JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

PID Graphs in Practice

Neil Byers

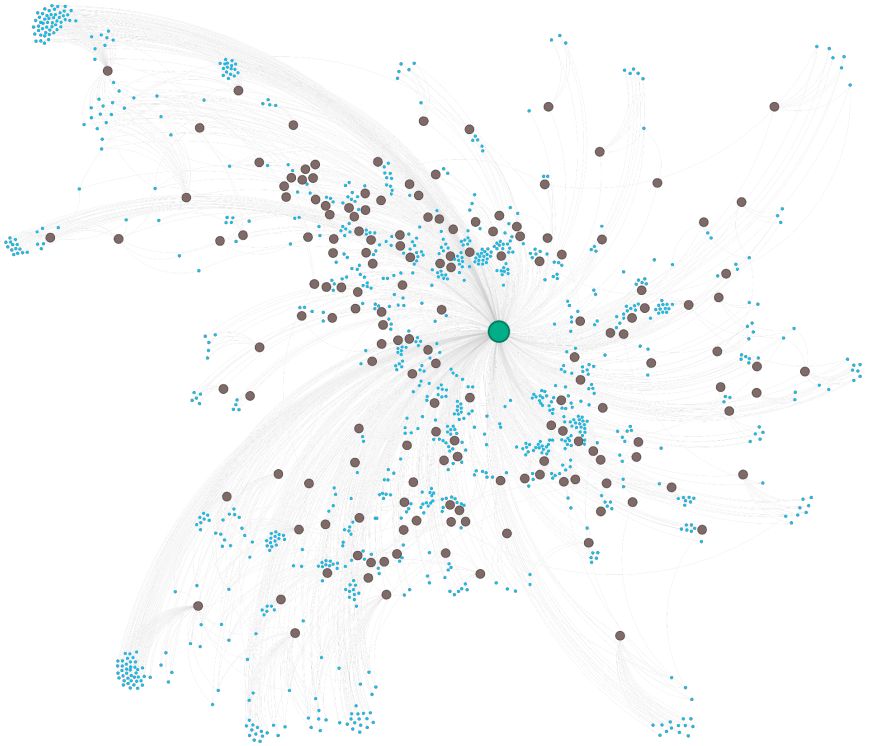
NSF NCAR FAIR Facilities and Instruments Workshop
September 23, 2025

