

# 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?

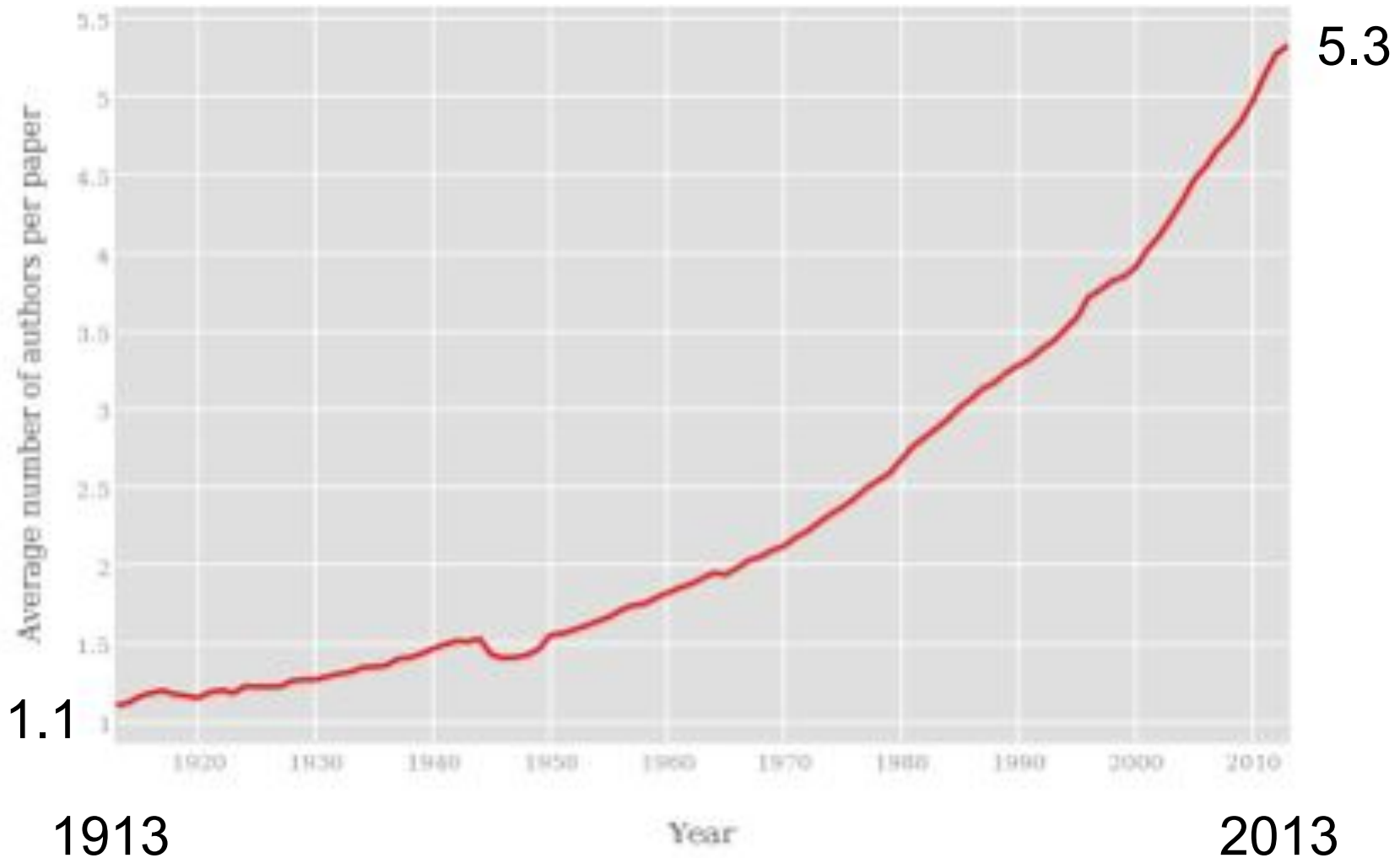
# A personal encounter with other rules

- My son's PhD work in neuroscience:
  - Simons, David L., Sanford L. Boye, William W. Hauswirth, and Samuel M. Wu. 2011. "Gene therapy prevents photoreceptor death and preserves retinal function in a Bardet-Biedl syndrome mouse model." *Proceedings of the National Academy of Sciences* 108(15): 6276-6281.
  - 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.”
    - **39.** Hauswirth WW, Lewin AS, Zolotukhin S, Muzyczk N (2000) Production and purification of recombinant adeno-associated virus. *Methods Enzymol* **316**:743–761
- They did not participate in the writing or research, but their intellectual contribution was indispensable and foundational

