LF AI & Data Technical Advisory Council (TAC)

Biweekly call - March 7, 2024

⊓LF AI & DATA

Antitrust Policy

› Linux Foundation meetings involve participation by industry competitors, and it is the intention of the Linux Foundation to conduct all of its activities in accordance with applicable antitrust and competition laws. It is therefore extremely important that attendees adhere to meeting agendas, and be aware of, and not participate in, any activities that are prohibited under applicable US state, federal or foreign antitrust and competition laws.

› Examples of types of actions that are prohibited at Linux Foundation meetings and in connection with Linux Foundation activities are described in the Linux Foundation Antitrust Policy available at http://www.linuxfoundation.org/antitrust-policy. If you have questions about these matters, please contact your company counsel, or if you are a member of the Linux Foundation, feel free to contact Andrew Updegrove of the firm of Gesmer Undergone LLP, which provides legal counsel to the Linux Foundation.

LF AI & DATA

Recording of Calls

Reminder:

TAC calls are recorded and available for viewing on the TAC Wiki

LF AI & DATA

LF AI & Data Useful Links

› Web site: lfaidata.foundation
› Wiki: wiki.lfaidata.foundation
› GitHub: github.com/lfai
› Landscape: https://landscape.lfaidata.foundation
› Mail Lists: https://lists.lfaidata.foundation
› Slack: https://slack.lfaidata.foundation
› Youtube: https://www.youtube.com/channel/UCfasaeqXJBCAJMNO9HcHfbA
› LF AI Logos: https://github.com/lfai/artwork
› PPT Template: https://drive.google.com/file/d/1eiDNJvXCqSZHT4Zk_-czASlz2GTBRZk2/view
› Events: https://lfaidata.foundation/events/
› Events Calendar   https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=12091544
› Event Wiki   https://wiki.lfaidata.foundation/pages/viewpage.action?pageId=10518553

□LF AI & DATA

Agenda

› Roll Call (1 mins)

› Approval of Minutes from previous meeting (2 mins)

› Foundation Updates (5 mins)

› Generative AI Commons Updates - Matt White (15 mins)

› Egeria project updates (30 mins)

› Open Discussion

LF AI & DATA

TAC Voting Members

Note: we still need a few designated backups specified on [wiki](#)

Member Company or Graduated Project	Membership Level or Project Level	Voting Eligibility	Country	TAC Representative	Designated TAC Representative Alternates
4paradigm	Premier	Voting Member	China	Zhongyi Tan	
Microsoft	Premier	Voting Member	USA	Ali Dalloul	
Amazon Web Services	Premier	Voting Member	USA	Brian Granger	Mark Atwood
Ericsson	Premier	Voting Member	Sweden	Rani Yadav-Ranjan	
Huawei	Premier	Voting Member	China	Howard (Huang Zhipeng)	Charlotte (Xiaoman Hu), Leon (Hui Wang)
IBM	Premier	Voting Member	USA	Susan Malaika	Beat Buesser, Alexandre Eichenberger
OPPO	Premier	Voting Member	China	Jimmy (Hongmin Xu)	
SAS	Premier	Voting Member	USA	*Nancy Rausch	Liz McIntosh
ZTE	Premier	Voting Member	China	Wei Meng	Liya Yuan
Adversarial Robustness Toolbox Project	Graduated Technical Project	Voting Member	USA	Beat Buesser	Kevin Eykholt
Angel Project	Graduated Technical Project	Voting Member	China	Jun Yao	
Egeria Project	Graduated Technical Project	Voting Member	UK	Mandy Chessell	Nigel Jones, David Radley, Maryna Strelchuk, Ljupcho Palashevski, Chris Grote
Flyte Project	Graduated Technical Project	Voting Member	USA	Ketan Umare	
Horovod Project	Graduated Technical Project	Voting Member	USA	Travis Addair	
Milvus Project	Graduated Technical Project	Voting Member	China	Xiaofan Luan	Jun Gu
ONNX Project	Graduated Technical Project	Voting Member	USA	Alexandre Eichenberger	Andreas Fehlner, Prasanth Pulavarthi, Jim Spohrer
Pyro Project	Graduated Technical Project	Voting Member	USA	Fritz Obermeyer	
Open Lineage Project	Graduated Technical Project	Voting Member	USA	Julien Le Dem	Michael Robinson, Mandy Chessell
Marquez Project	Graduated Technical Project	Voting Member	USA	Willy Lulciuc	TBD

Minutes approval

LF AI & DATA

Approval of February 22, 2024 Minutes

Draft minutes from the February 22, 2024, TAC call were previously distributed to the TAC members via the mailing list

Proposed Resolution:

› That the minutes of the February 22, 2024, meeting of the Technical Advisory Council of the LF AI & Data Foundation are hereby approved.