JGI Entity Graph

Facility *n=1*

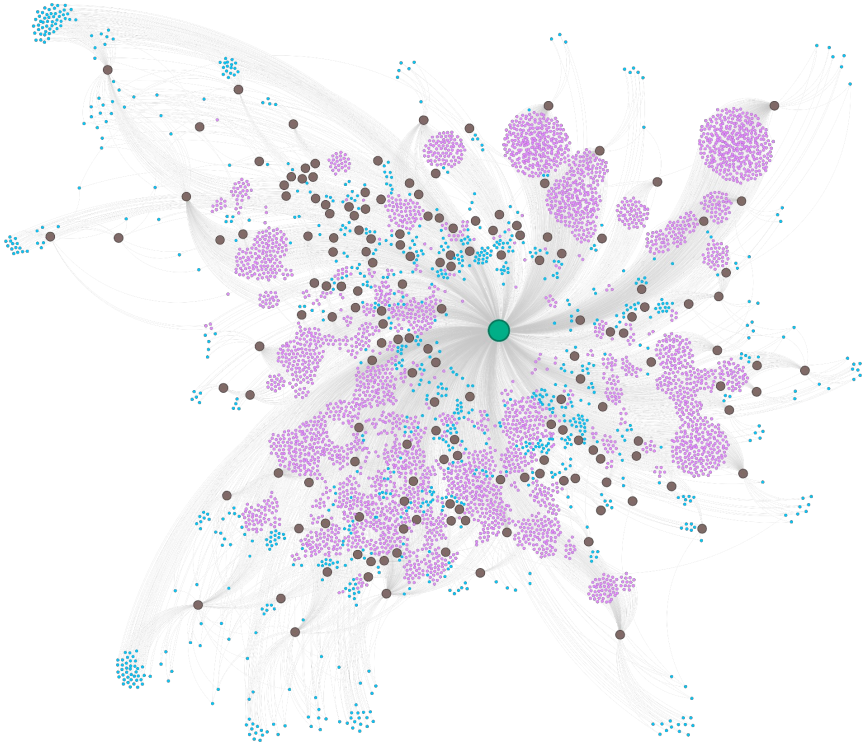
JGI Entity Graph

- **Facility** *n=1*
- **Proposal** *n=163*
- **Researcher** *n=993*



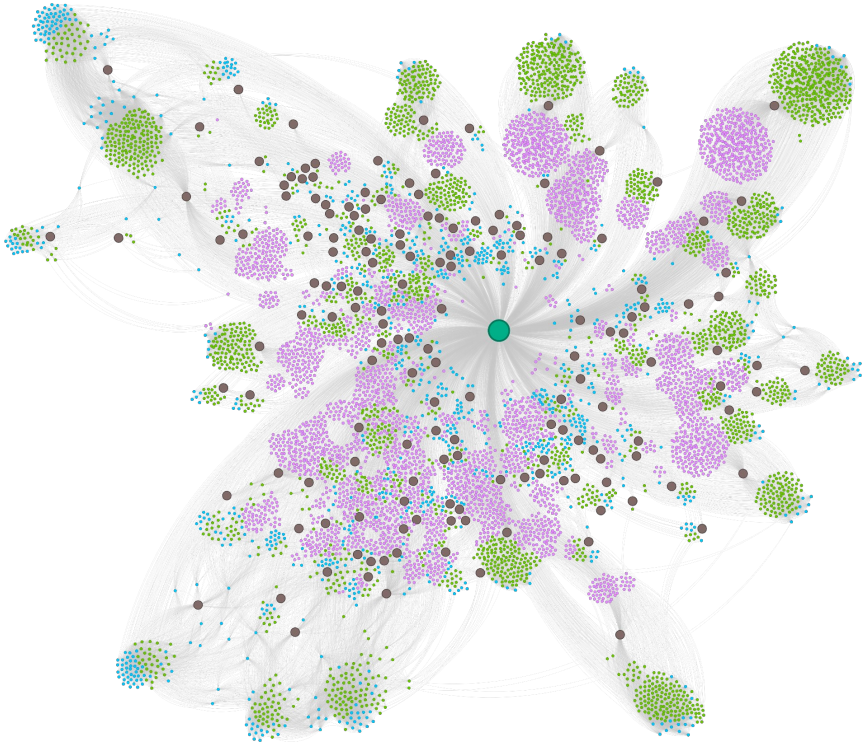
JGI Entity Graph

- **Facility** *n=1*
- **Proposal** *n=163*
- **Researcher** *n=993*
- **Sample** *n=4,097*



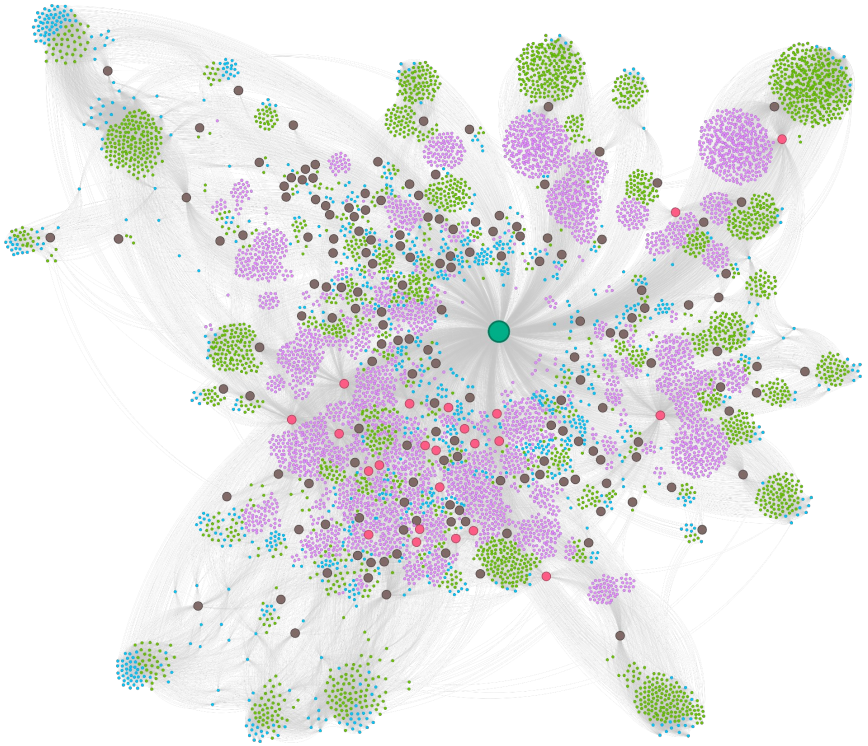
JGI Entity Graph

- **Facility** *n=1*
- **Proposal** *n=163*
- **Researcher** *n=993*
- **Sample** *n=4,097*
- **Dataset** *n=2,970*



