CFDE Knowledge Graph Working Group - Agenda & Notes

File location: *Knowledge Graph Working Group Folder*

Who: *Please contact the CFDE Help Desk (support@cfde.atlassian.net) if you do not have a calendar invite, or Jonathan Silverstein (j.c.s@pitt.edu) and Deanne Taylor (Taylordm@chop.edu) with any additional questions.*

What: The goal of the Knowledge Graph Working Group (KGWG) is to bring together Common Fund DCCs interested in establishing standards and model(s) for utilizing knowledge graphs for data integration.

When: *Third Tuesday of every month from 3 to 4 PM EST*

Where:
https://us06web.zoom.us/j/81043977847?pwd=dEdMWnpiZGVUdmJ2QnhaMWVSYkQ1dz09

Important Docs/Links:

Charter
Roster

Contacts for use cases:

- <Use case name>
  - <contact>

TEMPLATE <DATE>

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	https://us06web.zoom.us/j/81043977847?pwd=dEdMWnpiZGVUdmJ2QnhaMWVSYkQ1dz09
Participants: SIGN IN: Name & Affiliation	Not here… scroll down ; ; ; ; ; ; ;d ;		

Agenda Item	Action Items	Owner

Notes:

Upcoming Agenda Topics

Please add upcoming agenda topic ideas here.

Date	Agenda Topic	Who
Upcoming meetings...	● Visualizations and Knowledge Graphs	

Meeting Agendas

---Newest meeting on top—

2024-11-19

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom Link
Participants: SIGN IN: Name & Affiliation	; ;		

Agenda Item	Action Items	Owner

Notes:

2024-10-22

Objective	Knowledge Graph Working Group Break-out	Time	3:30pm EST
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Bethesda, MD
Participants: SIGN IN: Name & Affiliation	; ;		

Agenda Item	Action Items	Owner

Notes:

- Session #1
  - Review KGWG Charter
    - KG WG Charter_DRAFT
    - Aleks: Should we consider the impact on discovery as a measure of success? Should be in the first paragraph.
      - AI readiness should also be a key focus of the next phase of the KGWG.
    - Jonathan: Focus on knowledge interoperability rather than a specific tool
      - Making sure that KGs within the CFDE have a particular export format so that they can then be imported in a standard way to other graphs – is it the DDKG format? Probably not, but could be used as a starting point.
    - Deanne:
      - Could produce a set of standard CFDE triple files that anyone can use to start a new KG.
    - Jonathan:
      - Source of every assertion has to be tracked.
      - All characteristics on the nodes and edges have to be well-defined enough that they are not uncertain.
      - Can specify schema for import and export files.
      - Can’t specify what the schemas are for how data is related within individual knowledge graphs.

        - Revisit the charter in the next KGWG meeting, send a note to the channel for everyone to review the charter prior to the meeting for discussion.
    - Discussion of UBKG Data Model https://ubkg.docs.xconsortia.org/datamodel/
    - Standards: Integration of / ingestion many knowledge and graphs and setting standards. Survey work as number of new KGs and use cases grow.
    - Revisit variant schema
        - DDKG schema expansion for variants - node types and property choices could include fields from VCF 4.3 as well as suggested fields from GA4GH variation WG:
            - https://www.ga4gh.org/product/variation-representation/
            - https://vrs.ga4gh.org/en/stable/schema.html#overview
            - GA4GH genotype work published in https://www.worldscientific.com/doi/abs/10.1142/9789811270611_0035
        - Adam: What's the version of a knowledge graph product that DCCs can just generate as a part of their normal work that could be in alignment with the rest of CFDE? Get the barrier low enough that they come in relatively clean. How do we pipeline this process rather than making it a one-off event to generate "knowledeg graph ready data"?
        - Deanne shares proposed variant schema for DDKG
            - Should variant type be a property on Variant or its own node type?
            - Pathogenicity scores: Would this be binned? Own node type or property?
            - Population frequency: Would this be binned? Would it be its own node type or a property?
            - If some of these become properties on the nodes rather than their own node types, then any variant node missing this data has a NULL rather than just lacking a connection to a separate node.
            -
- Session #2
    -

2024-08-20

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	[Zoom Link]()
Participants: SIGN IN: Name & Affiliation	; ; ; ; ; ; ; ;Deanne Taylor (KF) ; ; ; Stephanie Olaiya (LINCS, DRC); ; ; ; ; ; Avi Ma'ayan (LINCS, DRC); ; ; ; Chris Nemarich (KF);Heesu Kim (LINCS, DRC); Sherry Jenkins (LINCS, DRC); Daniel Clarke (LINCS, DRC); Anna Byrd (LINCS, DRC) ;Jake Chen (ICC) ; ; ;Raja Mazumder (GlyGen); ; ; ; Mano Maurya (MW) Swathi Thaker (ICC-Admin); ;		

Agenda Item	Action Items	Owner
Introduction		

Notes:

2024-07-16

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom Link
Participants: SIGN IN: Name & Affiliation	; Mano Maurya (MW); ;Sherry Jenkins (LINCS, DRC) ;Avi Ma'ayan (LINCS, DRC) ; Jonathan Silverstein (SenNet/HuBMAP); ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; ;John Erol Evangelista (LINCS, DRC) ; ; ; ; ; ;Srini Ramachandran (MW), Swathi Thaker (ICC-Admin) ; Matt Roth (ERCC); David Chen (ERCC); ; Jimmy Zhen (MoTrPAC); Wei Wang (ICC-SC@UCLA) ; ; ; ; Peipei Ping (ICC-SC@UCLA) ; Henning Hermjakob (EMBL-EBI/ICC-SC); ; ; ; ; ; ;Nia Lingam (LINCS, DRC)		

Agenda Item	Action Items	Owner
Introduction		
Semantic typing in KGs		
Embeddings out of the graph		
Fall Meeting - Preliminary Agenda		

Notes:

Introduction

-

Semantic typing in KGs

- Deanne: semantic typing are categorizations that UMLS uses. Definition. Use several for Data Distillery. Put in by UMLS. Tend to be used on standards on ontologies possibly on some data. Have not yet "typed" data. Pros and Cons to this.
  - For our Data Distillery purposes, semantic types have not been added to data types.
  - Would require a link.
  - Per Deanne, Alan is putting in reference nodes to the datasets that include the info in the CFDE Data Distillery Data Dictionary. Putting those in as nodes with properties. References stand alone and put in items to support this: dataset type, analysis type. The data type can be categorized.
  - Searching reference nodes for the property. Have the list of SABs for query.

  - Deanne: The data is source data (will have calculation method), not necessary to connect all this data to a semantic type. It is technically a data type. Opening the floor for discussion:
    - Jonathan: sharing the URL in the chat. https://ontology.api.hubmapconsortium.org/semantics/semantic-types?skip=0&limit=10
    - Complete semantic type tree from UMLS. Fully connected tree. From one semantic type has a link to all of its parents. Framework to label everything. May add new ontologies that come in. **Not suggesting assigning a semantic type to all new things that come in**.
    - There are data types. Use the things that are there. We should associate it with each ontology. A list of data types associated with a SAB. The node that describes the source.
    - Deanne: We don't have to connect every point to that type. Go to the reference nodes and make sure they are labeled correctly.
    - Allowable entries for the data type. Jonathan notes that ingests of data are two things:
      - Nodes
      - Edges
    - List the types of concepts being ingested is useful.
    - Deanne: may want an analysis type as well as a data type. The type of analysis we did collect in methods should be there. Deanne states it is not searchable, not coded. Analysis type and data type according to a standard on each of the SABs is what we are suggesting, per Jonathan.