LF AI & DATA

General Foundation Updates

LF AI & DATA

AI_Dev Events

19-20 June | Paris:
https://hubs.la/Q02l67Yg0

14-15 Oct | Seattle:
https://hubs.la/Q02l63_w0

Save the date:
23 Aug | Hong Kong
28-29 Oct | Tokyo

Election Results

General Member Election:

› Election winners were announced: Wed. evening Mar. 6th

› Appointees: **Anni Lai (Futurewei) and Junping Du (Datastrato).**

Outreach Chair Election:

› Election winners were announced: Wed. evening Mar. 6th

› Appointee: **Richard Sikang Bian (The Ant Group)**

Board Treasurer is Still Open; we've also reopened nominations for Gen AI Commons Vice Chair in light of Alexy Khrabrov's departure from IBM.

Growing Community of Projects
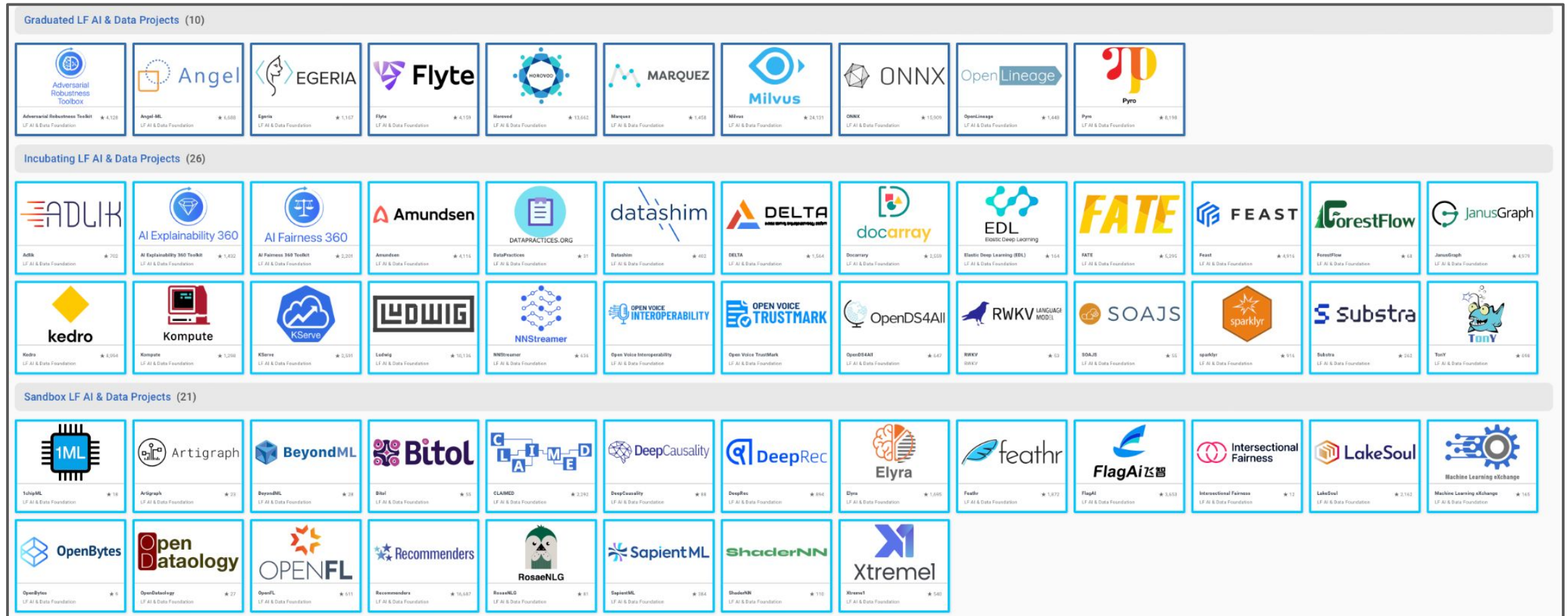
Companies hosting projects in LF AI & Data