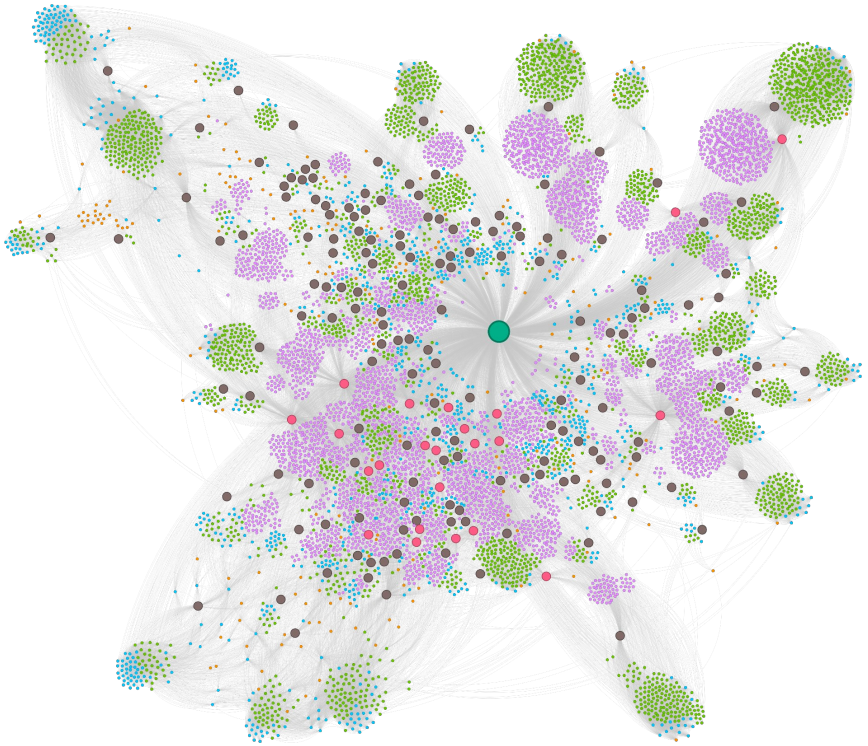
JGI Entity Graph

- Facility *n=1*
- Proposal *n=163*
- Researcher *n=993*
- Sample *n=4,097*
- Dataset *n=2,970*
- Instrument *n=23*

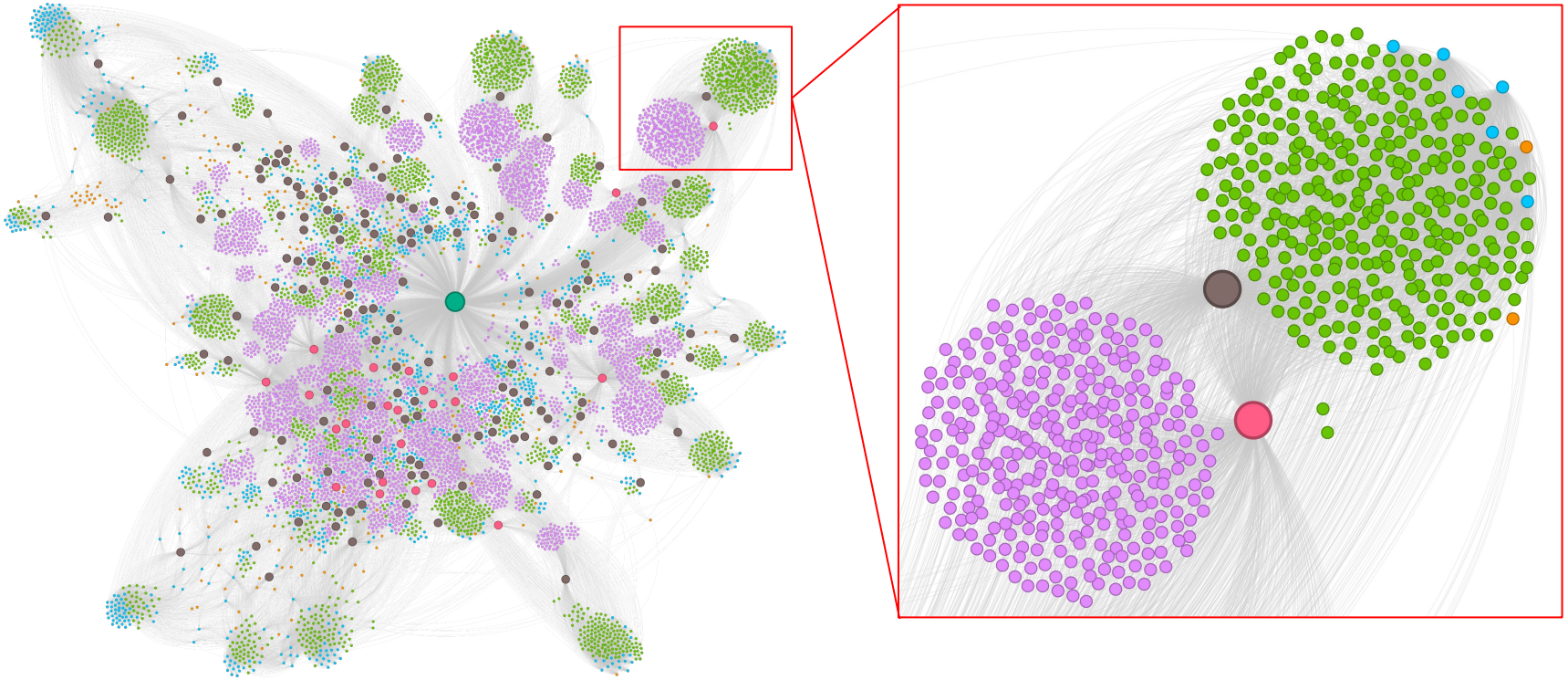


JGI Entity Graph

- Facility *n=1*
- Proposal *n=163*
- Researcher *n=993*
- Sample *n=4,097*
- Dataset *n=2,970*
- Instrument *n=23*
- Publication *n=226*



