

## Data from Docs: Final Report

March 17, 2023 // CS 206: Exploring Computational Journalism

Mahammad Shirinov, Justin Tinker, Cooper Reed, and Tianyu Fang

*It is no news that journalism is confronted with more challenges than ever.* While national newsrooms are struggling to monetize and adapt to new digital landscapes, local news—arguably with fewer technical resources—is struggling to survive. According to University of North Carolina researchers, there are [some 200 “news deserts”](#)—counties without a local newspaper—in the United States. And the number of localities without local news is growing fast: the US has [lost](#) more than 2,500 local newspapers since 2005. Digital alternatives, too, aren’t sufficient substitutes: most of the 545 digital-only state and local news sites [employ](#) no more than six full-time reporters.

National news is no substitute for local journalism. Reporters who cover city halls, police departments, and public schools provide citizens with important resources and hold government officials accountable; strong local news is the backbone of our democracy.

How can we make local government documents more accessible to journalists? We had two ideas, after interviewing technical and non-technical journalists from national, local, and independent publications.

First, with fewer local reporters, *sourcing* has become harder. In a bygone era, it was possible for journalists to go through agendas and attend meetings at the government department of their beat, in a city that has its own news agency. A local reporter in 2023, however, is more likely to cover several cities in the region. It’s important to know which events, meetings, documents, or bills are deserving of attention and coverage.

Second, local government documents aren’t sufficiently *networked*. What happens in Palo Alto doesn’t stay in Palo Alto. Reporters we’ve talked to want to draw connections between cases in different localities: if there’s someone suing municipalities across the state for ADA noncompliance, or a nonprofit that gets procurement contracts from several cities, it’s difficult to search through the government websites of every town in California to know where to look. We wanted to help reporters identify new insights across databases.

There are already tools that address these problems—well, sort of. We’re inspired by [AgendaWatch](#), a project that was born in this course, which allows journalists to subscribe to topics and regions they’re interested in; the relevant government documents are delivered to their inboxes. There’s also Datashare, Pinpoint, TK.




Among existing solutions, we found Datashare and Pinpoint to have great breadth in their ability to comb content, but poor depth in their query capabilities. While common search terms could be found and cross-referenced, both tools lacked the capability to generate and search semantically related queries. Additionally, their interfaces were visually imposing and

information-poor. Also included is a shot of OCCRP's Aleph, which we were unable to use but which lacks the same capabilities as Datashare and Pinpoint.


The screenshot displays the Aleph search interface. On the left is a dark sidebar with a 'FILTERS' section. The 'Contextualize filters' toggle is turned on. Under 'Languages', the 'Exclude' button is highlighted. The language counts are: All (1,086), English (994), French (80), German (10), and Italian (2). Other filter categories include Projects, Tags, Recommended by, File types, Creation dates, People, Organizations, and Locations.

The main search area has a search bar containing the text "beneficial owner". Below the search bar, there are tabs for 'beneficial owner' (selected), 'English', and 'Clear all filters'. The results are sorted by '1 - 43 on 43 documents'. The results are displayed in a grid of document thumbnails. Each thumbnail shows a document title, a snippet of text, and a 'BUREAU D'ENQUÊTE' stamp. The documents are related to various entities and projects, including 'Permira\_projet Nucleus\_18030...', 'ITS Umbrella\_K\_TMD\_100310...', 'ITS Umbrella\_TMD\_100310.pdf', 'Orion Income Finance - ATC 20...', 'Credit Suisse\_Bristol\_Leeds\_19...', 'tigerglobal\_1.pdf', 'ITS Umbrella\_Next Group\_1003...', and 'ITS Umbrella\_Ace Group\_1003...'. The interface also shows a '6.5.2' version number in the bottom left corner.


