

Crowdsourcing Cultural Heritage

Mia Ridge and Ben Brumfield

HILT 2014

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

Welcome!

Wednesday's Agenda

Session 9: The Text Talk/OCR Correction

Session 10: Handwriting and Manuscripts

Session 11: Tool Selection

Any questions from yesterday?

Introductions

Poll

Let's talk about text

The problem with “scan-and-dump”

One Volunteer's Story

- Nat Wooding
 - Retired data analyst
 - 100 pages of Julia Brumfield's diaries transcribed and indexed in six months
 - No relation to diarist

One Volunteer's Story

- Nat Wooding
 - Retired data analyst
 - 100 pages of Julia Brumfield's diaries transcribed and indexed in six months
 - No relation to diarist
 - Great-uncle was diarist's letter carrier, also named **Nat Wooding**

1801 Well hope I won't get
very worse that wooding
carried the mail today
I am very glad that he
was able to go - I got
a letter from him this

FromThePage

Julia Brumfield Diaries — 1920

page

transcribe

versions

[Previous Page](#)

Wednesday, February 25, 1920

A colde day. The ground was covered in snow this morning. The boys here did no try to work out.

Josie got dinner.

I sowed on my quilt most of the time.

I am not well. Hope I wont get any worse.

Nat Wooding carried the mail today. I am very glad that he was able to go. I got a

OCR Under the Hood

OCR Under the Hood



OCR Under the Hood

OCR Text Output:

