is able to execute codes orders of magnitude faster than is possible on a desktop computer.
Get results

<table>
<thead>
<tr>
<th>File</th>
<th>Time</th>
<th>Frame</th>
<th>applause</th>
<th>non-applause</th>
</tr>
</thead>
<tbody>
<tr>
<td>182728</td>
<td>1</td>
<td>2</td>
<td>0.00832098</td>
<td>0.9917902</td>
</tr>
<tr>
<td>182728</td>
<td>1.5</td>
<td>3</td>
<td>0.03584029</td>
<td>0.96415972</td>
</tr>
<tr>
<td>182728</td>
<td>2</td>
<td>4</td>
<td>0.16780267</td>
<td>0.83211739</td>
</tr>
<tr>
<td>182728</td>
<td>2.5</td>
<td>5</td>
<td>0.54615493</td>
<td>0.45384309</td>
</tr>
<tr>
<td>182728</td>
<td>3</td>
<td>6</td>
<td>0.20988462</td>
<td>0.79011538</td>
</tr>
<tr>
<td>182728</td>
<td>3.5</td>
<td>7</td>
<td>0.06240148</td>
<td>0.91799852</td>
</tr>
<tr>
<td>182728</td>
<td>4</td>
<td>8</td>
<td>0.40240347</td>
<td>0.59759659</td>
</tr>
<tr>
<td>182728</td>
<td>4.5</td>
<td>9</td>
<td>0.35543555</td>
<td>0.84445845</td>
</tr>
<tr>
<td>182728</td>
<td>5</td>
<td>10</td>
<td>0.0579943</td>
<td>0.9682017</td>
</tr>
<tr>
<td>182728</td>
<td>5.5</td>
<td>11</td>
<td>0.05481813</td>
<td>0.94180313</td>
</tr>
<tr>
<td>182728</td>
<td>6</td>
<td>12</td>
<td>0.102857</td>
<td>0.897143</td>
</tr>
<tr>
<td>182728</td>
<td>6.5</td>
<td>13</td>
<td>0.09698122</td>
<td>0.9030678</td>
</tr>
</tbody>
</table>
Get results

<table>
<thead>
<tr>
<th></th>
<th>audiolD</th>
<th>time</th>
<th>frame</th>
<th>instr</th>
<th>spoke</th>
<th>sung</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>179591</td>
<td>0.062</td>
<td></td>
<td>2 0.00214606</td>
<td>0.02794373</td>
<td>0.00223372</td>
</tr>
<tr>
<td>3</td>
<td>179591</td>
<td>0.093</td>
<td></td>
<td>3 0.00255539</td>
<td>0.0452456</td>
<td>0.00416794</td>
</tr>
<tr>
<td>4</td>
<td>179591</td>
<td>0.125</td>
<td></td>
<td>4 0.00192335</td>
<td>0.03774013</td>
<td>0.00416037</td>
</tr>
<tr>
<td>5</td>
<td>179591</td>
<td>0.156</td>
<td></td>
<td>5 0.00151301</td>
<td>0.02508593</td>
<td>0.00448945</td>
</tr>
<tr>
<td>6</td>
<td>179591</td>
<td>0.187</td>
<td></td>
<td>6 0.00122285</td>
<td>0.01972235</td>
<td>0.00216901</td>
</tr>
<tr>
<td>7</td>
<td>179591</td>
<td>0.218</td>
<td></td>
<td>7 0.00154858</td>
<td>0.01753468</td>
<td>0.00376334</td>
</tr>
<tr>
<td>8</td>
<td>179591</td>
<td>0.25</td>
<td></td>
<td>8 0.00158939</td>
<td>0.02207039</td>
<td>0.00399368</td>
</tr>
<tr>
<td>9</td>
<td>179591</td>
<td>0.281</td>
<td></td>
<td>9 0.00233506</td>
<td>0.02998627</td>
<td>0.00277988</td>
</tr>
<tr>
<td>10</td>
<td>179591</td>
<td>0.312</td>
<td></td>
<td>10 0.00201615</td>
<td>0.02638504</td>
<td>0.0021225</td>
</tr>
<tr>
<td>11</td>
<td>179591</td>
<td>0.343</td>
<td></td>
<td>11 0.00213254</td>
<td>0.03295553</td>
<td>0.00505241</td>
</tr>
<tr>
<td>12</td>
<td>179591</td>
<td>0.375</td>
<td></td>
<td>12 0.00183349</td>
<td>0.03933096</td>
<td>0.00523635</td>
</tr>
<tr>
<td>13</td>
<td>179591</td>
<td>0.406</td>
<td></td>
<td>13 0.00236624</td>
<td>0.03903677</td>
<td>0.0067799</td>
</tr>
<tr>
<td>14</td>
<td>179591</td>
<td>0.437</td>
<td></td>
<td>14 0.00323295</td>
<td>0.03617664</td>
<td>0.00351125</td>
</tr>
<tr>
<td>15</td>
<td>179591</td>
<td>0.468</td>
<td></td>
<td>15 0.00259572</td>
<td>0.02791928</td>
<td>0.00427845</td>
</tr>
<tr>
<td>16</td>
<td>179591</td>
<td>0.5</td>
<td></td>
<td>16 0.00237613</td>
<td>0.03938892</td>
<td>0.00605176</td>
</tr>
<tr>
<td>17</td>
<td>179591</td>
<td>0.531</td>
<td></td>
<td>17 0.00233118</td>
<td>0.03637224</td>
<td>0.00967597</td>
</tr>
<tr>
<td>18</td>
<td>179591</td>
<td>0.562</td>
<td></td>
<td>18 0.00201296</td>
<td>0.03603429</td>
<td>0.0053496</td>
</tr>
</tbody>
</table>
Visualize results

55 John Alan Lomax recordings 1926-194
Blue = sung; green = spoken; red = instrumental
Publish

http://soundstudiesblog.com/
What do we talk about when we talk about sound?

- Language dynamics: tempo, pitch, tone/timbre, volume, pace, laughter, silence, applause, moans, screams, dialects, changing speakers, gender, age, changing genres

- Environment: fan hums, car horns, chickens, train whistles, bird calls, frogs mating

- Materiality: recording noises, needle drops, feedback, the electronic grid, changing tracks
What do we talk about when we talk about sound?

What does this look like as data?
What do we talk about when we talk about audio?

Damping ratios, gain, frequencies, spectra, energy, and pitch energy
What do we talk about when we talk about audio?

What does this look like as data?
Thank you!

- tclement@ischool.utexas

Special thanks to HiPSTAS team and Eleanor Dickson from HathiTrust Research Center