## **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 <u>some 200 "news deserts"</u>—counties without a local newspaper—in the United States. And the number of localities without local news is growing fast: the US has <u>lost</u> 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 <u>employ</u> 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.

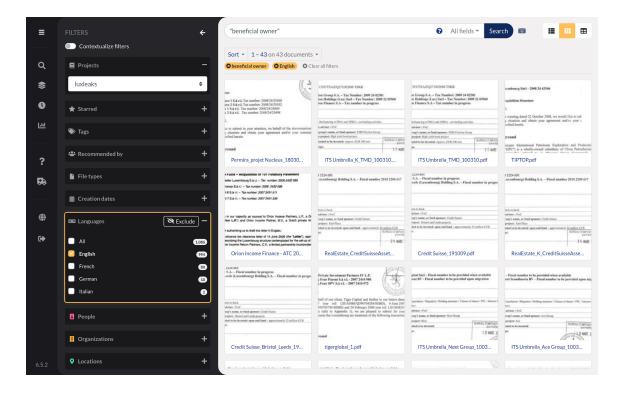
How can we make local government documents more accessible to journalists? We had two ideas, after interviewing technical and non-technical journalists from national, local, and independent publications.

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 <u>AgendaWatch</u>, 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.



# Datashare.

Q	moo	n	× (?)
	Docume	ents 1 - 29 of 29	
	PDF	datacard.jpg.pdf 1:30 REMOVE COVER POINT ANTENNA 3:00 NEAR FIELD DECAL 9:00 S-B TO LUNAR STAY 10:30 REST/PHOTO S-BD MOD - FM 12:00 TV CB - IN	D
	PDF	Apollo11_flightPlan_6734365.jpg.pdf PERFORM LUNAR CONTACT CHECKLIST FDAI : STAY/NO STAY RO PO ST NO STAY YO STOP 16mm CAMERA INITIATE DPS VENTING ASCENT	FAY/
	PDF	star_chart.jpg.pdf 193 240° SUN VENUS MOON 900 NUNKI 37 MARS SPICA 26 REGULUS ALDEBARAN 10° ANTARES -10° 1500 • TROP 24 OGIENAH 2100 • MENKA	

# Pinpoint.

	Search documents - Invest ×							Fried
	C Attps://data.occrp.org/search?q=&	offset=0&entity=cabd6	3c440f0b4eec889513ff24eb	8778		🔒 🛈	) 💟 😳 🐋	0
H	Investigative Dashboard Search	Document Search	People and Companies	Help	🛦 Admin	Settings	Alerts 🕞	Logo
Se	arch across documents and data sources			Search	↓å Sort <del>+</del>	🛆 Upload	🌲 Set ale	ert
	U.S. Blacklists Russian Officials Accused				Document colle	ctions		
	News reports Organized Crime & Co Reporting June 26, 2016	Project · @ from www.	occrp.org · @ in English · ⊘ as	of Ecollections •	Organized Crime		48 (	
	"quietly slapped travel bans on 60 Russian offi Magnitsky. The" on page <u>1</u>	cials accused of assisti	ng in the arrest of deceased F	Russian lawyer <mark>Sergei</mark>	Kleptocracy Archi	loud)	19	
					GlobalWitness W		2	
	Two doctors charged in Russia corruption		RISE Project Rom		2			
	June 26, 2016 "the 2009 death of Russian lawyer and whistle	al spokesman for the	The Proxy Platfor		2			
	committee investigating" on page 1				Persons and Cor	npanies	+7	Nev
	Officials Implicated in Magnitsky Fraud				🛔 Sergei Magnit	sky		79
	June 26, 2016 "former Hermitage Capital lawyer <mark>Sergei Magn</mark>	i Hermitage Capital Management			42			
	country last May, according a report" on page	Nomirex Trac		20				
	Magnitsky Relatives Complain of Haras	sment 1300-magnitsk	y-relatives-complain-of-harassr	ment	Bristoll Expor	t Limited		18
	News reports Organized Crime & Co Reporting June 26, 2016	wa reports Organized Crime & Co Reporting Project - C from www.occrp.org - 14 in English - O as of Collections -						18
	"Magnitsky Relatives Complain of Harassment Capital lawyer Sergel" on page 1	Magnitsky Relatives Co	omplain of Harassment Relati	ives of Hermitage	Dates			(