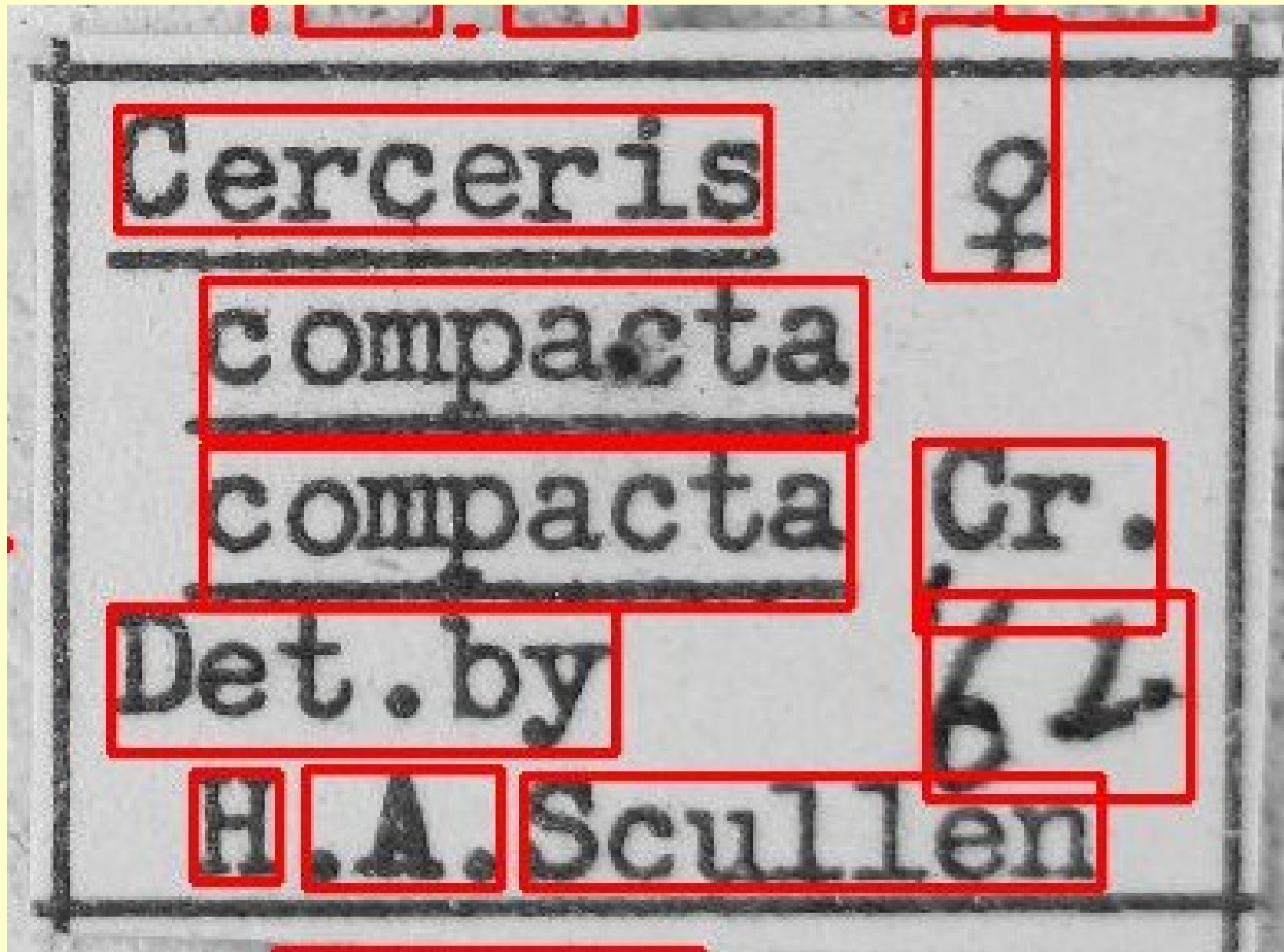
Cerceris compacta

OCR Under the Hood

HOCR output:

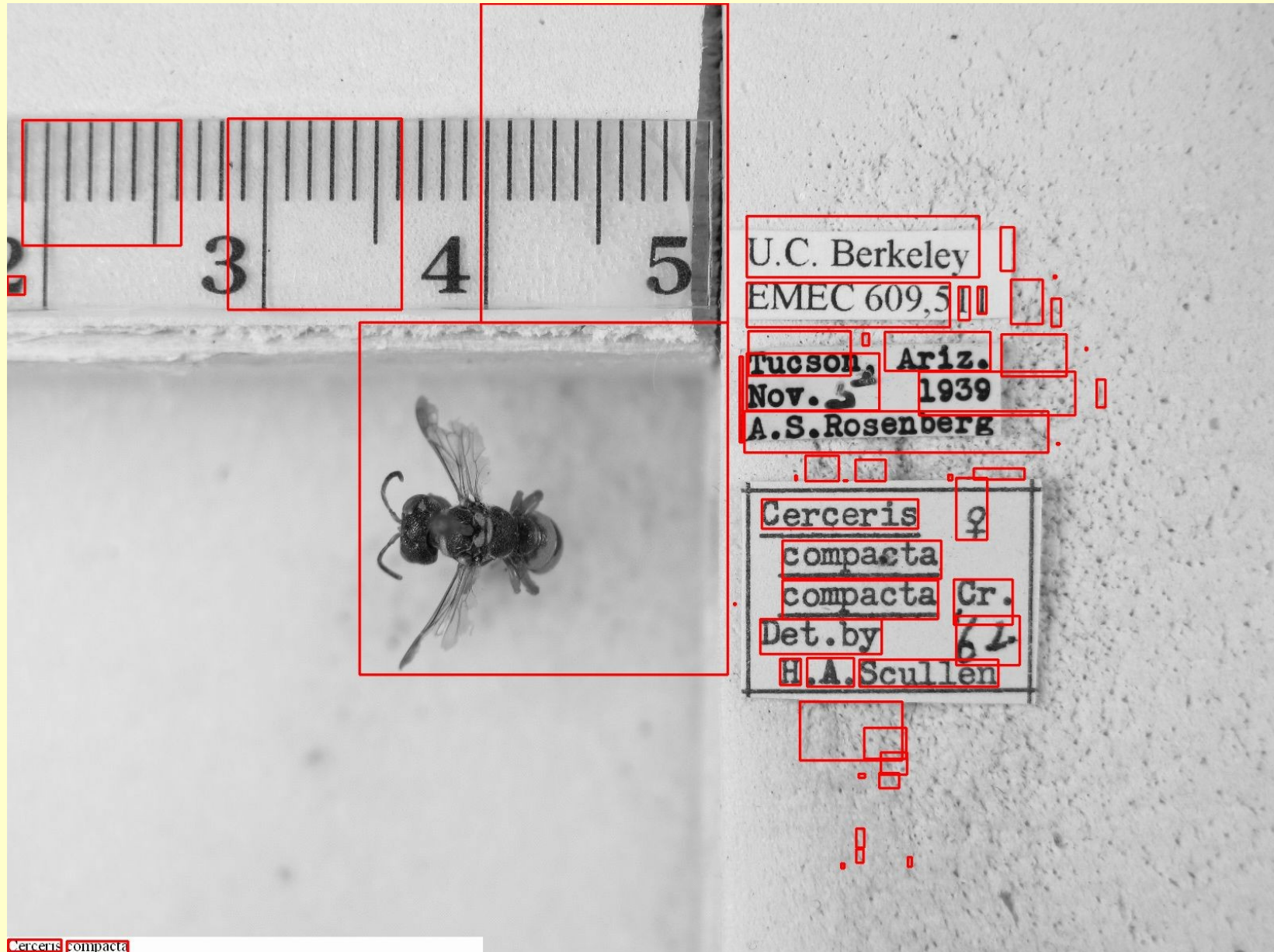
```
<span class='ocr_word'  
      id='word_7'  
      title="bbox 75 1182 152 1199">  
compacta  
</span>
```


