NAVARI FAMILY
CENTER for
DIGITAL
SCHOLARSHIP

Distant Reading for Quick Insights

Fighting COVID-19 by Mining Insights from Heterogeneous Datasets

Natalie Meyers | @nkmeyers
Eric Morgan | @ericleasemorgan

UNIVERSITY OF NOTRE DAME | HESBURGH LIBRARIES

Distant Reading for quick insights

Text mining for this project was performed using distantreader.org.

Morgan, Eric Lease. (2020, April 10). Distant Reader (Version Alpha). http://doi.org/10.5281/zenodo.3747777

More Info: https://osf.io/dtzc8/

NAVARI FAMILY CENTER for DIGITAL SCHOLARSHIP

Text mining to explor simple survey responses

Question 17

If available, website providing information for library users on library policies and procedures in light of COVID19

US Academic Library Response to COVID19 Survey conducted by Lisa Janicke Hinchliffe and Christine Wolff - Eisenberg

Text mining to explor simple survey responses

we are closed	11
library is closed	44
closed	435
available	862

NAVARI FAMILY
CENTER for DIGITAL SCHOLARSHIP

We Are CLOSED

Why don't most Library websites come right out and "Say it" ?

What's the real/big message?

Available?

Distant Reading Question #17

Open Data doesn't always mean FAIR data and that's okay . . . . but

- . . . Manual Download
- Preprocessing
- 86% Webcrawling success rate
- Capturing Org Names (retrieve from survey IPEDS name content or site?)
- NER tuning

Meyers, N. K., Molik, D., & Morgan, E. L. (2020, April 13). Using the Distant Reader to Text Mine the US Academic Library Response to COVID19. Retrieved from osf.io/cruzd

Text mining Q#17: Issues & FAIRifying

FAIRification of our gateway's output is a priority but how to decide what first ?

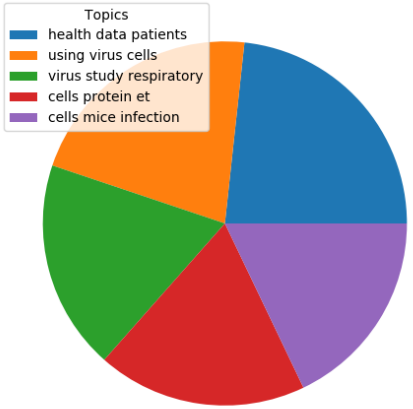
We have to make [More] FAIR gateway ouputs 

Need more FAIR Metadata

Credit for carrel Creators
Leverage ORCIDs for Carrel Creators
Present resolvable PIDs for all Carrels
Offer DOIs for public carrels

Resolvable Credit information for carrel content
License Checking for Carrel content
Make Carrel content license values visible

API for interacting with carrel content
RDF representation of Carrel content



CORD19

Open Research Dataset Challenge (CORD-19). https://[illegible]
institute-for-ai/CORD-19-research-challenge

Text mining COVID19 resources

Distant Reader & CORD19 Issues

- Scale Vs Portability
- Efficiently consuming new versions into the corpus
- Modularize to include SciSpaCy and other text processors
- Strengths and Use cases for Distant Reading
- Tuning NER & Treatment of Proper Nouns
- Doesn’t resolve Ngrams etc back to “text string in context “ [Yet]

Morgan EL, Molik D, and Meyers NK. “Distant Reader CORD19” (March 17, 2020): doi:10.17605/OSF.IO/AH37Q, Available at osf.io/ah37q

Distant Reader & CORD19 Issues

Performance Vs Portability

The Distant Reader gateway even though running on performant HPC doesn't scale well to the CORD19 corpus

The non parallel nature of the reduction step exposed a tension between two competing priorities of the Distant Reader

Service Unit estimation (1 SU = cores x days x 24 hours)

nodes	cores	days	sus
http server	2	90	4320
index server	2	90	4320
database server	2	90	4320
distant reader	48	12	13824
database reduce	6	12	1728
indexer	6	12	1728
http server (sustained)	2	270	12960
index server (sustained)	2	270	12960
database server (sustained)	2	270	12960
		Total SU's requested	69120

