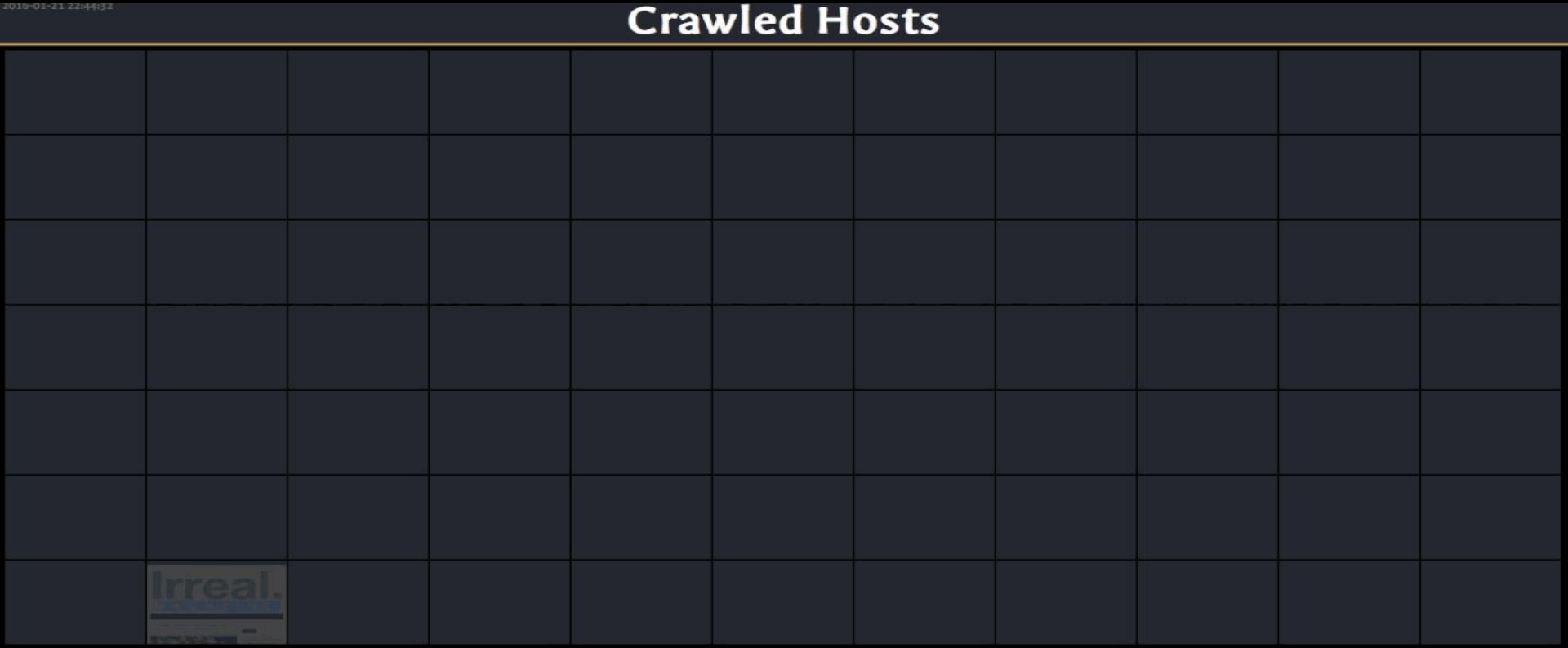
Tales from the Trenches:
Building Petabyte-scale Computational Research Services



Jefferson Bailey, Director, Archiving & Data Services, Internet Archive
Lowering the Barriers to Computational Access for Digital Archivists | DPC 2022 | jefferson@archive.org