OCR Under the Hood



The Problem With OCR

The Problem With OCR



The Problem With OCR

Character Accuracy vs. Word Accuracy

California Digital Newspaper Collection

Character accuracy = 89%

Courtesy: Frederick Zarndt

The Problem With OCR

Character Accuracy vs. Word Accuracy

California Digital Newspaper Collection

Character accuracy = 89%

Word accuracy = character accuracy^{word length}

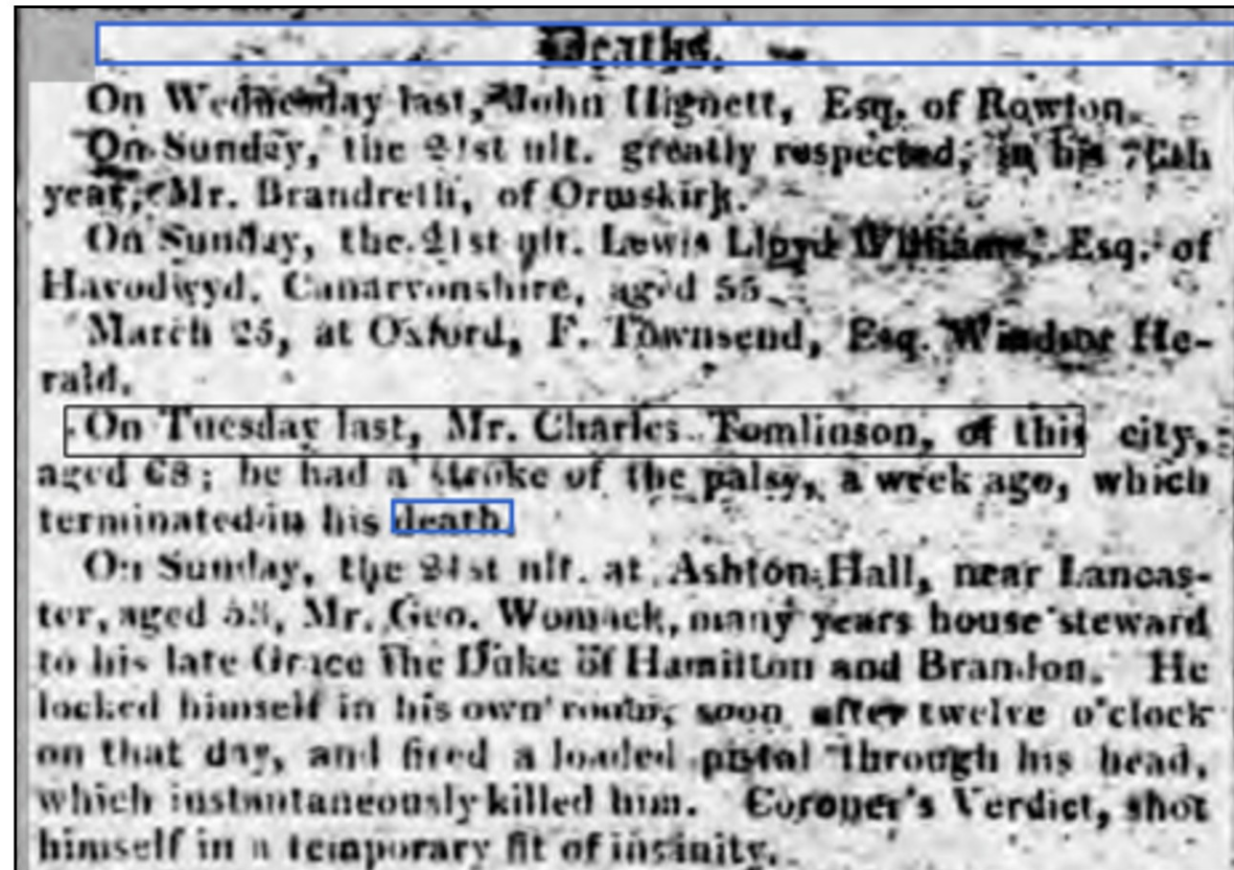
$$.89^5 = 55.8\%$$

Courtesy: Frederick Zarndt

raw OCR text

Deaths. **Illyrieff**, Esq. of <c .. Qn. Sunday, the till. greatly **Drandrellt**, of **Orms4\irJi.** ~ ; ;√ ' • * On **ijfr r inn ljjjil F iij '11 f Havodivyd**, **Carnarvonshire**, S ; *"* *- ' « ' March **Oxford, F. Tfovmeud, Uerald.** » • V . • On Tncsdav last, Mr. **Charles. IWilinson**, this 8 ; had vf thesis#, , a week ago, which terminate<i'iu his death. . / ' ■ O'i Sunday, dJst nit. at. **AsbtCnvHall**, mar **Lancaster**, Mr., **Geo. Worn ick**, many years house'steward hit late Once The **Hamilton** and **Brandon**. He locked himself h»oWn'r«wte<: soon. twelve o'clock" that dny, and fii»-d a loaded pistol "through Ins bead, 1 which instantaneously killed him. Coronet's Verdict, shot himself in a temporary fit of Friday week,

newspaper image



Excerpt from The British Newspaper Archive, Chester Courant, Tuesday 6-Apr-1819, page 3.