## 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.

R	oto From Doco						
U	ata From Docs	Up	load Sear	ching Santa Rosa, CA 🛞 🔶			
Q Di	sability benefits			Search			
	Entities: Supplemental	Security	Income (+)	(SSDI () (SNAP () (Medicaid ()			
			Relevance So				
Filt	Cocument Type		(elevance So	core 🖸 (Co-Occurrences 💽)			
Wik			its are funds provided from public or private sources to a				
		_					
#	Document Title	Date	Relevance	Query-Related Topics			
1	Advisory Board 📊	5/10/2016	91	SSI, SSDI, SNAP, Medicaid, Insurance			
2	City Council 🕰	8/19/2017	88	Disability, Advocacy, Rights			
3	Mar Financial Audit 🟁	5/10/2016	82	Budgeting, Resources, Underserved			
4	Planning Commission 🕰	5/10/2016	74	SSI, SSDI, SNAP, Medicaid, ADA			
5	Zoning Admin 👷	8/19/2018	71	Accessibility, SSI			
6	Public Utilities 🗼	5/10/2016	65	Design Accessibility, ADA			
7	Staff Report 📊	5/10/2016	59	Taxes, Bonds, Mayor's Office			
-		(		,			

Filter       Document Type       Relevance Score       Co-Occurrences         Agenda/Memo       Image: Magenda/Memo       Imagenda/Memo       Imagenda/Memo       Imagenda/Memo         WikiData       Meeting Minutes       Imagenda/Memo       Imagenda		isability b <mark>  Entities:</mark>			dicaid 🕀 (	Medicaid
1 Advisory Board 🖬 5/10/2016 91 SSI, SSDI, SNAP, Medicaid, Insurance			Agenda/Memo Meeting Minut	tes 🕥 ome ren,	(SSI) is a <u>mean</u> disabled adults	s-tested program that provides cash a, and individuals aged 65 or older who
	#	Documen	t Title	Date	Relevance	Query-Related Topics
2 Planning Commission 8 8/19/2017 74 SSI SSDI SNAP Medicaid ADA	1	Advisory	Advisory Board 📊		91	SSI, SSDI, SNAP, Medicaid, Insur
	2	Planning	Commission 🕰	8/19/2017	74	SSI, SSDI, SNAP, Medicaid, ADA
3 Zoning Admin 👷 5/10/2016 71 Accessibility, SSI	3	Zoning Ac	lmin 👷	5/10/2016	71	Accessibility, SSI

		0		
	Data From Docs	Upload Sear	ching Santa Rosa, CA 🛞 🕒	
	Q Disability benefits (SSI @	0	DT: Agenda 11 💌 🛞 Search	
	Add Entities: SSDI  SNA		Medicaid	
	Filter (Relevance Sco			
	WikiData payments to disat	curity Income (SSI) is a <u>mean</u> bled children, disabled adult: als of the <u>United States.[1] Re</u>	<u>s-tested</u> program that provides cash , and individuals aged 65 or older who are ad <u>More &gt;</u>	
	# Document Title	Date Relevance	Query-Related Topics	
	1 Advisory Board 🖬	5/10/2016 91	SSI, SSDI, SNAP, Medicaid, Insurance	
and the second second		(	,	
ſ				
	Data From Docs	Upload Sear	ch 🕀	
	Q	Upload Sear	ch 🕑	
	Q Add Entities:		Search	
	Q Add Entities: Filter Document Type	e 💌 Relevance S	Search	
	Q Add Entities: Filter Document Type		Search	
	Q Add Entities: Filter Document Type WikiData will out	e 💌 Relevance S	Search	
	Q Add Entities: Filter Document Type WikiData will out	e 💌 Relevance S	Search	
	Q Add Entities: Filter Document Type WikiData will out	e 💌 Relevance S	Search	
	Q Add Entities: Filter Document Type WikiData will out	Put open-source data here or	Core Co-Occurrences Contraction of the search term is entered.	
	Q Add Entities: Filter Document Type WikiData will out	e 💌 Relevance S	Core Co-Occurrences Contraction of the search term is entered.	
	Q Add Entities: Filter Document Type WikiData will out	Put open-source data here or	Core Co-Occurrences Contraction of the search term is entered.	