The Internet Archive

Non-Profit Library

Founded in 1996 by Brewster Kahle

Universal Access to All Knowledge

Total Digital Items

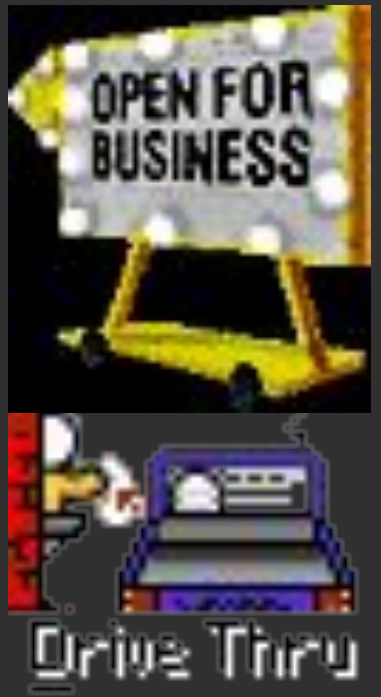
800,000	Software Titles
5,000,000	Moving Images
15,000,000	Audio Recordings
2,000,000	Hours of Television
6,000,000	eBooks
900,000,000,000	Web Pages
90	Petabytes

Program Considerations: Categories of Research Use

- Documentary
- Social/Political Scientists
- Web Science
- (Digital) Humanities
- Computer Science
- Data Analysts

Program Considerations: Categories of Services

- Bulk Data Model
- Cyberinfrastructure Model
- Roll Your Own Model
- Middleware Model
- Prepackaged Model
- Community & Support Model



Program Considerations: Technical Issues

- Format Complexities (.what?)
- Volume Complexities (Size)
- Processing Complexities (Tools)
- Breadth Complexities (Visibility)
- Collection Complexities (Accession)
- Heterogeneity Complexities (Content)

Program Considerations: Conceptual Issues

- Provenance can be opaque
- Acquisition has dependencies
- Scope and borders are porous
- Quantity does not equal quality
- Gaps and elisions can be undefined
- Questions may not map to collections

Program Considerations: What does this mean for libraries

- Where's the front door for data requests?
- What's your service model or workflow?
- What tech resources do/don't you have?
- What is your capacity for support?
- How and in what form will you deliver data?
- Can you leverage services internally?

Flexible Services

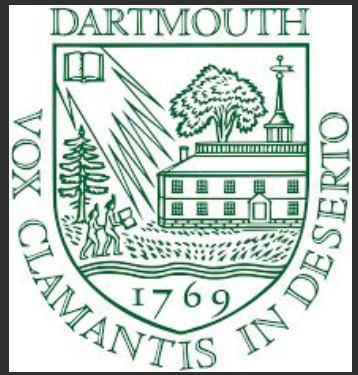
Researchers do not necessarily need huge sets of data to do interesting work… they do need flexible data delivery services…. Different formats based on different searches for different kinds of research at different times.

V.E. Varvel Jr. & A. Thomer, *Google Digital Humanities Awards recipient interviews report*

Case Study #1

Web-based Textual Network Industry Classification

- 800,000 corporate homepages
- Every change monthly 1996-current
- Access to private, not just public companies
- ML classifiers for product/market
- Required large, complex extraction



Case Study #2

Immersive Web Observatory

- Multi-PBs aggregated web archive collection
- Dedicated cluster and tool development
- Multi-institutional, nation-scale infrastructure
- ML/AI approaches to temporal analysis
- https://webis.de/

