Rooting a DH Project in a 1930s WPA Project: *The Annals of Cleveland*

A work-in-progress data extraction and public website derived from the Depression-era *Annals of Cleveland* newspaper digest series, tracing the dependence of each project on the labour of previous projects.

Peter Binkley
University of Alberta Library
peter.binkley@ualberta.ca @pabinkley

DOI: https://doi.org/10.7939/r3-5shd-gp89

"Works Progress Administration Newspaper Indexing at the Akron Beacon Journal, 1938", reproduced with permission. **https://cdm17124.contentdm.oclc.org/digital/collection/ABJarchives/id/655**

Demonstration website: https://wallandbinkley.com/projects/2019/annals-of-cleveland/
Code: https://github.com/pbinkley/annals-of-cleveland

The WPA is best known today for its infrastructure projects – roads, bridges, dams – but it encompassed a wider range of projects. The *Annals of Cleveland* newspaper project was a response to the call for projects that would employ **clerical and white collar workers**. The plan was to use their office skills to create a research tool for local history: a digest of Cleveland newspapers from 1818 forward. By the time the digest project closed in 1938 the *Annals of Cleveland* newspaper series had covered the years 1818-1876, comprising approximately 29,000 multigraphed pages (including indexes). It employed between 400 and 500 workers: abstractors, editors, typists, supervisors, as well as support staff such as carpenters and time keepers. The purpose of this part of the project is to expose various facets of the labor that produced this resource and projected it through eighty years of changes in information technology.

This poster attempts to **trace the workflows of the WPA project** and of subsequent projects which derived from it: the microfilming and digitization of the newspapers and of the WPA publications, down to my own data extraction project. The first column represents the workflow of the WPA project. Clerical workers wrote abstracts of items that dealt with local (i.e. Ohio) news and opinion (it was assumed that the *New York Times Index* provided adequate access to national and international news). The abstracts moved up an editorial chain, being proofread and accumulating metadata, and then being typed. The printers followed the layout of the typed forms precisely, so my parsing of the data from the OCR text depends on the **skilled labour of the typists in preparing documents** as much as on the training of the abstractors and editors.

The project operated under contradictory requirements. The federal priority was to move as many people as possible off the relief rolls and into employment; the budget for non-labour expenses therefore had to be kept to a minimum. Like other New Deal programs, though, the WPA was vulnerable to attack from the political right, which saw such projects as boondoggles at best, and creeping socialism at worst. The project therefore had to strike a balance between employing whatever workers were assigned by the WPA, and actually making progress on the digest. The project's publicity boasts of the **academic utility** of the product but also of the **Fordian efficiency** of the process. Labour relations in WPA projects are an interesting topic on which this project can shed light. The Workers Alliance of America, a **"union of the unemployed"**, attempted to represent WPA workers across the nation, and succeeded in having an influence on the *Annals of Cleveland* project. Most of the evidence for the running of the project is in the papers of the Joint Committee on Materials for Research in the Library of Congress. These enable a tantalisingly incomplete reconstruction of the **working methods** of the project as they evolved. At the end of 1937 the digest project was replaced by a new approach based on indexing alone, supplemented by publication on microfilm of the source newspapers – one of the earliest large-scale microfilming projects.

The *Annals of Cleveland* volumes were the subject of a previous datafication project by Samuel Kernell and Gary C. Jacobson in the mid-1980s, dealing with newspaper coverage of congress and the presidency in the 19th century. Data for ten sample years was coded by hand and transferred to **punch cards** for analysis.

The digitization of the published volumes by the Google Books project and by the library of Case Western opened the way for an attempt at a comprehensive approach. A Ruby script uses **regular expressions** to parse the data elements out of the OCR text; this involves an iterative process, first identifying gaps in the parsed page and abstract numbers, fixing the OCR or enhancing the regexes, running it again, etc. Once the abstracts are detected, a similar process extracts the date, page, column, etc. Other processes handle the indexes and subject headings. The complexity in this script lies in the regular expressions, which have to account for the layout of the original printed abstracts and the variants introduced by the OCR process. The regex that looks for a "see also" references provides a simple example: `/^See [Aa]l[s§][Qo]/`. The resulting ability to process in bulk brings the required human labour down to the point where **a single person with a laptop** could conceivably handle the extraction of the data encoded by the labour of hundreds of WPA workers in 1930s Cleveland.

