# Alternatives to Searching and Browsing a Massive Collection of Documents

## The Case of the Opioid Industry Documents Archive

Kevin Hawkins


UCSF
University of California
San Francisco

Johns Hopkins University

1. What is the Opioid Industry Documents Archive (OIDA)?
2. Why is the web interface to this digital collection insufficient?
3. How do we aim to support use of computational methods by humanities scholars and others?

# What is the Opioid Industry Documents Archive (OIDA)?

**The Opioid Industry Documents Archive serves as a groundbreaking digital archive of documents from the opioid industry that advances understanding of the root causes of the opioid epidemic and helps address corporate behavior that is harmful to the public.**

# The Opioid Industry Documents Archive…

**Consolidates information into an easily accessible digital archive for use by researchers, journalists, policy-makers, community members, and other stakeholders**

- **Aids in the search, retrieval and analysis of documents arising from opioid litigation including:**
  - Court and litigation documents (expert depositions, complaints and amendments, internal company communications)
  - Documents from federal, state and local government agencies
  - Testimony from advocacy organizations (accounts of people with lived experience, addiction, and impact of the epidemic)
- **Offers reference and technical assistance to those wishing to use materials**
- **Serves as a repository for new information from future cases**
- **Provides evidence to support changes in policy and practice**

# Massachusetts Governor Maura Healey

*A top focus of our opioid litigation has always been about ensuring transparency and telling the story of the harm done to families across this country. Making this evidence available to the public will pave the way for important reforms that will save lives and let the families, survivors, and advocates who have been affected by the opioid epidemic see the truth. We thank UCSF and Johns Hopkins University for launching this digital document archive as we continue the fight to make evidence in our opioid cases public.*

**Maura Healey**
*Attorney General of Massachusetts (2021)*

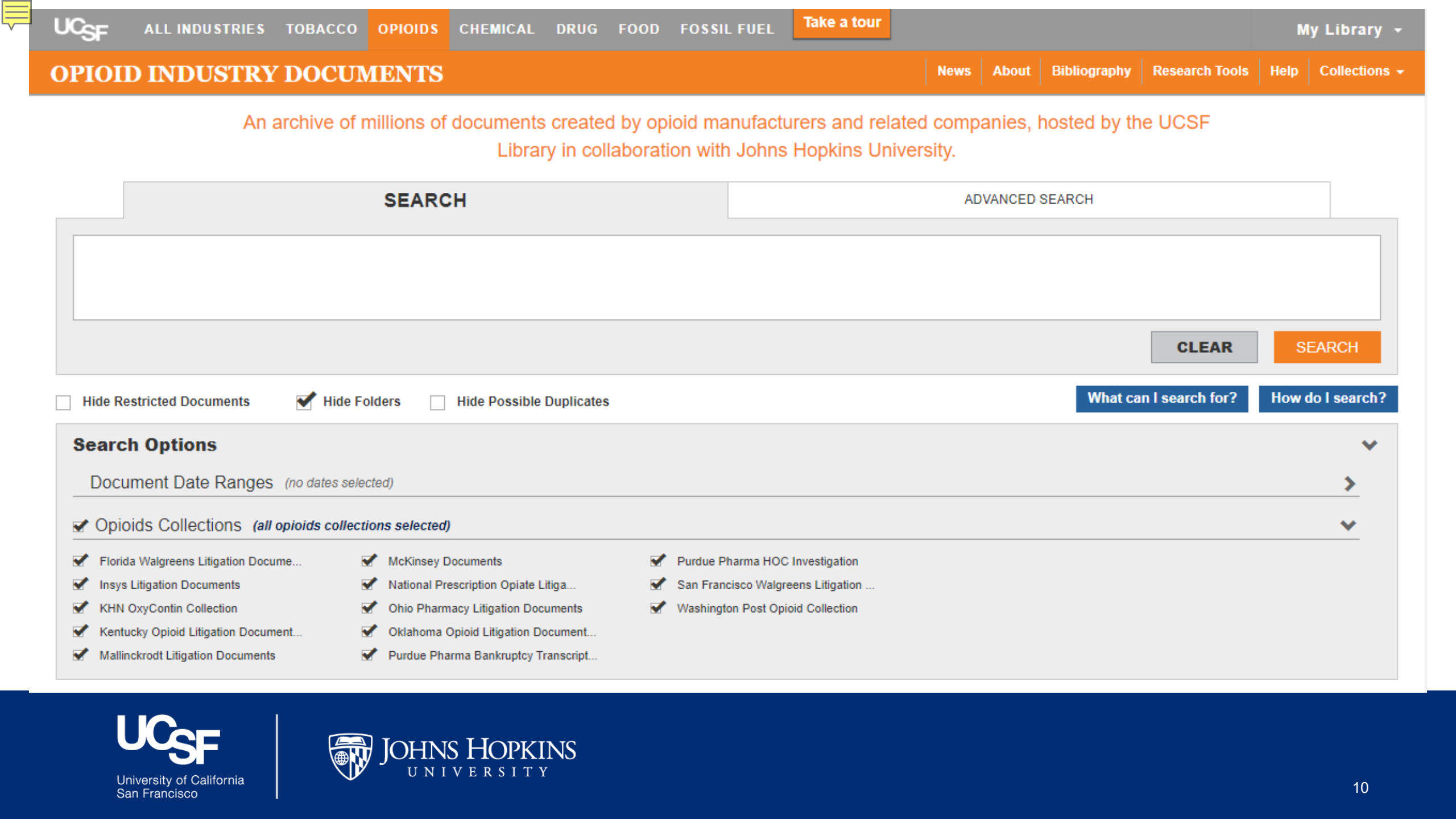
# OIDA Addresses Many Important Questions