## Datashare.

 moon  


Documents 1 - 29 of 29

 **datacard.jpg.pdf**

1:30 REMOVE COVER POINT ANTENNA 3:00 NEAR FIELD DECAL 9:00 S-BD  
TO LUNAR STAY 10:30 REST/PHOTO S-BD MOD - FM 12:00 TV CB - IN

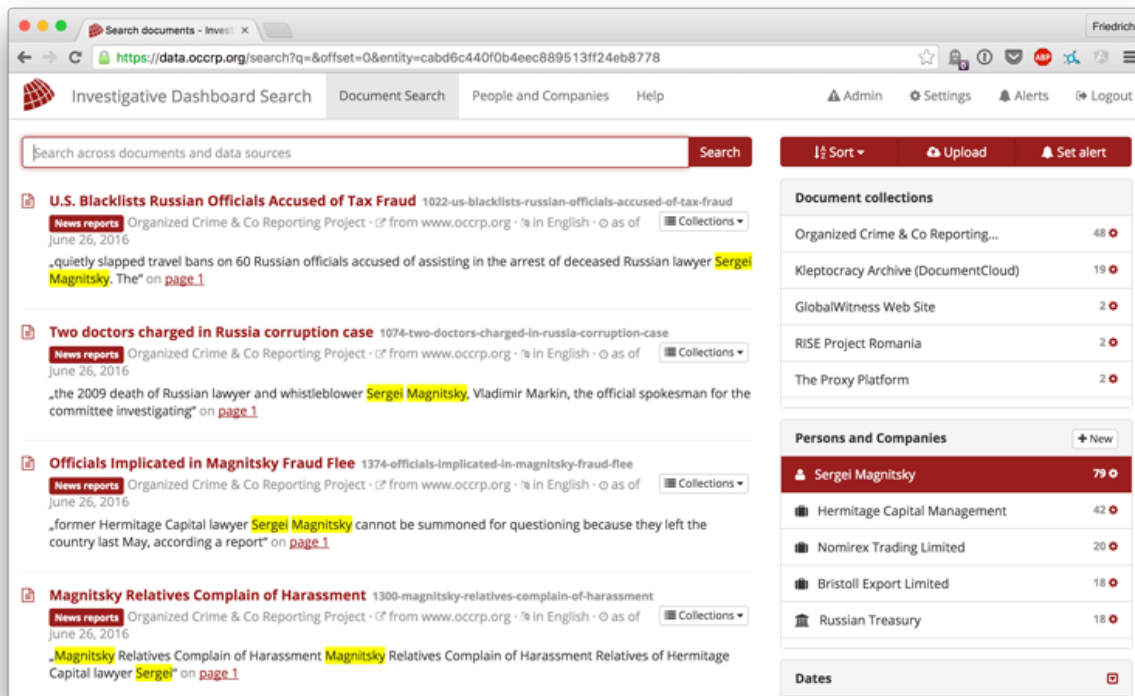
 **Apollo11\_flightPlan\_6734365.jpg.pdf**

PERFORM LUNAR CONTACT CHECKLIST FDAI : STAY/NO STAY RO PO STAY/  
NO STAY YO STOP 16mm CAMERA INITIATE DPS VENTING ASCENT

 **star\_chart.jpg.pdf**

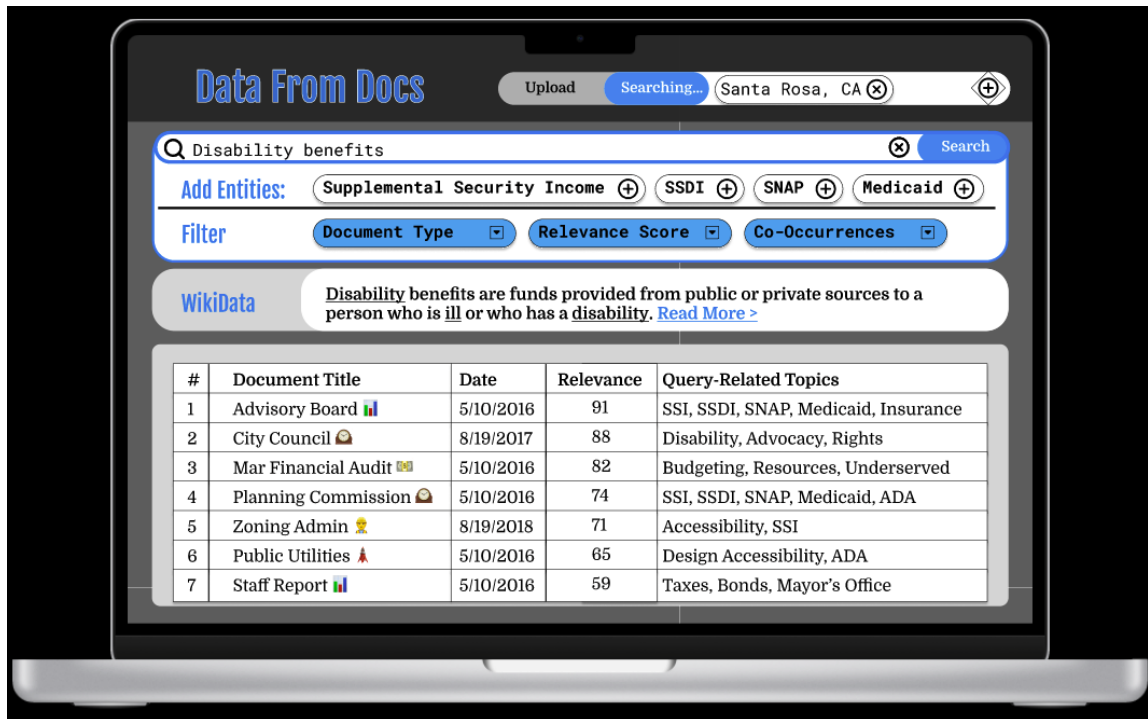
193 240° SUN VENUS MOON 900 NUNKI 37 .- MARS SPICA 26 REGULUS 22  
ALDEBARAN 10° ANTARES -10° 1500 • TROP 24 OGIEHAH 2100 • MENKAR

