Crowdsourcing Cultural Heritage

Mia Ridge and Ben Brumfield

HILT 2014
August 4-8, Maryland Institute for Technology in the Humanities

Welcome!

Session 16: Friday, 9:00 – 10:30 am

- Making the Case for Crowdsourcing
- Ethics of Crowdsourcing (discussion)
- Legal Issues

Any questions from yesterday?

The Financial Case

Outsourced OCR Correction

- Cost: $.35-$1.20 per thousand characters
- 25-50 Characters per newspaper column-line

CDNC: 2.6 million lines * 40 characters/line * $.0005 / character = $53,130

Trove: 129 million lines => $2,580,926 (USD)

Source: Frederick Zarndt, ALA 2014

The Financial Case

In-House OCR Correction

Cost: $10-$41.88 USD per hour

15 seconds/line

CDNC: 2.6 million lines @ $10/hour = $110,687

Trove: 129 million lines @ $41.88/hour = $22,518,579

Source: Frederick Zarndt, ALA 2014

The Financial Case

Biodiversity Volunteer Portal

- Funds for Digitization
- Funds for Software
- Funds for Outreach

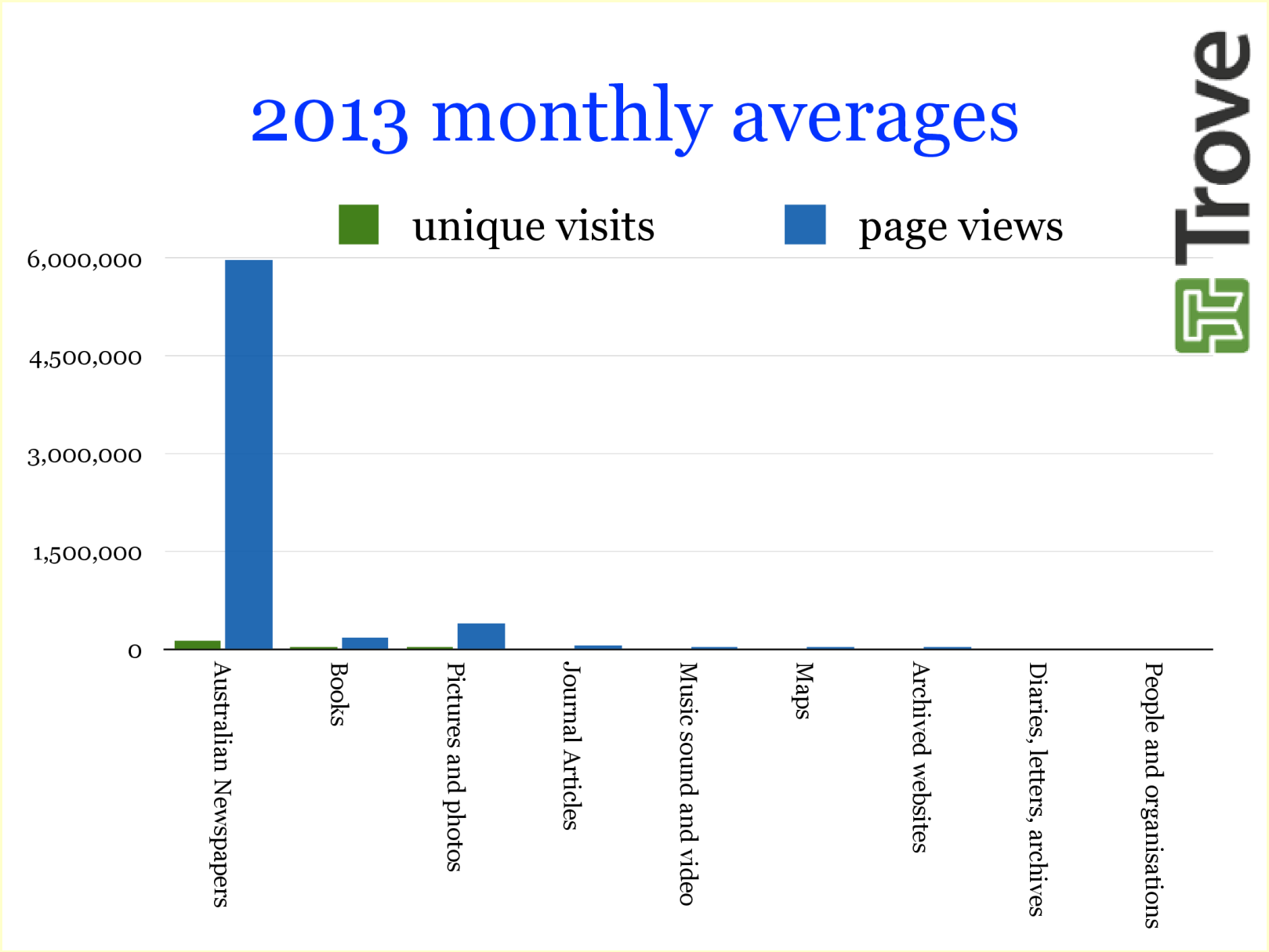
The Scholarly Case

Oliver Duke-Williams

“The potential of using crowd-sourced data to re-explore the demography of Victorian Britain”

DH2012

The Usage Case



Courtesy Frederic Zarndt

corrected OCR accuracy by newspaper title

Title	OCR character accuracy	~OCR word accuracy	Corrected accuracy	~Corrected word accuracy
PRP 1871 - 1922	92.6%	68.1%	99.3%	96.5%
SFC 1890 - 1913	92.6%	68.1%	99.6%	98.0%
LAH 1873 - 1910	88.7%	54.9%	99.1%	95.6%
LH 1877 - 1899	88.6%	54.6%	99.9%	99.5%
DAC 1841 - 1891	88.2%	53.4%	99.9%	99.5%
CF 1855 - 1880	86.5%	48.4%	98.3%	91.8%
SN 1885 - 1922	70.4%	17.3%	100.0%	100.0%

*Word accuracy assumes average word length is 5 characters

Ethics

Discuss:

Is Volunteer Crowdsourcing Exploitation?

“Citizen Scientists” or Unpaid Bottle-Washers?

Data Reciprocity

Crowdsourcing => Deprofessionalization?

Ethics

http://xkcd.com/1060

Law

Disclaimer:

- We are not lawyers
- This is not legal advice

Break!