Process for proposing a project for hosting in LF AI & Data

1. Contact ibrahim@linuxfoundation.org

2. Decide on a date to present to the TAC and request incubation

3. Ensure that your project implements these [recommendations](#)

4. Submit a formal request to incubate the project via a [GH](#) PR

5. Prepare deck and share with ED about 10 days prior to the presentation

6. Present to the TAC and get approval

7. Onboard the project with the LF AI & Data team and integrate the project with our services

8. Announce the project becoming hosted in LF AI & Data

□LF AI & DATA

Agenda

- Background

- MOF Components

- MOF Acceptable Licenses

- MOF Classes

- MOF Implementation

- MOF Benefits

- MOF Limitations

- Out of Scope

- Conclusion

Background

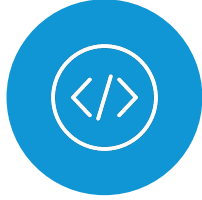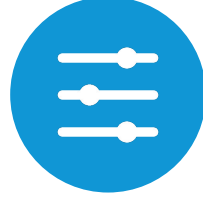| **BACKGROUND**	**INDUSTRY NEEDS**	**ADDED VALUE**
The rise in commercialization of General Artificial Intelligence (GAI) has brought forth numerous challenges.	Some "open" models lack essential components for full understanding and reproduction. Some use confusing licenses, a practice known as "openwashing."	The framework offers guidance to researchers and developers seeking to enhance model transparency and reproducibility while allowing permissive usage.
These include mounting apprehensions regarding reproducibility, transparency, ethics, and appropriate usage, among others.	To address this, we propose the Model Openness Framework (MOF), evaluating models based on completeness and openness with principles from open science.	For enterprise and industry, it provides a clear indication of models suitable for commercial use without restrictions.
A significant number of AI systems function as black boxes, posing challenges in comprehending their operations and ensuring fairness and impartiality.		

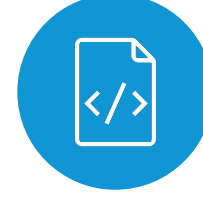# MOF Components

Datasets

Preprocessing Code

Model Architecture

Model Parameters

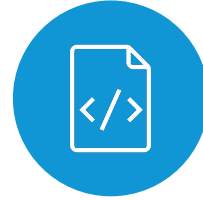Training Code

Inference Code

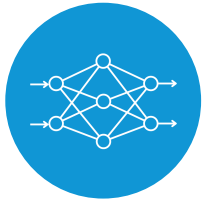Evaluation Code

Evaluation Data

Evaluation Results
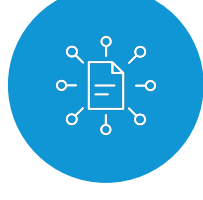
Supporting Libraries & Tools

Technical Report

Model Card

Data Card

Research Paper

Sample Model Outputs

MOF Configuration file

MOF Acceptable Licenses

COMPONENT	DOMAIN	CONTENT TYPE	ACCEPTED OPEN LICENSE
Datasets	Data	Data	CDLA-Permissive-2.0, CC-BY-4.0 or equivalent permissive license
Data Preprocessing Code	Data	Code	OSI-approved
Model Architecture	Model	Code	OSI-approved
Model Parameters	Model	Data	CDLA-Permissive-2.0, CC-BY-4.0 or equivalent permissive license
Model Metadata	Model	Data	CDLA-Permissive-2.0, CC-BY-4.0 or equivalent permissive license
Training Code	Model	Code	OSI-approved
Inference Code	Model	Code	OSI-approved
Evaluation Code	Model	Code	OSI-approved
Evaluation Data	Model	Data	CDLA-Permissive-2.0,CC-BY-4.0 or equivalent permissive license
Evaluation Results	Model	Documentation	CC-BY-4.0, CC0 or equivalent permissive license
Supporting libraries and Tools	Model	Code	OSI-approved
Model Card	Model	Documentation	CC-BY-4.0, CC0 or equivalent permissive license
Data Card	Data	Documentation	CC-BY-4.0, CC0 or equivalent permissive license
Technical Report	Model & Data	Documentation	CC-BY-4.0, CC0 or equivalent permissive license
Research Paper	Model & Data	Documentation	CC-BY-4.0, CC0 or equivalent permissive license
Sample Model Outputs	Model	Data or Code	OSI-Approved for Code, CDLA-Permissive-2.0 for Data, CC-BY-4.0, CC0 or equivalent permissive license for content