Courtesy: Frederick Zarndt

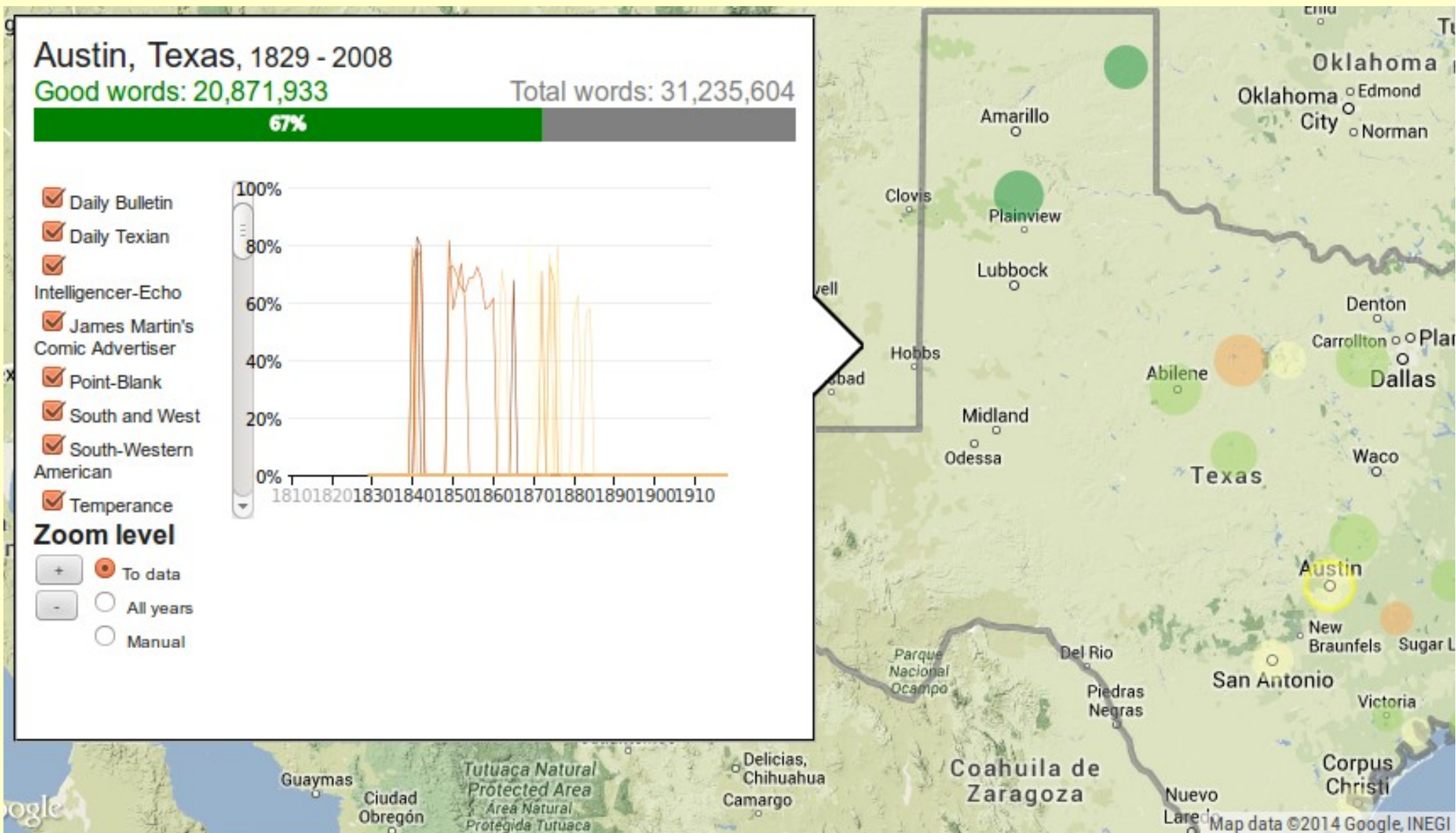


uncorrected OCR accuracy by newspaper title

Title	OCR character accuracy	~OCR word accuracy
PRP Pacific Rural Press 1871 - 1922	92.6%	68.1%
SFC San Francisco Call 1890 - 1913	92.6%	68.1%
LAH Los Angeles Herald 1873 - 1910	88.7%	54.9%
LH Livermore Herald 1877 - 1899	88.6%	54.6%
DAC Daily Alta California 1841 - 1891	88.2%	53.4%
CFJ California Farmer and Journal of Useful Sciences 1855 - 1880	86.5%	48.4%
SN Sausalito News 1885 - 1922	70.4%	17.3%