Open Discussion:

- Matthew Roth asking for Pros and Cons:
  - Doing it means, if someone wants to use the Data Distillery (example single cell data analysis) there isn't a way to search for a data type: single cell type for example. This would be the con and the pro is that you would.
  - Proposing: create a new canonical field for CFDE: Data Type and Analysis Type. Those slots would be an individual reference node.
  - Deanne: would be able to search the data type property with information on the analysis type. It would be on the reference nodes. Users of the data dictionary can go into the graph itself and should get the list of SABs. Would need to know to query the reference nodes.
  - Deanne: adding two new ontology slots to the reference nodes that define your dataset.

- John Erol in chat: How would this be different with the node label in neo4j? Or do we want to keep the label as 'Concept'
  - Jonathan: The new type of node is an info node, this is where these data types would be. A different label than the concept.

- Hermjakob Henning: resource is there for querying and do simulations, querying
    - Jonathan: did not do the experiment to prove that is resource sufficient. Certain types of queries will be faster. What is in the source is faster. This is a reasonable thing to go testing. We can make things with big memory and fast.
    - Hermjakob: collect a few use cases and phrase the queries, from a user point of view - how complicated is my query, how hard to phrase and get a feeling what does this look like from the user side.
    - Deanne: states it should be pretty simple. Should only be an extra line or two; “find me all these types of nodes and get me these properties from them” If we want to test amongst the alternatives would require labor and time.
    - Jonathan: asking for the specific use cases that we are solving is important.
    - Deanne: The use cases would be people who are exploring and trying to access data types in the Data Distillery.
    - Jonathan: which of the things beyond data type and analytic type should be classified at the source. More eyes could look at that.
    - **Action:** look at other things to classify. For example, platforms, per Deanne. Deanne notes Geo has examples, faceted searches on the left. For example, chemistry use, what organism is it (mouse, human). Jonathan notes assay types or instrument types, the list could get long. The main things are the ones we think people will come in looking for. Jonathan is requesting folks to look at the Data Dictionary and provide feedback.

Embeddings out of the graph

- Deanne: CFDE KG, large collection of nodes with edges. PCA, different dimensions, node embedding methods. Multi dimensional space, Open AI uses 3k dimensions to represent a data point. We could do it for KG, as a product or representation for people that want to do clustering. Can use it for categorization and other things that use multi dimensional matrices. Different embedding methods can be used. There are different choices on how embeddings are set up. For the CFDE, we can offer with the graph, different representations - a different way of looking at the graph. Run several embedding methods. and leave that as a KG product, this will give people the ability to use the data to do hypothesis testing. All these nodes have labels.
- Taha: do we want to make embeddings a part of what we offer people that come into the CFDE on any KG that we work on? Deanne is asking for feedback from the teams:
    - Jake Chen: notes he is a user of multiple embedding methods, each method will create a different result. Noting the need to publish.
    - Wei Wang: did research on different types of embeddings, notes what Jake says - there are many embedding methods and some are better than others. Define a few number of tasks that will benefit the community and train and recommend the

embedding models that optimize for the tasks. Very important to publish this, agrees with Jake. People will need a good understanding of how and when to use this.

  - Deanne: would love to hear use cases from other DCCs as well. This could be published in the Data Distillery paper we are working on as a section.
  - Jonathan: regarding Jake and Wei's comments, thinks that we have the graph and methods and use case, and the embeddings themselves. API as a table for fast lookup. You have KG and any set of embeddings - tables reference back to the concepts. List of external references, concept, and set of embeddings - embedding tables that have the code lookup, concept and the embeddings. People could distribute and make their own embeddings without affecting the graph at all. Embedding doesn't always apply to the entire graph, could be a sub graph. Could be in a subset of the graph.
  - Deanne: choosing an embedding that we would design.
  - **Henning in chat:** You could do a demonstrator embedding based on a subset of the data, but matching a (set of) realistic use case(s), and then decide on if/how to scale up to the full graph. Based on feedback.
  - Taha: agrees with Jonathan. Not adding the vectors to the nodes directly.

Fall Meeting - Preliminary Agenda

- Jake notes the need for an agenda.
- Deanne: Working group does not have an agenda yet for the meeting. Deanne asking if we should do this async to see if anyone has agenda items for the in-person meeting.
- Jonathan: where does the KG and Data Distillery extend - translator project. We could invite people in the other areas and ask their thoughts.
- Peipei Ping: has someone on their team that could share ideas, working on the translator. Professor in the bioinformatics division, TSRI: Suggested speaker at the upcoming fall meeting: Dr. ChunLei Wu, TSRI.
- Jake Chen: NCATS and the Common Fund.
- Christy Kano: can circulate the link to the proposal (not final): https://d1dth6e84htgma.cloudfront.net/NIH_Reform_Report_f6bbdca821.pdf
- Deanne: GA4GH, they do have ontologies and standards for medical records. Do we want to have conversations with GA4GH? Deanne can talk with contacts to see if there is any advantage in looking for an overlap.
- Deanne: NCBI as data curators or at least invite them and also UMLS.
- Jake Chen: Bridge2AI team, will propose names.
- Feedback on what to see: Deanne notes can review async, need the list of people and groups we have talked about.
- Lu from NCBI could share the workflows, Peipei Ping heard his talk recently.
- **Action:** gather a list of names that were suggested to capture, by email or chat.
  - Peipei recommends: Zhiyong Lu, NCBI
  - Jake recommends: Ying Ding UT Austin

Notes:

- □ Semantic typing in KGs: look at the Data Dictionary and provide feedback on what to classify.

- □ Embeddings out of the graph: obtain use cases from the DCCs.

- □ Fall Meeting - Preliminary Agenda: inviting people, gather a list:
    - □ Peipei recommends: Zhiyong Lu, NCBI
    - □ Jake recommends: Ying Ding UT Austin

2024-06-18

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	[Zoom Link](#)
Participants: SIGN IN: Name & Affiliation	Swathi Thaker (ICC-Admin); Chris Nemarich (KF); ;Deanne Taylor (KF/CHOP) ; ;Sherry Jenkins (LINCS, DRC) ; ; ; ; ; Jeffrey Grethe (SPARC); ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; ; ;Henning Hermjakob (EMBL-EBI, ICC-SC/UCLA) ; ;Peipei Ping (ICC-SC/UCLA) ; ; ;Jake Chen (ICC-AC) ; ; ; ; ; ; Dean Wang (ICC-SC); ;		

Agenda Item	Action Items	Owner
Introduction		Chris
Data Provenance		Deanne

Notes:

- Data Provenance
  - Example from DDKG
    - Included fields currently:
      - Name
      - SAB
      - Citations
      - Contexts (e.g. base context, etc. – structural for the UBKG)
      - Description
      - Download date
      - Home URLs
      - License(s)
      - Source ETL
      - Source Type (e.g. OWL)
      - Source Version
    - Does not have GitHub links for code that was used to generate the dataset
- There have been discussions on integrating the knowledge graph data with C2M2 - right now they are separate
  - Trying to have two-way linkages