## Pinpoint.



## Aleph.

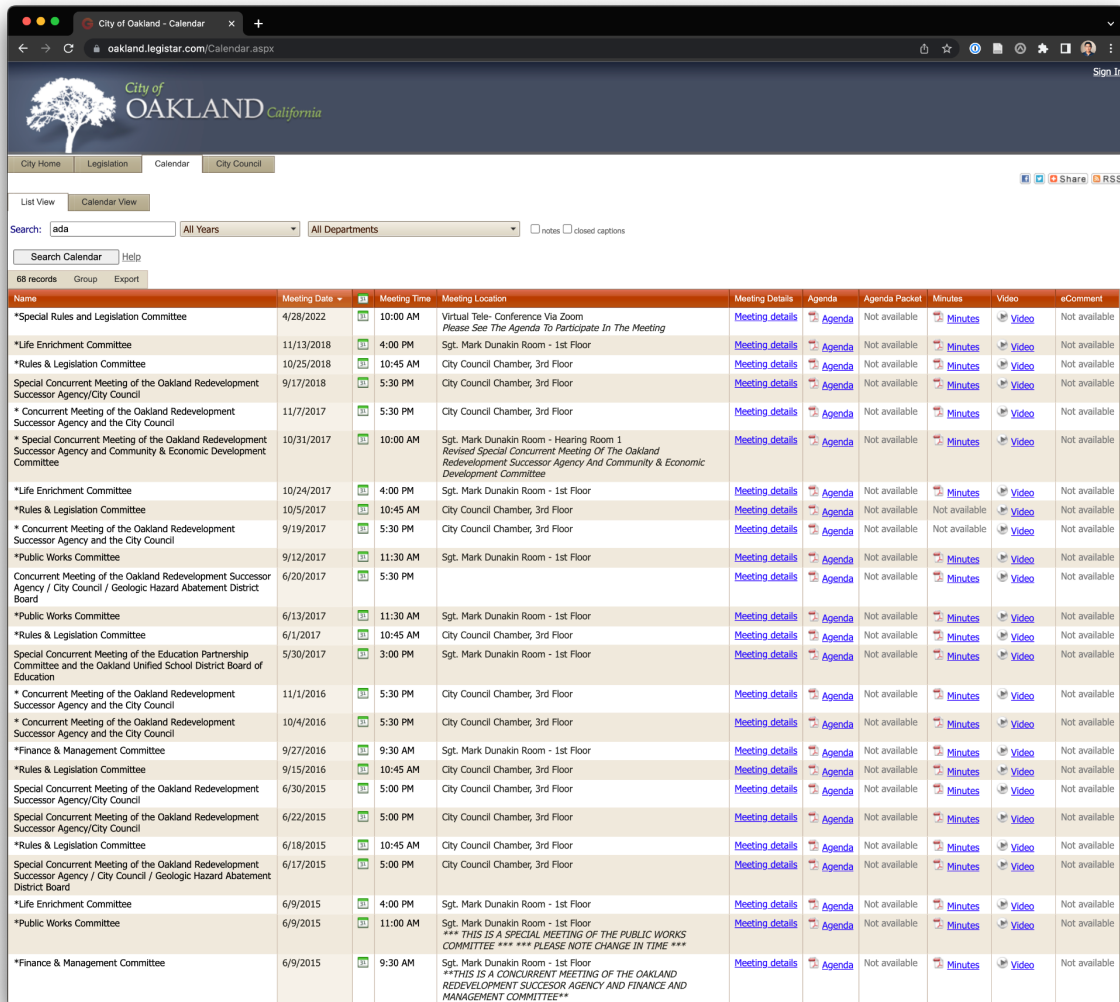
Our two main improvements on these capabilities, a search engine for text passages and a semantic cross-reference function, would become the foundation for our tool. In order to address sourcing issues from other tools, and obviate the need for journalists to upload their own data, we utilized a set of APIs from Legistar (more details below). In order to address networking issues from other tools, and obviate the need for journalists to search through libraries of documents from different jurisdictions, we cross-networked as many Legistar instances as we could, given our limitations. A long-term solution would permanently maintain and update these document repositories, in addition to fully streamlining the drilldown process during which documents are searched through with fine-toothed filters. In order to visualize this “ideal final iteration”, we developed a set of idealized user flows which would fit the updated database and search capabilities. These flows are below, but do not contain any backend data.





