JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

The Genome Citation Service:

Capturing JGI data citations for comprehensive impact assessment

Neil Byers

Scientific Impact Analyst

JGI: A DOE SC User Facility

JGI JOINT GENOME INSTITUTE

- Part of the **Lawrence Berkeley National Laboratory**
- Located in **Berkeley, California**
- **~400 Staff**, including graduate students and postdocs
- **2,000+** Active PIs/collaborators per year
- **10,000+** Online users per year

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Bringing Science Solutions to the World

BioSciences

UNIVERSITY OF CALIFORNIA

JGI User Programs

SCIENTIFIC PEER REVIEW

40%
Community Science Program (CSP)

10%
Facilities Integrating Collaborations for User Science (FICUS)

30%
Bioenergy Research Centers (BRC)

10%
Director's Science

10%
Biological and Environmental Research Support Science (BERSS)

Plant

Fungal, Algal

Metagenome

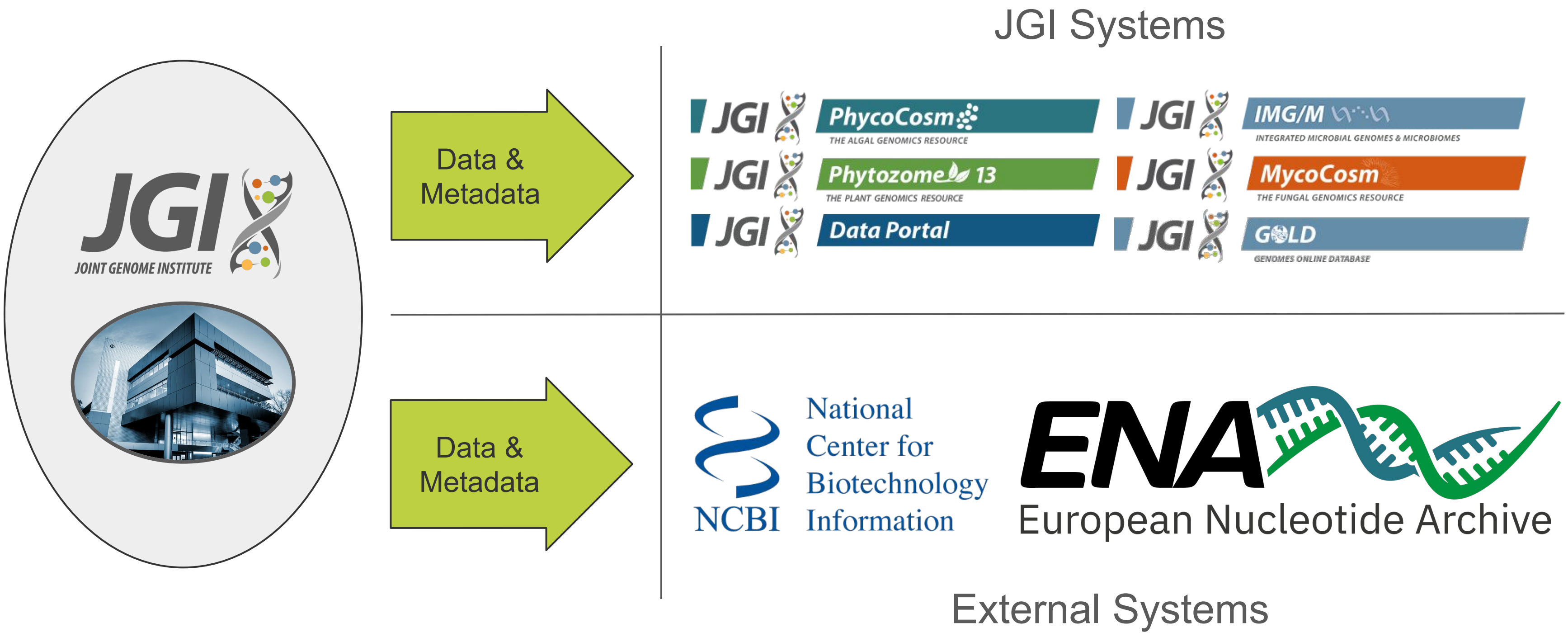
Microbial

DNA Synthesis

Metabolomics

2^{o} Metabolites

JGI Public Resources



JGI Systems

Data & Metadata

Data & Metadata

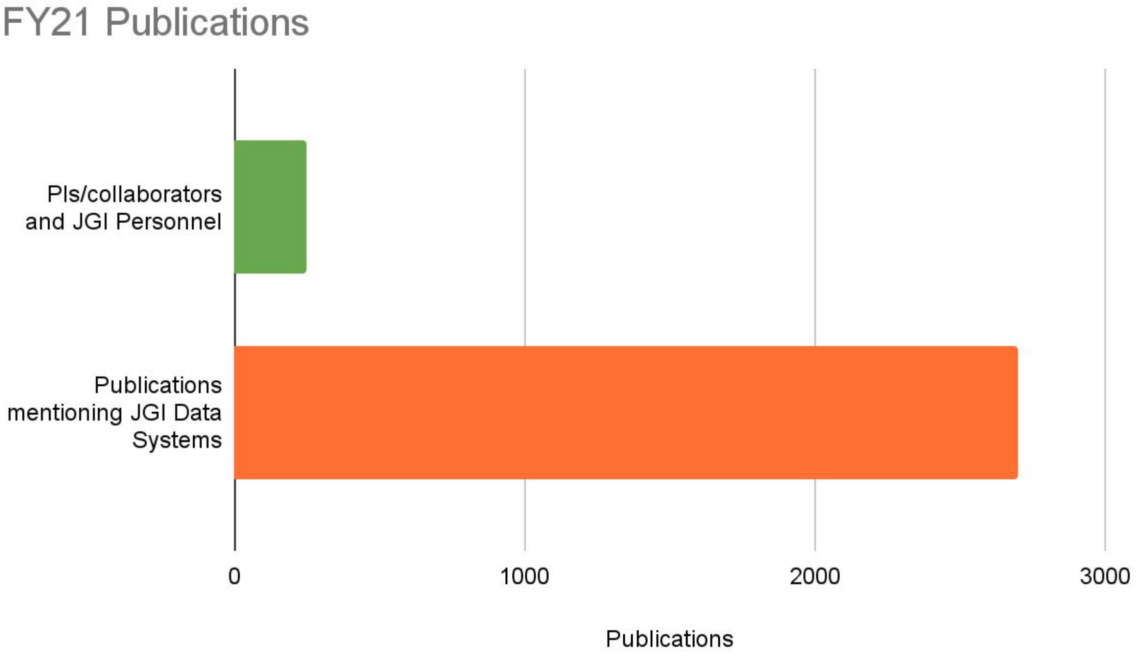
External Systems

JGI
JOINT GENOME INSTITUTE
JGI PhycoCosm
THE ALGAL GENOMICS RESOURCE
JGI Phytozome 13
THE PLANT GENOMICS RESOURCE
JGI Data Portal
JGI IMG/M
INTEGRATED MICROBIAL GENOMES & MICROBIOMES
JGI MycoCosm
THE FUNGAL GENOMICS RESOURCE
JGI GOLD
GENOMES ONLINE DATABASE
NCBI
National Center for Biotechnology Information
ENA
European Nucleotide Archive

JGI User Groups

How do JGI user groups compare?