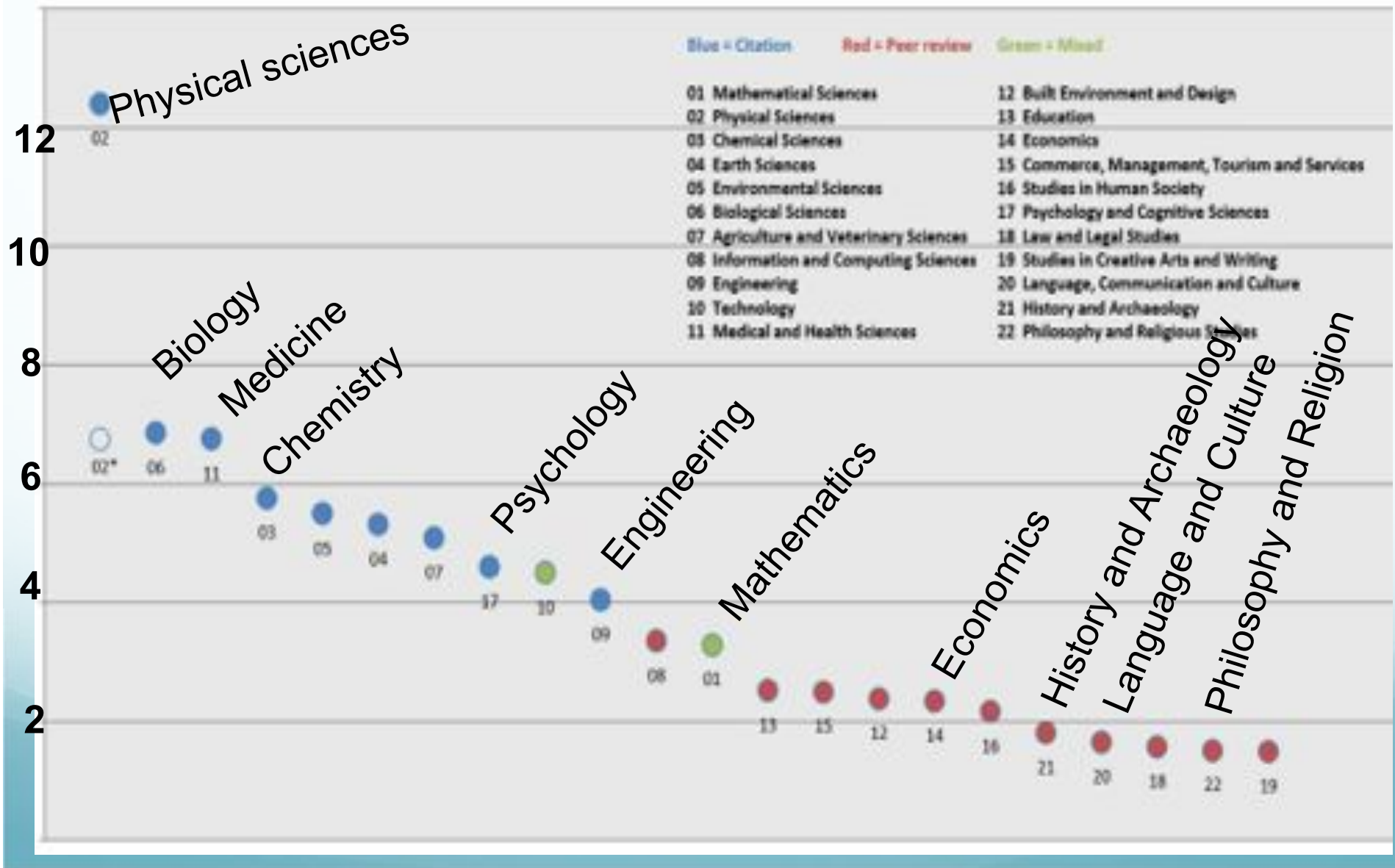
# The rise of co-authorship (from PubMed database)



<https://thewinnow.com/papers/the-rising-trend-in-authorship>

# Authors per publication by field

[http://archive.arc.gov.au/archive\\_files/ERA/2012/Outcomes/ERA\\_2012\\_National\\_Outcomes.pdf](http://archive.arc.gov.au/archive_files/ERA/2012/Outcomes/ERA_2012_National_Outcomes.pdf)





# Another story: closer to home (at least, disciplinarily)

- Headland, Thomas N., Janet D. Headland, and Ray T. Uehara. 2011. Agta Demographic Database: chronicle of a hunter-gatherer community in transition, version 2.0. *SIL Language and Culture Documentation and Description 2*. <http://www.sil.org/resources/publications/entry/9299>
- 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-McQueen1  
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

