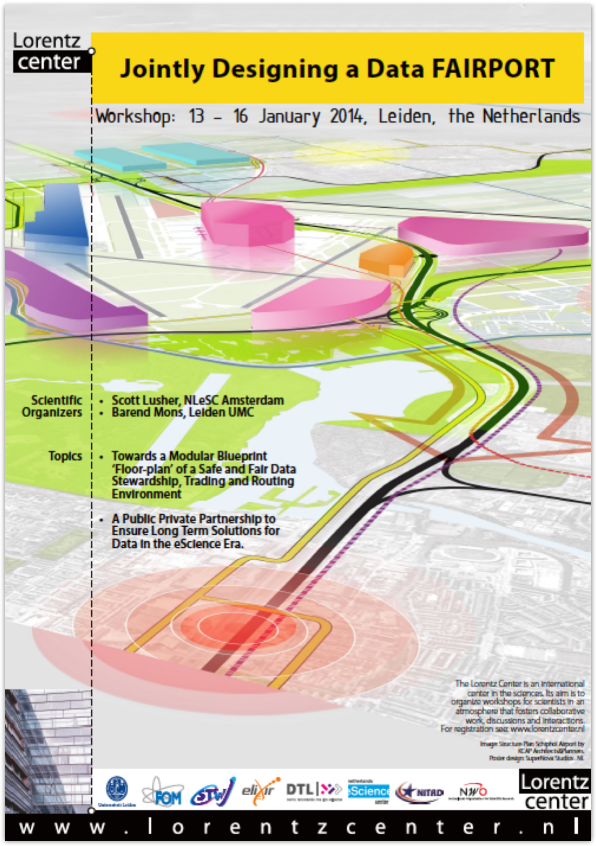
EOSC Symposium 2024
21 October (15:30 to 16:30)
Examples of EOSC scientific, societal and economic impact session (Part 1)

FAIR Data and data that are Fully AI Ready

Erik Schultes
http://orcid.org/0000-0001-8888-635X
eriks@gofair.foundation
These slides: https://osf.io/cr9sv