FY21 Publications



Raw data source:

Dimensions

The Problem

Scalability: Publications, Data, & Metadata

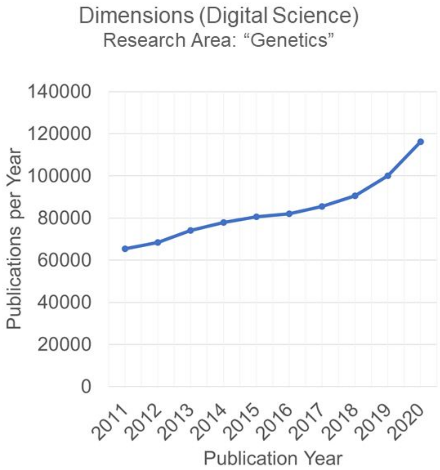
- Large amounts of data
- Rich and diverse metadata
- Growing body of literature

=

– Too many citations –
– Using too many means –
– Of too many products –

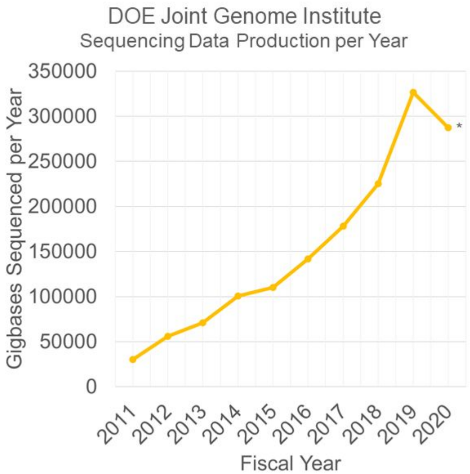
—

To identify manually



Net increase in yearly output: 2011-2020: 77.53%

Publication total 2011-2020: 840,780



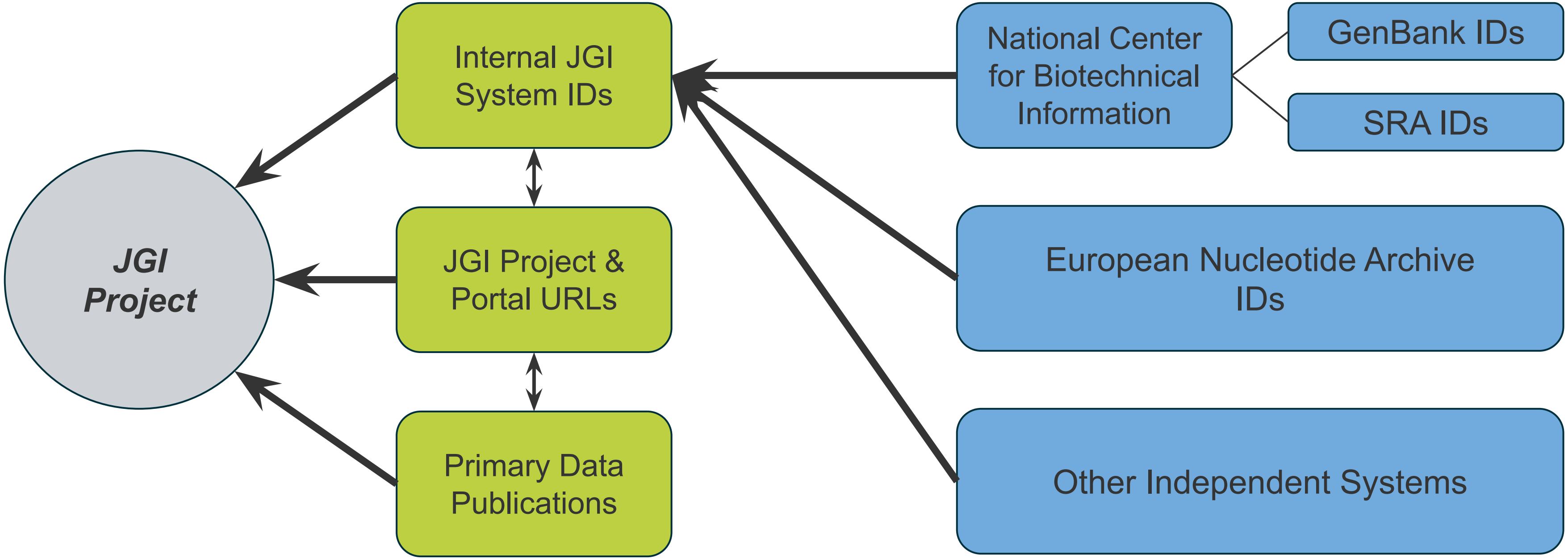
*FY2020 sequencing output affected by the COVID-19 pandemic.

Raw data source:

Dimensions

The Problem

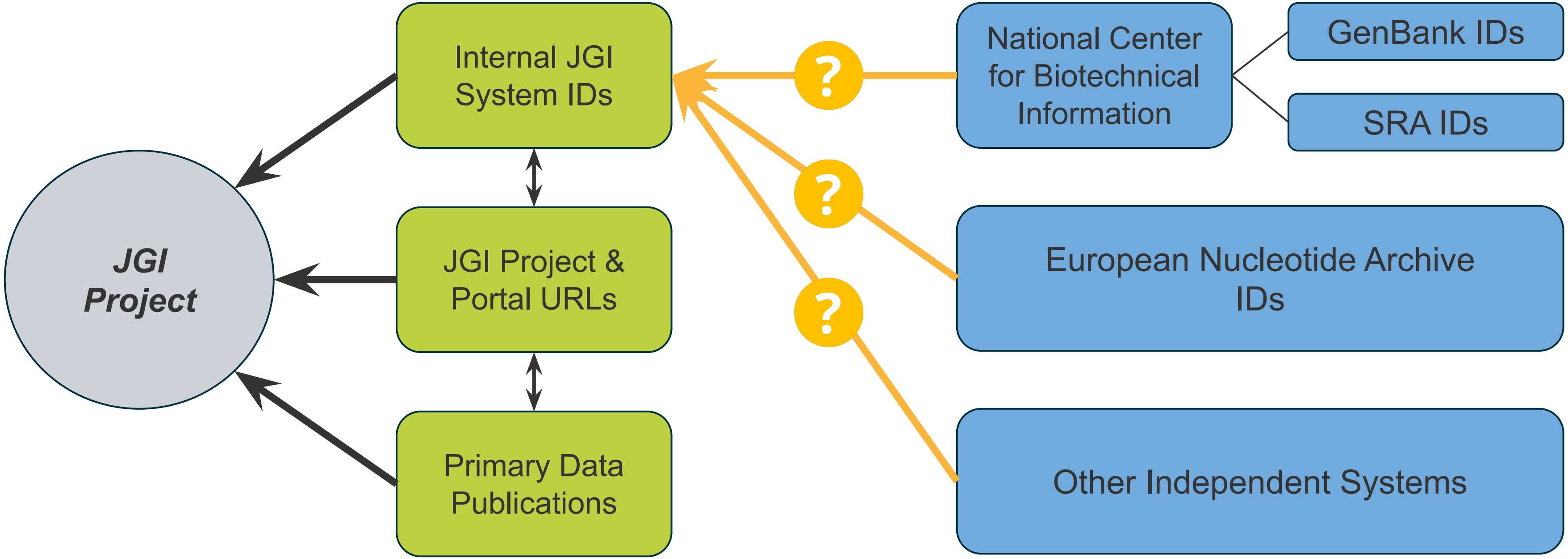
Incomplete Metadata



JGI Project
Internal JGI System IDs
JGI Project & Portal URLs
Primary Data Publications
National Center for Biotechnical Information
GenBank IDs
SRA IDs
European Nucleotide Archive IDs
Other Independent Systems

The Problem

Is a cited identifier linked to JGI? What project(s) is it linked to?


Internal JGI System IDs
JGI Project
JGI Project & Portal URLs
Primary Data Publications
?
?
?
National Center for Biotechnical Information
GenBank IDs
SRA IDs
European Nucleotide Archive IDs
Other Independent Systems

The Solution

Automation: The Genome Citation Service (GCS)

Problem:
Incomplete Metadata

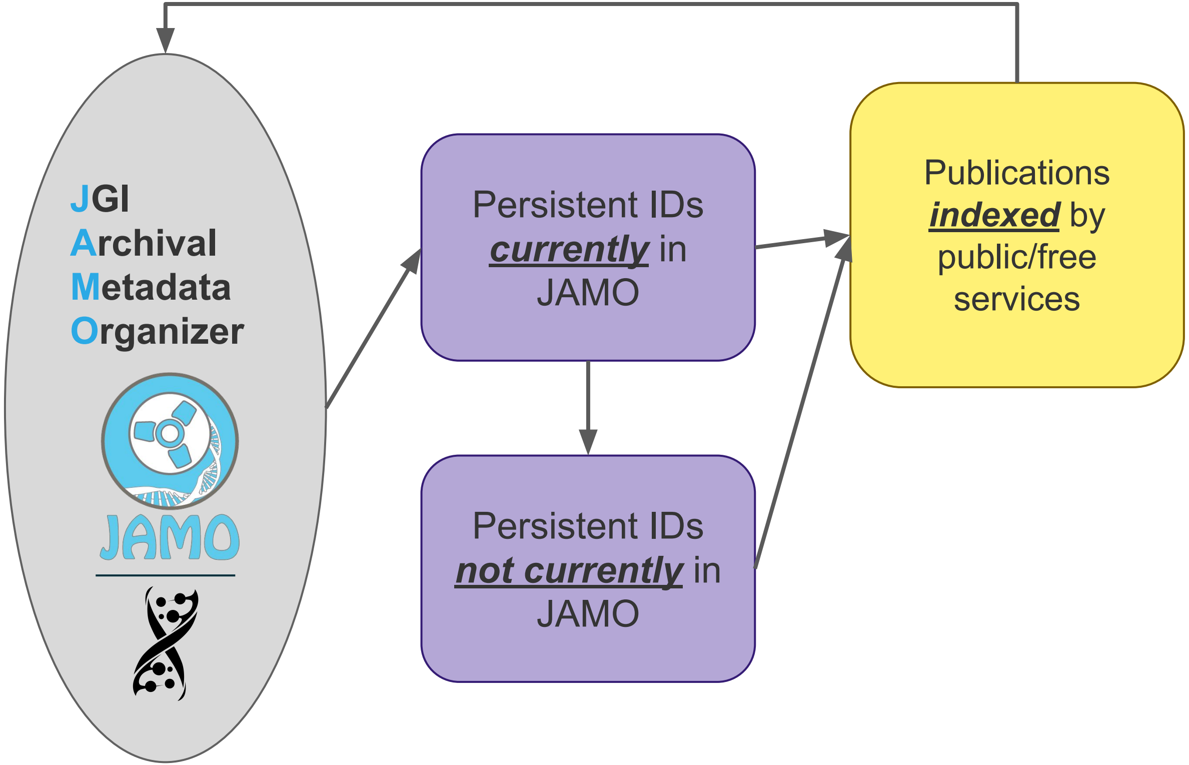
Solution:
Automatic, incremental discovery and traversal of additional linked data resources

Problem:
Scalability

Solution:
Automatic queries of public literature and cataloging of pathways back to JGI data resources

The Genome Citation Service

NamesforLife
Bringing meaning to life...