Our Attempt

Government agencies typically use Legistar to publish documents for public disclosure. These documents include event agendas, minutes of meetings, and government reports. Sometimes there are even videos for the meeting! Here's what the Legistar portal looks like in [Oakland, California](#):



City of Oakland - Calendar

oakland.legistar.com/Calendar.aspx

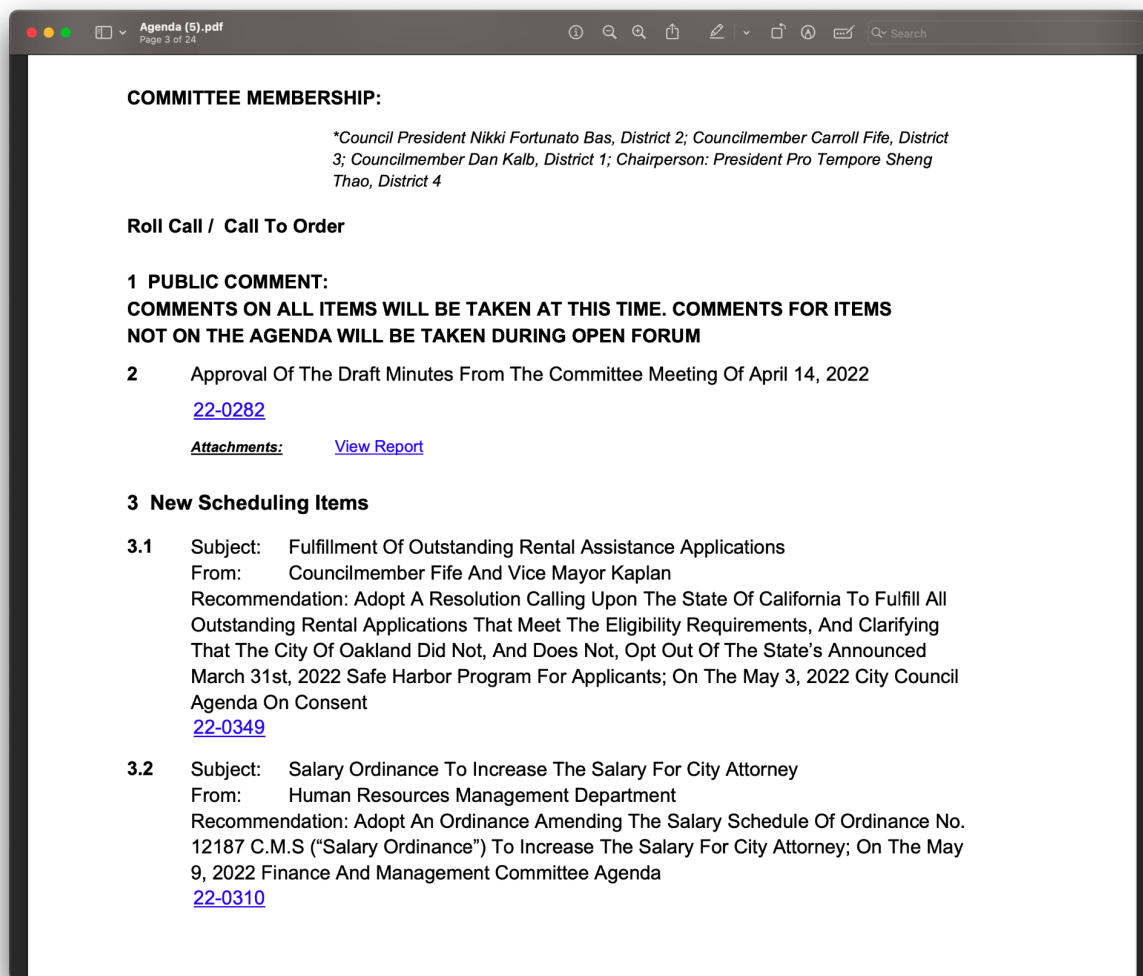
City Home Legislation Calendar City Council