MOF Classes

Model Openness Framework Class	Components Included
Class I – Open Science	- Research Paper - Datasets - Data Preprocessing Code - Model Parameters and Metadata (Intermediate Checkpoints and Optimizer States) - All Class II Components
Class II – Open Tooling	- Training Code - Inference Code - Evaluation Code - Evaluation Data - Supporting Libraries & Tools - All Class III Components
Class III – Open Model	- Model Architecture - Model Parameters and Metadata (Final Checkpoints and Optimizer State) - Technical Report - Evaluation Results - Model Card - Data Card

MOF Implementation

Preparing Distribution

Include a LICENSE file that describes the licenses used for the project.

In addition, the distribution must include an MOF.JSON file to describe the MOF class, included components and the licenses used.

Class Assignment

Model Openness Tool (MOT) will help project owners know how their project lines up with MOF classes.

http://isitopen.ai

Badging

The MOT will issue badges based on classified class and issue code for github README and record model openness to the MOC Scoreboard.

MOF benefits:

Clarity: The framework distinctly outlines the included components and the corresponding licenses, facilitating a clear understanding of acceptable use and indicating whether a project is genuinely open and complete.	**Accountability & Fairness**: Publicly available data and models can be audited for unwanted biases and harms, allowing the community to notify model producers of discovered issues.
Completeness and Openness: By categorizing models and their artifacts into increasing levels of completeness and openness, the MOF encourages model producers to create comprehensive and open models. This advancement supports open science, benefiting both academic and commercial applications.	**Continuous Improvement:** Open models facilitate building upon existing work, accelerating innovation and progress in AI by enabling model producers and consumers to collaborate and evolve collectively.
Reproducibility: Providing comprehensive access to data, code, and models enables independent replication of results, helping identify errors, biases, or disparities and enhancing scientific rigor.	**Collaboration**: Sharing open resources encourages collaboration among model producers and consumers across different fields and organizations, pooling knowledge and capabilities.
Transparency & Explainability: Opening model architectures, weights, training code, and documentation illuminates how models operate, fostering trust and enabling scrutiny for potential issues.	**Education & Learning:** Open access to data, code, and models supports teaching and learning about AI, making it easier for students, new researchers, and developers to enter the field.
Data Provenance: Releasing details about data origins and attribution allows for tracing bias in models and identifying sources of Personally Identifiable Information (PII) leakage.	**Regulation**: Openness in models enhances their suitability for oversight and governance, unlocking potential public policy options.

MOF Limitations

- The MOF is tailored for deep learning artifacts and may not directly apply to other learning forms, although a similar approach can be adopted for statistical machine learning or reinforcement learning.

- Model producers are urged to be transparent about the availability of released components, the openness of licenses for each, and the overall completeness of their models. Convincing producers to share their work openly, especially without initial restrictions, is a challenge.

- Openness objectives need to be balanced against considerations like privacy, intellectual property (IP), institutional policies, and commercialization pressures.

- Classifying models may not fully capture their functionality, and concerns about bias, safety, and other harms persist. However, openness facilitates external audits for quality and completeness.

- Simplicity in classification may overlook nuances, but continuous improvement of the rubric is possible.

- Does not address the use of copyrighted materials in training data, an issue currently navigated through legal and legislative processes. Using an open license for data is sufficient for qualification as "open," but researchers and developers are encouraged to respect copyrights and use authorized data in model training.

Out of Scope

- Bias and Fairness

- AI Safety

- Code Review

- Model Review

- Trustworthiness

- Performance Testing

- Red-Teaming

- Security and Privacy

Conclusion

- The **Model Openness Framework** offers a clear and actionable methodology to assess and enhance transparency in machine learning models. It provides a roadmap for model producers by specifying components, such as training data, code, models, and documentation, that should be openly released, promoting reproducible and ethical AI development.