JGI
Archival
Metadata
Organizer

JAMO

Persistent IDs ***currently*** in JAMO

Persistent IDs ***not currently*** in JAMO

Publications ***indexed*** by public/free services

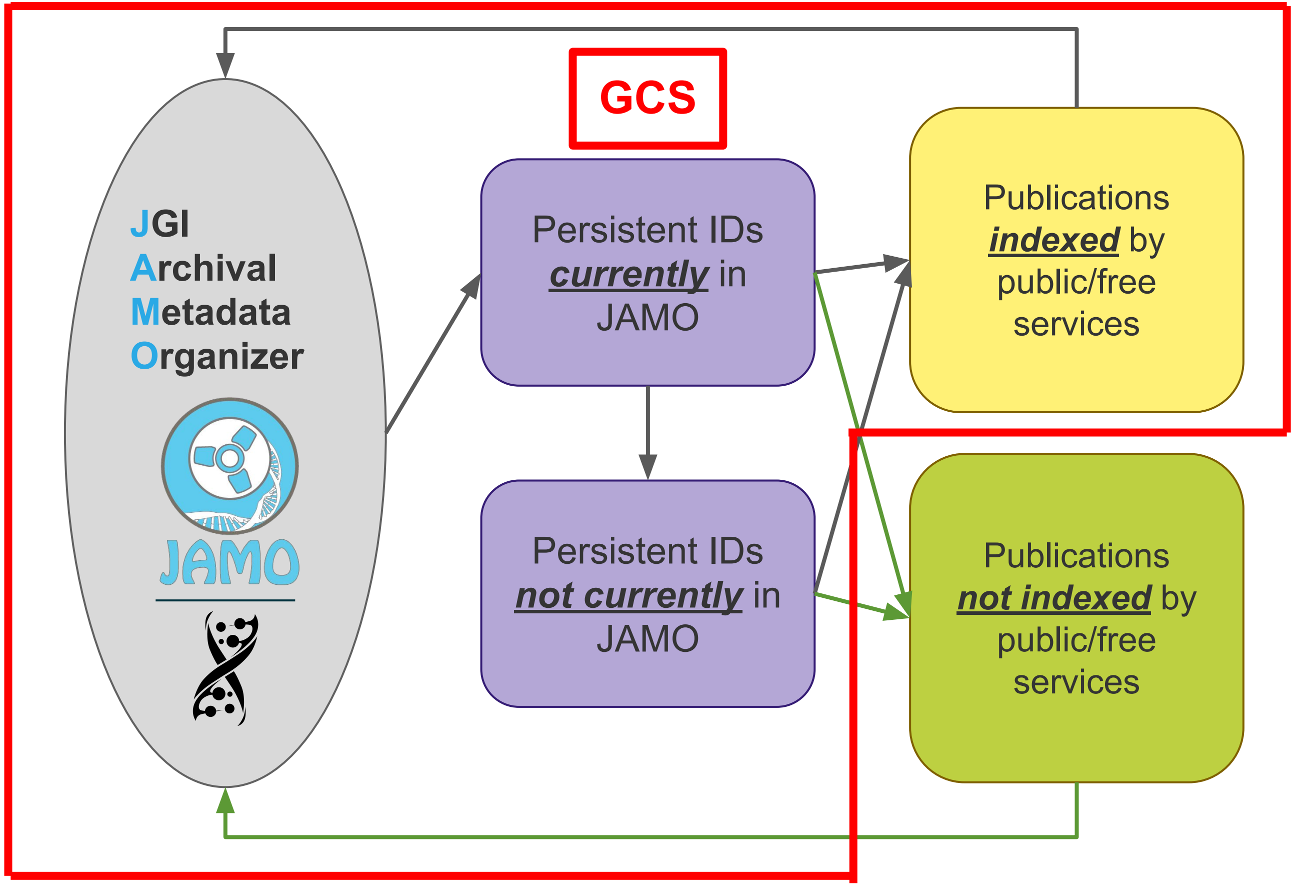
PubMed

PubMed Central

The Genome Citation Service

JGI JOINT GENOME INSTITUTE

NamesforLife
Bringing meaning to life...


GCS
JGI
Archival
Metadata
Organizer
JAMO
Persistent IDs *currently* in JAMO
Persistent IDs *not currently* in JAMO
Publications *indexed* by public/free services
Publications *not indexed* by public/free services

PubMed

PubMed Central

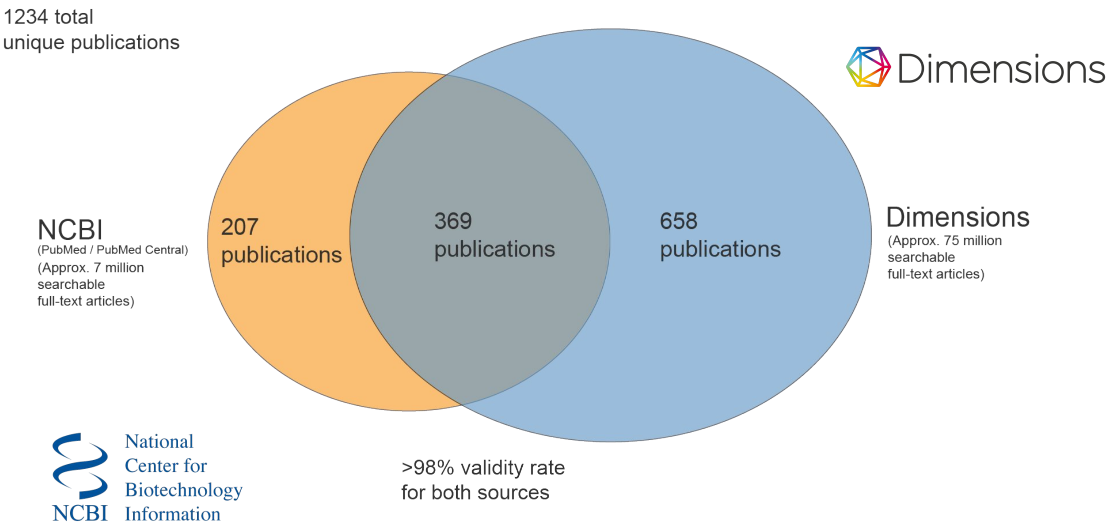
Dimensions

Evaluating the GCS

How well did the GCS perform?

- **Methodology**
  - 300 JAMO records (stratified, random)
  - Search public and subscription sources for linked publications
  - Manually evaluate hits
- **Results**
  - >98% Precision
  - 1234 total publications
  - 90% previously unidentified

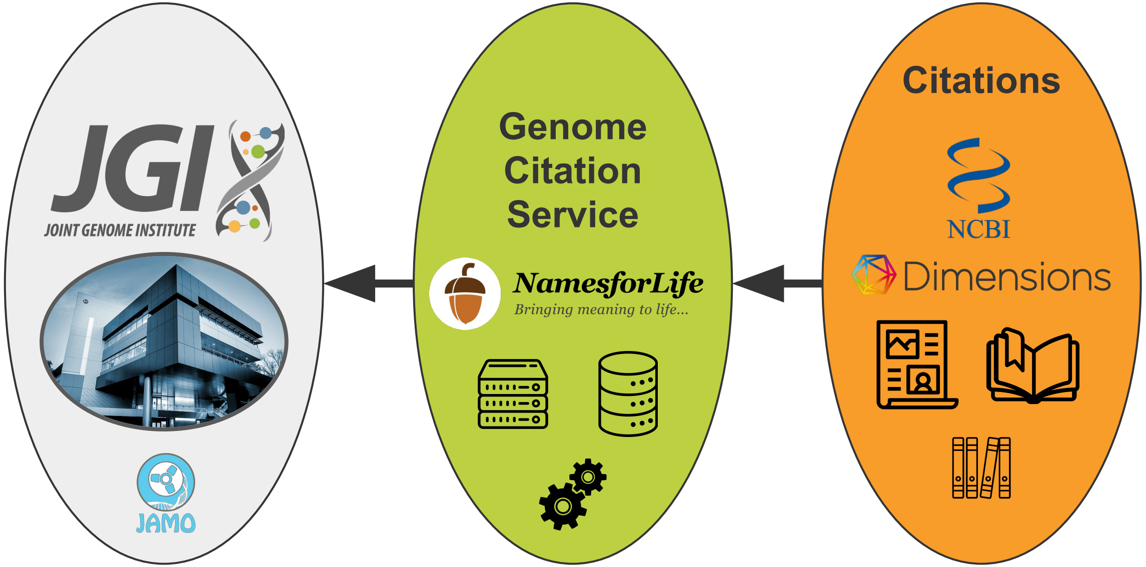
Publications identified through the NamesforLife / JGI "Genome Citation Service" using two distinct publication databases and a stratified random sample of 300 JAMO records



GCS Results: Implications

How does JGI benefit from the Genome Citation Service?