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 <u>Oakland, California</u>:

← → C a oakland.legistar.com/Calendar.aspx								) 🖈 🛈	• • •	
City of OAKLAND CALCULATION CONTINUES CALCULATION CALC	alifornia									Sign O Share) 🙆 R
earch: ada All Years	<ul> <li>All Depart</li> </ul>	tment	š	notes Closed captions						
Search Calendar Help										
68 records Group Export										
Name	Meeting Date 👻	31	Meeting Time	Meeting Location	Meeting Details	Agenda	Agenda Packet	Minutes	Video	eComment
*Special Rules and Legislation Committee	4/28/2022	31	10:00 AM	Virtual Tele- Conference Via Zoom Please See The Agenda To Participate In The Meeting	Meeting details	🔁 <u>Agenda</u>	Not available	🔁 <u>Minutes</u>		Not availabl
Life Enrichment Committee	11/13/2018	31	4:00 PM	Sgt. Mark Dunakin Room - 1st Floor	Meeting details	🔁 Agenda	Not available	🔁 Minutes	Video	Not availab
Rules & Legislation Committee	10/25/2018	31	10:45 AM	City Council Chamber, 3rd Floor	Meeting details	🔁 Agenda	Not available	Dinutes Minutes	Video	Not availab
Special Concurrent Meeting of the Oakland Redevelopment Successor Agency/City Council	9/17/2018	31	5:30 PM	City Council Chamber, 3rd Floor	Meeting details	🔁 Agenda	Not available	D Minutes	Ideo Video	Not availabl
* Concurrent Meeting of the Oakland Redevelopment Successor Agency and the City Council	11/7/2017	31	5:30 PM	City Council Chamber, 3rd Floor	Meeting details	🔁 Agenda	Not available	🔁 Minutes		Not availab
Special Concurrent Meeting of the Oakland Redevelopment successor Agency and Community & Economic Development Committee	10/31/2017	31	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	Meeting_details	🔁 <u>Agenda</u>	Not available	D Minutes	. € <u>Video</u>	Not availab
*Life Enrichment Committee	10/24/2017	31	4:00 PM	Sgt. Mark Dunakin Room - 1st Floor	Meeting details	🔁 Agenda	Not available	🔁 Minutes	Video	Not availab
*Rules & Legislation Committee	10/5/2017	31	10:45 AM	City Council Chamber, 3rd Floor	Meeting details	🔂 Agenda	Not available	Not available		Not availab
* Concurrent Meeting of the Oakland Redevelopment Successor Agency and the City Council	9/19/2017	31	5:30 PM	City Council Chamber, 3rd Floor	Meeting details	🔁 Agenda	Not available	Not available		Not availab
*Public Works Committee	9/12/2017	31	11:30 AM	Sgt. Mark Dunakin Room - 1st Floor	Meeting details	🔁 Agenda	Not available	🖘 Minutes	Video	Not availab
Concurrent Meeting of the Oakland Redevelopment Successor Agency / City Council / Geologic Hazard Abatement District Board	6/20/2017	31	5:30 PM		Meeting details	🔁 Agenda	Not available	Dinutes	Video	Not availab
*Public Works Committee	6/13/2017	31	11:30 AM	Sgt. Mark Dunakin Room - 1st Floor	Meeting details	D Agenda	Not available	Dinutes	Video	Not availab
*Rules & Legislation Committee	6/1/2017	31	10:45 AM	City Council Chamber, 3rd Floor	Meeting details	Agenda	Not available	Minutes	Wideo     Video	Not availab
Special Concurrent Meeting of the Education Partnership Committee and the Oakland Unified School District Board of Education	5/30/2017	31	3:00 PM	Sgt. Mark Dunakin Room - 1st Floor	Meeting details	🕏 Agenda	Not available	D Minutes	Video	Not availab
* Concurrent Meeting of the Oakland Redevelopment Successor Agency and the City Council	11/1/2016	31	5:30 PM	City Council Chamber, 3rd Floor	Meeting details	🔁 <u>Agenda</u>	Not available	🗊 <u>Minutes</u>		Not availab
* Concurrent Meeting of the Oakland Redevelopment Successor Agency and the City Council	10/4/2016	31	5:30 PM	City Council Chamber, 3rd Floor	Meeting details	🕏 Agenda	Not available	Dinutes	<sup>™</sup> Video	Not availab
*Finance & Management Committee	9/27/2016	31	9:30 AM	Sgt. Mark Dunakin Room - 1st Floor	Meeting details	🖘 Agenda	Not available	Dinutes	€ <u>Video</u>	Not availab
*Rules & Legislation Committee	9/15/2016	31	10:45 AM	City Council Chamber, 3rd Floor	Meeting details	D Agenda	Not available	Dinutes	Video	Not availab
Special Concurrent Meeting of the Oakland Redevelopment Successor Agency/City Council	6/30/2015	31	5:00 PM	City Council Chamber, 3rd Floor	Meeting details	🔁 <u>Agenda</u>	Not available	🗊 <u>Minutes</u>		Not availab
Special Concurrent Meeting of the Oakland Redevelopment Successor Agency/City Council	6/22/2015	31	5:00 PM	City Council Chamber, 3rd Floor	Meeting details	🕏 <u>Agenda</u>	Not available	Dinutes	♥ <u>Video</u>	Not availab
*Rules & Legislation Committee	6/18/2015	31	10:45 AM	City Council Chamber, 3rd Floor	Meeting details	🔁 Agenda	Not available	🔁 Minutes		Not availab
Special Concurrent Meeting of the Oakland Redevelopment Successor Agency / City Council / Geologic Hazard Abatement District Board	6/17/2015	1	5:00 PM	City Council Chamber, 3rd Floor	Meeting details	🕏 <u>Agenda</u>	Not available	Dinutes	Video Video	Not availab
*Life Enrichment Committee	6/9/2015	31	4:00 PM	Sgt. Mark Dunakin Room - 1st Floor	Meeting details	🗊 Agenda	Not available	🖘 Minutes		Not availab
*Public Works Committee	6/9/2015	53	11:00 AM	Sgt. Mark Dunakin Room - 1st Floor *** THIS IS A SPECIAL MEETING OF THE PUBLIC WORKS COMMITTEE *** *** PLEASE NOTE CHANGE IN TIME ***	Meeting details	🕏 Agenda	Not available	Dinutes	Video	Not availab
*Finance & Management Committee	6/9/2015	31	9:30 AM	Sgt. Mark Dunakin Room - 1st Floor **THIS IS A CONCURRENT MEETING OF THE OAKLAND	Meeting details	D Agenda	Not available	D Minutes		Not availab