- Adopting open licenses, as recommended by the framework, fosters collaboration, community engagement, and the freedom to use, modify, and distribute models and components within the license terms. The tiered classification system encourages the release of models with increasing levels of openness, promoting collective innovation and ensuring fairness, safety, and public oversight.

- Achieving this vision requires a collaborative effort from all AI stakeholders, including researchers, developers, institutions, companies, and governments, to embrace openness as a fundamental principle. The substantial benefits for science, business, and society justify the challenges of pursuing model transparency.

- With well-designed incentives, policies, and community norms, the ideals of open source and open science can become the standard in AI, rather than the exception. Through cross-domain collaboration, we can shape the progress of AI to be as open, ethical, and empowering as possible. The Model Openness Framework provides practical guidance for this journey towards trustworthy and democratized artificial intelligence.

Questions?

Thank you!

LF AI & DATA

Egeria Project Update

Mandy Chessell

LF AI & DATA

A bit of background …

› Egeria was started about 5 years ago by IBM, ING and SAS.

› It provides:
  › An open metadata type system (~1000 types) that describe a knowledge graph.
  › Metadata exchange services to be able to incrementally synchronize metadata between different tools and technologies.
  › Governance processes to improve and repair metadata.

› Its communication style is peer-to-peer, avoiding a centralized metadata repository.

Piecing together knowledge from many sources
to build trust and confidence in your data and analytics

Who

How

What

Why

Where

Which

When

"who owns?"

"who uses?"

"who permits use?"

"what"

"which license?"

"how created?"

DataSet

"how used?"

"when last updated?"

"where useful?"

"where from?"

"why needed?"