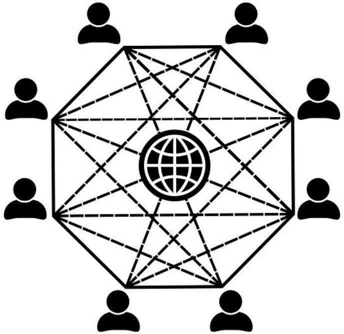
1. Thousands of JGI data citations now potentially identifiable
2. Little or no manual effort required

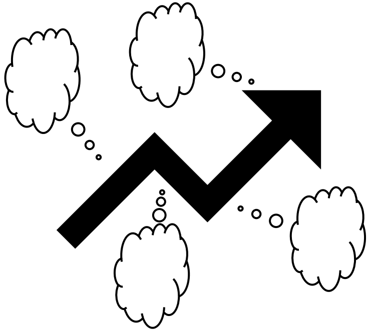


Data Citations for Impact Assessment

How can we use this information for impact assessment?

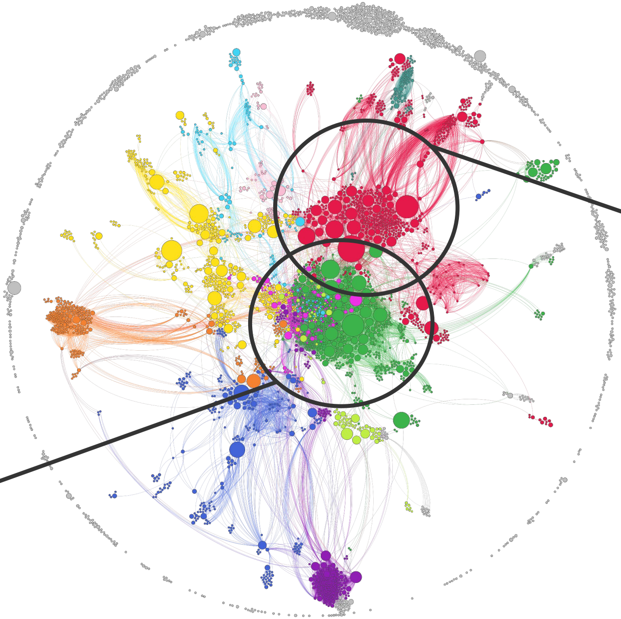
- **Community Analysis**
  - Coauthor and JGI user networks
- **Topic Analysis**
  - Data use trends
  - DOE goal alignment
- **New researcher metrics**
  - Equitable credit attribution





Community Analysis

Which PI/Collaborator groups are producing data of scientific interest?



JGI Proposal Contributors

Group A

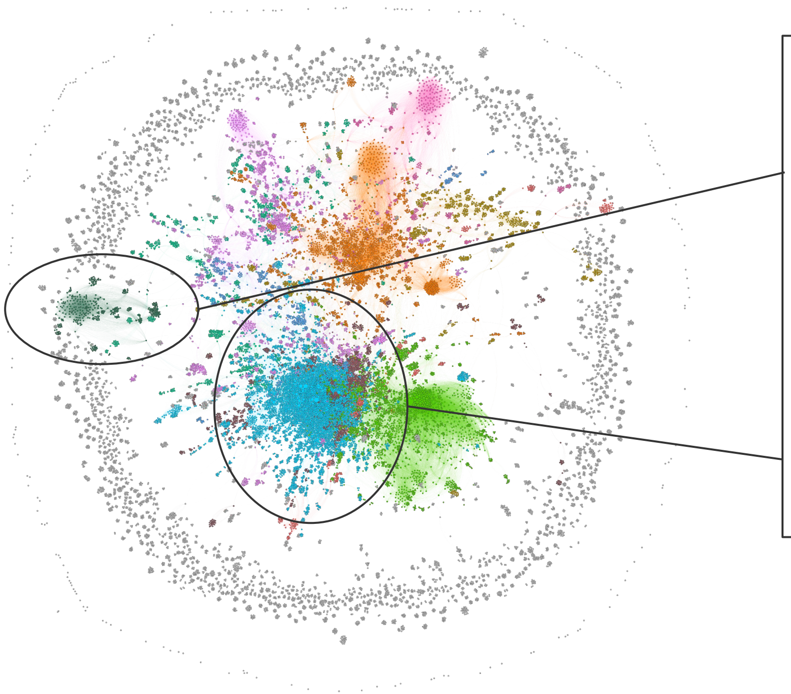
- X proposals with
- Y datasets receiving
- Z total data citations
- Consists largely of DOE Bioenergy Research Center (BRC) personnel

Group B

- X proposals with
- Y datasets receiving
- Z total data citations
- Contains many contributors to JGI's 1000 Fungal Genomes Project

Community Analysis

Can we identify distinct communities among users of public JGI resources?



Blue Group* vs. Teal Group

- Total Authors:
  - ○ 5015
  - ○ 378
- Top Institutions:
  - ○ DOE Labs, American & European Universities
  - ○ European universities
- Top concepts:
  - ○ Evolution, Pathogens, Biomass degradation
  - ○ Bioinformatics, Ontology
- Publications
  - ○ 1774 publications (~18 average citations)
  - ○ 123 publications (~30 average citations)

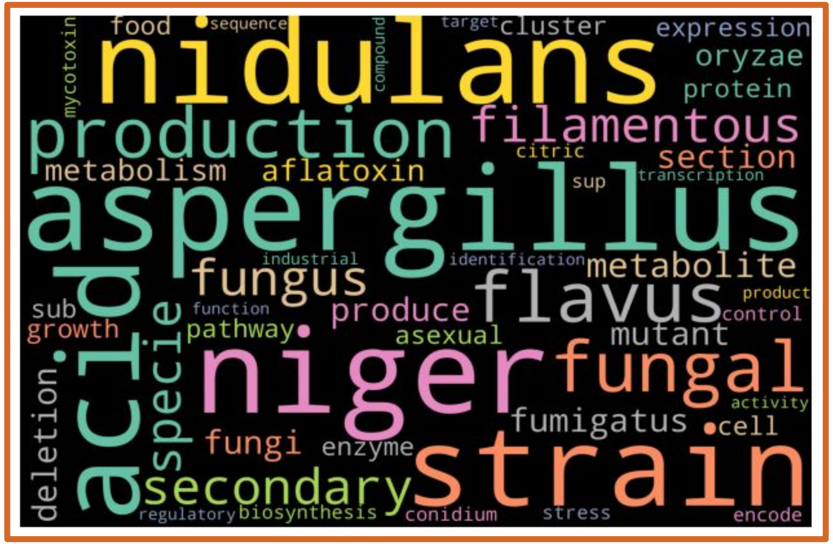
JGI Users
(mock subset)