One Example

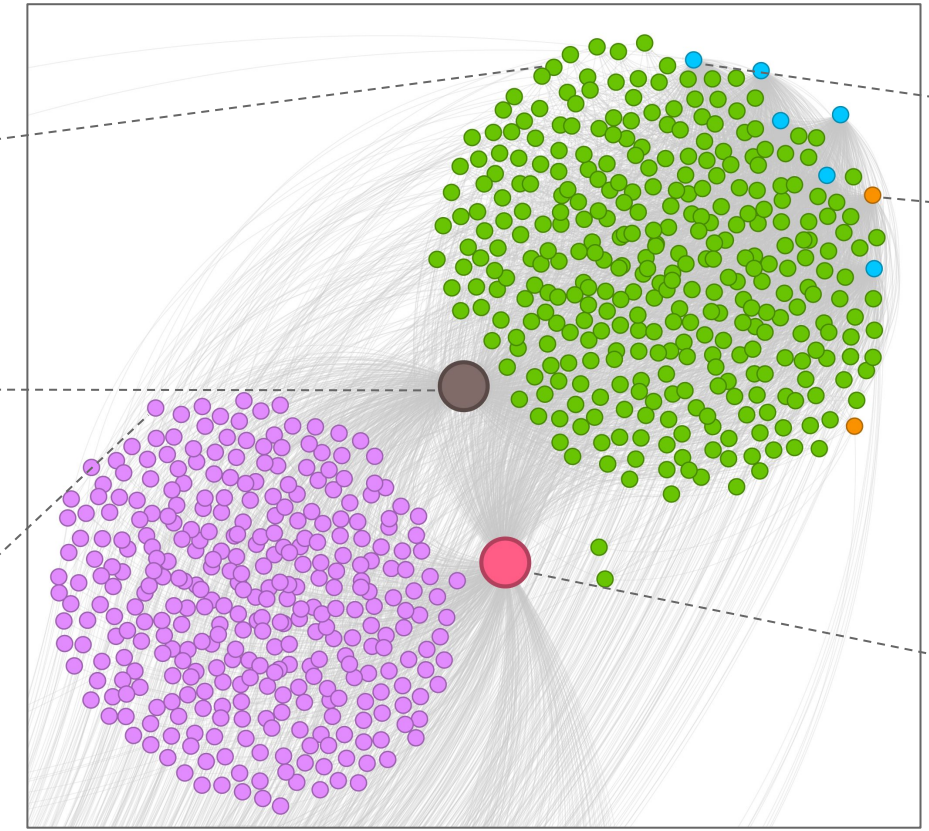


One Example

Dataset:
Enriched cells from coal slurry in the Powder River Basin, Montana, United States
Total cells FG11 rep3
HSBNCT.FG11.300.03.J15
3300034212**

Proposal:
Multi-omic Sequencing of Sulfate Transition Zones in the Terrestrial Subsurface with Recalcitrant Carbon. (2017)
10.46936/10.25585/60000992*

Sample:
Slurry from a subsurface sampler placed in a Montana coal seam
216252***
T0FGP***
Gb0210694***



Co-PI/Author:
M. Fields
0000-0001-9053-1849*

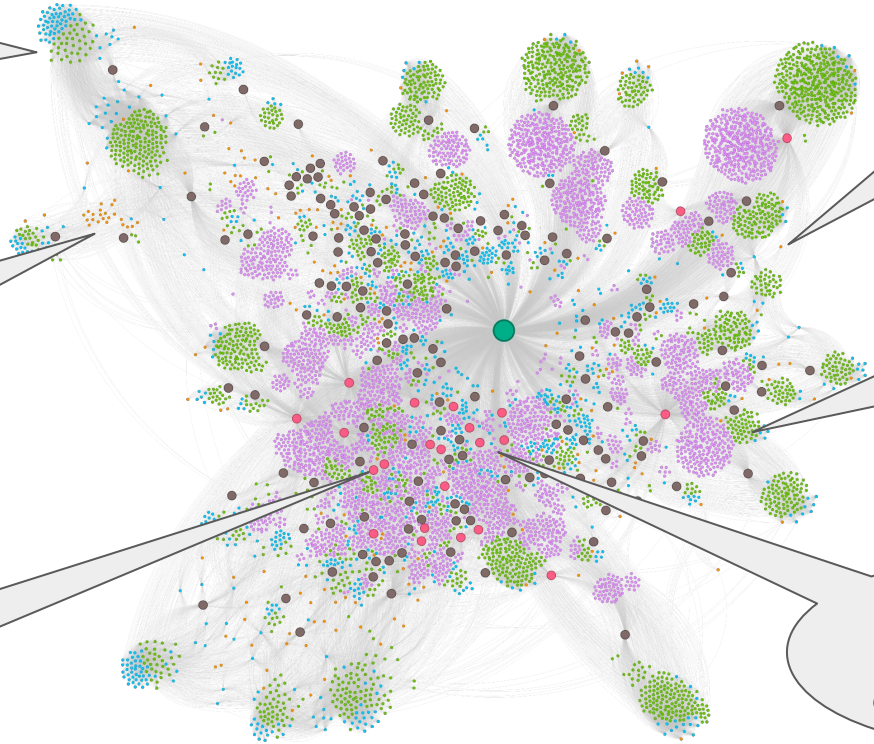
Publication:
10.1038/s41522-022-00267-2*