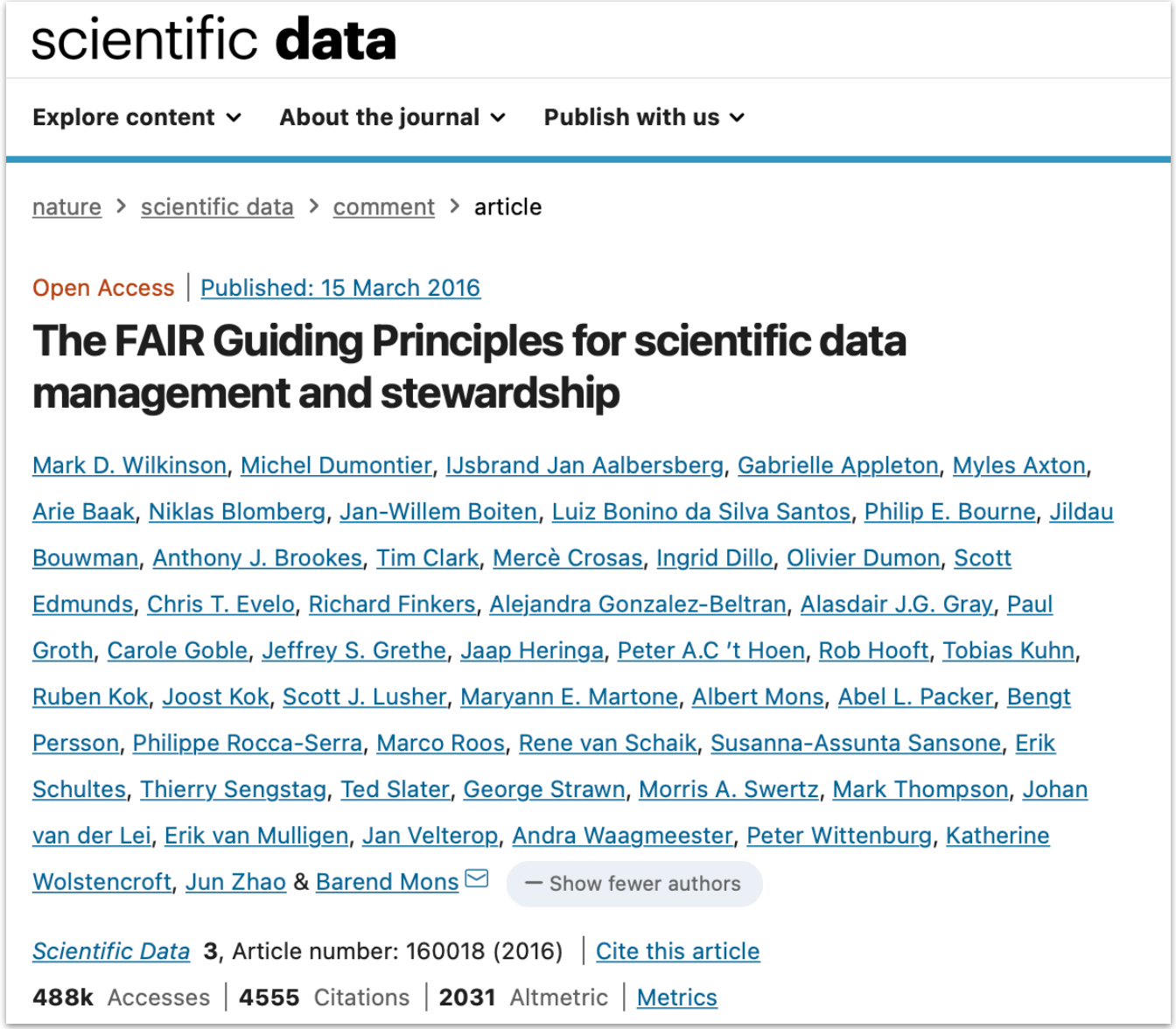


2014

2024

- Day 1 (Monday) Reflection and looking ahead
- Day 2 (Tuesday) Focus area 1: Machine actionability
- Day 3 (Wednesday) Focus area 2: Equitability
- Day 4 (Thursday) Focus area 3: Fully AI Ready
- Day 5 (Friday) Workshop conclusion

2016



scientific **data**

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › scientific data › comment › article

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons — Show fewer authors

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

…the FAIR Principles put specific emphasis on **enhancing the ability of machines to automatically find and use the data**, in addition to supporting its reuse by individuals.

2016

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

2016

Box 2 | The FAIR Guiding Principles

Computability

- *FAIR data* use knowledge representation languages and controlled vocabularies that reduce/eliminate ambiguity.

Trustworthiness

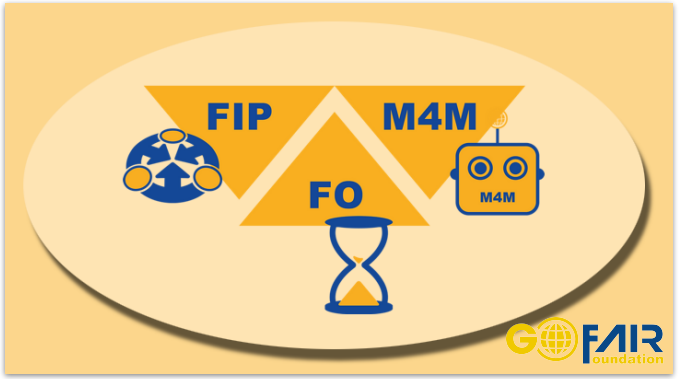
- *FAIR data* have (rich) provenance which provides evidence for the source of data. Provenance can include uncertainties and error estimation. Trust also requires large amounts of metadata.