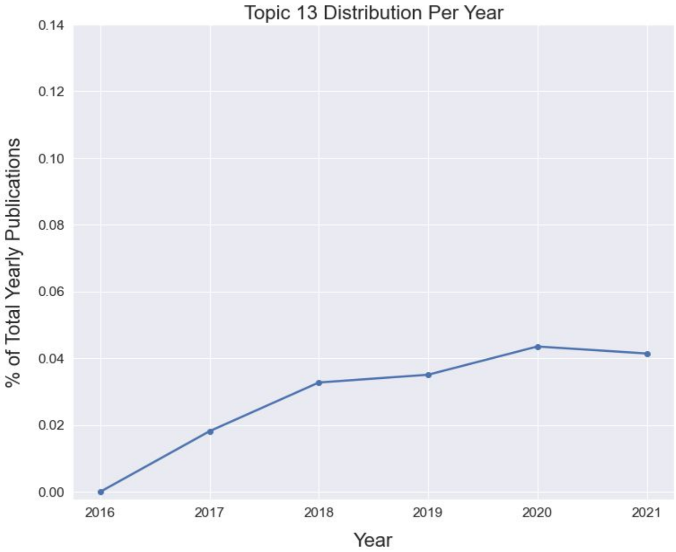
Raw data source:
Dimensions

*Top authors in Blue Group are JGI PIs & Collaborators

Topic & Concept Analysis

What 'topics' are trending amongst our user base?

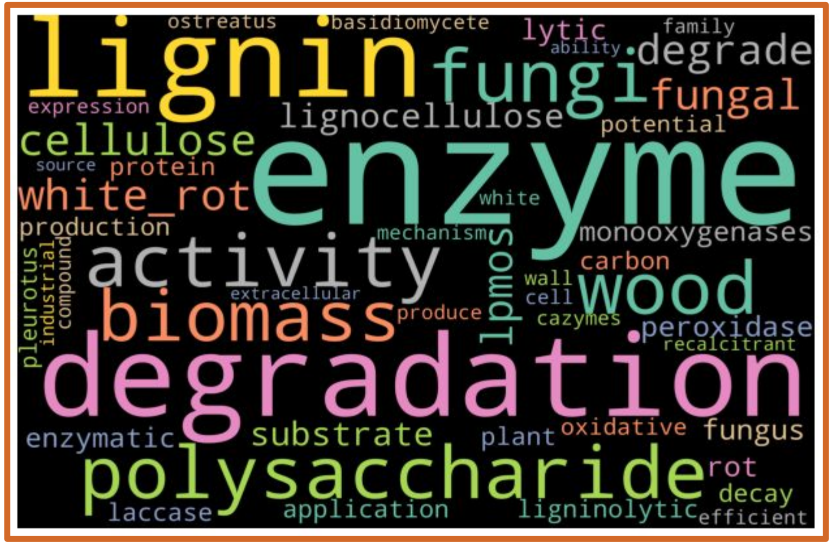


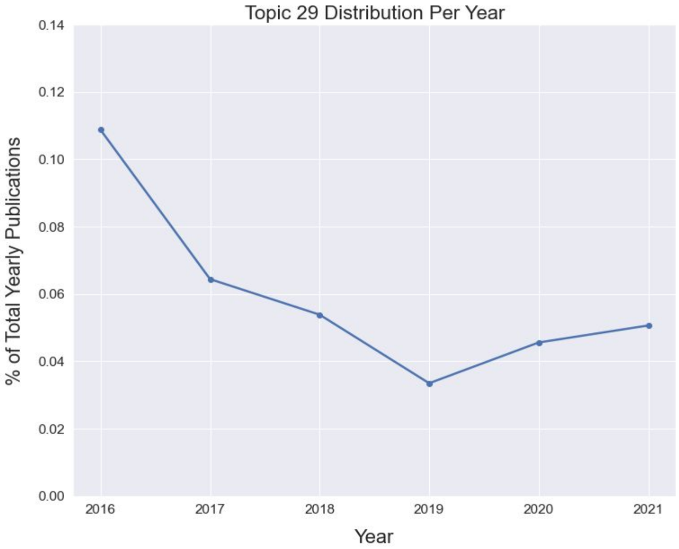


This 'topic' seems to be on the rise among citing publications. It consists of 235 publications from our test set, and is increasing in total publication share over time.

Topic & Concept Analysis

What 'topics' are trending amongst our user base?



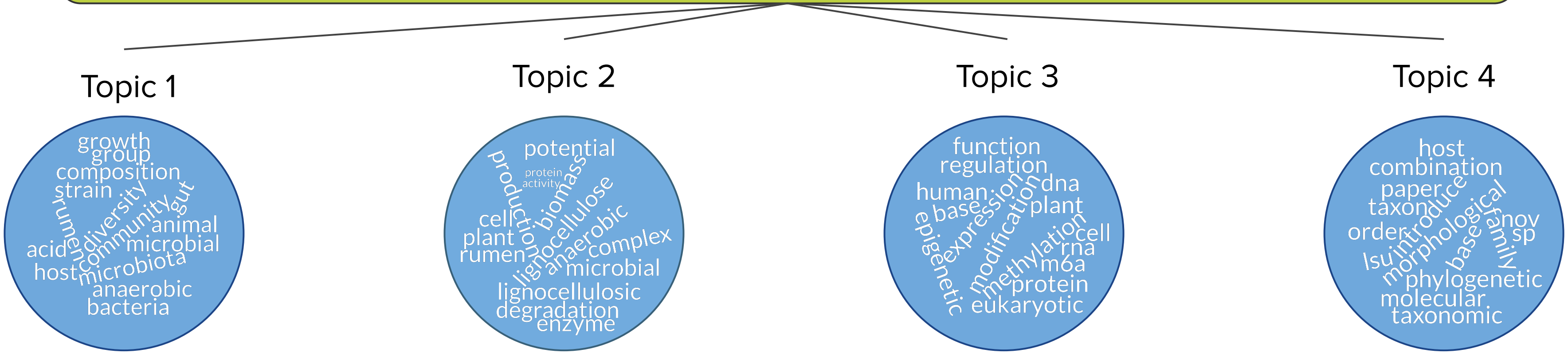


This 'topic' seems to be losing prominence among citing publications. It consists of 341 publications from our test set, and is decreasing in total publication share over time.

Topic & Concept Analysis

Are there counter-intuitive use cases for JGI data?