Morgan EL, Molik D, and Meyers NK. "Distant Reader CORD19" (2020): doi:10.17605/OSF.IO/AH37Q, Available at osf.io/ah37q

Text mining CORD19: Issues & FAIR

- The dataset is open but not FAIRified
- Processing Metadata meaningfully alongside bagofwords in heterogeneous text documents demands attention to workflow and sorting/ranking features
- How to output FAIR analysis of the CORD19 dataset ?
- Similar problems as with Library survey but amplified at Scale

Other Approaches

Rakesh Thoppaen's Kaggle collaboration via NLP notebook

Q ***What if specifically, we want to know what the literature reports about the range of incubation periods for the disease in humans ?***

A

1. **Incubation Period :**

Key HighLight in Search

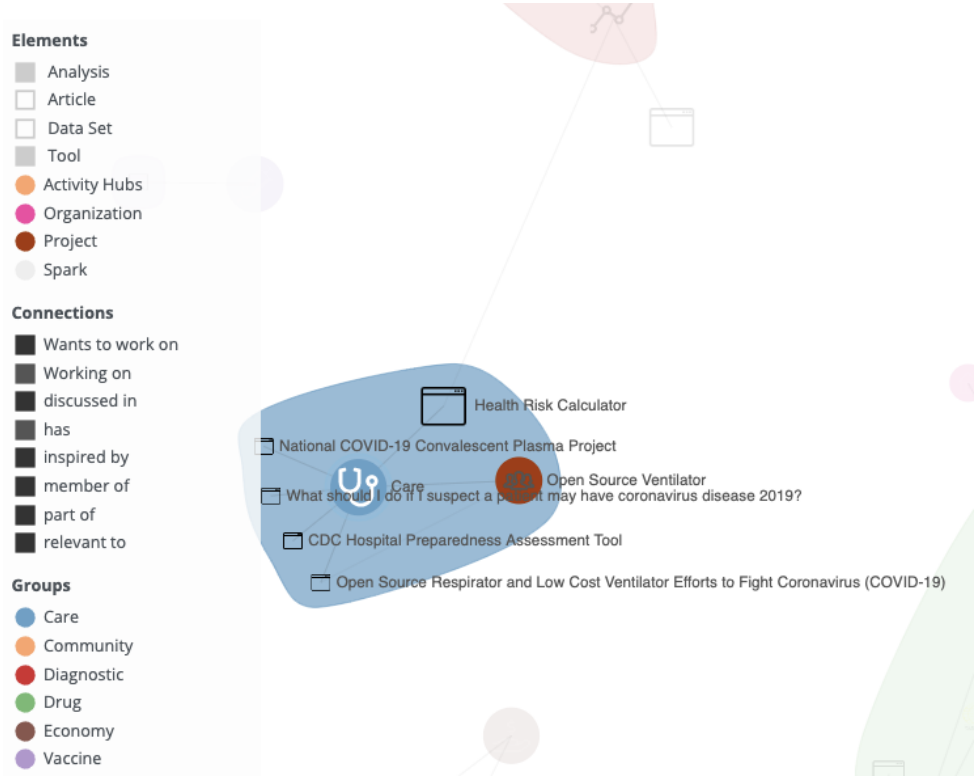
- Mean Incubation Period is 5days across various papers
- Range of Incubation Period is upto 14 days
- Median Incubation Period is reported to be 3 days
- There are evidences where incubation period is mentioned as 20 days

1. Mean Incubation Period is 5days across various papers

https://www.kaggle.com/fabulinus/covid-19-challenge-word2vec-keyword-search/notebook