Equitabilty

- *FAIR data* make explicit the conditions for reuse. Equitabilty can be technologically ensured, when data and metadata are FAIR

First decade of FAIRification…

FAIR Practices & Training



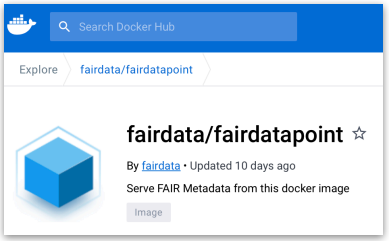
CEDAR Workbench

FIP Wizard

FAIR Connect

WorldFAIR

FAIR Orchestration



NANODASH

CEDAR Workbench

SRDC

REDCap
Research Electronic Data Capture

SeSaMe
SemanticSerchMetadata

MOLGENIS

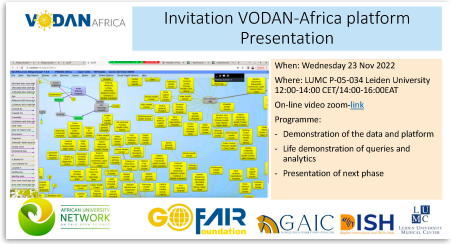
castor

FAIR Solutions

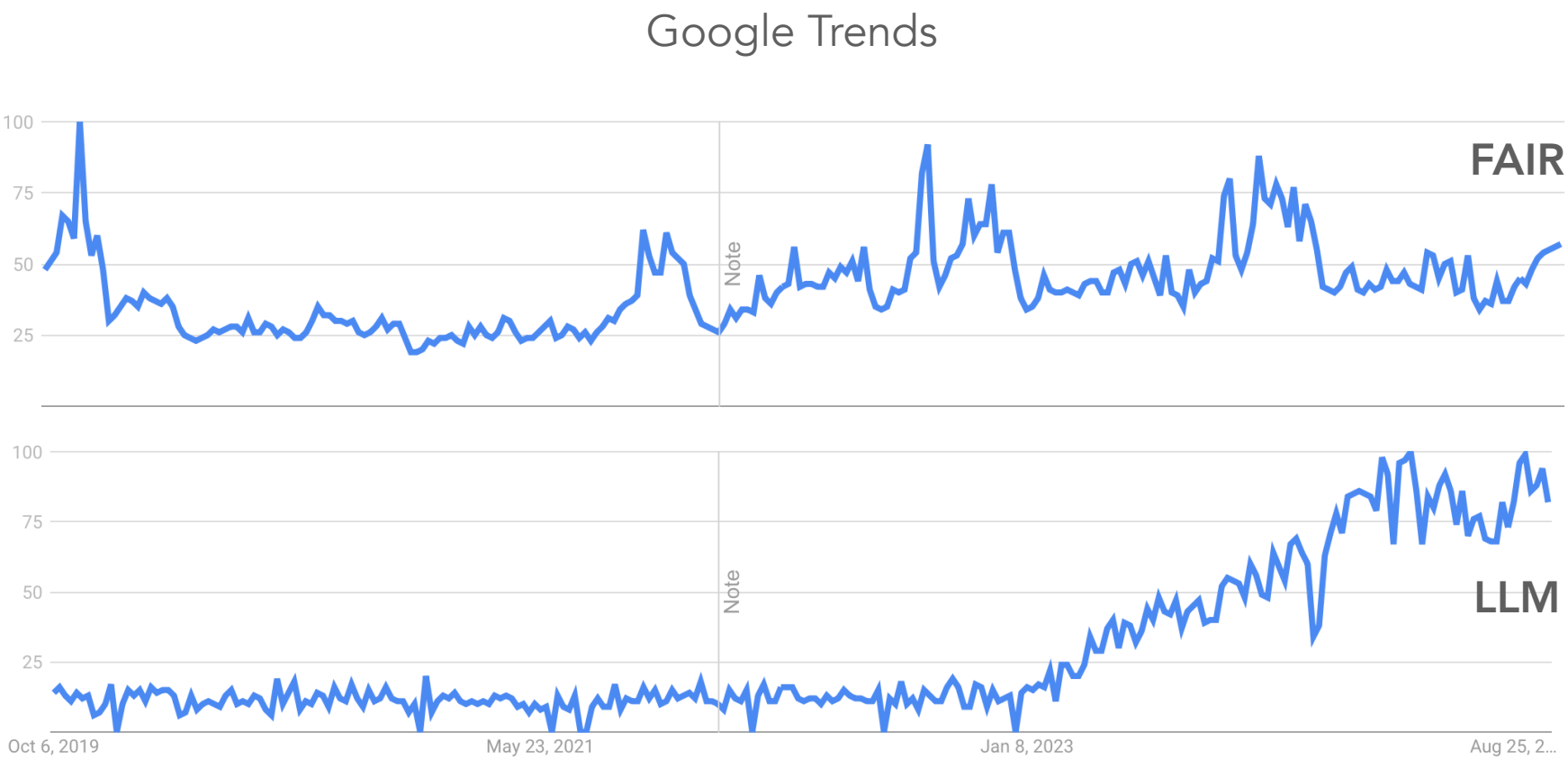
DESCRIBO
DESCRIBE EVERYTHING

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures

FAIR Data Visiting

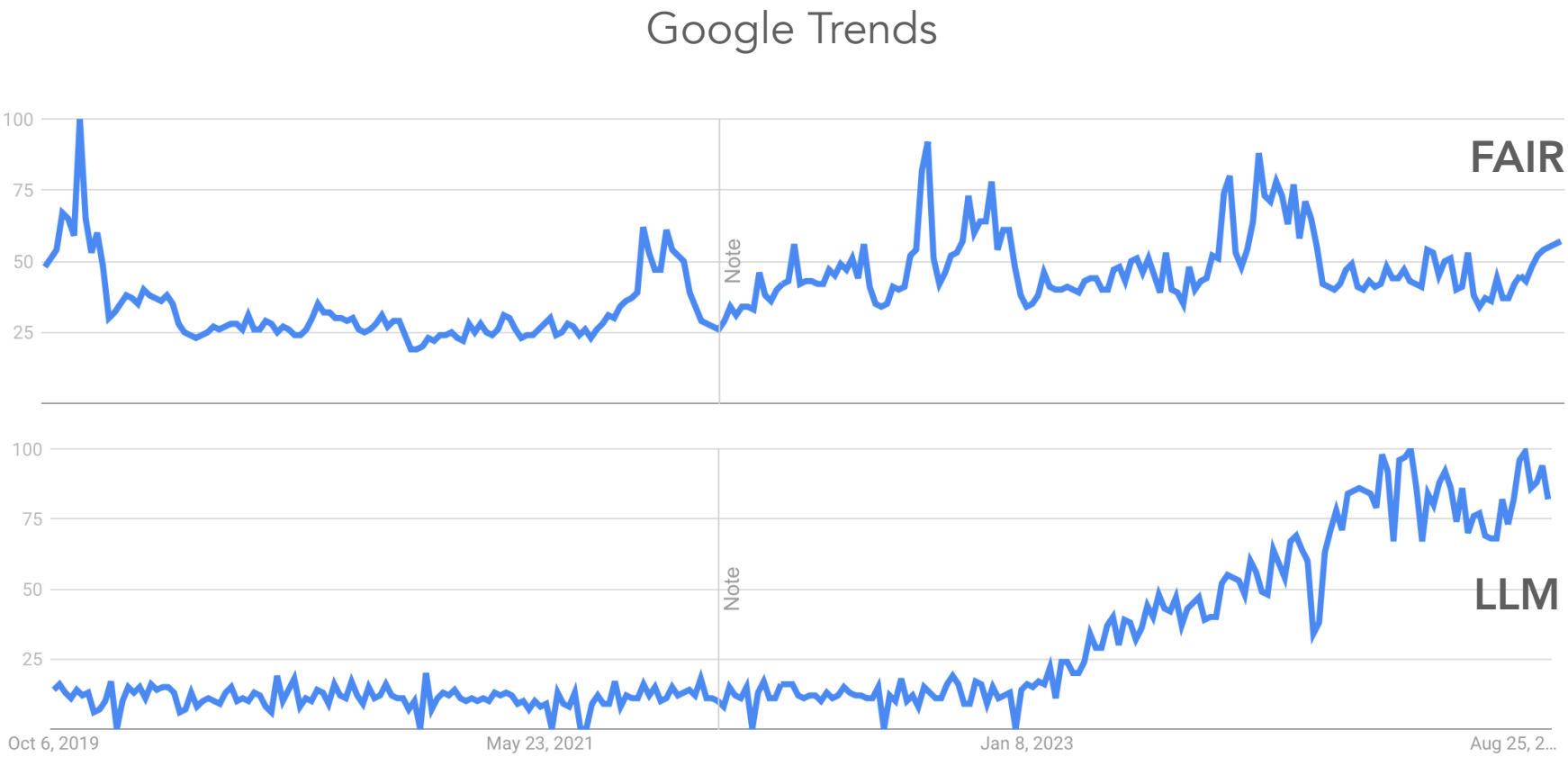


Then came (2023) the LLMs

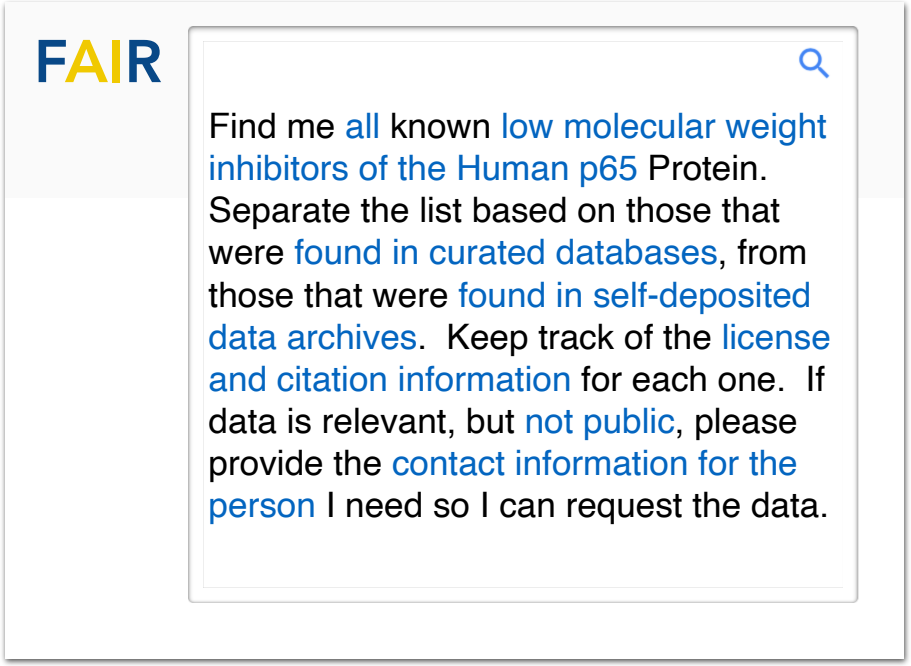


Then came (2023) the LLMs

Emphasis on “Large” = ChatGPT-4 has 1.76 x 10^{12} parameters and cost $100M



Question 1: LLMs and the original vision of FAIR…



Hypothetical search application on FAIR data.
Mark Wilkinson, CBGP-UPM/INIA, Madrid
2017

Question 1: LLMs and the original vision of FAIR…

ChatGPT follows a complex, natural language command.

ChatGPT resolves ambiguity:

- *p65* - RelA, NF-κB
- *Inhibit* - target, modulate

Reality check:

It seems that Chat is good at identifying potential data sources, but isn't going much further than that... APIs will still be pretty opaque to an LLM! Data needs to be well-modeled and linked.