*Word accuracy assumes average word length is 5 characters

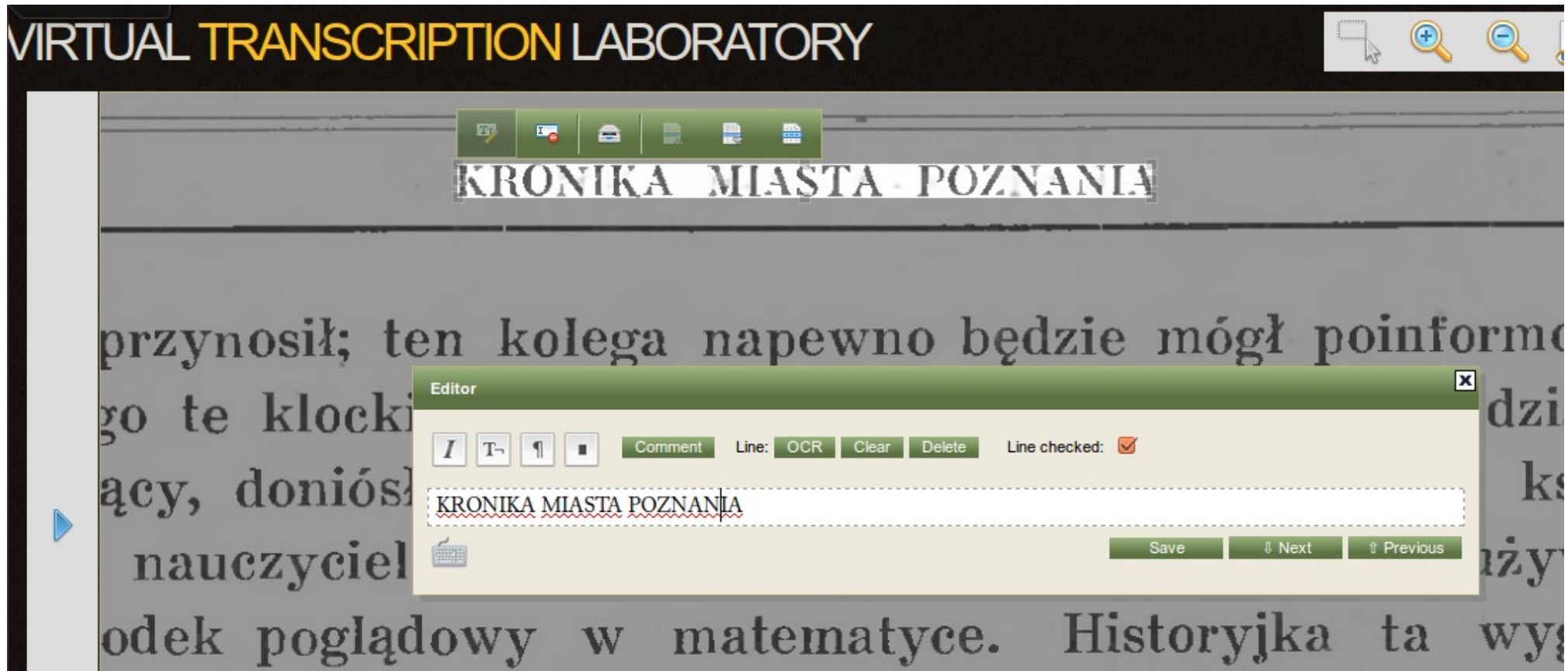
The Problem With OCR



Lab Session

Trove

Virtual Transcription Laboratory



Virtual Transcription Laboratory

VIRTUAL TRANSCRIPTION LABORATORY

SŁOWO WSTĘPNE.

Myślistwo istnieje tak dawno, jak świat. Dla pierwszych ludzi - naszych prapradziadów - myślistwo nie było jeno zabawą, dawny człowiek polował na zwierzy-
nę po to, żeby mieć co jeść i czem się przyodziać.
Z biegiem czasu jednakże, kiedy człowiek się cywi-
lizował, łowiectwo nabierało charakteru godziwej rozryw-
ki dla ludzi osiadłych na roli. Nic też dziwnego, że rol-
nik każdy czuje pociąg i wielkie zamięłowanie do myś-
liwstwa. Utarło się jednak - zwłaszcza w Polsce -
mniemanie, że polowanie może być zabawą dla wyższych
tylko sfer. To też otrzymanie od władzy pozwolenia na
posiadanie broni palnej i prawo polowania było do
ostatnich czasów połączone z wielkimi trudnościami dla
chłopa. Na tysiąc - jeden znalazł się taki, któremu
pozwolenie udało się wydostać. Kiedy przyszła wolna
i niepodległa Polska - władze administracyjne polskie
również powodowały się tem, że polować może duży
rolnik, polować może urzędnik czy jakiś mieszczech-
kupczyk, ale nigdy chłop. Prawie że żadnych pozwoleń
chłopom na posiadanie broni palnej i prawo polowania
nie udzielano. Chcąc tej krzyczącej niesprawiedliwości

3

SŁOWO WSTĘPNE.

Myślistwo istnieje tak dawno, jak świat. Dla pierw-
szych ludzi — naszych prapradziadów — myślistwo nie
było jeno zabawą, dawny człowiek polował na zwierzy-
nę po to, żeby mieć co jeść i czem się przyodziać.

Z biegiem czasu jednakże, kiedy człowiek się cywi-
lizował, łowiectwo nabierało charakteru godziwej rozryw-
ki dla ludzi osiadłych na roli. Nic też dziwnego, że rol-
nik każdy czuje pociąg i wielkie zamięłowanie do myś-
liwstwa. Utarło się jednak — zwłaszcza w Polsce —
mniemanie, że polowanie może być zabawą dla wyższych
tylko sfer. To też otrzymanie od władzy pozwolenia na
posiadanie broni palnej i prawo polowania było do
ostatnich czasów połączone z wielkimi trudnościami dla
chłopa. Na tysiąc — jeden znalazł się taki, któremu
pozwolenie udało się wydostać. Kiedy przyszła wolna
i niepodległa Polska — władze administracyjne polskie
również powodowały się tem, że polować może duży
rolnik, polować może urzędnik czy jakiś mieszczech-
kupczyk, ale nigdy chłop. Prawie że żadnych pozwoleń
chłopom na posiadanie broni palnej i prawo polowania
nie udzielano. Chcąc tej krzyczącej niesprawiedliwości

Broń myśliwska — gdzie najlepiej nabywać — porad
udziela redakcja „Nowinek Myśliwskich” w Warszawie.

Lab Session

Wikisource