    - Knowledge graph is built on the unified biomolecular data - underpins canonical information resource for HuBMAP and SenNet operations (multiple use general biomedical knowledge source framework)
- Good place to start for entry into the Knowledge Graph: https://github.com/nih-cfde/data-distillery
- This project ends in September so not sure what will happen after that
- Is this compatible with the data translators model?
    - It turns out that the group that is working in the Knowledge Center has connections with these models, but a formal connection has not been made
- Could think about having folks from the Data Translator attend the October CFDE meeting
    - Maybe there are other outside groups that could also be beneficial
    - In previous meetings there were breakout sessions
    - There have also been cross breakout meetings where working groups came together
    - Could think about inviting UMLS, NCBI
    - Have been working on developing a pediatric atlas from Kids First to work with HuBMAP

2024-05-21

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	https://uab.zoom.us/j/84863925833?pwd=MnI5WjNBT3UzU2s4VHhsN3lBZVhMdz09
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); ; Jake Chen; ; ; ; ;Daniel Clarke (DRC, LINCS) ; ; ; ; ; ;Sherry Jenkins (DRC, LINCS) ; ; ;John Erol Evangelista (DRC, LINCS) ; ; ; Avi Ma'ayan (DRC, LINCS); ; ; ; ; Mano Maurya (MW, DRC); ; ; ;Jeffrey Grethe (SPARC) ; ; ; ; Taha M. Ahooyi (KF); ; ; Stephanie Olaiya (DRC, LINCS); ; Nasheath Ahmed(LINCS); ; Heesu Kim (DRC, LINCS); ; ; ;Raja Mazumder (GlyGen) ; ; ; ; ; ;Jake Chen (ICC-AC) ; Swathi Thakar (ICC-AC); Sean Davis (ICC-EC) ; Wei Wang (ICC-SC) ;Dean Wang (ICC-SC) ; ;Keyang Yu (ERCC) ; ; ; ; ;Sherry Xie (LINCS, DRC) ; ; ; ; ; ; ; ;		

Agenda Item	Action Items	Owner
Introduction		Chris
Database Versioning		Jonathan
UMLS API		Jonathan
SIO (DisGeNET ontology)		Deanne
Update about fall meeting and Infrastructure		Jake

Notes:

- Database Versioning
  - For UBKG-based knowledge graphs (including but not limited to DDKG), the proposal is to update on a 6-month cycle of July and January
    - Based on the release cycle of UMLS and a few other key base datasets
- UMLS API
  - Have made updates to Smart API that provides graph returns for DDKG so that it will shortly verify users based on a valid UMLS key
  - Should the API be versioned or just always return the latest version of the graph?
    - This depends on how many users we have, if we have a lot of users on a regular basis that may necessitate versioning the API, but don't really need that for now. Users will be able to download old versions of the database for themselves and rebuild the standard API for their own use.

        - For now will just maintain the current version, not multiple versions through the DDKG API.
- SIO (DisGeNET Ontology)
    - https://www.disgenet.org/rdf
    - OWG to look into its utility for C2M2-relevant aspects
    - Sequence Ontology is a more "location" focused ontology that might be worth exploration as well, can include in future discussions: http://www.sequenceontology.org/browser
        - Could help with post-translational modification locations
        - Has been (semi-)adopted for bulk processing for gene transfer format (gff - describes features on a sequence and a common bulk dump format from databases)
        - Also a G4GH working group thinking about standardizing certain file formats and controlled vocabularies within the columns of those files
        - Could also look at linking the Sequence Ontology and HSCLO Ontology [https://github.com/TaylorResearchLab , https://www.biorxiv.org/content/10.1101/2024.02.15.580505v1] (?)
- Update about fall meeting and Infrastructure
    - Dates finalized for Fall meeting: Oct 22-23 (Tues-Wed)
    - Co-chairs will be determined, can nominate
    - Then will determine program committee, volunteer basis
    - Breakout discussions for working groups if desired
    - Phasing out the paid version of Slack for CFDE (will only hold 60 days of archived messages, but will archive existing old messages on Google Drive for additional searchability. Can also archive specific channels or discussions on a periodic basis as needed.)
    - Would like to create targeted channels on Slack for certain purposes, such as a #news channel for newsworthy announcements/events/etc that can be added to the newsletter.

2024-02-20

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	https://us06web.zoom.us/j/81043977847?pwd=dEdMWnpiZGVUdmJ2QnhaMWVSYkQ1dz09
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); Avi Ma'ayan (LINCS); ; Daniel Clarke (LINCS); Sherry Xie(LINCS); Heesu Kim(LINCS); Sherry Jenkins(LINCS); ; Jimmy Zhen (MoTrPAC); ; ; ; Mano Maurya (MWA); ; ; ; ;Matt Roth ; ; Keyang Yu (ERCC); ; ; ;Bosko Jevtic ; ; ; ; ; ; ; ; ; ; ; ; ; ;		

Agenda Item	Action Items	Owner
Introduction		Chris
Accessory Datasets for Data Distillery Knowledge Graph		Deanne
Updates on DDKG Deployment		Chris
Updates on UBKG / DDKG API		Chris

Notes:

- Follow-up related to KEGG licensing
    - Still unclear about where this falls, seems to be in a gray area in terms of derivative products and what counts as "small" vs. "large" usage.
    - Solution from Deanne: Ben could on the side look into what it would take to ingest KEGG, then write up some instructions for people to be able to walk through the steps themselves for ingestion for local usage.
    - As an example, instructions for MetGENE (from Metabolomics Workbench) include specific language regarding KEGG: https://github.com/metabolomicsworkbench/MetGENE
- Updates on DDKG deployment
    - A Dockerized Data Distillery neo4j instance running on a VM - deployed
    - A Dockerized instance of the Data Distillery API, based on the UBKG API - deployed
    - AWS API Gateway configuration - In progress. Estimated completion end of month.
    - SmartAPI registration of specification to document the endpoints of the DD API. In progress. Source ready; waiting on 3.

2023-11-21

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	https://us06web.zoom.us/j/81043977847?pwd=dEdMWnpiZGVUdmJ2QnhaMWVSYkQ1dz09
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); Deanne Taylor (KF); ; Mano Maurya (MW/DRC); ; ; ; ;Christy Kano ; Julia Markowski (4DN); ; ; ; Jeffrey Grethe (SPARC); ; Taha M. Ahooyi (KF); ; ; Michelle Giglio; ; ; Keyang Yu (ERCC); ; ; ; ; ;Noel Burtt (KC) ; ; ; ; ; ; ;		

Agenda Item	Action Items	Owner
Introduction		Chris
Accessory Datasets for Data Distillery Knowledge Graph		Deanne

Notes:

- Accessory Datasets for DDKG
  - SPARC is working on Reactome dataset for this year
    - Cannot use KEGG as a database because of licensing, this will be removed before adding it to DDKG
- KEGG licensing
  - https://www.kegg.jp/kegg/legal.html
  - “Academic users who utilize KEGG for providing academic services are requested to obtain an academic service provider license, which is included in the KEGG FTP academic subscription.”
  - https://www.pathway.jp/en/academic.html
- Criteria for evaluating accessory datasets for use in CFDE KGs
  - FAIR dataset
  - Open license with open distribution
- Metabolomics Workbench (MW) dataset leverages KEGG to develop their assertions for the UBKG and as a derivative product may break licensing issues. Mano will investigate this and return with an answer from the KEGG licensing team.