npj | biofilms and microbiomes
Article | Open access | Published: 17 February 2022
Subsurface hydrocarbon degradation strategies in low- and high-sulfate coal seam communities identified with activity-based metagenomics
Hannah D. Schweitzer, Heidi J. Smith, Elliott P. Barnhart, Luke J. McKay, Robin Gerlach, Alfred B. Cunningham, Rex R. Malmstrom, Danielle Goudeau & Matthew W. Fields

Instrument:
Illumina
NextSeq
NS500302***

*PID **Local ID

What do you want to know?



How are user communities shaped by available product types?

How do users bring together outcomes of multiple proposals for single questions?

Which users are most prolific in publishing on our data and why?

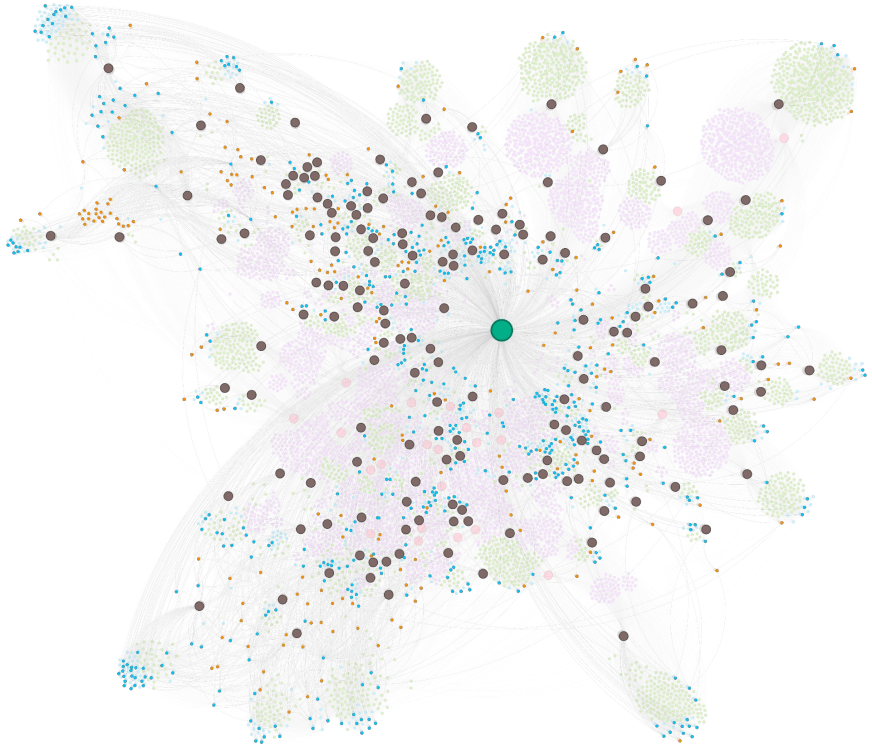
How many JGI datasets were never used to support user findings?

Do users really care about which instruments produced their deliverables?

Are users bringing together different product types from different instruments?

JGI Entity Graph (PIDs)

- Facility
- Proposal
- Researcher**
- ~~Sample~~
- ~~Dataset~~
- ~~Instrument~~
- Publication



JGI Entity Graph (PIDs)