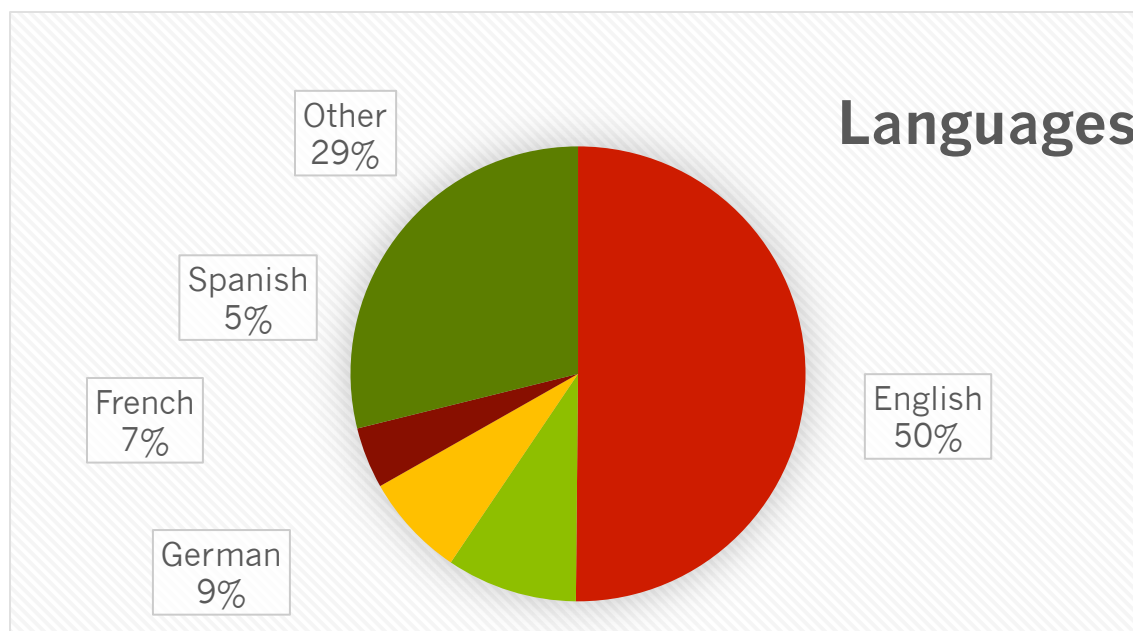


# HathiTrust



# HathiTrust Digital Library

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

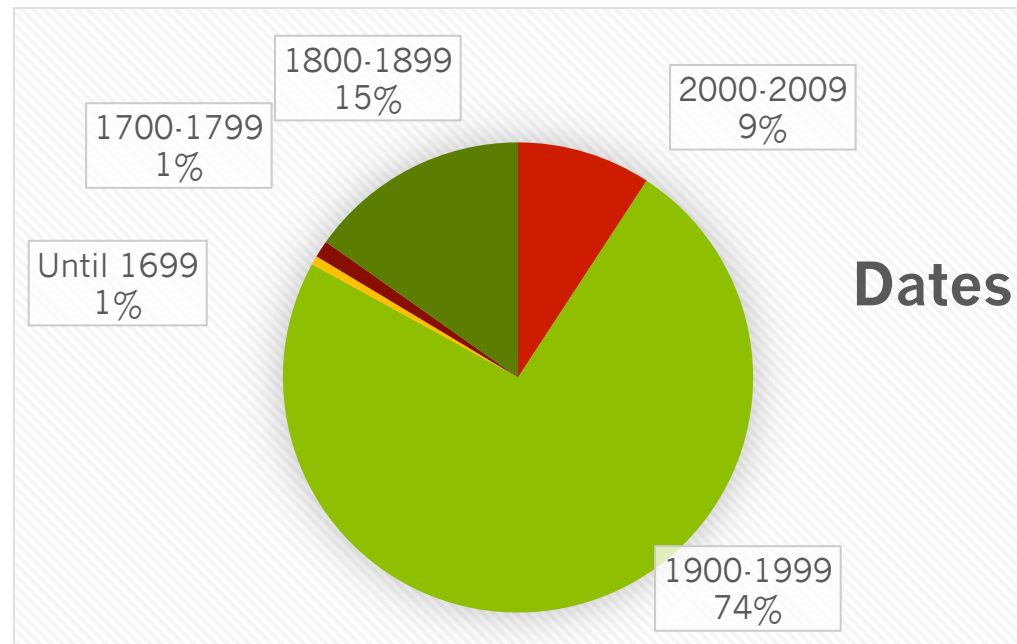


Around 50%  
of volumes  
are in English  
| Many other  
languages  
included as  
well