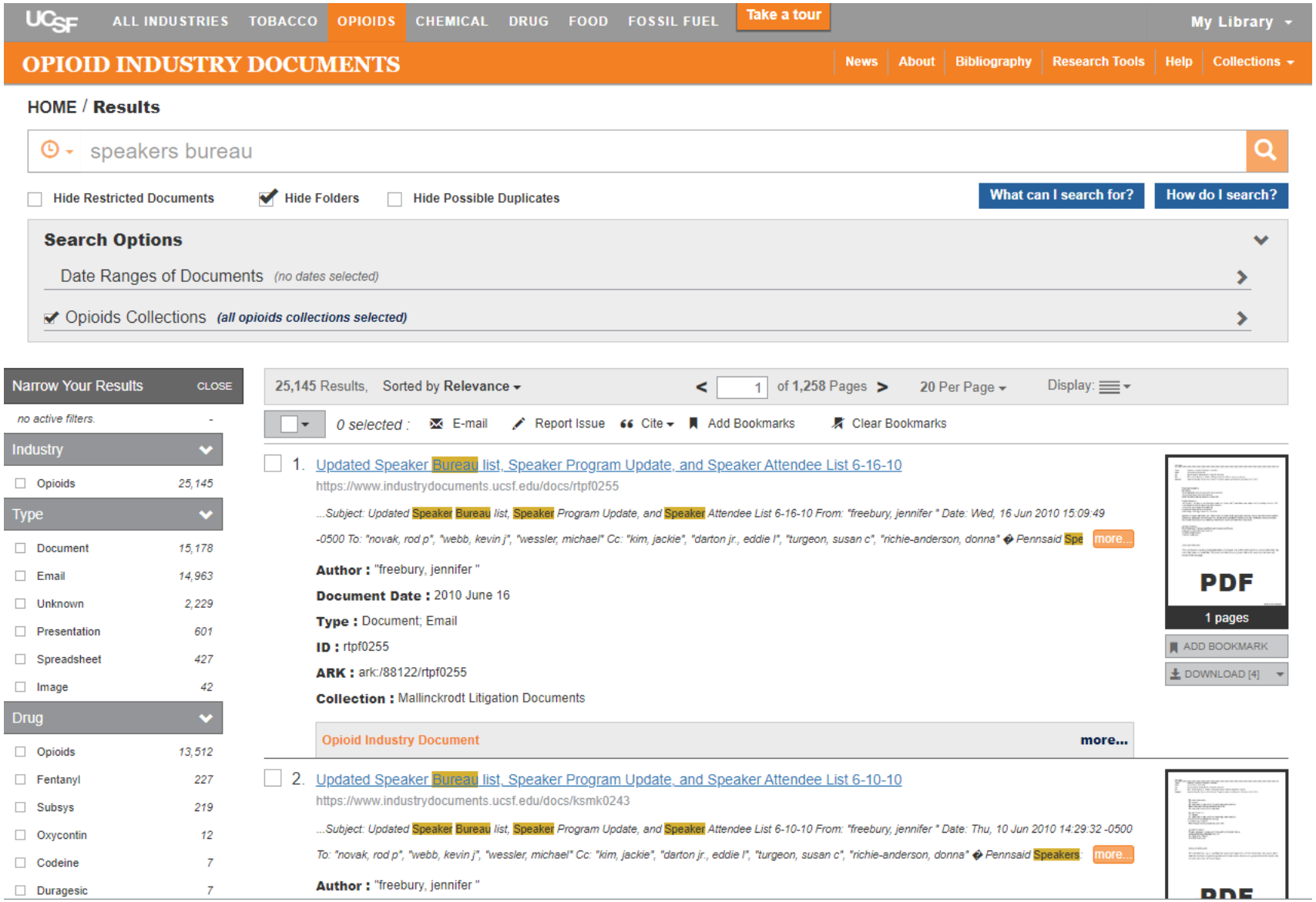
1. How have **pharmaceutical companies** driven the current opioid epidemic?
2. What was **known within these companies, at what time, by whom**, regarding potential risks of addiction and overdose?
3. How have **marketing and promotional practices deviated** from FDA standards?
4. How have **intermediary organizations, such as advocacy organizations,** been used by the industry to promote messaging that misconstrues the risks and benefits of opioids?
5. What role have **other stakeholders (e.g., wholesalers, pharmacies)** played in driving opioid oversupply?
6. What **changes in policy** are needed to prevent such practices in the future?
7. What information about the **risks and benefits of opioids** should be corrected among prescribers and patients?
8. What are the most **evidence-based ways to abate further harms**, and at what cost?

