Knowledge mining from Research Graph and Open Scholarly Works

OASPA 2021 Conference

A/Prof Amir Aryani
orcid.org/0000-0002-4259-9774
Part 1: Background
Problem: tracking the footprint of research data

- Similar datasets in other repositories?
- Related publications?
- Potential global collaborators?
- Related projects, grants, activities?
Initiated from the Data Description Registry Interoperability Working Group

Goal: enabling cross-platform discovery between research data infrastructures

https://www.rd-alliance.org/groups/data-description-registry-interoperability.html
IDMM Immunome Database for Marsupials and Monotremes

The University of Sydney

Associate Professor Katherine Below (Managed by) & Emily S W Wong (Managed by)

IDMM Immunome Database for Marsupials and Monotremes

Full description

IMDD is a database of marsupial and monotreme immune gene sequences. The IMDD supports the work undertaken by researchers working on the tammar wallaby genome and transcriptome project. The main goal of IMDD is to fill gaps in the curation of immune genes by providing immune gene sequence annotations for published genome-wide expression studies and sequence information for divergent immune genes that are not hosted on any other public database.

Database content

In February 2012, total of 2,935 genes, 602 expressed (538 tammar wallaby, 24 opossum, 16 platypus, 11 echidna, 6 red-necked wallaby, 4 brushtail possum and 3 bandicoot) and 2,333 predicted (1,639 opossum, 694 platypus), were stored in the database. In February 2012, the database included 1,985 published genome sequences. This includes protein and DNA sequences. The database may be added to over time. Data has been integrated from various published resources, which include expressed and predicted genes from opossum (1,843), tammar (37), brushtail possum (4) echidna (11), bandicoot (3), red-necked wallaby (6) and platypus (261). Manually annotated gene families include: major histocompatibility complex.


One of the 105 articles …

Sequence Analysis

A schematic of the bioinformatics workflow is presented in Figure 2. A Tachyglossus aculeatus venom gland transcriptome was sequenced on an Illumina GAIIx instrument using proprietary methods to describe methods for sample preparation and sequencing [7]. The venom gland was provided by Frank Grutzner under University of Adelaide Animal Ethics Committee protocol number S-032-2008. The sequence reads have been deposited under accession number SRP027593 in the SRA database at NCBI. Quality filtered reads were assembled with Velvet-Oases pipeline (kmer length = 31bp and –ins_length = 260) [12], [13]. To improve transcriptome assembly, the Scaffolding using Translation Mapping (STM) strategy [14] was used to scaffold Oases contigs using Ensembl predicted proteins, an experimentally defined dataset of platypus venom proteins [5], [7], [15], and all human Refseq proteins as reference. Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.4g00v.

[16] searches were performed against a database containing all Tox-Prot proteins [17] and additional platypus venom toxins that have not been yet included in the Tox-Prot database and human Refseq sequences using scaffolds derived from STM to identify possible homologies between echidna and other venom toxin proteins.
Background
We found another dataset from the same author...
We found another dataset from the same author… ✅
Toward automating the process
Research Data Switchboard

Data Providers
- ands.org.au
- inspirehep.net
- datadryad.org
- ORCID
- da-ra.de
- dlibservice.research-infrastructures.eu

Switchboard components
- CrossRef Integration
- DataCite Integration
- Google Search API
- Node Linking Process
- Export Harmonised Research Graph

Graph Database
- Compute Optimised Machine
  - 36 Cores, 132 (ECU), 60 GB RAM

Data Consumers
Graph nodes and relationships are accessible via
- Research Graph JSON-LD
- Neo4j Graph Database
Leveraging open data, open publications and open metadata
Connection Types

Type A: $\alpha$: PID $\rightarrow$ [metadata] $\rightarrow$ PID

Type B: $\beta$: Text (e.g. PDF) $\rightarrow$ [DM & NLP (%)] $\rightarrow$ PID

Type T: $\Theta$: Text (e.g. PDF) $\rightarrow$ [DM & NLP (%)] $\rightarrow$ Text
Type A

Explicit link between PIDs in their metadata
Connecting (Text → PIDs (DOI, ORCID, PubMed))

- Text mining,
- Fuzzy search, and
- Graph modeling.

dataset are an accurate representation of reality as some of the interventions capture a governmental request that might not be followed by the entire population. Thus, it might be useful to integrate the WTNTRAC dataset with other publicly available data sources that can provide information regarding the level of compliance with an intervention, such as mobility information as exemplified by the NPI Index discussed above. Lastly, one other interesting use case is to estimate the economic impact of NPIs, for example, by relating unemployment rates and jurisdictional debt with NPIs. Estimating the effect of NPIs on non-COVID-19 health problems, such as late cancer detection due to missed screening tests, will also be useful.

**Code availability**

The source code for the WTNTRAC automated NPI curation system, including the data processing pipeline, WTNTRAC Curator tool and NPI data browser is available in a public GitHub repository at https://github.com/IBM/wntrac/tree/master/code alongside the up-to-date version of the dataset https://github.com/IBM/wntrac/tree/master/data. Please refer to the README file in repository for further instructions on using the code.