2023-09-19

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); Deanne Taylor (KF); Jonathan Silverstein (HuBMAP); Ben Stear (KF); Mano Maurya (MW); Christy Kano (NIH); Haluk Resat (NIH); George Papanicolaou (NIH); Aleks Milosavljevic (ERCC); Sherry Jenkins (LINCS); Taha M. Ahooyi (KF); Sherry Xie (LINCS); ; Erol Evangelista (LINCS); ; Avi Ma'ayan (LINCS); ; ; ; ; ; ; ; ; ; ; Julia Markowski (4DN); ; ; ; ; ; Daniall Masood (GlyGen); ; Keyang Yu (ERCC); Varduhi Petrosyan (ERCC); Srini Ramachandran (MW); ;		

Agenda Item	Action Items	Owner
Introduction		Chris
Reviewing updates to Data Distillery - GitHub site w/ User Guide and Data Dictionary		Chris

Notes:

- Reviewing updates to GitHub site for Data Distillery w/ User Guide and Data Dictionary
    - Avi noted that the CMAP dataset included in the Data Dictionary is potentially conflictual with other LINCS data and should be removed from future distributions

2023-06-20

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	https://us06web.zoom.us/j/84006612795
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); Deanne Taylor (KF); ; ;Sherry Jenkins (LINCS) ; ; ; ; ; ; ; ; ; ; ; ; Srini Ramachandran (MW); ; ; ; ; ; Owen White; ; ; ; ;Emmanuel Esquivel (ERCC) ; ; ; ;Avi Ma'ayan (LINCS) ; ; ; ;Keyang Yu (ERCC) ; Taha M. Ahooyi (KF) ;Jeffrey Grethe (SPARC) ; ; Michelle Giglio (CC); ;; ;Julia Markowski (4DN) ; Clara Bakker (4DN); ; ; Matt Roth (ERCC); ; ; ; ; ; Sherry Xie (LINCS); ; ; ;John Erol Evangelista (LINCS) ; ;Andy Schroeder (4DN) ;Varduhi Petrosyan (ERCC) ;		

Agenda Item	Action Items	Owner
Introduction		Chris
GENCODE Update		Alan Simmons
Demonstration of the Data Distillery Graph		

Notes:

- Gencode updates
    - Latest Gencode version is included in the UBKG, and by extension in the Data Distillery
- https://www.ebi.ac.uk/intact/home
- https://maayanlab.cloud/archs4/
- https://thebiogrid.org/
- https://bioplex.hms.harvard.edu/interactions.php
- Data Distillery team will provide a set of test queries and ways of knowing that your graph has built appropriately for a given set of CSVs in the README.txt file.
- Topic for next discussion: numerics on nodes vs. edges, pros and cons and for which sources

2023-05-23 CANCELED

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	; ;		

Agenda Item	Action Items	Owner
- CANCELED		Chris

Notes:

-

2023-04-18

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Eric Wenger (KF); Deanne Taylor (KF) ; Avi Ma'ayan (LINCS); ; Daniel Clarke (LINCS); ; ; ; ; ; ; ; ; ; ;Sherry Jenkins (LINCS) ; ;Sherry Xie (LINCS) ;Erol Evangelista (LINCS) ; ; Michelle Giglio(CC); ;Tom Gillespie (SPARC) ; ; Jeffrey Grethe (SPARC); ; ; ;Srini Ramachandran (MW) ; Mano Maurya (MW); Julia Marowski (4DN); Clara Bakker (4DN); Andy Schroeder (4DN); Jimmy Zhen (MoTrPAC); ;Keyang Yu (ERCC);Varduhi Petrosyan (ERCC) ; Ben Stear (KF); Taha M. Ahooyi (KF); Matt Roth (ERCC)		

Agenda Item	Action Items	Owner
Introduction: - Additional agenda or updated items?		Eric
(tentative) Docker container usage with the Core KG	Jonathan/Deanne: move forward in implementing a core KG version that does not have UMLS license restrictions	Jonathan
Other discussion: visualization opportunities for KG	Avi: share JSON node/edge model for subgraphs as a potential basis for standardizing KG Smart API output	

Notes:

Docker and Core KG

- As part of the core KG efforts was to make available all of the source files for the KG and the accompanying code to ingest the data into the KG
- Use of Docker locally provides an efficient mechanism for downloading and testing the core KG
- US UMLS licensees can use this without restriction in the United States

Q: Can we have a UMLS free version of the graph or is UMLS the backbone?
A: Not currently implemented license-free, but is a future opportunity. The current download will have some missing connections (i.e. some clinical limitations).

Unlicensed version of KG and opportunity for generalizable framework

- When implemented, this will miss only licensed codes and relationships
- If ingest code is designed to layer in licensed content on top of unlicensed content, this could support a “layer-based” / generalizable framework that still allows the core KG to be distributed without required licensing
- Short-term solution: download the core KG without the licensed content

New APIs

- Future pending (~month) updated SmartAPIs with underlying details on cypher queries provided

Future discussion topics of interest - and exploratory discussion today

- **Visualizations with KG’s**
    - Opportunity to explore best practices and direction for CFDE
    - We can tap into the insights and experience of DCC’s on the call, such as LINCS and others, who have experience and have explored options in the visualization arena. Taha has also recently identified some visualization integration options.
    - LINCS/Erol developed a visualization layer on top of neo4j. Code is open source, and can be further customized as requirements require. Some examples:
        - Enrichr-KG is one prototype example.
            - Supports multiple layouts and includes the option to add missing nodes
        - CFDE Gene Set Knowledge Graph is an additional example
            - Includes search for shortest path, etc.
    - Neo4j also has additional visualization tools which could also be explored in an upcoming meeting
    - There are opportunities to create an on-the-fly projection (a KG equivalent of a db view) that subject performant subgraph queries.