Black labels indicate the output of a processing step

Newspaper Files

WPA Project
(1935-1938)

Abstracting
(by 200 Abstractors)
Abstract

Date Sort

Copy Reading
Corrections

Filing under broad terms
Broad classification

When a year is complete:

Editing
(by Final Editors)
Corrections

Filing

Keywording
(by Keywording Editors)
Keywords, Classification, Cross-refs, etc.

Typing
(by 50 Typists)
Layout

Recheck
(by Final Editors)
Corrections

Multigraph
Pages

Binding
Volumes

Storage

Distribution

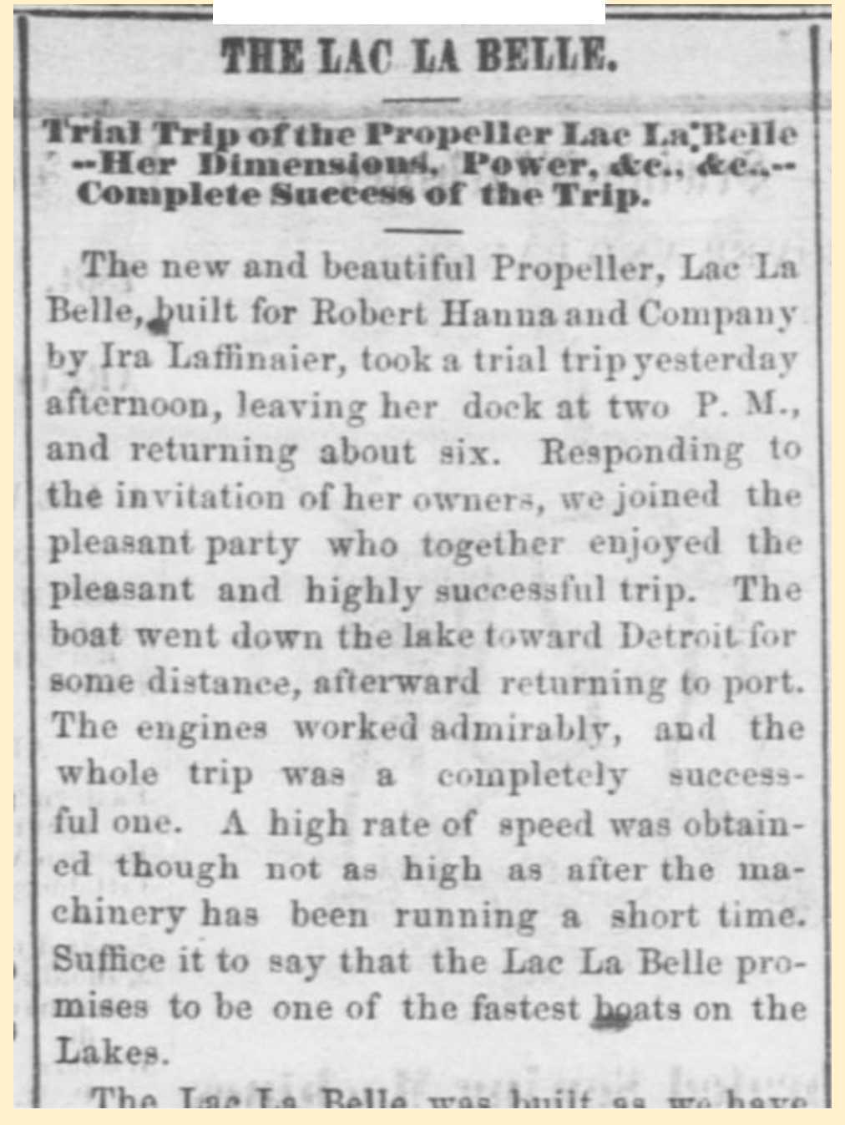
Newspaper Digitization
(2000s?)