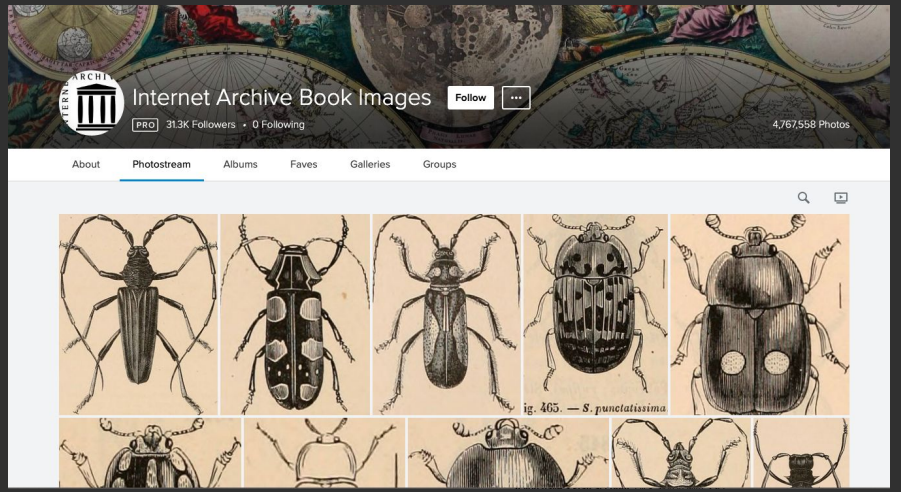
Case Study #3

Broader Web-Scale Provision of Parallel Corpora for EU Languages

- Petabyte-scale data mining
- Parallel corpora English to all EU languages
- 1 billion translated words for 7 languages, 100 million for 16 languages
- Custom extractions for Icelandic, Croatian, Norwegian, and Irish
- Mozilla Project Bergamot

Case Study #4

Book Images Extraction



- 4.7M images in Flickr
- Automated extraction
- Manifest & API access
- https://www.flickr.com/photos/internetarchivebookimages/

Case Study #5

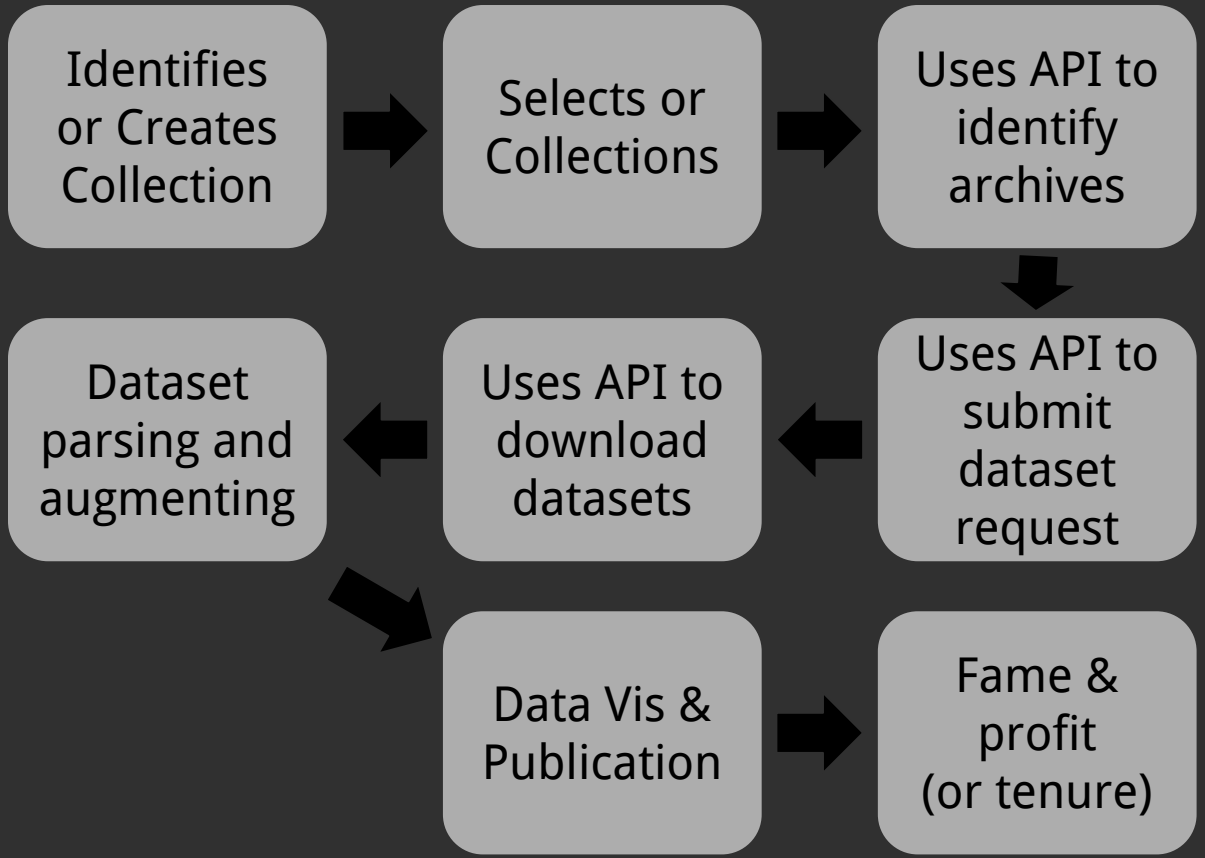
News Measures Research Project

- 663 local news sites from 100 communities
- Snapshot & ongoing monthly crawls
- Public collections & open datasets
- “Journalism History, Web Archives, and New Methods for Understanding the Evolution of Digital Journalism”
- https://doi.org/10.1080/21670811.2018.1510293