Other Approaches

Exaptive's COVID19 Cognitive City

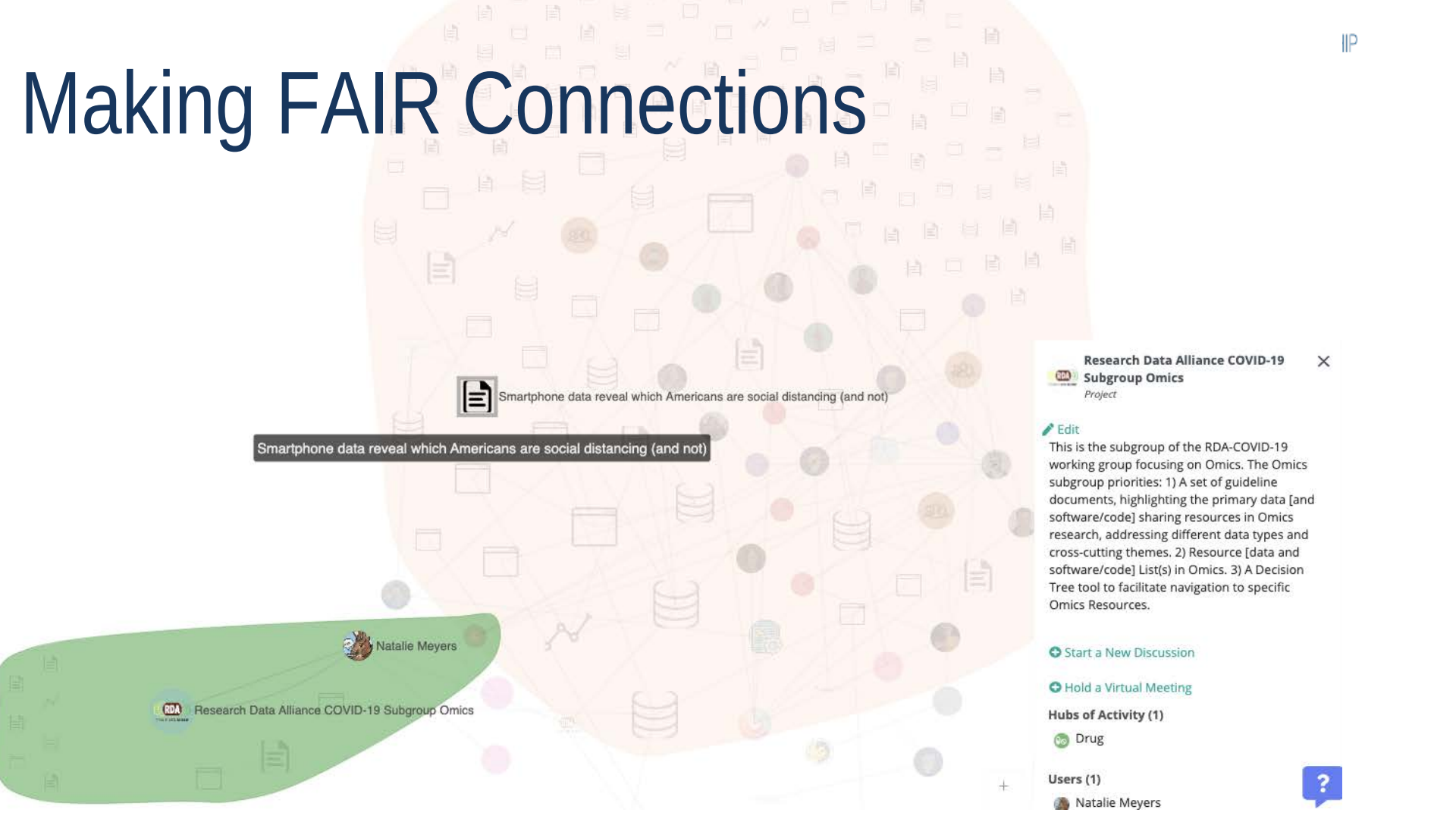


Graph DB for facilitated innovation bcs collaboration is network based

covid-19.cognitive.city

NAVARI FAMILY
CENTER for DIGITAL SCHOLARSHIP

Making FAIR Connections



Thank you!

These Slides

https://osf.io/dtzc8/

Eric Morgan

000-0002-9952-7800

Natalie Meyers

0000-0001-6441-6716

David Molik

0000-0003-3192-6538

cds.library.nd.edu

NAVARI FAMILY
CENTER for DIGITAL SCHOLARSHIP