- Mark Wilkinson (30 September 2024)

Finding known low molecular weight inhibitors of the Human p65 protein (also known as RelA, part of the NF-κB complex) can be sourced from curated databases and self-deposited data archives.

Curated Databases:

1. **ChEMBL**: A well-established database of bioactive molecules, ChEMBL includes several small molecule inhibitors targeting p65/RelA. These molecules are indexed based on their interaction data, activity, and relevance in drug discovery. You can access and download compound data under an open license from the ChEMBL site (EMBL-EBI).

2. **PubChem**: This public resource houses chemical compounds, including those with inhibitory action on p65. Detailed molecular properties, including structure and bioactivity, are available. Licensing for data is open for academic use.

3. **SMMDB**: The SMMDB (Small Molecule Modulator Database) contains detailed information on small molecules modulating targets like p65 in neurological and other diseases. It is a manually curated resource, providing PubMed links and compound details, including molecular weight, 3D structures, and inhibition data【15†source】【18†source】.

Self-deposited Data Archives:

1. **ZINC Database**: This includes user-submitted and curated molecules, many of which target NF-κB, including p65. Researchers can deposit their findings here, making it a mixture of community-contributed data. Usage of the database is open, but some entries may require contacting the depositor for further licensing details.

2. **IPAD-DB**: Although primarily focused on Alzheimer's disease, the IPAD-DB contains data on molecules that modulate related inflammatory pathways, including NF-κB. Licensing details vary, but access is generally open【17†source】.

Citation and License Information:

- Most curated databases like **ChEMBL** and **PubChem** offer open access with clear academic use licenses. **SMMDB** is similarly freely accessible for non-commercial research, while data in **ZINC** may have different licensing based on individual submissions.

- For direct inquiries or requesting non-public data, each database typically provides the depositor's contact, especially in **ZINC** and **IPAD-DB**, where self-deposited entries may have restricted access.

For the most comprehensive set of inhibitors targeting p65, I recommend starting with **ChEMBL** and **PubChem** for licensed, peer-reviewed data, and exploring **ZINC** or **IPAD-DB** for community-deposited resources.

Content generated using OpenAI's ChatGPT-4

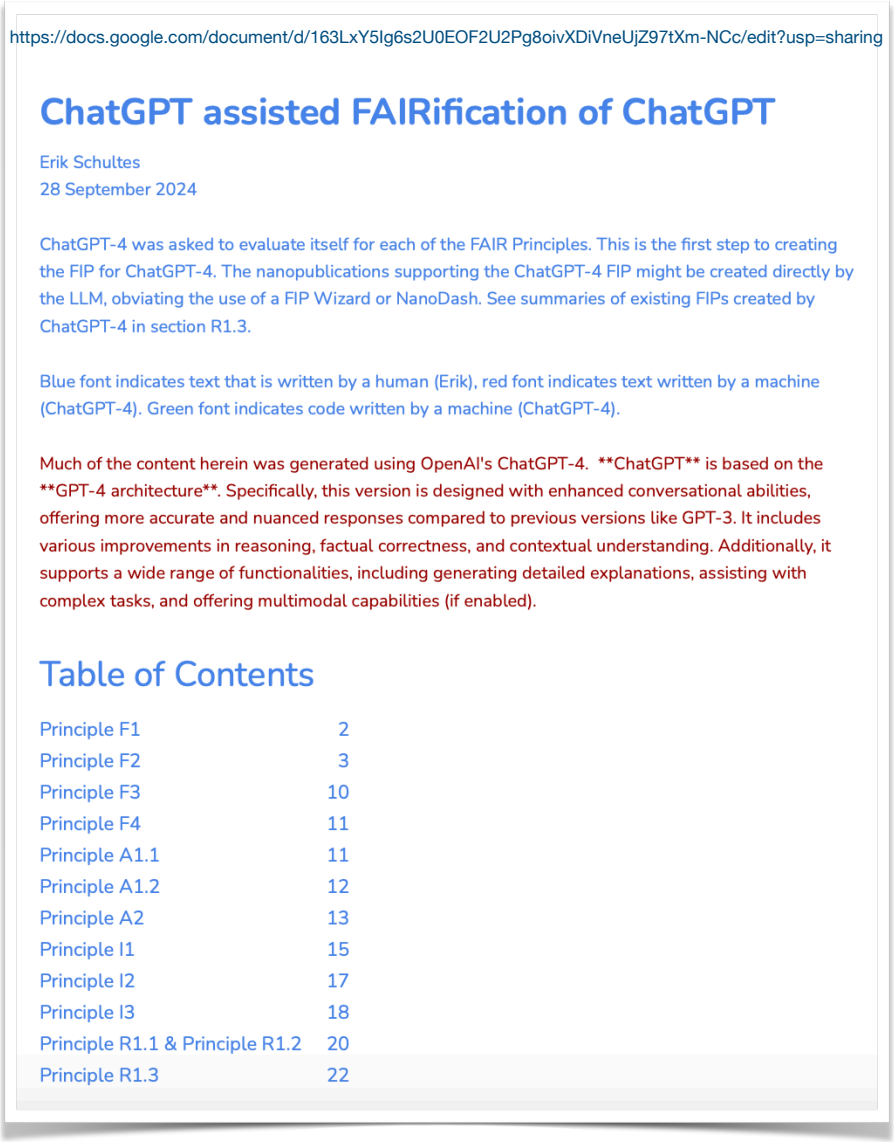
Question 2: Is ChatGPT-4 FAIR?