(Microfilming)

Digitization

Online Publication

Newspaper



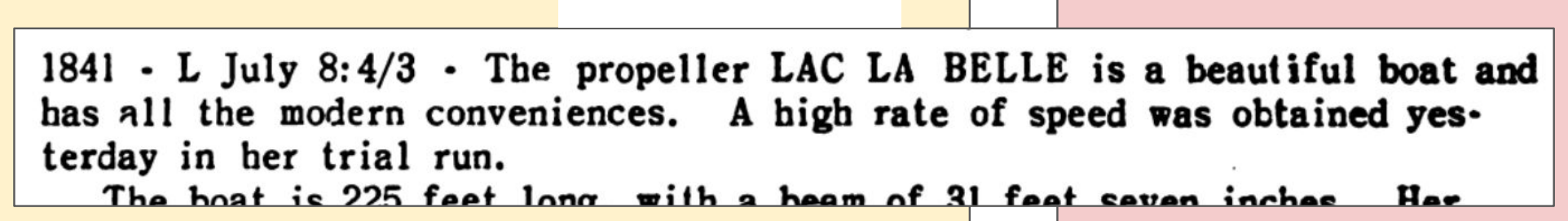
THE LAC LA BELLE.

Trial Trip of the Propeller Lac La Belle --Her Dimensions, Power, &c., &c.-- Complete Success of the Trip.

The new and beautiful Propeller, Lac La Belle, built for Robert Hanna and Company by Ira Laffinaier, took a trial trip yesterday afternoon, leaving her dock at two P. M., and returning about six. Responding to the invitation of her owners, we joined the pleasant party who together enjoyed the pleasant and highly successful trip. The boat went down the lake toward Detroit for some distance, afterward returning to port. The engines worked admirably, and the whole trip was a completely successful one. A high rate of speed was obtained though not as high as after the machinery has been running a short time. Suffice it to say that the Lac La Belle promises to be one of the fastest boats on the Lakes.

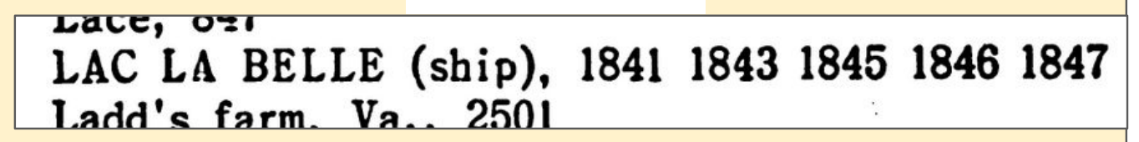
The Lac La Belle was built as we have

Abstract



1841 - L July 8:4/3 - The propeller LAC LA BELLE is a beautiful boat and has all the modern conveniences. A high rate of speed was obtained yesterday in her trial run.
The boat is 225 feet long, with a beam of 31 feet seven inches. Her

Keyword



Lace, 647
LAC LA BELLE (ship), 1841 1843 1845 1846 1847
Ladd's farm, Va., 2501

Consistent layout is critical for data extraction from OCR text

Annals of Cleveland volumes in Kelvin Smith Library, Case Western Reserve University (taken in 2014). The red volumes are the newspaper series.

Annals of Cleveland Print Volumes

Microfilm
(1960s?)

Disbinding

Microfilming

Distribution

Destruction of paper original
(often)

Digitization
(2000s?)

