Harnessing the Data Revolution (Track A): **Open Knowledge Network**

https://www.nsf.gov/od/oia/convergence-accelerator
NSF Open Knowledge Network: Project KONQUER

KONQUER: Knowledge Open Network and Queries for Research

Principal Investigator

Bio & Health Science

Geo-science

Data repository

Domain specific OKN

Climate-science

(2.34 mil dataset records from 75 sources)

(1.67 mil dataset records from 40+ sources)

PANGEO

dataMED

NSF
NSF Open Knowledge Network: COVID-19 Efforts

- Local regulations: social distancing, quarantine, ...
- Infrastructure resilience: transportation, supply chains
- Population characteristics: from perceptions to genomics
- Health data: diseases, treatments, prevention, social conditions
- Biomedicine: genes, proteins, body, diseases, drugs
- Environmental factors
- Pathogens: COVID viral strains, pathogen-host interactions, SARS, MERS

Open Knowledge Network forecasts, models, dashboards
1 attaaaggtt tataacctgcc caggtaacac accaaccacac tttcgatctc ttgtagatct
61 gttctctaaa cgaactttaa aatctgtgtg gctgtcactc ggctgcatgc ttagtgcact
121 caacgcatat aattaataac taatctctgt cgttgacagg acacagtaac ctgctctatc
181 ttctgacagg tcgctctacc ttcgctgcag gcagcgcagc tcctgacgac acctaggttt
241 cgcggcggtg tgcacagaaag tgcctgtgaa gtcgctgtgc cctggttccca cggagaaacc
301 acaaagctcc caacagctccc catttaacct cctctggttt gctgttgcgg
361 agctccttcg gaggagcttc actcagagcc aaacacatg ttagaagct 3ccttgcgg
421 ctagtctgaa gttgaaaaag gcgttttgcc tcaacttgaa cagccctatg tgttcatcaag
481 accctcctgt gcgctctcag tccctctctg tcatatttac gtctgctact ttaggagaact
541 cgaagctttt cctctcaggtg gtagcgtgcttg gctcttgccct ccatcgtcttgg
601 cgaataacca gggctctaco gcaaggtgtt ctctctcgaag accgcattg tggagctttg
661 tggcagatt gtgcgctagct attctatta gctggcctgg cggcagtctg tggcactgta
721 cctctgagtt gattctaaa aacaaggtgg caacaagctct tagctgtcag ttcctcttcg
781 attctctctgt gcgctctcag gcggcctata cactgctatg ctgataaaaa acctctctcc
841 cctgataggtg tagactgttgc gttgcttttga aagaaaaa gcacggtgctg gttagatcc
Goal: Link Heterogeneous Datasets to about COVID-19

Host-Pathogen Interactions
- Protein-protein
- Immune response

Animal and Human Strains
- Genomes
- Genes
- Proteins

Previous Outbreaks
- SARS
- MERS

Animal to Human Transmission
- Animal reservoirs
- Animal diversity/hot spots
- Food chain

Weather
- Seasonal effects
- Temperature
- Precipitation

Transmission & Migration
- Travel
- Transportation
- Venues
- Education
- Work
- Sports

Cases
- Outcomes
- Demographics

Patients
- Demographics
- Symptoms

COVID-19

Host

Pathogen
Prototype COVID-19 Graph

COVID-19 Outbreak:
Cities, States/Provinces, Countries

Pathogen:
Strains

Genome:
Genes, Proteins
Community spread of A2a strain: A. San Diego County, B. San Francisco County, C. Sao Paulo; **Imported cases**: D. A2a strain imported from Italy into Sao Paulo; All cases carry the A2a strain, a strain that is prevalent in Italy and can be traced back to China. This subgraph integrates data from Nextstrain.org and GISAID.
COVID-19-Community Project - Crowdsourcing

Automated Reproducible Community Contributions
Interoperability for Knowledge Graphs

**CURIEs**: Compact URIs as unique and interoperable identifiers

Example: ncbiprotein:YP_009725311 *(prefix: accession number)*

**Identifiers.org**: resolves CURIEs to persistent URLs:

http://identifiers.org/resolve?query=ncbiprotein:YP_009725311

**Biolink**: Biolink-Model: https://biolink.github.io/biolink-model,

Knowledge Graph Exchange: https://github.com/NCATS-Tangerine/kgx, Prefix Commons: https://github.com/prefixcommons
Metagraph for SARS-CoV-2 Proteins