# HathiTrust Digital Library

Dates span the  
15<sup>th</sup> -  
21<sup>st</sup> centuries



70% in copyright or undetermined | 30% out of copyright

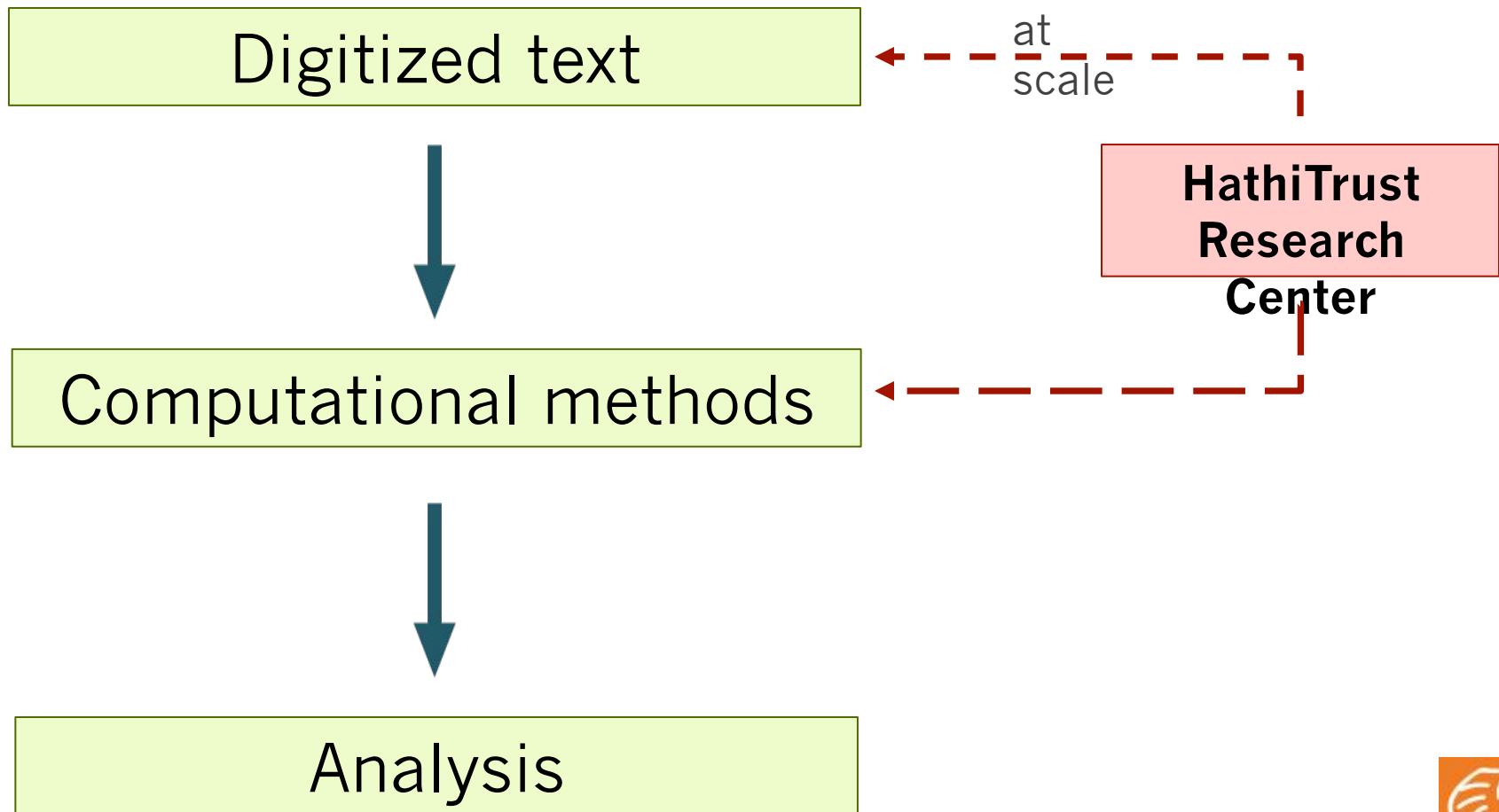


# HathiTrust





# “Reading” with computers



# 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



# HIGH PERFORMANCE SOUND TECHNOLOGIES FOR ACCESS AND SCHOLARSHIP

[About](#) [Resources](#) [Call for Participation](#) [Organizers](#) [Readings](#) [Meetings](#)

## Welcome to HIPSTAS

By TANYA CLEMENT | Published: NOVEMBER 14, 2012 | Edit

**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).

### CONTACT

hipstasinfo[at]urlists.umass.edu

### SEARCH

### RECENT POSTS

- Welcome to HIPSTAS

### META

<http://www.hipstas.org>

Developing Standards for Data Citation & Attribution in Linguistics | Mini-presentation Session 3: Digital Humanities  
; CO | September 18-20 2015



# 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

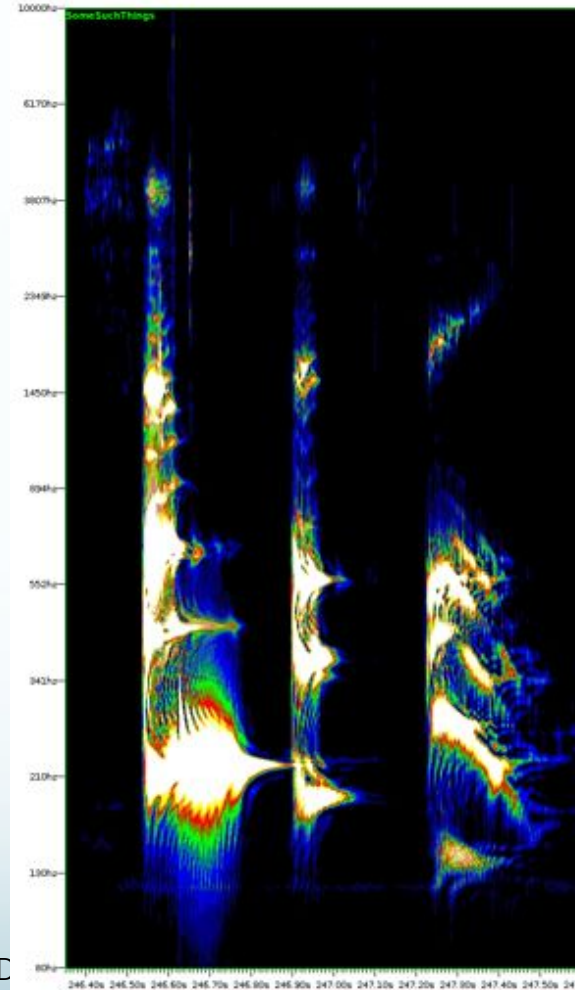
To develop a virtual research environment in which users can better access and analyze spoken word collections of interest to humanists through:

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**

**Blue**

**Black** – coolest, least intense

**Black**

**Black** – coolest, least intense

# Choose parameters

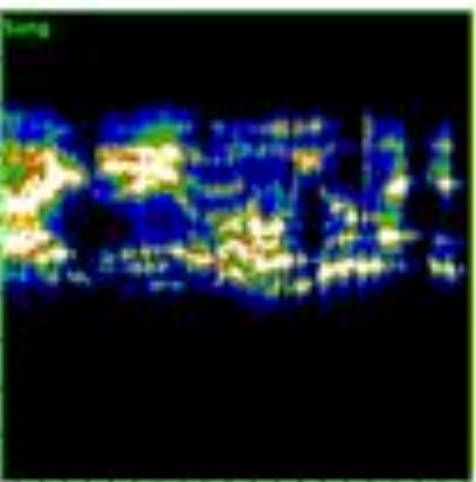


# Label examples

ARLO Version 1.0

Home Projects paulo parker **Folklore** Instruments WaveForms Speech Sounds Background Noise folklore2 Ellison-R

Half-Speed (Chrome Only)



Prev Trial || Next Trial ( Random Window job Name: tonyb-simplified-tagging )

[Back Half](#) [Forward Half](#)

Mediafilepath:

Mediafileid:

startTime:

endTime:

Gain:

Num Frames Per Second:

Create Tag  ↓

Tag Duration:

Tagset:  ↓

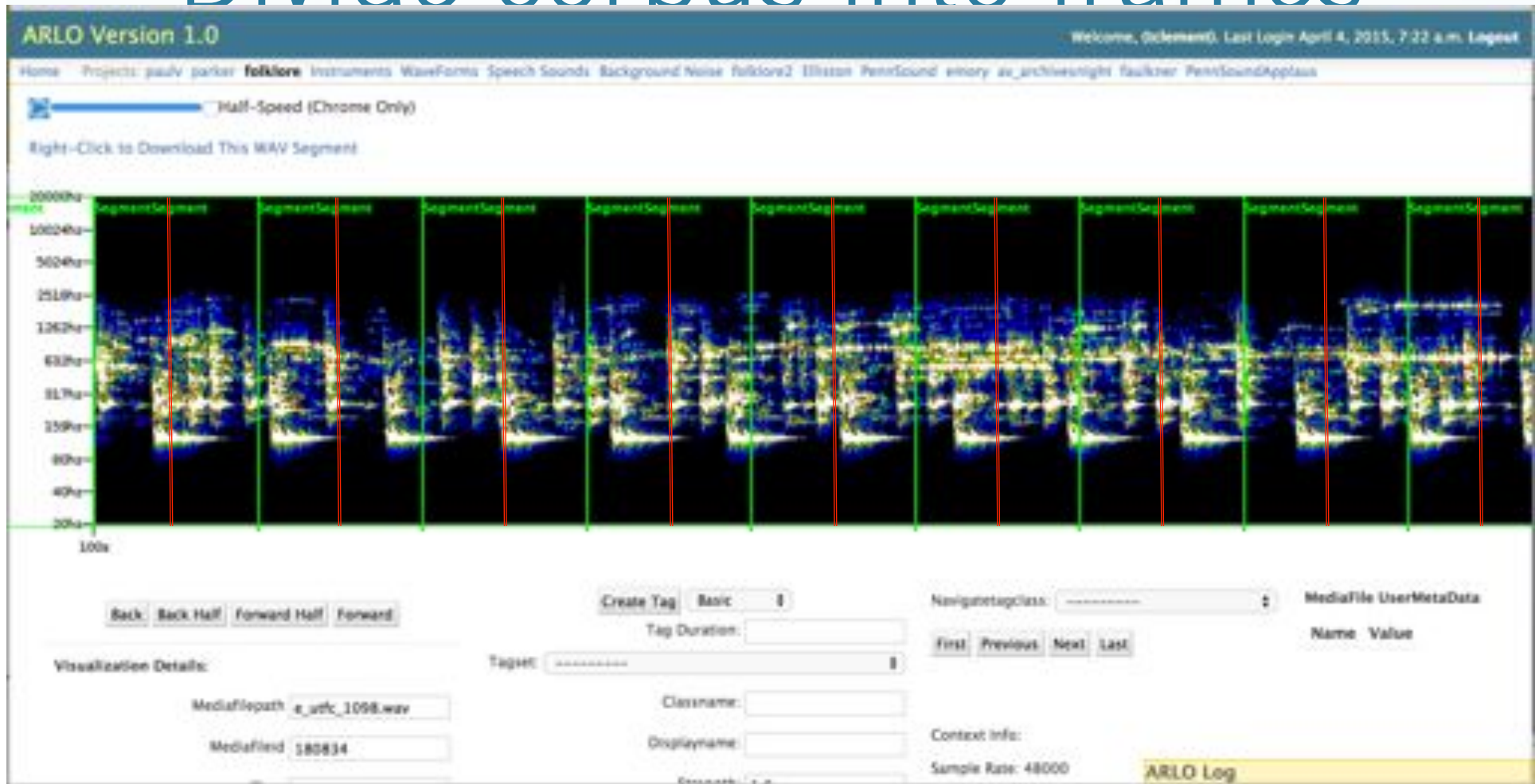
Classname:

Displayname:

Strength:

Tagstarttime:

# Divide corpus into frames



Developing Standards for Data Citation & Attribution in Linguistics  
Workshop 1 | Boulder, CO | September 18-20 2015

Mini-presentation Session 3: Digital Humanities  
Tanya E. Clement

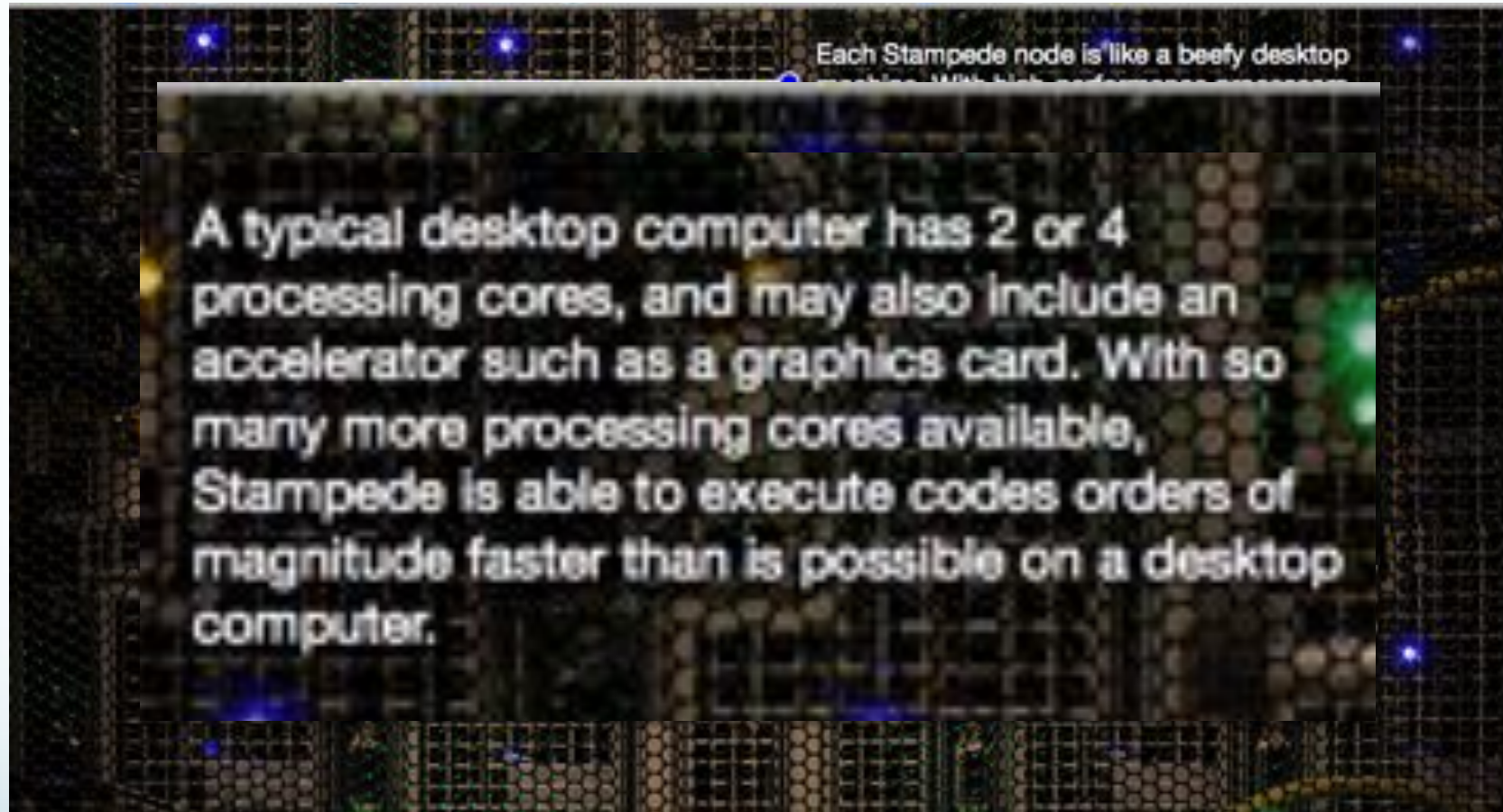
## .5 second time slices in ARLO

# Create the model

	Band1	Band2	Band3	Band4	FileID	FrameID	Instr	Speak	Sing
1	5269	30272	697668	61495	180674	8107	yes	no	yes
2	7945	46834	750671	59708	180674	8108	yes	no	yes
3	4856	42833	566044	50430	180674	8109	yes	no	yes
4	6091	36418	461895	45732	180674	8110	yes	no	yes
5	4082	34732	661432	43447	180674	8111	yes	no	yes
6	6281	39368	752475	55325	180674	8112	yes	no	yes
Total data entries/lines			Min band value			Max band value			
257830			1			1000000-20000000			