Disbinding

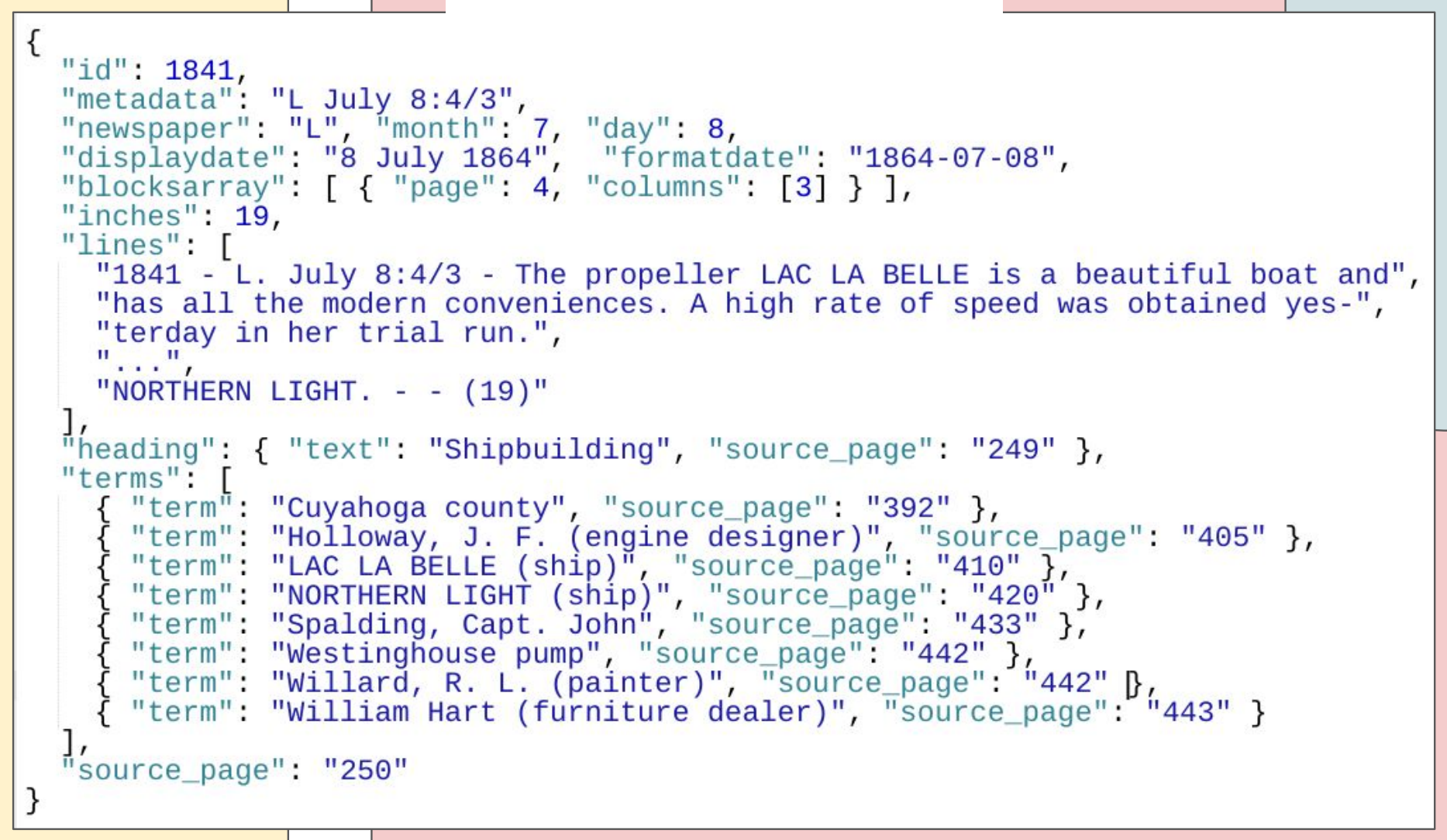
Scanning

OCR

Metadata

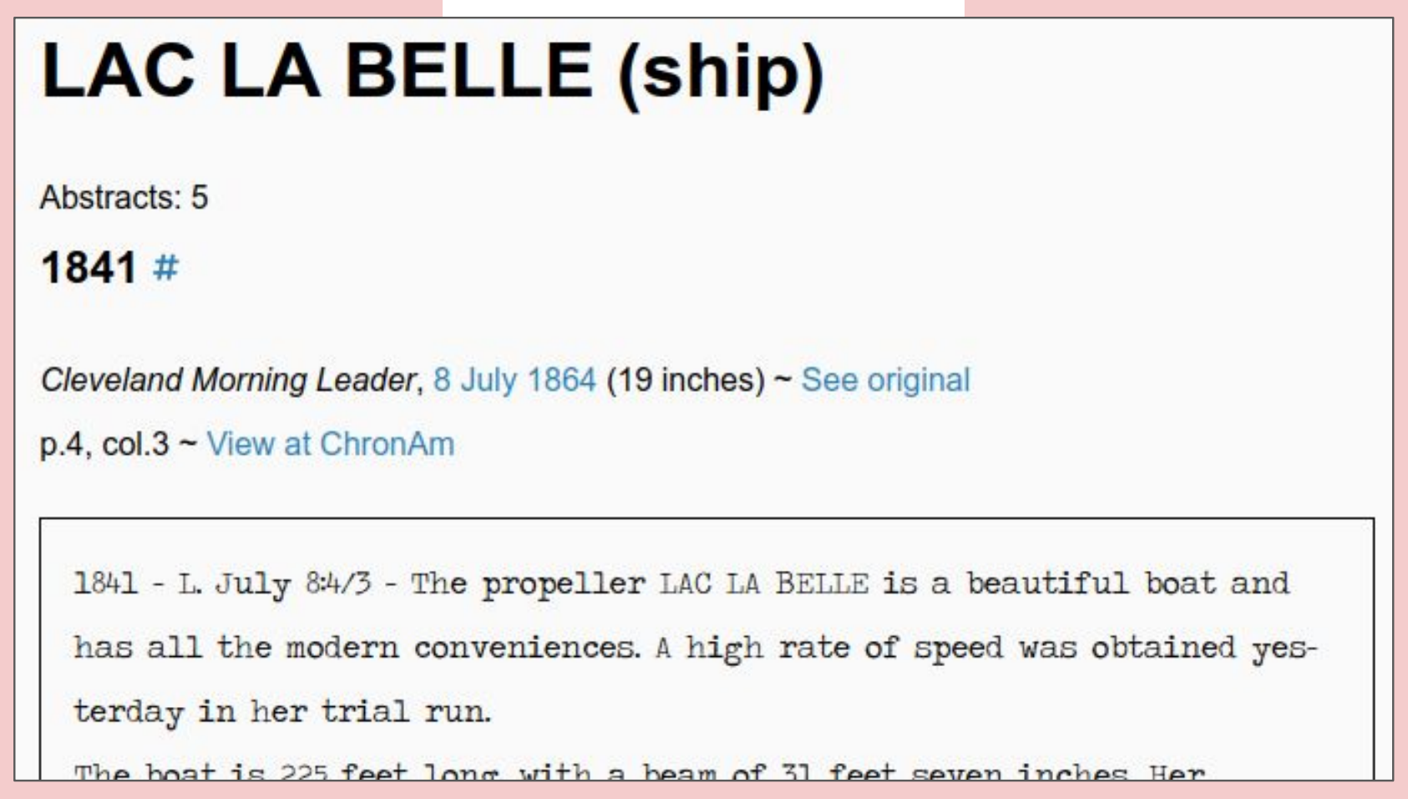
Online Publication

JSON (simplified)



```
{
  "id": 1841,
  "metadata": "L July 8:4/3",
  "newspaper": "L", "month": 7, "day": 8,
  "displaydate": "8 July 1864",  "formatdate": "1864-07-08",
  "blocksarray": [ { "page": 4, "columns": [3] } ],
  "inches": 19,
  "lines": [
    "1841 - L. July 8:4/3 - The propeller LAC LA BELLE is a beautiful boat and",
    "has all the modern conveniences. A high rate of speed was obtained yes-",
    "terday in her trial run.",
    "...",
    "NORTHERN LIGHT. - - (19)"
  ],
  "heading": { "text": "Shipbuilding", "source_page": "249" },
  "terms": [
    { "term": "Cuyahoga county", "source_page": "392" },
    { "term": "Holloway, J. F. (engine designer)", "source_page": "405" },
    { "term": "LAC LA BELLE (ship)", "source_page": "410" },
    { "term": "NORTHERN LIGHT (ship)", "source_page": "420" },
    { "term": "Spalding, Capt. John", "source_page": "433" },
    { "term": "Westinghouse pump", "source_page": "442" },
    { "term": "Willard, R. L. (painter)", "source_page": "442" },
    { "term": "William Hart (furniture dealer)", "source_page": "443" }
  ],
  "source_page": "250"
}
```

Online



LAC LA BELLE (ship)

Abstracts: 5

1841 #

Cleveland Morning Leader, 8 July 1864 (19 inches) ~ See original

p.4, col.3 ~ View at ChronAm

1841 - L. July 8:4/3 - The propeller LAC LA BELLE is a beautiful boat and has all the modern conveniences. A high rate of speed was obtained yesterday in her trial run.
The boat is 225 feet long, with a beam of 31 feet seven inches. Her

Datafication
(2019-22)

Design of processing script

Extract data from OCR text

Correcting OCR text
JSON

Generate source files
Markdown

Run Hugo

Online Publication