Screenshot of Oakland Legistar.

••• 🗈	Agenda (5). Page 3 of 24	pdf ③ Q Q ① L → C ②  Q · Search
	сомміт	TEE MEMBERSHIP:
		*Council President Nikki Fortunato Bas, District 2; Councilmember Carroll Fife, District 3; Councilmember Dan Kalb, District 1; Chairperson: President Pro Tempore Sheng Thao, District 4
	Roll Call	/ Call To Order
L .	COMMEN	C COMMENT: NTS ON ALL ITEMS WILL BE TAKEN AT THIS TIME. COMMENTS FOR ITEMS THE AGENDA WILL BE TAKEN DURING OPEN FORUM
	<b>2</b> Aj	pproval Of The Draft Minutes From The Committee Meeting Of April 14, 2022
	_	<u>2-0282</u>
	<u>At</u>	tachments: View Report
	3 New S	Scheduling Items
	Fr Ri O Th M Ag	ubject: Fulfillment Of Outstanding Rental Assistance Applications rom: Councilmember Fife And Vice Mayor Kaplan ecommendation: Adopt A Resolution Calling Upon The State Of California To Fulfill All utstanding Rental Applications That Meet The Eligibility Requirements, And Clarifying hat The City Of Oakland Did Not, And Does Not, Opt Out Of The State's Announced larch 31st, 2022 Safe Harbor Program For Applicants; On The May 3, 2022 City Council genda On Consent <u>2-0349</u>
	Fr Ri 12 9,	ubject: Salary Ordinance To Increase The Salary For City Attorney rom: Human Resources Management Department ecommendation: Adopt An Ordinance Amending The Salary Schedule Of Ordinance No. 2187 C.M.S ("Salary Ordinance") To Increase The Salary For City Attorney; On The May 2022 Finance And Management Committee Agenda 2-0310

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*:

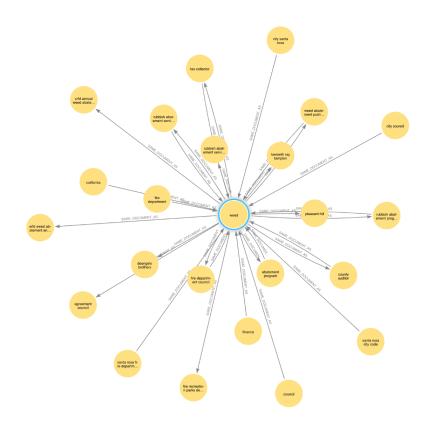
### ISSUE(S)

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?