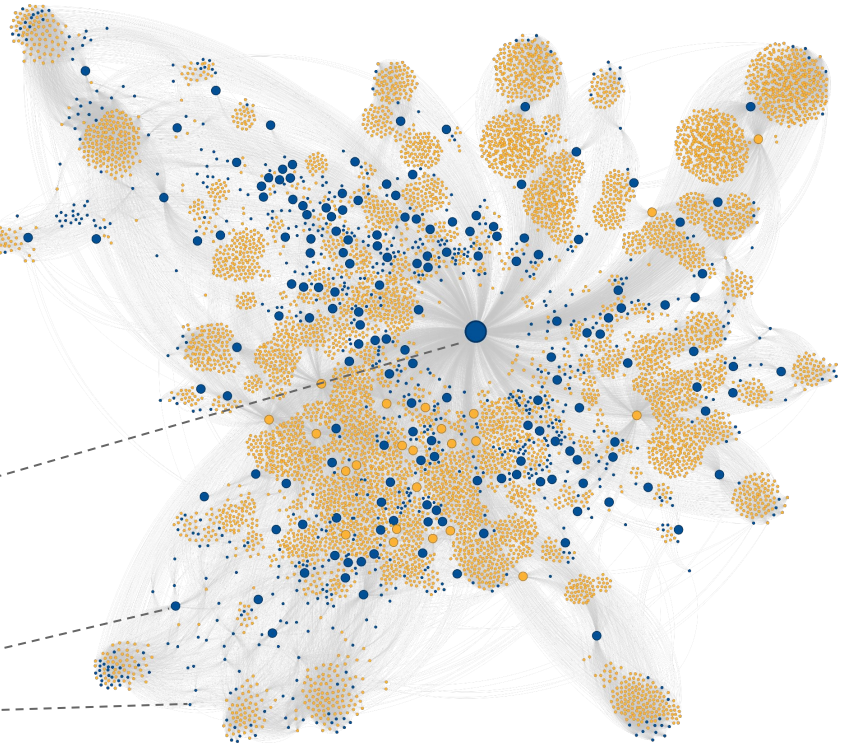
PID

Non-PID

ORCID

ROR

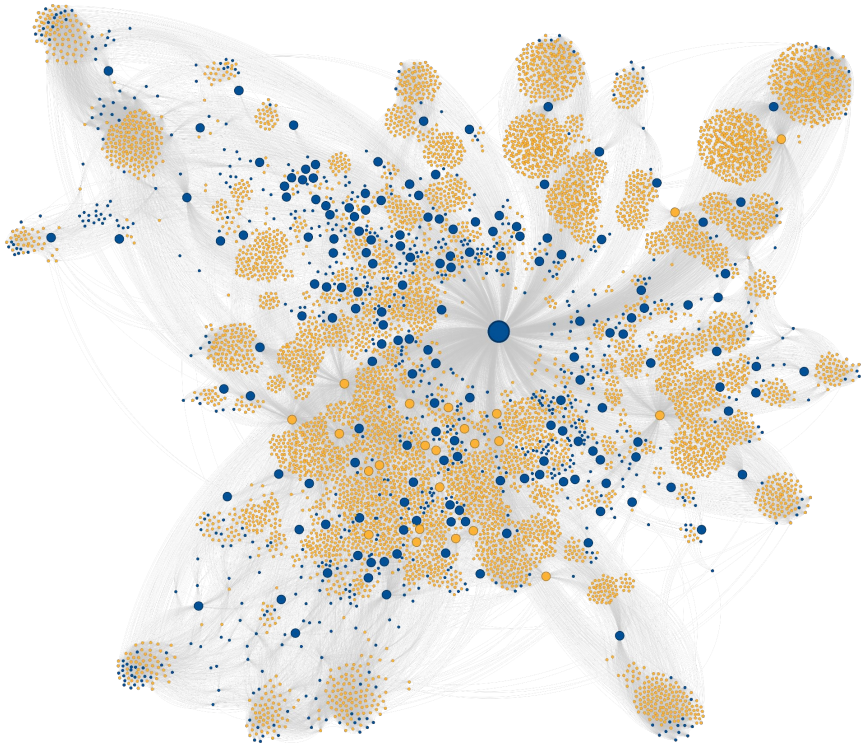
DOI



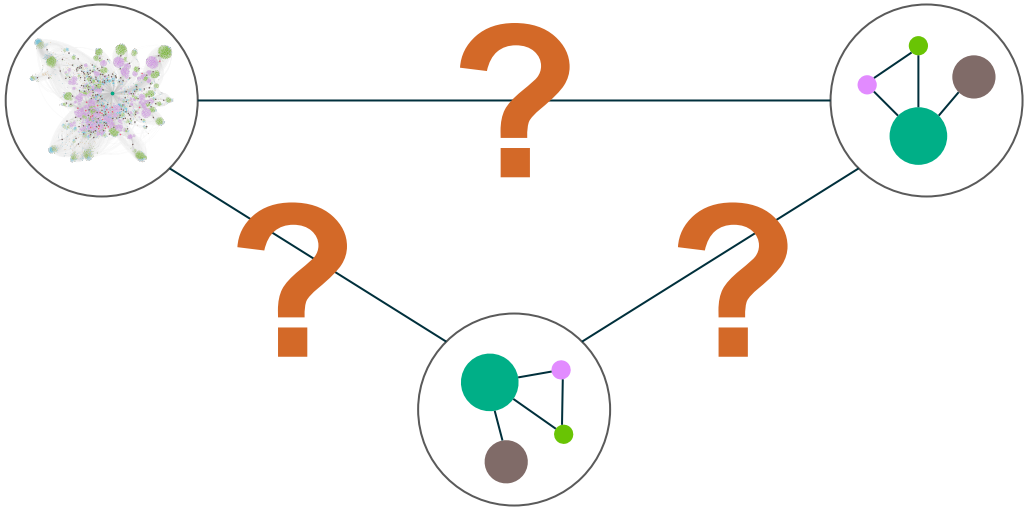
Who Maintains the Graph?