# Why is our web interface insufficient?

# A massive and growing collection

- Online as of May 2023: 12,462,061 pages in **3,119,568 documents**
- About **1.5 million documents in the pipeline** (most need to be redacted before they can go online)
- About **1.9 million documents** that we are likely to receive by mid-2024
- **Tens of millions** more documents that might become available for disclosure due to pending litigation


UCSF
ALL INDUSTRIES
TOBACCO
OPIOIDS
CHEMICAL
DRUG
FOOD
FOSSIL FUEL
Take a tour
My Library
OPIOID INDUSTRY DOCUMENTS
News
About
Bibliography
Research Tools
Help
Collections
An archive of millions of documents created by opioid manufacturers and related companies, hosted by the UCSF Library in collaboration with Johns Hopkins University.
SEARCH
ADVANCED SEARCH
CLEAR
SEARCH
Hide Restricted Documents
Hide Folders
Hide Possible Duplicates
What can I search for?
How do I search?
Search Options
Document Date Ranges (no dates selected)
Opioids Collections (all opioids collections selected)
Florida Walgreens Litigation Docume...
Insys Litigation Documents
KHN OxyContin Collection
Kentucky Opioid Litigation Document...
Mallinckrodt Litigation Documents
McKinsey Documents
National Prescription Opiate Litiga...
Ohio Pharmacy Litigation Documents
Oklahoma Opioid Litigation Document...
Purdue Pharma Bankruptcy Transcript...
Purdue Pharma HOC Investigation
San Francisco Walgreens Litigation ...
Washington Post Opioid Collection
UCSF
University of California San Francisco
JOHNS HOPKINS UNIVERSITY