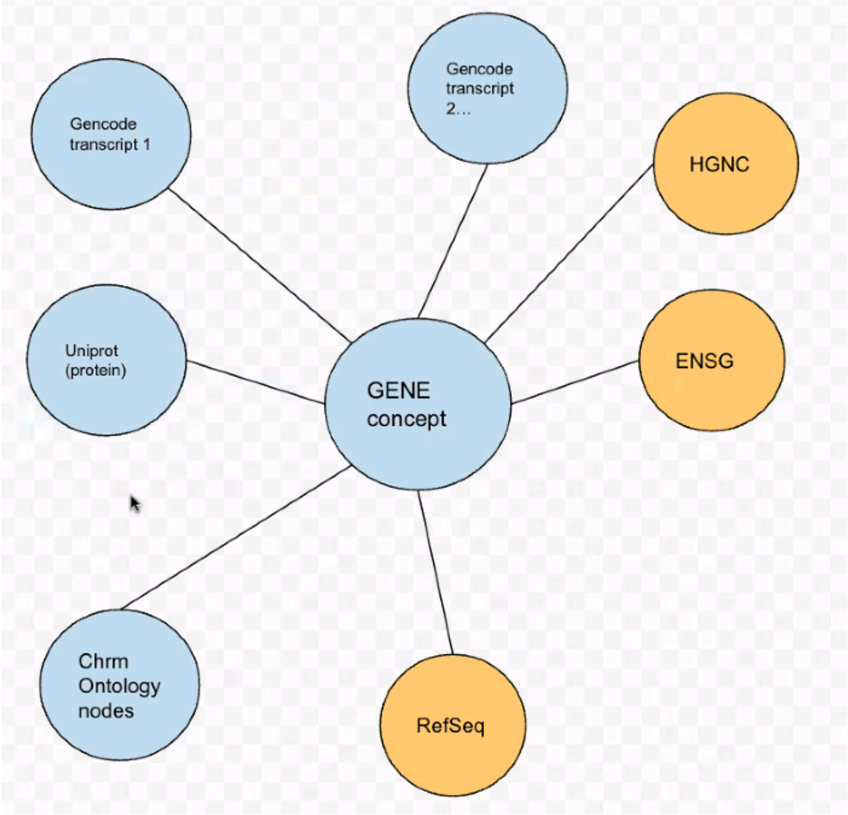
2023-03-21

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); Avi Ma'ayan (LINCS); ; Daniel Clarke (LINCS); ; ; ; ; ; ; ; ; ; ;Sherry Jenkins (LINCS) ; ;Sherry Xie (LINCS) ;Erol Evangelista (LINCS) ; ; ; ;Tom Gillespie (SPARC) ; ; Jeffrey Grethe (SPARC); ; ; ;Srini Ramachandran (MW) ; Mano Maurya (MW); Julia Marowski (4DN); Clara Bakker (4DN); Andy Schroeder (4DN); Jimmy Zhen (MoTrPAC); ; ;Varduhi Petrosyan (ERCC) ; Ben Stear (KF) Matt Roth (ERCC)		

Agenda Item	Action Items	Owner
Introduction: - Additional agenda items?		Chris
GENCODE Models		Deanne

Notes:

- GENCODE Models
  - How deep do we want to go for Data Distillery for the models? For knowledge graphs in general in CFDE?
  - Avi: Add proteins w/ Uniprot IDs
  - Mano: Add NCBI Entrez (GeneID)
  - Aliases will come off of HGNC nodes, etc.
  - Want to ensure that any gene transcripts that are brought into the graph are sufficiently distilled
  - Need to see if there is something in the relationship ontology for “transcript of…”


Gencode transcript 1
Gencode transcript 2…
HGNC
Uniprot (protein)
GENE concept
ENSG
Chrm Ontology nodes
RefSeq

2023-02-21

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); ; Shankar Subramaniam (MW); Deanne Taylor (KF); ; ; ; ;Sherry Jenkins (LINCS) ; Sherry Xie (LINCS); John Erol Evangelista (LINCS); Daniel Clarke (LINCS);Varduhi Petrosyan (ERCC) Mano Maurya (MW); ; ;Taha M. Ahooyi (KF) ; ; Clara Bakker (4DN); Julia Markowski (4DN) ;Rahi Navelkar(4DN) ; Matt Roth (ERCC); ;Avi Ma'ayan (LINCS) ; ; ; ; ; ;Tom Gillespie (SPARC) ;Andy Schroeder (4DN) ; ; ; ; ; ;		

Agenda Item	Action Items	Owner
Introduction: - Additional agenda items?		Chris
Discussion regarding Gencode: - Should we support Gencode for KGs in CFDE?		Deanne
KG / C2M2 subgroup: - Meeting scheduled for 11am EST this Friday - Clarify mandate / agenda for this group		Chris / Team
Items from last Ontology WG meeting: - Proposal to move C2M2 to Mammalian Phenotype Ontology - There is a federal govt. proposal to split Middle Eastern & North African from White in Race & Ethnicity controlled vocab following updated Census		Chris / Michelle / Team

Notes:

- Support for Gencode for KGs in CFDE
  - Gencode provides cross-referencing, annotations, feature mapping on the genome
  - However, updates frequently and has a number of versions
  - Should we pick a specific gencode version to support or support gencode broadly and annotate which version information comes from?
  - Link to Gencode release 41 files: https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_41/
  - This issue crosses a number of different groups
  - Jonathan: The current pattern for projects like DD is to update the KG as we go along, but largely around the 6 month basis on the UMLS release schedule,

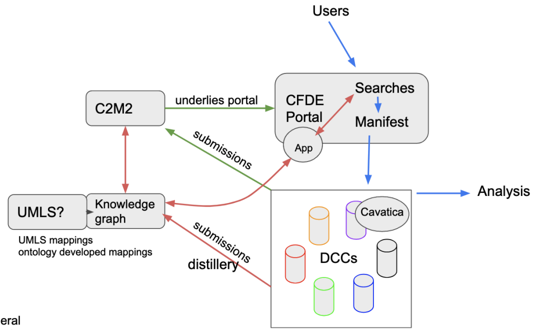
    which could be adopted as a general recommendation for these larger multi-group KGs.
  - Tom: We do quarterly releases with SPARC based on similar principles to what was outlined by Jonathan. Question of proper metadata to know which version you have loaded. Question of documentation for users. However, rapidly changing datasets can create conflicts.
    - "Schema changes and need to update queries are a known issue on our end." - Tom
  - Potential challenge with older data what different codes can mean if breaking changes are made to codes rather than additions. Deletions can also present a challenge.
  - Gencode uses Ensemble IDs. This has the possibility of changing the structure of the graph when it is updated (genes that are deleted, added). Stability only goes so far, e.g. when boundaries of exons change, etc.
  - "This isn't so much an ontology question so much as a sourcing question" - Jonathan
    - There's a version issue, but there is also a question of whether we trust Gencode is a reliable source for all of the other datasets that it brings along with it.
    - Potential issues utilizing something like Gencode are completeness and correctness.
  - Gencode is a community standard for usage.
  - This would be for mouse and human. If we wanted to include other species, this would have to go to other sources.
  - Choosing to use gencode now really incentivizes using it again in the future.
  - Several voices of support, no dissent on the usage of Gencode, so for now moving forward with the use of Gencode.
  - Implication in Data Distillery is that we would accept Ensemble IDs & NCBI Gene IDs. Canonical mappings would be coming from Gencode.
  - Majority of this will be in the form of links, since this is cross-references.
- KG / C2M2 Subgroup
  - First meeting this Friday will be used to discuss the scope and mandate of this group, will bring back a proposal to this group to discuss rather than discuss here now.
- Ontology WG Follow-up items
  - Mammalian Phenotype (MP) Onotology is quite large compared to human ontology.
    - Shouldn't impact CFDE in terms of performance.
    - May impact ability for users to find appropriate search terms.
    - Not mapped to HPO, though there are efforts underway to do that. For instance, are all mappings from animals to humans one-to-one. There was an RO3 awarded through CFDE to map birth defects in MP to HPO.
    - Correction – fairly equivalent, about ~13,000 terms in each.

- Had assumed that there was a mapping between MP and HPO already. Were trying to avoid needing to avoid supporting multiple ontologies for phenotypes.
- There was a previous Common Fund project COMP2 (?) for which data still exists and exclusively uses MP. If we wanted to include this data, we would have to support MP.
- Michelle may send out a survey to DCCs to see who is using what.
- By way of example, usage of MP for cross-walking research across species has been incredibly helpful in the Kids First space.