DeWitt Wallace Center for Media & Democracy

democracy fund

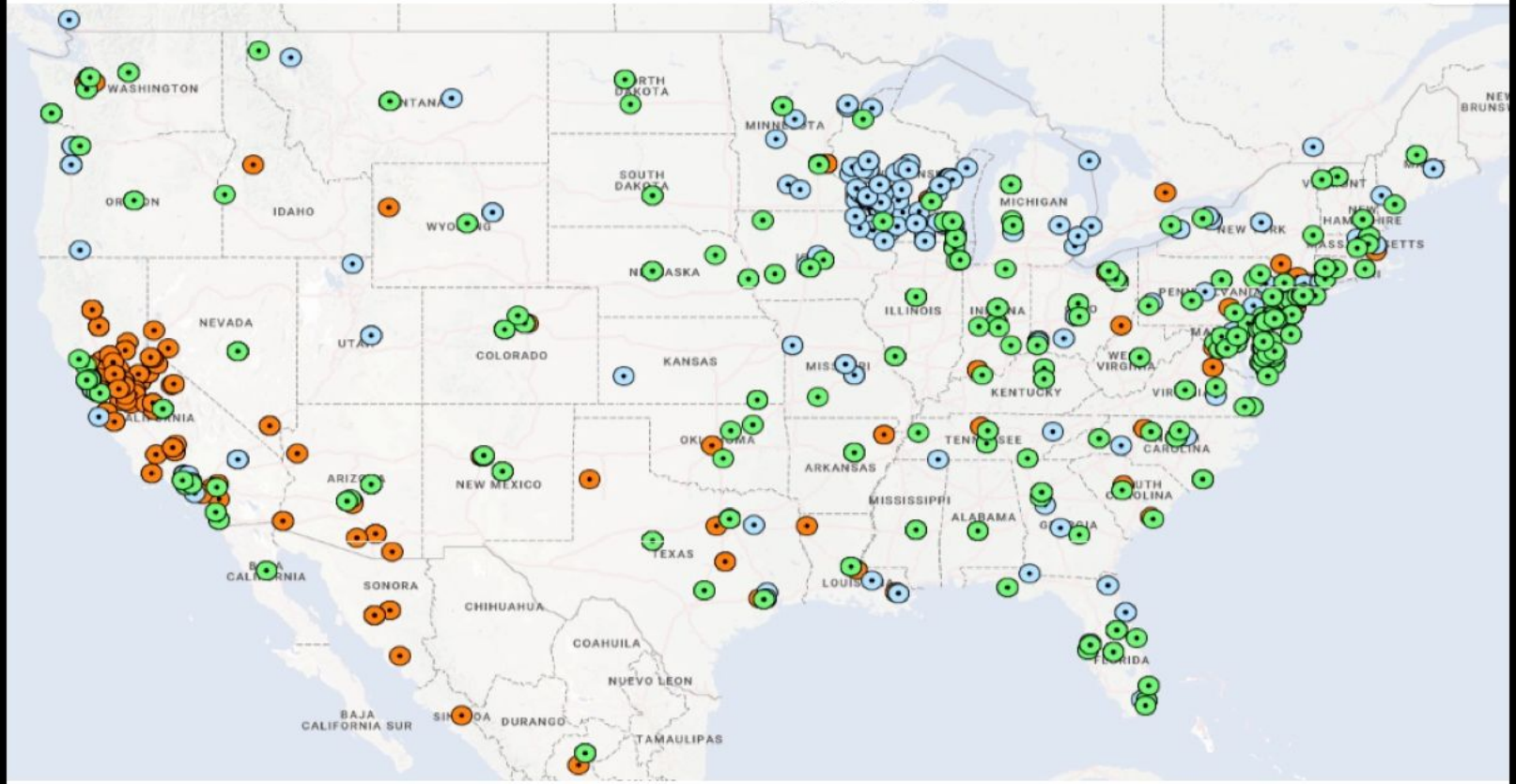
Case Study #5



DeWitt Wallace Center for Media & Democracy

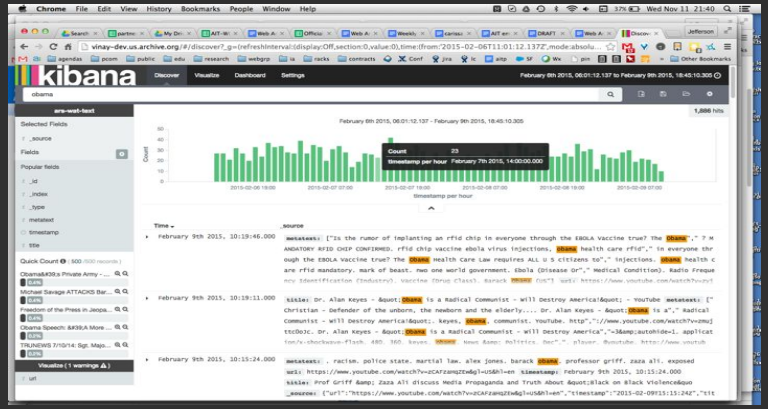
democracy fund

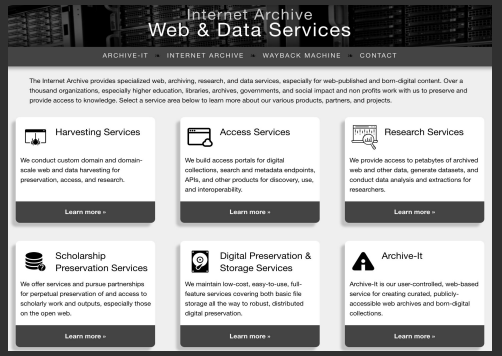
Location Identification and Mapping

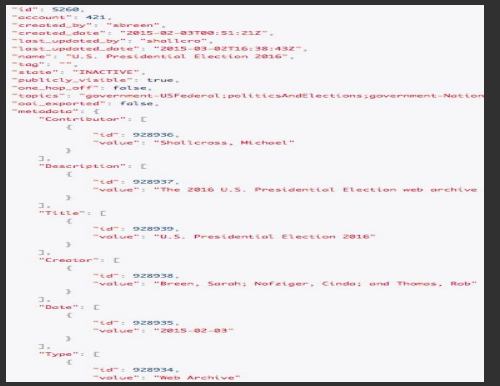


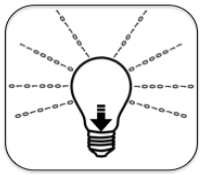
Mapping representation community coverage of local news in Stockton, CA (reddish orange), La Cross, WI (blue) and Newark, DE (greenish yellow). Data were extracted using the Location Identification API, converted to latitute / longitude and mapped using tools available in R.

Various Web Archive Data Services

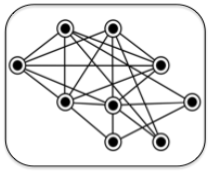








WAT Datasets
(Web Archive Transformation)
Key Metadata from Every Resource



LGA Datasets
(Longitudinal Graph Analysis)
What Links to What over Time

WANE Datasets
(Web Archive Named Entities)
Names of People, Places, Organizations

CEDWARC

ARCH

Reflections & Lessons Learned

- Computational use of digital archives is increasing
- Digital archives support research uses beyond documentary
- Digital archives have unique affordances for large scale use
- Digital archives are imperfect (physical are too, if less clearly)
- Greater complexity for libraries helping scope research question
- Greater complexity for libraries helping identify relevant collections
- Libraries need to deliver data in many way (bulk, APIs, middleware)
- Libraries need to deliver data in many forms (raw to derived)
- Libraries need to help mitigate technical & methodological issues
- Supporting computational use can be collaborative not transactional

THANKS!

Jefferson Bailey
Director of Archiving & Data Services
jefferson@archive.org | @jefferson_bail

Internet Archive, https://archive.org

Archive-It, https://archive-it.org