**Figure 2 Basic dataset statistics**

# Run the model



# Get results

	A	B	C	D	E	F
1	id	Time	Frame	applause	non-applause	
2	182728	1	2	0.0082098	0.9917902	
3	182728	1.5	3	0.03584029	0.96415971	
4	182728	2	4	0.16788367	0.83211733	
5	182728	2.5	5	0.54615891	0.45384109	
6	182728	3	6	0.20988462	0.79011538	
7	182728	3.5	7	0.08240148	0.91759852	
8	182728	4	8	0.40240347	0.59759653	
9	182728	4.5	9	0.13534355	0.86465645	
10	182728	5	10	0.03579983	0.96420017	
11	182728	5.5	11	0.05881813	0.94118187	
12	182728	6	12	0.102857	0.897143	
13	182728	6.5	13	0.09648122	0.90351878	

# Get results

1	audiolD	time	frame	instr	spoke	sung	
2	179591	0.062		2 0.00214606	0.02794373	0.00223372	
3	179591	0.093		3 0.00255539	0.0452456	0.00416794	
4	179591	0.125		4 0.00192335	0.03774013	0.00416037	
5	179591	0.156		5 0.00151301	0.02508593	0.00448945	
6	179591	0.187		6 0.00122285	0.01972235	0.00216901	
7	179591	0.218		7 0.00154858	0.01753468	0.00376334	
8	179591	0.25		8 0.00158939	0.02207039	0.00399368	
9	179591	0.281		9 0.00233506	0.02998627	0.00277988	
10	179591	0.312		10 0.00201615	0.02638504	0.0021225	
11	179591	0.343		11 0.00213254	0.03295553	0.00505241	
12	179591	0.375		12 0.00188349	0.03933096	0.00523635	
13	179591	0.406		13 0.00236624	0.03903677	0.0067799	
14	179591	0.437		14 0.00323295	0.03617664	0.00351125	
15	179591	0.468		15 0.00259572	0.02791928	0.00427845	
16	179591	0.5		16 0.00237613	0.03938892	0.00605716	
17	179591	0.531		17 0.00233118	0.03637224	0.00967597	
18	179591	0.562		18 0.00201296	0.03603429	0.0053496	



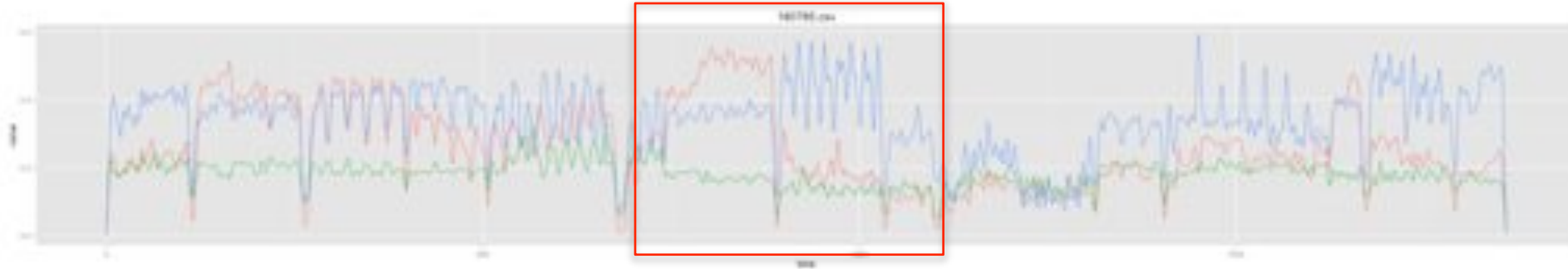
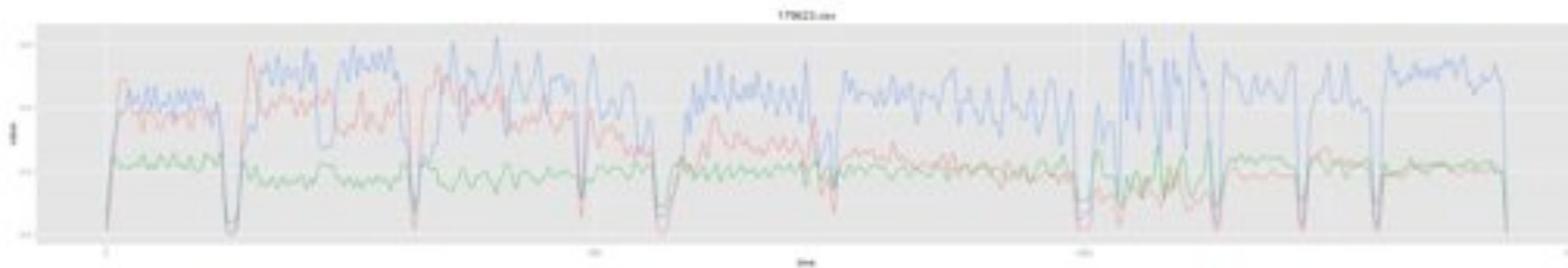
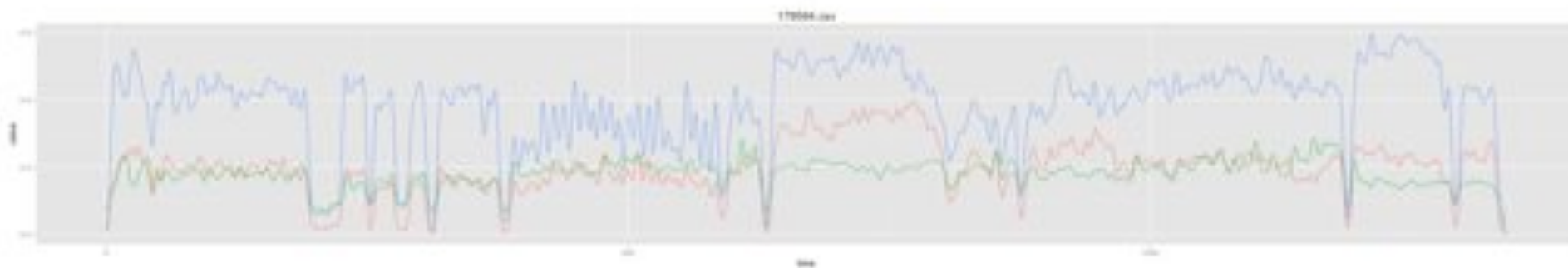
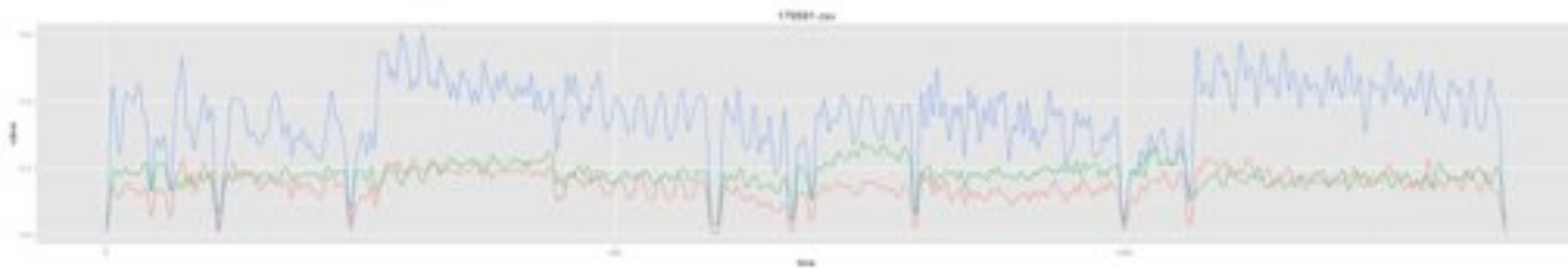
# Visualize results