Search: ada All Years All Departments

68 records

Name	Meeting Date	Meeting Time	Meeting Location	Meeting Details	Agenda	Agenda Packet	Minutes	Video	eComment
*Special Rules and Legislation Committee	4/28/2022	10:00 AM	Virtual Tele- Conference Via Zoom Please See The Agenda To Participate In The Meeting	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Life Enrichment Committee	11/13/2018	4:00 PM	Sgt. Mark Dunakin Room - 1st Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Rules & Legislation Committee	10/25/2018	10:45 AM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
Special Concurrent Meeting of the Oakland Redevelopment Successor Agency/City Council	9/17/2018	5:30 PM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
* Concurrent Meeting of the Oakland Redevelopment Successor Agency and the City Council	11/7/2017	5:30 PM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
* Special Concurrent Meeting of the Oakland Redevelopment Successor Agency and Community & Economic Development Committee	10/31/2017	10:00 AM	Sgt. Mark Dunakin Room - Hearing Room 1 Revised Special Concurrent Meeting Of The Oakland Redevelopment Successor Agency And Community & Economic Development Committee	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Life Enrichment Committee	10/24/2017	4:00 PM	Sgt. Mark Dunakin Room - 1st Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Rules & Legislation Committee	10/5/2017	10:45 AM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
* Concurrent Meeting of the Oakland Redevelopment Successor Agency and the City Council	9/19/2017	5:30 PM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Public Works Committee	9/12/2017	11:30 AM	Sgt. Mark Dunakin Room - 1st Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
Concurrent Meeting of the Oakland Redevelopment Successor Agency / City Council / Geologic Hazard Abatement District Board	6/20/2017	5:30 PM		<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Public Works Committee	6/13/2017	11:30 AM	Sgt. Mark Dunakin Room - 1st Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Rules & Legislation Committee	6/1/2017	10:45 AM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
Special Concurrent Meeting of the Education Partnership Committee and the Oakland Unified School District Board of Education	5/30/2017	3:00 PM	Sgt. Mark Dunakin Room - 1st Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
* Concurrent Meeting of the Oakland Redevelopment Successor Agency and the City Council	11/1/2016	5:30 PM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
* Concurrent Meeting of the Oakland Redevelopment Successor Agency and the City Council	10/4/2016	5:30 PM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Finance & Management Committee	9/27/2016	9:30 AM	Sgt. Mark Dunakin Room - 1st Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Rules & Legislation Committee	9/15/2016	10:45 AM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
Special Concurrent Meeting of the Oakland Redevelopment Successor Agency/City Council	6/30/2015	5:00 PM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
Special Concurrent Meeting of the Oakland Redevelopment Successor Agency/City Council	6/22/2015	5:00 PM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Rules & Legislation Committee	6/18/2015	10:45 AM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
Special Concurrent Meeting of the Oakland Redevelopment Successor Agency / City Council / Geologic Hazard Abatement District Board	6/17/2015	5:00 PM	City Council Chamber, 3rd Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Life Enrichment Committee	6/9/2015	4:00 PM	Sgt. Mark Dunakin Room - 1st Floor	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Public Works Committee	6/9/2015	11:00 AM	Sgt. Mark Dunakin Room - 1st Floor *** THIS IS A SPECIAL MEETING OF THE PUBLIC WORKS COMMITTEE *** PLEASE NOTE CHANGE IN TIME ***	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available
*Finance & Management Committee	6/9/2015	9:30 AM	Sgt. Mark Dunakin Room - 1st Floor ***THIS IS A CONCURRENT MEETING OF THE OAKLAND REDEVELOPMENT SUCCESSOR AGENCY AND FINANCE AND MANAGEMENT COMMITTEE**	<a href="#">Meeting details</a>	<a href="#">Agenda</a>	Not available	<a href="#">Minutes</a>	<a href="#">Video</a>	Not available

Screenshot of Oakland Legistar.