<table>
<thead>
<tr>
<th>Name</th>
<th>Prefix</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI Protein</td>
<td>ncbiprotein</td>
<td>The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB.</td>
</tr>
<tr>
<td>UniProt Knowledgebase</td>
<td>uniprot</td>
<td>The UniProt Knowledgebase (UniProtKB) is a comprehensive resource for protein sequence and functional information with extensive cross-references to more than 120 external databases. Besides amino acid sequence and a description, it also provides taxonomic data and citation information.</td>
</tr>
</tbody>
</table>
Location index for the knowledge graph

• Location is often the only way to connect disparate data
• But it’s tricky:
  • Locations can be referenced by coordinates (in various CRS); using different types of placenames; verbal descriptions
  • Quality of location assignments varies greatly
  • Mixture of location types, with different relationships between them (census geographies, zip codes, HUCs, weather stations, etc.)
  • Standardizing location identifiers: we use Geonames and various geocoders
  • Disambiguation is non-trivial, especially outside the geosciences (we applied DataDiscoveryStudio’s Spatial Enhancer to DataMed, with mixed results; additional ML training is needed)
• Spatial processing is typically done outside a knowledge graph, esp. for large data collections

We are building a pipeline for bringing location data into a knowledge graph
Making the data FAIR

While there is no shortage of coronavirus content on the internet, the American people need access to the most up-to-date public health guidance and most relevant information on Coronavirus Disease 2019 (COVID-19) testing facilities in their communities. The White House and key Federal agencies are working alongside Schema.org to help ensure these critical resources surface across online search engine results.

Schema.org is a collaborative community activity founded by tech companies such as Google, Microsoft, and Yahoo to create standard tags that can be added to any website’s code, making webpages containing particular information easier to find in search results. In response to the coronavirus pandemic, Schema.org has created a variety of new tags including information about COVID-19.

Schema.org may also provide a common vocabulary shared by multiple KGs

The OSTP directive mentions NSF OKN projects

https://www.whitehouse.gov/articles/connecting-americans-coronavirus-information-online/
Earlier work in Data Discovery Studio

900K DDStudio dataset descriptions in schema.org are indexed by Google
• Schema.org v. 7 includes COVID-related tags: SpecialAnnouncement; CovidTestingFacility; diseaseSpreadStatistics; CDC markup…

```json
{
  "@context": "http://schema.org",
  "@type": "SpecialAnnouncement",
  "name": "2019 Novel Coronavirus (COVID-19) in San Diego County",
  "text": "2019 Novel Coronavirus (COVID-19) in San Diego County",
  "encodingFormat": ["text/html", "image/png", "text/javascript", "text/css"],
  "inLanguage": "en",
  "datePosted": "2020-03-26T16:00:00+07:00",
  "keywords": "COVID-19",
  "publisher": {
    "@type": "GovernmentOrganization",
    "name": "San Diego County"
  },
  "category": "https://www.wikidata.org/wiki/Q81068910",
  "spatialCoverage": [
    {
      "@type": "AdministrativeArea",
      "name": "San Diego County, CA"
    },
    {
      "geo": {
        "type": "GeoShape",
      }
    }
  ]
}
```
ABOUT THIS COURSE

DSC198 is a workshop-style course focused on several types of data science applications, mostly in spatial data science but extending to other areas, from information visualization to recommender systems. It will be largely driven by problems that are made urgent by the pandemic. Because of this, a precise topic sequence is hard to lay out at the start. The situation changes rapidly, and we need to be flexible. We will identify several seed projects that teams of students will attack, working alongside the instructors. These projects may change as we go.

Prerequisites: DSC-170 (Spatial Data Science) or instructor's consent. We mostly expect that you can work independently to carry out the projects from planning through implementation.
What is a **Knowledge Graph**?

**Findable:** Registry of KGs?

**Accessible:** Standard APIs: GraphQL, SPARQL, Cypher?

**Interoperable:** Persistent identifiers, CURIEs, federated queries?

**Reusable:** Common vocabulary such as schema.org?