2023-01-17

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); ; ; Sherry Xie (LINCS); ;Sherry Jenkins (LINCS) ; ; ; ;Deanne Taylor (KF) ;Ben Stear (KF) ; Taha Mohseni (KF) ;Aditya Lahiri ;Aleks Milosavljevic (ERCC); ; ;Avi Ma'ayan (LINCS) ; Daniel Clarke (LINCS); Mano Maurya (MW); Srini Ramachandran (MW); Rahi Navelkar; Clara Bakker ; Julia Markowski ; John Erol Evangelista (LINCS); ; Tom Gillespie (SPARC); ; ; ; Michelle Giglio ;Varduhi Petrosyan (ERCC) ; Jesse Helfer ; ; Matt Roth (ERCC); ; ; ;		

Agenda Item	Action Items	Owner
Introduction - Additional agenda items?		Chris
Continuation of discussion from Ontology WG meeting last week: - How can we make use of the Knowledge Graph info within the context of C2M2 and the Portal? And vice versa, how can C2M2 and the Portal be made use of within the Knowledge Graph? Key notes from Ontology WG meeting: - "The Knowledge Graph effort will be capturing information coming in from the Distillery project that is not, and will not, be captured in C2M2. Also, the Knowledge Graph effort will allow linkages to vocabularies well beyond those chosen for harmonization within C2M2."		Deanne / Jonathan

- “First proposed use case that was explored: translation between vocabularies” - “An API could be built that would include within it all of the C2M2 associations as well as all of the KG associations where associations are linked to each through UMLS or ontology-provided mappings (which might be equivalence or other types of mappings).” - “If a user would like to search at the Portal with a term that is not part of a vocabulary used in C2M2, the App would be able to look up what C2M2 terms are linked to the query term via the mappings/linkages in the KG and return this information to the user. The App could also provide some ‘roll up’ or ‘roll down’ abilities to provide more general or more specific information as relevant.”  Users C2M2 underlies portal CFDE Portal Searches Manifest App submissions UMLS? Knowledge graph UMLS mappings ontology developed mappings submissions distillery Cavatica DCCs Analysis use cases specific to general general to specific translation - allow entry by any vocab - precalc/preingest or transit graph with back-and-forth API calls in an app attached to the portal; can do equivalence or limited roll ups		
AOB		All

Notes:

- The KG’s place in the CFDE Ecosystem
  - This is a continuation of the discussion from the Ontology WG meeting last week
  - Don’t want to rule out the possibility of DCCs providing more granular information to the Knowledge Graph than what they provide to C2M2, but don’t want to create redundant reporting and ingestion processes for DCCs that increase the burden on them
  - An additional question is whether or not the Knowledge Graph should be aligned to the standards of C2M2 and the Ontology WG?
  - Existing structure: DCCs submit data to C2M2 which underlies the CFDE Portal. Users enter through the CFDE Portal, search for data, obtain a manifest, go to a cloud environment or to the individual DCCs to obtain the data, then conduct their analysis.
  - Proposed structure: Also ingest data from the DCCs into the Knowledge Graph, as well as data from C2M2, which would allow additional translations between vocabularies

    - Knowledge Graph represents high-level, abstracted version of DCC data, while C2M2 is raw data. Adding assertions from lower-level data from C2M2 would potentially grow the knowledge graph unnecessarily and make it a mess.
        - This should be a selective inclusion of information from C2M2 into the Knowledge Graph
    - Should be careful in how the diagram included above is interpreted: this is not a data flow diagram.
    - Potential uses for querying C2M2 from Knowledge Graph: Look at assertions in the Knowledge Graph and then find data in C2M2 that is related to those assertions.
    - Short term will provide an ability to have assertions linked to DOIs or GitHub repos that contain code used to derive the assertion, but long term want to standardize how we connect assertions in the Knowledge Graph back to the evidence used to derive it.
    - There is an Evidence and Conclusion Ontology (ECO) that "contains terms (classes) that describe types of evidence and assertion methods. ECO terms are used in the process of biocuration to capture the evidence that supports biological assertions (e.g. gene product X has function Y as supported by evidence Z). Capture of this information allows tracking of annotation provenance, establishment of quality control measures, and query of evidence." This or other ontologies could be leveraged in the Knowledge Graph to provide useful results when querying data. https://evidenceontology.org/
        - For instance, could query graph based on assertions that have a very high level of evidence underlying them.
        - In other cases, you may want to query the knowledge graph using any sort of level of evidence for assertions when just doing early exploration.
        - ECO captures *types* of evidence rather than *quality* of evidence. Some people may use *type* of evidence as a proxy for *quality*, but this is not the same thing.
        - KG WG could introduce one or more new properties to capture *quality* of evidence to mature ideas around that or could look for an existing ontology that captures this information. Might require the creation of a smaller group to be tasked with creating a first draft of this standard. Would *not* be included in current distillery knowledge graph, since this project needs to deliver a number of specific things in the next ~8.5 months.
- Will follow up in the next meeting to clarify what exactly this new sub-group will be tasked with that will be working with the Ontology WG in building out a use case for connecting the CFDE C2M2 / KG / Portal resources
    -

2022-11-15

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom Meeting
Participants: SIGN IN: Name & Affiliation	Chris Nemarich (KF); Deanne Taylor (KF) ; Mano Maurya (MW); ; ; ; ; ; ; ;Avi Ma'ayan (LINCS); Andy Schroeder (4DN); ; ; Jeremy Yang (IDG) ; ;Taha M. Ahooyi (KF) ; Daniel Clarke (LINCS); Sherry Xie (LINCS); ;Sherry Jenkins (LINCS); ; ; John Erol Evangelista (LINCS); ; ; ; ;George Papanicolaou ; ; ;Jeffrey Grethe (SPARC) ;Srini Ramachandran (MW) ; ; ;		

Agenda Item	Action Items	Owner
Discussion: Linking chromosomal locations	Review of the chromosomal location "ontology"	Deanne, Taha
Continued discussion on the Distillery_Ingest_format and walk through a visualization of the KG components at https://github.com/dbmi-pitt/UMLS-Graph	Team: ● Review Data Distillery format recording here ● Review data format sheet from presentation: Distillery_Ingest_format ● Visualization from presentation here	Jonathan
Discussion: what external references to include as standard in the KG which are meaningful for CFDE	Team (last meeting) ● Ontologies: review Ontology WG list of ontologies leveraged for C2M2 and determine whether there are additional ontologies to consider for the KG ● Consider which standard datasets to include in the KG, you may find it helpful to see what DCC use case anchors currently exist for the Data Distillery Project here	Deanne
Types of edges in the KG	Ways to qualify types of edges (address, phenomena, etc) in a KG.	
Discussion: Follow-up from last week's CFDE DC Meetings		
Suggested future subjects for discussion: ● Types and levels of evidence ● Numerics on codes ● Numerics on relationships		

Notes:

- Chromosomal location ontology
    - Originally ingested into the KG as chromosomal bands, obtained from the MSigDB C1 dataset
    - 278 locations of variable length were introduced as new nodes
    - Relationships were defined as connections to HGNC nodes
    - Updated chromosomal location ontology allows for easy connection of different resolution datasets including but not limited to:
        - Physical contact regions (4DN Hi-C)
        - Chromatic accessibility (ATAC-seq)
        - Functional regions of DNA (GTEx genes, regulatory elements, etc.)
        - Genomic features within certain chromosomal regions (e.g. all KF SNVs in the regulatory elements)
    - Updated chromosomal location ontology has XX nodes and 12 million relationships
    - Should reduce the overall # of connections in the KG and simplify it overall by allowing you to avoid connecting dataset-to-dataset and instead connect dataset-to-scaffold-to-dataset
    - Question from Jonathan: It appears that there are edges that “skip levels” is this needed or can this be a true hierarchy (many advantages to not fully connect I think …?)?
        - Diagram on the right of the slide is not accurate to the intention, diagram on the left shows the actual ontology structure, which has each node only connecting to the immediate level above it
- Distillery ingest format: Updates from Jonathan
    - Took some shortcuts for distillery, some open questions as to how we want to deal with nodes and edges in future cases
    - Could continue to let people import evidence as they need/see fit, or could create a set of standards to how to
    - Will come back with a proposal of several ideas to seed discussion in a future session
    - Mano asked about how relationships that are not directional will be handled
    - If you are using the relations ontology, the reverse relationship will be added automatically
    - If users enter relationships that are not in relationship ontology, will define inverse relationship with `inverse_`
    - An example of discussion around inverse relationships previously was coexpression, where the final decision was to define the inverse of `co-expression` as `co-expression`
    - Also have the ability to maintain a separate side table that can be used to define some of these specific relationships (e.g. `coexpression`, `associated_with` etc.)
- What external references to include as default in KG:
    - Need to get this on the agenda of the Ontology WG