A sample meeting agenda from Oakland.

As you can see, not only does the catalog look messy, the sheer amount of text in these government documents makes it difficult for reporters to search for relevant files and identify what's worth spending time on. But if we look more closely, we can see that these documents contain many *named entities*:



### ISSUE(S)

Shall the Council, by motion, approve the Third Amendment to General Services Agreement F000096 for a one-year extension for weed and rubbish abatement services, with no increase in fees, to Kenneth Ray Tamplen, Pleasant Hill, California, in the amount of \$120,000, subject to approval as to form by City Attorney?

DOCUMENT

PERSON CITY

### BACKGROUND

1. On June 7, 2011, as a result of a competitive bidding process, Contract No. F000096 was awarded by Council to Kenneth Ray Tamplen for weed and rubbish abatement services in the amount of \$200,000 for a two-year period, with three one-year renewal options. The First and Second Amendments extended the contract term and added funds. The current contract term expires April 30, 2014.
2. The weed and rubbish abatement services contract provides abatement of weeds and other ground cover fire or health hazards as determined by the Fire Department. The properties upon which abatement work is done consists of streets, parkways, sidewalks, parks, commercial or residential property, vacant or otherwise, upon which weeds or refuse have become a fire or health hazard.

DATE

DOCUMENT PERSON

AGENCY

These names of individuals, cities, government agencies, documents, and dates that are connected to one another. We know that Kenneth Ray Tamplen lives in Pleasant Hill; they were awarded the contract F000096; they are in the weed and rubbish abatement business. That means we can extract these entities and turn them into a knowledge map that looks something like this:



# Keyword Lookup

Browse all possible connections of any two entities in the minutes of a given municipality.

## Search

City

Entity 1

Entity 2



A screenshot from our double entity search page.

Double entity search answers the question “What’s the relationship between ENTITY\_\_\_\_ and ENTITY\_\_\_\_ as recorded in CITY\_\_\_\_’s municipality meetings?”. Again, the entities can be any Named Entity as defined in the traditional NLP literature; the most relevant for the journalism use-cases are usually people, places and organizations. For example, we can ask our service to bring up the relationship of “Columbus Avenue” (an avenue in San Francisco) and “Planning Code”, in the city San Francisco. It will then search and uncover any documents where the two entities have been mentioned together, and return them nicely in a list, as shown in figure below:

# Keyword Lookup

Browse all possible connections of any two entities in the minutes of a given municipality.

## Results

- Approval of a 90-Day Extension for Planning Commission Review of an Ordinance (File No. 131120) Establishing the Broadway Alcohol Restricted Use District | Apr 22, 2014

Attachments:

- Board\_Pkt\_042214 | Nov 06, 2015 [url](#)
  - Leg Ver1 | Nov 06, 2015 [url](#)
  - Leg Final | Nov 06, 2015 [url](#)
- Approval of a 180-Day Extension for Planning Commission Review of an Ordinance (File No. 131120) Establishing the Broadway Alcohol Restricted Use District | Jul 15, 2014

Attachments:

- BOS File No. 131120 | Nov 06, 2015 [url](#)
  - Board Pkt 071514 | Nov 06, 2015 [url](#)
  - Leg-Ver1 | Nov 06, 2015 [url](#)
  - Leg Final | Nov 06, 2015 [url](#)
- Petitions and Communications | Aug 02, 2016

Attachments:

- Board Pkt 080216 | Aug 16, 2017 [url](#)
- Approval of a 60-Day Extension for Planning Commission Review of an Ordinance (File No. 131120) That Would Establish the Broadway Alcohol Restricted Use District | Mar 04, 2014

Attachments:

- Leg Ver1 | Nov 06, 2015 [url](#)
- Board\_Pkt\_030414 | Nov 06, 2015 [url](#)
- Leg\_Final | Nov 06, 2015 [url](#)

[back to search](#)

A screenshot from the results page of our double entity search.