JGI Proposal:
"Transcriptomic Characterization of Anaerobic Gut Fungi", 2014
549 Downstream Citations



Topic 1
- **109 (20%) Publications**
- 1.8k citations
- Top Journals: *Frontiers in Microbiology; Microorganisms*

Topic 2
- **137 (25%) Publications**
- 2.3k citations
- Top Journals: *Biotechnology for Biofuels; Bioresource Technology*

Topic 3
- **96 (17%) Publications**
- 2.2k citations
- Top Journals: *Nucleic Acids Research; Genome Biology; Scientific Reports*

Topic 4
- **154 (28%) Publications**
- 3.4k citations
- Top Journals: *Fungal Diversity; Mycological Progress*

Topic & Concept Analysis

Are there counter-intuitive use cases for JGI data?

JGI Proposal:

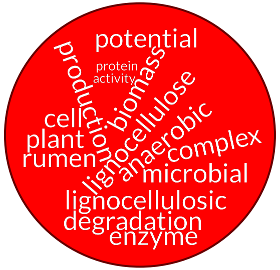
"Transcriptomic Characterization of Anaerobic Gut Fungi", 2014

549 Downstream Citations

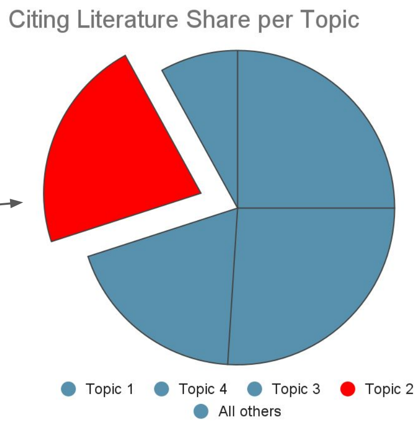
Proposal Text:

"The overall goal of our efforts is to develop the tools to engineer anaerobic gut fungi as novel platform organisms for ***biofuel production*** *from* ***lignocellulosic biomass****."*

Topic 2

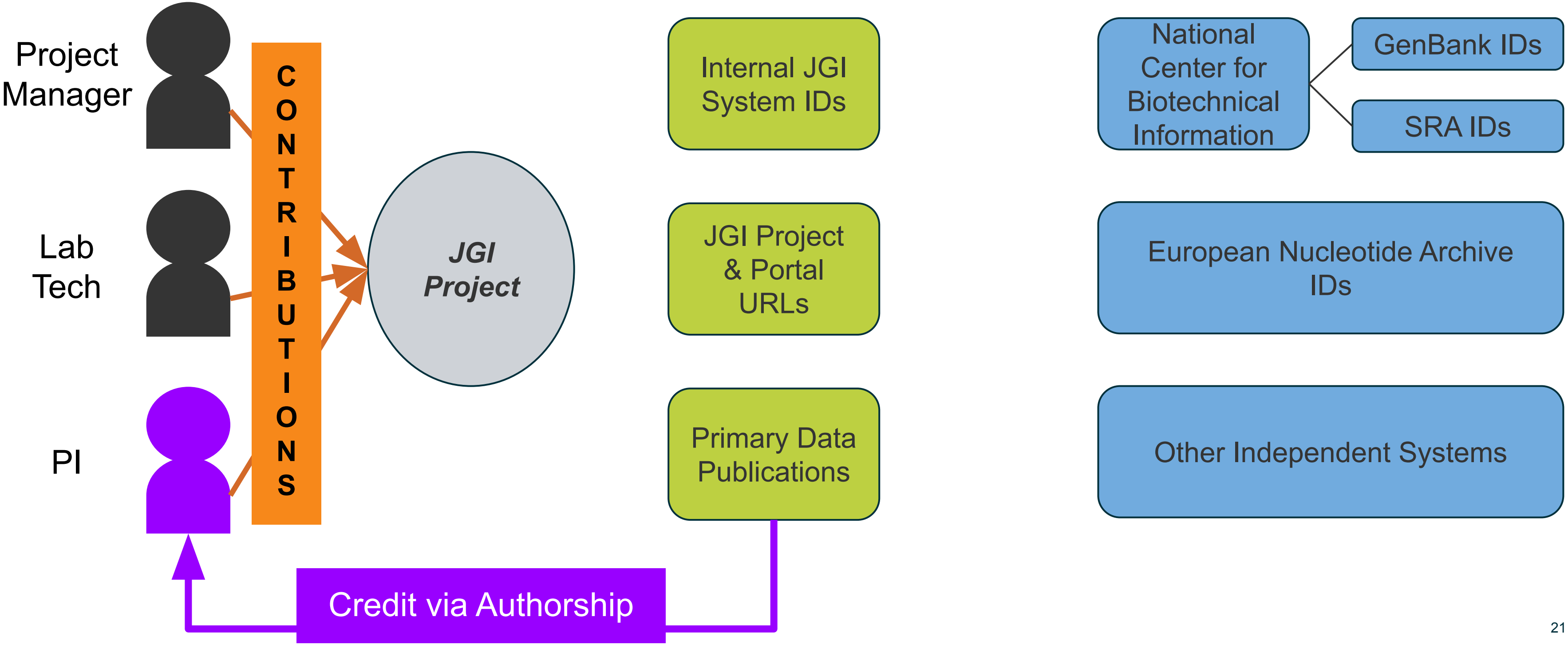


- 137 (25%) Publications
- 2.3k citations
- Top Journals: *Biotechnology for Biofuels; Bioresource Technology*