Received: 4 November 2020; Accepted: 19 February 2021;
Published online: 23 March 2021

**References**

Type B (Text to WikiData)

- Melbourne University
- The University of Melbourne
- UMelb
- Melbourne Uni
- UniMelb
- جامعة ملبورن
- მელბურნის უნივერსიტეტი
- Университет у Мелбурну

Q319078
Nodes: 333.26M
Relations: 743.14M
Part 2: Augmented Intelligence
**Mental Health and Social Resilience in Glen Eira**

**Data Repositories:**
- data.gov.au
- ada.edu.au
- wikitdata.org
- auric.org.au
- icpsr.umich.edu
- data-archive.ac.uk
- who.int/gho/database
- google.com/publicdata

**Academic and Policy Documents**
- apo.org.au
- ssrn.com
- arxiv.org
- openaire.eu
- trove.nla.gov.au
- ncbi.nlm.nih.gov/pubmed
- springernature.com/scigraph

**Private Collection**
- articles
- technical reports
- datasets

---

**Artificial Intelligence (AI) Engine**
- Data Enhancement
- Geospatial Mapping
- Clustering
- Statistical Modeling
- Machine Learning
- Predictive Analytics
- PDF Content Extraction
- Document Linking

**Data Publishing and Visualisation**
- Linked Data Package
  - Data (2010-20):
    - Domestic Violence Hotlines
    - Journal Article (2019):
      - Insight into family violence.
    - Report (2017):
      - Poverty and Social Exclusion
  - Social Services
  - Loneliness
  - Mental Health
  - Poverty
  - Family Violence
  - Crime

**Data CO-OP**
- Communities
- Policy Makers
- Researchers

**Tableau and Power BI Data Visualisation**

**Metadata Dissemination**
- APO
- Trove
- ADA
- Wikidata
Areas of research

- Public Health
- Social Wellbeing
- Sustainable Community
- Information Resilience
- Public Policy
Studying Coronaviruses Research Collaboration Network

Research Graph Project (2020-2)
Coronaviruses Collaboration Network

Dec 2009

1,685 Research Organisations
based on ORCID, and DOI links
Coronaviruses Collaboration Network

Dec 2009

1,685 Research Organisations
based on ORCID, and DOI links

Great Britain

United States
Coronaviruses Collaboration Network

Dec 2019

3,543 Research Organisations
based on ORCID, and DOI links

Spain
Coronaviruses Collaboration Network

Feb 2020

10,162 Research Organisations
based on ORCID, and DOI links
Coronaviruses Collaboration Network

April 2020

17,483 Research Organisations
based on ORCID, and DOI links

China
Coronaviruses Collaboration Network

June 2020

29,473 Research Organisations
based on ORCID, and DOI links

Germany
Coronaviruses Collaboration Network

Dec 2020

49,773 Research Organisations
based on ORCID, and DOI links

<table>
<thead>
<tr>
<th>Country</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>11.42%</td>
</tr>
<tr>
<td>BR</td>
<td>6.55%</td>
</tr>
<tr>
<td>IN</td>
<td>6.54%</td>
</tr>
<tr>
<td>GB</td>
<td>5.36%</td>
</tr>
<tr>
<td>IT</td>
<td>4.23%</td>
</tr>
<tr>
<td>ES</td>
<td>4.23%</td>
</tr>
<tr>
<td>CN</td>
<td>3.58%</td>
</tr>
<tr>
<td>FR</td>
<td>2.88%</td>
</tr>
<tr>
<td>DE</td>
<td>2.5%</td>
</tr>
<tr>
<td>AU</td>
<td>1.75%</td>
</tr>
<tr>
<td>PT</td>
<td>1.72%</td>
</tr>
<tr>
<td>RU</td>
<td>1.64%</td>
</tr>
<tr>
<td>TR</td>
<td>1.51%</td>
</tr>
</tbody>
</table>
About 290,000 research articles related COVID-19 were extracted from Research Graph as of July 2021, filtered by ‘COVID19’, ‘COVID’, ‘Coronavirus’, or ‘SARS-CoV-2’ in titles, abstracts or subjects.
Clustering Papers using Topic Modelling

- 290,000 papers
- AI: Topic Modeling
- Collection 0
- Collection 1
- ...
Timelapse

infection, syndrome, control: 46
antibody, response, immunology: 31
case study, transition, model: 22
vaccination: 11
treatment: 9
disease, severity, management: 6
time, education, research: 3
review, children, literature: 3
health care: 2
patient, cancer, outcome: 1
socio-economic impact: 1
pandemic, study, management: 0

2002-07-01
Benefits Realization Categories and Proximal Measurements

Identification

Records identified through database searching (n = 14698)

Additional records identified through other sources (n = 4675)

Records after duplicates removed (n = 7100)

Screening

Records screened (n = 1945)

Records excluded (n = 1305)

Eligibility

Full-text articles assessed for eligibility (n = 640)

Full-text articles excluded, with reasons (n = 306)

Included

Studies included in qualitative/quantitative synthesis (n = 332)

Atlas of Health Services

Other AI projects
Access, Scalability & Sustainability

- Access to metadata beyond current open access services
- Challenge of scalability for large data linkage
- Cost and effort for metadata harmonisation
How to get involved?

To access projects’ data and collaboration please contact

Amir Aryani: aaryani@swin.edu.au