LF AI & DATA
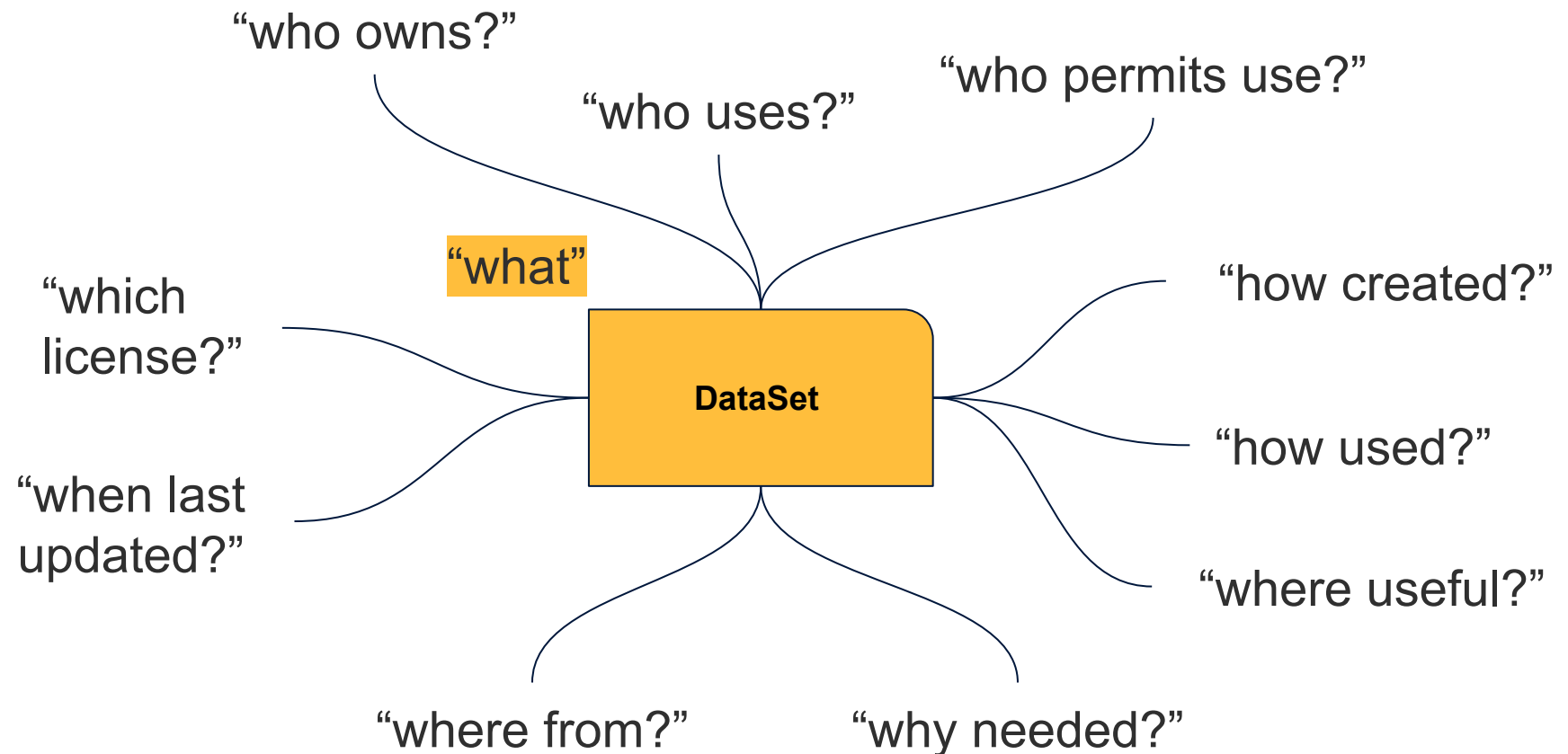
Project changes

› June/July 2023
  › After 5 years of valuable contribution, both IBM and ING announced that they were no longer going to contribute to Egeria.
  › Effectively, this changed the number of people who were full-time on the project from about 20 to 2 people.

› Pragmatic Data Research Ltd wished to continue, but it was obvious that things needed to change…

Taking stock …

› The project organization was designed to allow for multiple parallel streams of work - how could this be simplified:
  › 20+ git repositories
  › 5 different community calls per week
  › Monthly releases
› The software included multiple partially completed, uncommented/undocumented modules - what should be completed and what should be removed or deprecated?
› What steps should be taken to attract new contributors and partners?

The plan …

› Managing the workload
  › Reduced community calls to 1 per week
  › Releases now thematic rather than driven by the calendar

› Reviewing the project scope, focus and next steps
  › How easy is it to get up and running?
  › What features are needed to get started and deliver value quickly?
  › What is the long-term vision for the software - has it changed?
  › How do we make it easier to contribute to the project?

Release 5.0 (due late March 2024)

› Consumability
  › Install image includes samples, utilities, pre-configured servers, connectors and utilities.
  › Consolidating key components into main egeria.git repository.
  › Renaming components with confusing names.
› New features to reach new audiences
  › New python clients
  › New survey framework
  › New APIs for data products

https://egeria-project.org/release-notes/next/

Demo

Next Steps

› Building the community is a key priority
  › In discussions with a large technology company and an international bank. Both interested in both using and contributing to the project.
  › Egeria's education and tutorials need to be updated to reflect the simplification and usability changes made in release 5.0.
  › New UIs to allow us to demonstrate the capabilities of Egeria are in progress
  › We plan to create a series of short videos and blogs to publicise Egeria - see

https://getting-started-with-egeria.pdr-associates.com/introduction.html

Open Discussion

LF AI & DATA

Upcoming TAC Meetings

➔ March 21, 2024

➔ April 4, 2024

If you have a topic idea or agenda item, please send agenda topic requests to
tac-general@lists.lfaidata.foundation

LF AI & DATA

TAC Meeting Details

TAC Biweekly Meeting LF AI & Data

Ways to join meeting:

1. Join from PC, Mac, iPad, or Android

https://zoom-lfx.platform.linuxfoundation.org/meeting/95332329356?password=c708f2ee-fb78-4a12-91a3-47daa19b708f

2. Join via audio

One tap mobile:

US: +12532158782,,95332329356# or +13462487799,,95332329356

Or dial:

US: +1 253 215 8782 or +1 346 248 7799 or +1 669 900 6833 or +1 301 715 8592 or +1 312 626 6799 or +1 646 374 8656 or 877 369 0926 (Toll Free) or 855 880 1246 (Toll Free)

Canada: +1 647 374 4685 or +1 647 558 0588 or +1 778 907 2071 or +1 204 272 7920 or +1 438 809 7799 or +1 587 328 1099 or 855 703 8985 (Toll Free)

Meeting ID: 95332329356

Meeting Passcode: 040721

◻LF AI & DATA

Legal Notice

LF AI & DATA