Equitable Contributor Metrics

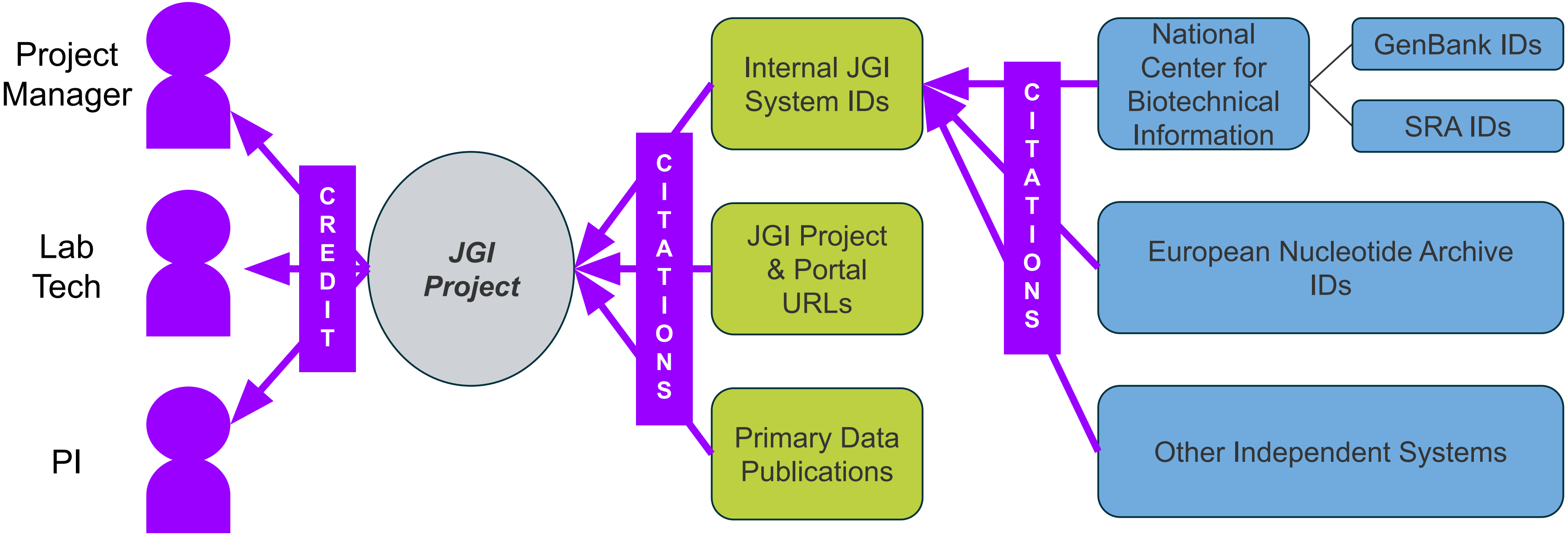
Authorship Model - Often arbitrary and uneven credit attribution



Project Manager
Lab Tech
PI
CONTRIBUTIONS
JGI Project
Internal JGI System IDs
JGI Project & Portal URLs
Primary Data Publications
Credit via Authorship
National Center for Biotechnical Information
GenBank IDs
SRA IDs
European Nucleotide Archive IDs
Other Independent Systems

Equitable Contributor Metrics

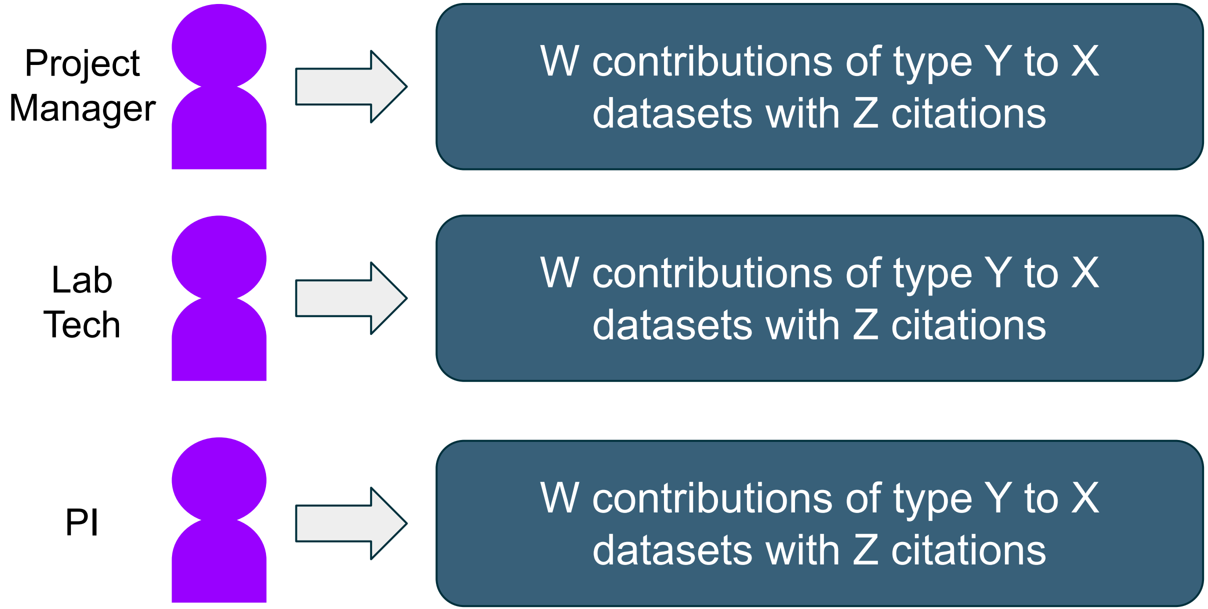
Data Citation Model - Equitable attribution of credit



Project Manager
Lab Tech
PI
CREDIT
JGI Project
CITATIONS
Internal JGI System IDs
JGI Project & Portal URLs
Primary Data Publications
CITATIONS
National Center for Biotechnical Information
GenBank IDs
SRA IDs
European Nucleotide Archive IDs
Other Independent Systems

Equitable Contributor Metrics

Data Citation Model - Equitable attribution of credit



Project Manager

W contributions of type Y to X datasets with Z citations

Lab Tech

W contributions of type Y to X datasets with Z citations

PI

W contributions of type Y to X datasets with Z citations

Example:

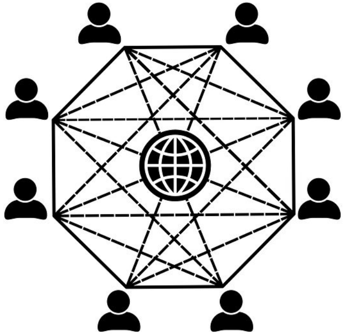
Hood & Sutherland,

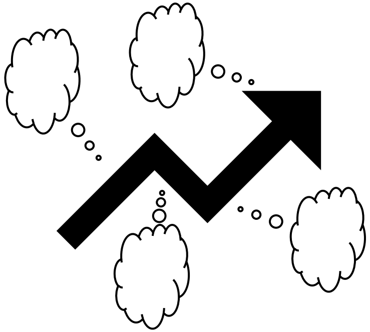
"The data-index: An author-level metric that values impactful data and incentivizes data sharing"

Ecology and Evolution
Oct. 2021

Conclusions

- **GCS Goals**
  - Link metadata and capture data citations at scale
- **Results**
  - Many JGI data citations identified with high validity rates
- **Impact Implications**
  - More comprehensive picture of JGI community impact
  - Equitable contributor metrics





Thank you!

JGI JOINT GENOME INSTITUTE

Genome Citation Service Team:

JGI JOINT GENOME INSTITUTE

Chris Beecroft
Hugh Salamon
Kjiersten Fagnan
TBK Reddy
Neil Byers

NamesforLife *Bringing meaning to life...*

Charles Parker

MICHIGAN STATE UNIVERSITY NamesforLife *Bringing meaning to life...*

George Garrity

Kjiersten Fagnan

Chris Beecroft

Neil Byers

Hugh Salamon

TBK Reddy

Charles Parker

George Garrity

Contact: Neil Byers, npbyers@lbl.gov