  - The Ontology WG chooses which references to favor
  - The Ontology WG will likely be interested in the Chromosomal Location ontology
  - Will this be discussed in a joint WG session, just the Ontology WG forum, or via Slack? Sounds like it will be taken care of in the next Ontology WG meeting for at least next steps to pursue.
  - Need to provide the long list of ID formats of ingestion to this WG via Slack
  - If DCCs find that there are codes that are not included in the KG that are favorable for their purposes, can address how to handle those or include them
- Types of edges in KG
  - Some edges are facts, such as physical location of chromosomes
    - This requires a different way of thinking about edges compared to other assertions that have previously been included
  - Need a system for qualifying edges
  - How do we define the quality of the edges?
  - Weighting edges - how do you weight a fact?
  - Creating colors or flavors of edges for future analysis work
  - Jeremy: this also brings in questions of provenance, etc.
- Topics for future discussions:
  - Should try some things that involve numerics in the Data Distillery context and then bring back findings from that to the KG WG afterwards
    - Develop a list of advantages and disadvantages based on different ideas that are generated in that context
  - Create a sub-group for working on evidence
    - Deanne, Tom

Upcoming Agenda Topics

Date	Agenda Topic	Who
Upcoming meetings...	●	

2022-10-18

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom https://zoom.us/j/4286167066?pwd=Z1M4SFRjL2pUWjJVbGt0S3NTR2hjZz09
Participants: SIGN IN: Name & Affiliation	Sherry Jenkins (LINCS); Avi Ma'ayan (LINCS); Daniel Clarke (LINCS); John Erol Evangelista (LINCS); Mano Maurya (MW); Eric Wenger (KF); Taha M. Ahooyi(KF), Varduhi Petrosyan (BCM); Srini Ramachandran (MW)		

Agenda Item	Action Items	Owner
LINCS DCC–initial try at providing assertions based on the specs provided by @Jonathan Silverstein (file shared in the KG WG Slack here)	Numerics on nodes and relationships. Review with new hire (Silverstein).	Avi
Discussion: Linking chromosomal locations	Deanne, Ben, Taha: develop initial version of proposed addition, share w/ Jonathan for structural review and then share with larger WG for feedback	Deanne
Present overview of Distillery_Ingest_formaand walk through a visualization of the KG components at https://github.com/dbmi-pitt/UMLS-Graph **Note: presentation deferred – given that a recording of an overview presentation was created**	**Team:** • **Review Data Distillery format recording here** • **Review data format sheet from presentation: Distillery_Ingest_format** • Visualization from presentation here	Jonathan
Discussion: what external references to include as standard in the KG which are meaningful for CFDE	**Team:** • **Ontologies: review Ontology WG list of ontologies leveraged for C2M2 and determine whether there are additional ontologies to consider for the KG** • **Consider which standard datasets to include in the KG, you may find it helpful to see what DCC use case anchors currently exist for the Data Distillery Project here**	Deanne
Suggested future subjects for discussion:		

• **Types and levels of evidence** • **Numerics on codes** • **Numerics on relationships**		

Notes:

Discussion: linking chromosomal locations

- Suggestion to start with 10 to 1 KB window
- Need to define nodes, and naming conventions

Discussion: what external references to include as standard in the KG which are meaningful for CFDE

- Standard datasets, (for example: for variants)
    -
- Ontologies
    - Starting point is the list of ontologies that were chosen in the Ontology WG for C2M2
        - Ontologies and CVs used for C2M2 submissions:
            - Ontology for Biomedical Investigations (for assay type, analysis type, sample prep method)
            - EDAM (for both data type and file format)
            - Disease Ontology
            - PubChem
            - Human Phenotype Ontology
            - Uberon
            - NCBI taxon
            - internal CVs for race (based on federal terms), sex (a selection of SMOMED terms), ethnicity (based on federal terms)
    - Further team feedback on any additional ontologies to consider

---

Upcoming Agenda Topics

Please add upcoming agenda topic ideas here.

Date	**Agenda Topic**	**Who**
Upcoming meetings...	•	

2022-08-16

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Deanne Taylor (KF); ;Daniel Clarke (LINCS) ; ; Sherry Xie (LINCS) ; Sherry Jenkins (LINCS) ; ; Avi Ma'ayan (LINCS); ; Jeffrey Grethe (SPARC); Tom Gillespie (SPARC); ; Jonathan Silverstein (HuBMAP); Taha M. Ahooyi (KF); Mano Maurya (MW); ;Srini Ramachandran (MW) ; ;Raja Mazumder (GlyGen) ; Michelle Giglio; Shankar Subramaniam. Matt Roth (exRNA)		

Agenda Item	Action Items	Owner
Review of Charter and next steps	Identifying data sets from each DCC for next month More specific use cases for next month	
Ideas for use cases	Researchers/Labs as nodes connected to genes, diseases, cell types, drugs, glycosylations and other KG entities (?) (Avi)	
Interface and ingestion scripts	Interface to KG underway (Avi/LINCS/Reprotox KG) and a prototype was developed for the Gene and ReproTox partnerships with the same UI (https://maayanlab.cloud/gene-kg; https://maayanlab.cloud/reprotox-kg). Can drive any neo4j instance, flexible. Extended schema beyond turtle files that also supports attributes for edges that neo4j (JSON extension for specification for ingestion for all properties on edges/nodes).	

Notes:

Upcoming Agenda Topics

Date	**Agenda Topic**	**Who**
Upcoming meetings...	●	

2022-06-21

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Deanne Taylor (KF); ; ;Christy ; George ; ; ; ;Daniel Clarke (LINCS) ; ; Sherry Xie (LINCS) ; Sherry Jenkins (LINCS) ;Taha M. Ahooyi (KF) ; Avi Ma'ayan (LINCS); ; Jeffrey Grethe (SPARC); Tom Gillespie (SPARC); ; Jonathan Silverstein (HuBMAP); ; Mano Maurya (MW); ; ; ;Raja Mazumder (GlyGen) ; Michelle Giglio;		

Agenda Item	Action Items	Owner
Discuss status of budget of partnership	Submit modified budget once decision is made by NIH	Deanne
Discuss charter	File for Cole	Jonathan
Discuss current schema for the KG	WG: Knowledge Graph WG Next Steps Michelle Giglio to Everyone (3:54 PM) Full disclosure - I run the Evidence Ontology - happy to chat about any of these issues any time. Using just GO codes is fine too, they all map to the Evidence Ontology. Raja Mazumder to Everyone (3:55 PM) Really hard to come up with "strength" IT is contextual	Jonathan