### BACKGROUND

- 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.
   The weed and rubbish abatement services contract provides abatement of weeds and other ground cover fire or health hazards as determined by the Fire AGENCY
- weeds and other ground cover fire or health hazards as determined by the Fire AGENCY 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.

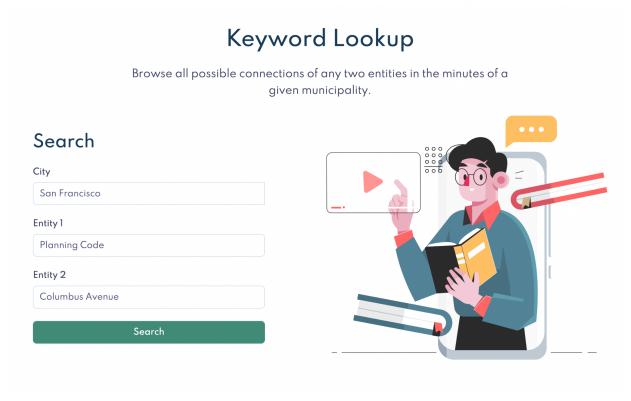
These names of individuals, cities, government agencies, documents, and dates that are connected to one another. We know that Kenneth Ray Tamplen lives in Pleasant Hill; they were awarded the contract F000096; they are in the weed and rubbish abatement business. That means we can extract these entities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



To "extract knowledge" from documents in such a way, we rely on two very interesting tools. The first one is <u>Named-entity recognition (NER)</u>, a Natural Language Processing (NLP) algorithm that identifies people, organizations and places (and much more) from documents. It works in a domain-agnostic way: it will extract the named entities from any input text.

This in and of itself is quite powerful, since most interesting stories revolve around specific people and organizations. What's much more powerful though is being able to automatically extract *relationships* among entities in such a way as to be able to search for them and auto-discover common themes not visible to the human eye at first sight. For this, we employ a special database system called <u>Knowledge Graphs</u>. A knowledge graph is built for the purpose of storing relationships between entities and makes it much easier to discover, as well as search for similarities in relationships. For example, if a document mentions some celebrity pushing their local government to enact zoning laws, this will be stored in a knowledge graph as "(CELEBRITY - PUSHES - ZONING LAW)", and so if you parse thousands of documents and a similar instance occurs somewhere else, the parallels will be easy to spot using the graph.

We've developed two functionalities that bring this vision closer to reality: a double entity search, and document search.



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

```
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 <u>url</u>
     • Leg Verl | Nov 06, 2015 url
      ∘ Leg Final I Nov 06, 2015 <u>url</u>
  • 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 <u>url</u>
     • Leg-Verl | Nov 06, 2015 url
      • Leg Final I Nov 06, 2015 url
  • Petitions and Communications | Aug 02, 2016
    Attachments:
      • Board Pkt 080216 | Aug 16, 2017 <u>url</u>
  • 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 Verl | Nov 06, 2015 <u>url</u>
      • Board_Pkt_030414 | Nov 06, 2015 url
      ∘ Leg_Final | Nov 06, 2015 <u>url</u>
back to search
```

Approval of a 90-Day Extension for Planning

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**

# Input Text

Entity: Mary Traverso Open Space (frequency: 1)

 Name Change of Creekside Open Space to Mary Traverso Open Space | Aug 31, 2021 Attachments:

Staff Dependents.

- Staff Report | Aug 25, 2021 <u>url</u>
- Attachment 1- Creekside Open Space
   Proposed Renaming Report | Jul 28, 2021
   url
- Attachment 2- City Council Policy 000-25 | Jul 28, 2021 <u>url</u>
- Resolution | Sep 01, 2021 <u>url</u>
- Presentation | Jul 28, 2021 url
- Attachment 3 Location Map | Aug 19, 2021 <u>url</u>
- Late Correspondence (Uploaded 8-31-2021) | Aug 31, 2021 url
- Late Correspondence (Uploaded 9-2-2021) | Sep 02, 2021 <u>url</u>
- (No title available) | Dec 09, 2020

#### Attachments:

- Creekside Proposed Renaming PPT | Dec 03, 2020 <u>url</u>
- (No title available) | May 26, 2021

#### Attachments:

Creekside Renaming PPT | May 20, 2021

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** 

A screenshot from the results page of our document search.

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:

	JOHN APPLESEED (person)	
$\frac{1}{2}$	Voted for	
	AMENDMENT TO OAKLAND MUNICIPAL	
	✓ Voted agains	t
	ALICE ANDERSON (person)	$\mathcal{I}$

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.