UCSF
ALL INDUSTRIES
TOBACCO
OPIOIDS
CHEMICAL
DRUG
FOOD
FOSSIL FUEL
Take a tour
My Library
OPIOID INDUSTRY DOCUMENTS
News
About
Bibliography
Research Tools
Help
Collections
HOME / Results
speakers bureau
Hide Restricted Documents
Hide Folders
Hide Possible Duplicates
What can I search for?
How do I search?
Search Options
Date Ranges of Documents (no dates selected)
Opioids Collections (all opioids collections selected)
Narrow Your Results
CLOSE
no active filters.
Industry
Opioids 25,145
Type
Document 15,178
Email 14,963
Unknown 2,229
Presentation 601
Spreadsheet 427
Image 42
Drug
Opioids 13,512
Fentanyl 227
Subsys 219
Oxycontin 12
Codeine 7
Duragesic 7
25,145 Results, Sorted by Relevance
1 of 1,258 Pages
20 Per Page
Display:
0 selected :
E-mail
Report Issue
Cite
Add Bookmarks
Clear Bookmarks
1. Updated Speaker Bureau list, Speaker Program Update, and Speaker Attendee List 6-16-10
https://www.industrydocuments.ucsf.edu/docs/rtpf0255
...Subject: Updated Speaker Bureau list, Speaker Program Update, and Speaker Attendee List 6-16-10 From: "freebury, jennifer " Date: Wed, 16 Jun 2010 15:09:49 -0500 To: "novak, rod p", "webb, kevin j", "wessler, michael" Cc: "kim, jackie", "darton jr., eddie l", "turgeon, susan c", "richie-anderson, donna" Pennsaid Spe more...
Author : "freebury, jennifer "
Document Date : 2010 June 16
Type : Document; Email
ID : rtpf0255
ARK : ark:/88122/rtpf0255
Collection : Mallinckrodt Litigation Documents
Opioid Industry Document
more...
PDF
1 pages
ADD BOOKMARK
DOWNLOAD [4]
2. Updated Speaker Bureau list, Speaker Program Update, and Speaker Attendee List 6-10-10
https://www.industrydocuments.ucsf.edu/docs/ksmk0243
...Subject: Updated Speaker Bureau list, Speaker Program Update, and Speaker Attendee List 6-10-10 From: "freebury, jennifer " Date: Thu, 10 Jun 2010 14:29:32 -0500 To: "novak, rod p", "webb, kevin j", "wessler, michael" Cc: "kim, jackie", "darton jr., eddie l", "turgeon, susan c", "richie-anderson, donna" Pennsaid Speakers: more...
Author : "freebury, jennifer "

# Needing to support computational methods

# Research questions that require computational methods

- **Social network analysis**: Which employees within a company communicated with each other most frequently, and during which time periods? What does this social network tell us about how communication and decision-making work in a large organization?
- **Sentiment analysis**: Did employees in certain roles use language (euphemisms, tone, commands, etc.) in certain ways, or did it change over time? What does this tell us about how language is used in the modern workplace and in relation to specific steps in the wrongdoing of these companies?
- Other forms of **text mining**:
  - At what point were key decisions made, and by whom? How did these patterns differ from one company to another?
  - After manually locating a few instances of deception or misconstrual of evidence in the OIDA corpus, can we detect more instances automatically using machine-learning models?

# Supporting computational methods by humanities scholars and others

# What functionality do we think we need to support?

- Querying and retrieving documents by metadata, by full text, and by format
- Processing documents remotely (including on cloud platforms) or locally
- Building visualizations and other tools on top of our data without requiring people to download everything
- Integration of our content with other document corpora

*And we need to support users of different technical abilities!*

# What are we exploring?

- Making data and Jupyter notebooks available in SciServer

  As my colleagues presented yesterday, we're ready for people to try using our metadata now; we're still preparing the full content for sharing.
- Making data available in the Registry of Open Data on AWS
- Creating custom datasets of particular types of documents
- Creating an API for querying and retrieving full content

  Currently we have an API for metadata only.
- Working with the developers of the Archives Research Compute Hub (ARCH) to allow for OIDA documents to be searched and queried alongside web archives

# To receive updates on OIDA, including our support for computational methods

purl.org/oida/subscribe