Notes:

Upcoming Agenda Topics

Date	Agenda Topic	Who
Upcoming meetings...	●	

2022-05-17

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Christophe Lambert (IDG); Avi Ma'ayan (LINCS); Sherry Jenkins (LINCS); Daniel Clarke (LINCS); Erol Evangelista (LINCS); Sherry Xie (LINCS); Deanna Taylor (KF); Jeffrey Grethe (SPARC); Srini Ramachandran (MW)		

Agenda Item	Action Items	Owner
Deanne presentation of how they ingested the HuBMAP data Jonathan documentation and presentation of the schema for the Knowledge Graph		
Hackathon recap Tutorial: https://github.com/TaylorResearchLab/CFDIKG/blob/master/Tutorials/CFDE_Hackathons/Tutorial_CFDE_Hackathon_5-2022.md		
Schema	●	
Use cases		

Notes:

Upcoming Agenda Topics

Date	**Agenda Topic**	**Who**
Upcoming meetings...	● Deanne presentation of how they ingested the HuBMAP data ● Jonathan documentation and presentation of the schema for the Knowledge Graph	

2022-04-19

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Diane E. (PSC), Deanne T., Shankar S. Haluk R., Ben S., Christy K., Jeffrey G., Lynn S., Mano M., Owen W., Srini R., Suman S., Sherry X. (LINCS), Tom G., Avi M., Taha A., Jonathan S., Keyang Y., Raja M., Daniel C., Sherry J.		

Agenda Item	Action Items	Owner
Application Proposal	Update calendar invites (DE)	
Data model		

Notes:

- JS - UMLS Graph presentation
- BS - GTEx Ingest presentation *(slide deck to be linked - tech difficulties)*
- TG - added wording about evidence in charter
- DE will be updating Zoom link on calendar invite and pinning it in Slack channel

2022-03-15

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Jonathan Silverstein; Deanne Taylor; Kevin Hanshaw; Avi Ma'ayan (LINCS); Sherry Xie (LINCS); Daniel Clarke (LINCS); Sherry Jenkins (LINCS); Shankar Subramaniam; Mano Maurya (MW); Christophe Lambert (IDG); Suman Sirimulla (UTEP); Manuel Aguilar (UTEP); Michelle Giglio; Keyang Yu (ERCC); Haluk Resat; Jeffrey Grethe (SPARC); Tom Gillespie (SPARC)		

Agenda Item	Action Items	Owner
Finalize the Charter		
• Data Distillery LOI		
• Next steps • If time -- Owen has a question for group	•	

Notes:

- Concerns about how the knowledge graph can lose its purpose without the correct data and the mapping defined properly
- Suggestion of using filtering mechanisms.
- CL suggests an API or GUI for interaction with the knowledge graph
    - JS suggests not rewriting the the APIs and GUIs but find the solutions that works best and build on those
- Governance of the Data Distillery involvement
    - Important question that is worthy discussion
    - The KGWG should define the requirements
        - Used the participation of the working group of who was included in the LOI
    - Make sure we have each of the major bio categories?
    - The ones that are already prepared as assertions
        - Review the ones that have very mature assertions already to focus on those
    - Data that goes will empower the queries that can happen
- For data being ingested, model the data in a way that works where the gating takes place
- AM Pasted the assertions into the Charter that can be focused on to shape into the data for the Knowledge Graph.
- Specifications would be making sure the appropriate data is in the spreadsheets.

    - Documentation and demonstration of the schema
    - edge types will explode, especially if you try to eliminate blank nodes
- Write scripts that could do the plumbing of the data to make sure they are mapped properly.
    - C2M2 should help with keeping the data standardized with common entities
    - HuBMAP Ontology API is already implemented in Smart API and that is automated from the discussed ingest code into a full research fully automatic in Docker container on AWS, etc…so that part is done
- Remember that the knowledge graph is “distilled” (summarized data to assertions) data oriented, whereas C2M2 is dataset cardinality
- CL: I'm still getting my bearings. How expressive are the inferences over these knowledge graphs? My comment about computation was coming from the place of contemplating that there are increasingly more powerful logics with more expressiveness that we may or may not use.
    - Knowledge Graphs of this design explicitly is NOT for reasoning, but for the uses in the documents (both distillery and the WG charter per se state this) - so we’re not thinking logics but rather queries, which to “sort of” answer your question can be highly expressive
- OW: A bunch of effort into making summary reports. Need to be vigilant about what types of data that is in the portal that will be valuable.

Action items:

- Everyone review the charter for it to receive group approval next month
    - Tom Gillepsie add text about assertions in the Charter

2022-02-15

Objective	Working Group call	Time	3 PM EST every 3rd Tuesday
Leader(s):	Deanne Taylor Jonathan Silverstein	Where	Zoom
Participants: SIGN IN: Name & Affiliation	Kevin Hanshaw; Jeffrey Grethe (SPARC); Michelle Giglio; Matt Roth; Tom Gillespie (SPARC); Taha Ahooyi; Christy Kano; Jessica Binder (IDG); Christophe Lambert (IDG); Eric Wenger (KF/CHOP); Deanne Taylor (KF); Srinivasan Ramachandran (MW); Mano Maurya (MW); Haluk Resat		

Agenda Item	Action Items	Owner
Introductions of attendees	none	Leaders
● Review of Charter KG WG Charter_DRAFT	● **Member comments**	leaders

Notes:

- AM highlighted that the variant WG is looking to produce knowledge related to their work and included in the KG discussion
    - JS agreed and referenced the Ontology WG line in the charter that also highlights the same interest. AM can offer up the change he would like.
- JB is currently writing a review on heterogeneous knowledge graphs for genes to disease predictions.
    - https://journals.plos.org/ploscompbiol/article/figure?id=10.1371/journal.pcbi.1004259.g001
- JS thinks we should hold demonstrations on what they think is valuable to contribution
    - Each DCC provide their work that is going to be used with a knowledge graph form
- Early meetings will be sharing their data models to figure out how to combing for a collection then work towards a common goal by the end of this working group.
- AM defined the assertions being used in the Variant WG and how we may be utilizing the term in the wrong
    - Examples discussed about how to highlight the assertions without overloading the graph.
- JS indicated the graph can not provide reasoning to avoid creating bias
    - Loops and redundancy will happen but the APIs will help be used in different ways to mitigate this
- JB asked if the plan is to the use the portal as a database? Then the researcher would select the different relationships?
    - Data that is being added is gene-centric and not disease centric
    - The sourcing for the data will indicate where it came from

- Taking in Mondo would bring in overlapping information that is asserted by the other ontologies. The Mondo statistical information would be hard to integrate properly.
    - TG chatted mondo has mappings to snomed now iirc. kg as hypothesis vs kg as "truth" is a critical distinction in the approach
- TG chatted there are still lingering issues with the disease ontology being used in a full owl setting
    - A way around it is to remove the semantics
- New use case?
    - JG thinks it fits with the data analysis use case already
- Next steps
    - How are we going to track the deliverables?
        - Github
    - Who is going to be respresentable from the DCCs?
    - Design the actions driven outside of this WG as a partnership per Haluk.

Action Items:

- TG is going write some text about the assertions
- Action plan defined on how to apply and collaboration
    - The data each DCC will deliver, the volunteers who will be the engaged person for providing data to the knowledge graph. Survey