Most of these connections are **JGI-internal**.

For those PIDs that are present, there is **no single external system** establishing linkages between them.



Graph Extensibility & Interoperability



Facility X

Facility Y

Some Citation/Mention Examples

LightCycler480 real-time PCR instrument. Quantified libraries were multiplexed and prepared for sequencing on the Illumina HiSeq 2500 platform utilizing a TruSeq paired-end cluster kit, v4, and Illumina's cBot instrument to generate a clustered flow cell for sequencing. Sequencing was performed using HiSeq TruSeq SBS sequencing kits, v4, following a 2 × 150 indexed run recipe. Raw reads were filtered

Instruments

ACKNOWLEDGMENTS

The work (proposal DOI: https://doi.org/10.46936/10.25585/60001024) conducted by the US Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the US Department of Energy operated under contract no. DE-AC02-05CH11231. This announcement

Facility, Proposal

Genome sequences of key bacterial symbionts of entomopathogenic nematodes: *Xenorhabdus cabanillasii* DSM17905, *Xenorhabdus ehlersii* DSM16337, *Xenorhabdus japonica* DSM16522, *Xenorhabdus koppenhoeferii* DSM18168, and *Xenorhabdus mauleonii* DSM17908

Authors: Raegan Robertson, Katie Conrad, Baarik Ahuja, Markus Göker, Richard L. Hahnke, Alex Spunde, Natalia N. Ivanova, Rekha Seshadri, Craig Stephens | AUTHORS INFO & AFFILIATIONS

Researchers

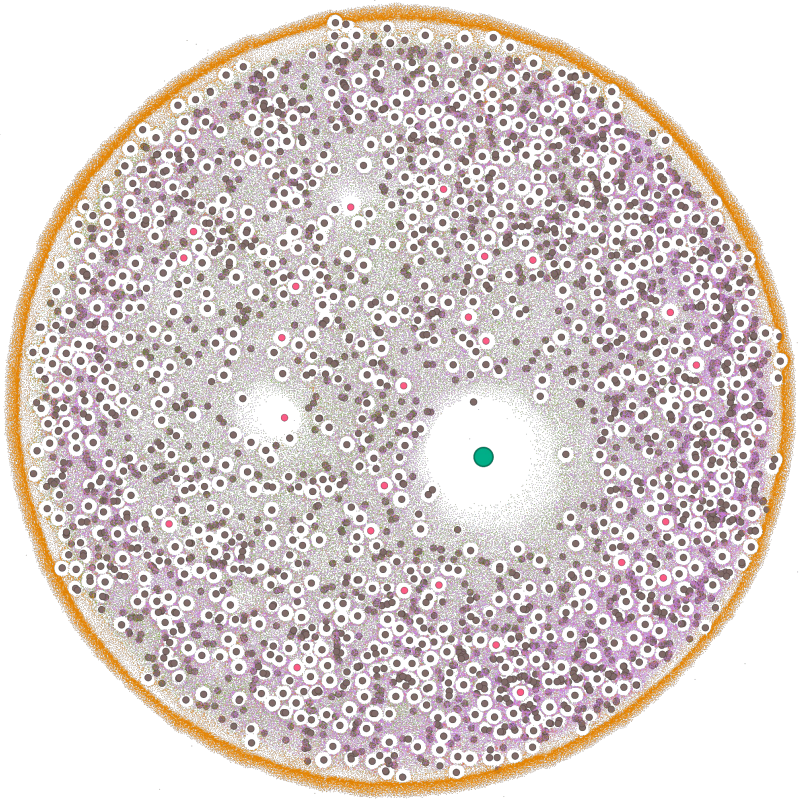
GenBank accession number	NZ_RAQI00000000.1	NZ_QTUB00000000.1	NZ_FOVO00000000.1	NZ_FPBJ00000000.1
NCBI SRA accession number	SRX3886576	SRX3785650	SRX2156714	SRX2156721
JGI IMG/G taxon ID	2778260932	2772190835	2684622846	2684622845