Every hit comes with a list of Attachments relevant to the municipal meeting that the match was found in, so the journalist can now follow those links and dig deeper into the relationship.

Our second feature, document search, allows one to upload a text or a document that they have and which they're interested to learn more about. For example, a journalist can upload a transcript of a conversation they had with someone, or even the notes they took, and the system will automatically recognize the important entities in that text and generate leads for them to take up. The output of a sample document search looks like this:

# Document Parse & Reference

Browse top references of entities in your documents.

## Matches

### Entity: Mary Traverso Open Space (frequency: 1)

- Name Change of Creekside Open Space to Mary Traverso Open Space | Aug 31, 2021  
Attachments:
  - Staff Report | Aug 25, 2021 [url](#)
  - Attachment 1- Creekside Open Space Proposed Renaming Report | Jul 28, 2021 [url](#)
  - Attachment 2- City Council Policy 000-25 | Jul 28, 2021 [url](#)
  - Resolution | Sep 01, 2021 [url](#)
  - Presentation | Jul 28, 2021 [url](#)
  - Attachment 3 - Location Map | Aug 19, 2021 [url](#)
  - Late Correspondence (Uploaded 8-31-2021) | Aug 31, 2021 [url](#)
  - Late Correspondence (Uploaded 9-2-2021) | Sep 02, 2021 [url](#)
- (No title available) | Dec 09, 2020  
Attachments:
  - Creekside Proposed Renaming PPT | Dec 03, 2020 [url](#)
- (No title available) | May 26, 2021  
Attachments:
  - Creekside Renaming PPT | May 20, 2021

A screenshot from the results page of our document search.

## Input Text

Mary Traverso Open Space **FAC** was visited by workers from Sonoma County Homeless System of Care **ORG** and the U.S. Department of Housing and Urban Development **ORG**.

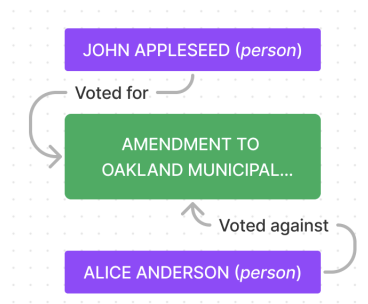
Here, we extract the entities, sort them by importance in the input text by counting how many times they were repeated, and search for mentions for them elsewhere. It's easy to extend this functionality to rather search for pairs or triplets of entities, and follow them together in other documents and places, but this is left for future work.

## Challenges, and Looking Ahead

Throughout the project we've run into several technical challenges. First, off-the shelf NER tools are too generic and can't always detect entities in a useful format. You might think "ADA," "Americans with Disabilities Act," and "Disabilities Act" all refer to the same subject, but it's much harder for algorithms to pick up on that. One potential solution is to fine-tune our own model to pick up important terms, training it to learn what entities should be merged. Secondly, our tool incurs high computing cost given the sheer amount of data; if deployed for even larger datasets, organizations might not have the computing resources for it.

Looking ahead, there are a few more features we would like to implement beyond what we have done in the past ten weeks.

First, topic-based retrievals. While the entity “fire department” can appear in all sorts of documents, what if we wanted it only in the context of forest fires? Second, relationships between entities. Currently, any link between entities is a generic, “Entity A is related to Entity B.” We want to reflect more nuanced relationships in our knowledge graph, to know that Steph Curry isn’t just “related to” the City of Atherton, but he actually “lives” there. Or Gavin Newsom isn’t just “related to” California; he “is the governor of” it! Something like this:



And last but not the least, our current model doesn’t support entity linking. In documents and reports, you’d see something like this: “The Currys’ acknowledgement of their reluctance to ‘add to the “not in our backyard” (literally) rhetoric’ shows they realized the narrative is basically a layup.” Clearly “The Currys” in this sentence is a reference to Steph Curry and Ayesha Curry, but our current NER does not link “The Currys” to Steph Curry and Ayesha Curry.

We’d also like to see our tool integrate the subscription feature from AgendaWatch. For AgendaWatch, a reporter could sign up for updates in specific localities pertaining to their topics of interest. In the next step, they should be able to look up their entities of interest in our knowledge graph and get an email alert should there be any future updates—a useful way to keep track of developments across regions for longer-term projects.