Developing Standards for Data Citation & Attribution in Linguistics  
Workshop 1 | Boulder, CO | September 18-20 2015

Mini-presentation Session 3: Digital Humanities  
Tanya E. Clement

## 55 John Alan Lomax recordings 1926-1941



# Publish



**Sounding Out!**

May 14, 2015

by gowdlatiner

In 100 Years of Lomax.  
American Studies.  
Archival, Article.  
Cultural Studies.  
Custom, Digital  
Media, Ethnographic.  
Field Recording.  
History, Listening.  
Music, Recording.  
Sound, Sound Studies,  
Technology

1 Comment

## Machinic Ballads: Alan Lomax's Global Jukebox and the Categorization of Sound Culture

ISSN 2333-0309

\*Indexed by the Modern Language  
Association Bibliography



Developing Standards for Data Citation & Attribution in Linguistics  
Workshop 1 | Boulder, CO | September 18-20, 2015

Mini-presentation Session 3: Digital Humanities  
Tanya E. Clement

<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

The screenshot displays a software interface for audio analysis. At the top, there is a frequency spectrum plot with a vertical axis ranging from 0 to 20000 Hz and a horizontal axis from 0.00 to 0.10 seconds. Below the plot, the interface is divided into several sections:

- Navigation:** Buttons for "Back", "Back Half", "Forward Half", and "Forward".
- Media Info:** Fields for "Media/Source" (Elision, Martha, Cell), "Media/ID" (180962), "Starttime" (0), and "Endtime" (4.7).
- Gain:** A field for "Gain" set to "L0".
- Spectrum Parameters:** Fields for "Spectrum/responsesec" (128), "Spectrum/frequencybands" (156), "Spectrum/dampingfactor" (0.02), and "Spectrum/minbandfrequency" (20.0).
- Tagging Controls:** A "Create Tag" section with a "Name" field, "Tag Duration", "Tagset", "Classname", "Displayname", "Strength" (0.0), "Tagstarttime" (0.0), "Tagendtime" (1.0), "Tagfrequency" (60.0), and "Tagmaxfrequency" (300.0).
- Context Info:** A section on the right showing "Sample Rate" (48100) and "Duration" (70.34775130004082).

# What do we talk about when we talk about audio?

## What does this look like as data?

# Thank you!

- [tclement@ischool.utexas](mailto:tclement@ischool.utexas)

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