Datasets

https://doi.org/10.1128/MRA.00548-23

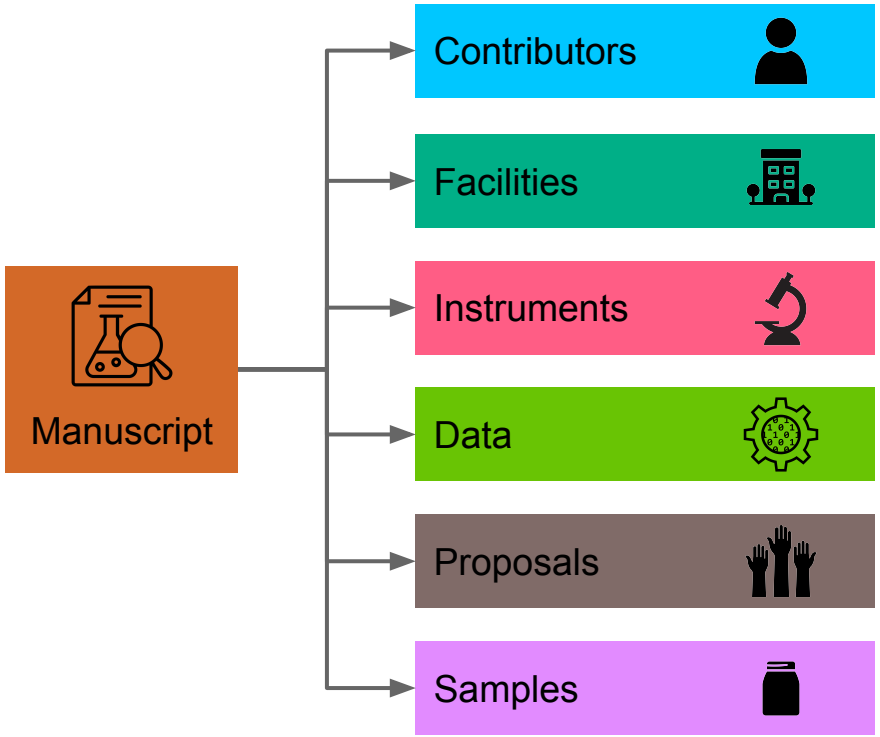
JGI Entity Graph

- **Facility** *n=1*
- **Proposal** *n=2,615*
- **Researcher** *n=4,858*
- **Sample** *n=361,740*
- **Dataset** *n=168,363*
- **Instrument** *n=25*
- **Publication** *n=232,379*

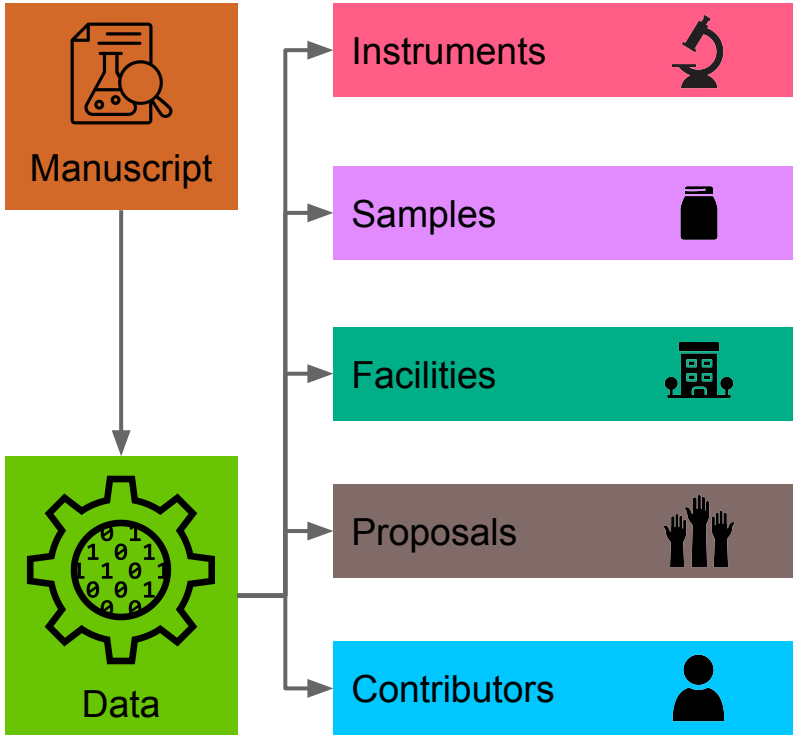


How to cite new PIDs?

Model A: Manuscript → All Entities



Model B: Manuscript → Data → All Other Entities



Parting Thoughts

- **What questions do we want to answer?**
  - Can each of us articulate research questions worth building the infrastructure required to answer them?
- **Where/how do PIDs fit into our graphs?**
  - Are PIDs worth the lift for this use case?
  - Other purposes beyond external connections?
- **Are graph structures generalizable?**
  - Do the connections at my organization make sense for yours?
- **Who maintains a multi-organization graph?**
  - Do we want a big third party graph to which everything is exported?
- **How do we want PIDs for instruments/facilities treated in the literature?**
  - Do we want people citing instruments/facilities directly?
  - Are solid data citations enough?

BioSciences

BERKELEY LAB

U.S. DEPARTMENT of ENERGY | Office of Science

Thank You!