I dialogued with ChatGPT-4 to assess the FAIRness of the tool, much as we do with many other resources in GO FAIR.

Note: in this capacity, ChatGPT-4 is both a resource to be evaluated, as well as an automated FAIR assemsment tool.

Reality check:
The answer for F1 fails the relevance test fairly quickly. A lot of detail but not answering your question. F2 tends to do the same thing, listing criteria but not always responding to them, instead just defining them.

- John Graybeal (X October 2024)



https://docs.google.com/document/d/163LxY5Ig6s2U0EOF2U2Pg8oivXDiVneUjZ97tXm-NCc/edit?usp=sharing

ChatGPT assisted FAIRification of ChatGPT

Erik Schultes
28 September 2024

ChatGPT-4 was asked to evaluate itself for each of the FAIR Principles. This is the first step to creating the FIP for ChatGPT-4. The nanopublications supporting the ChatGPT-4 FIP might be created directly by the LLM, obviating the use of a FIP Wizard or NanoDash. See summaries of existing FIPs created by ChatGPT-4 in section R1.3.

Blue font indicates text that is written by a human (Erik), red font indicates text written by a machine (ChatGPT-4). Green font indicates code written by a machine (ChatGPT-4).

Much of the content herein was generated using OpenAI's ChatGPT-4. **ChatGPT** is based on the **GPT-4 architecture**. Specifically, this version is designed with enhanced conversational abilities, offering more accurate and nuanced responses compared to previous versions like GPT-3. It includes various improvements in reasoning, factual correctness, and contextual understanding. Additionally, it supports a wide range of functionalities, including generating detailed explanations, assisting with complex tasks, and offering multimodal capabilities (if enabled).

Table of Contents

These resutls lay the ground work for the FAIRification of the LLM itself. For example:

- I asked also for JSON-LD regarding metadata (and got it without hesitation).
- Can we can use this docuemnt to build a FIP for ChatGPT-4?
- Can we ask ChatGPT-4 to FAIRify itself?
- Can we ask ChatGPT-4 to FAIRify itself, aligned with a given FIP?
- With all FIPs?

Might we conclude that AI has obviated the need for FAIR?

- In the web of 1.76 trillion parameters, does ChatGPT-4 contain, if only **implicitly**, all the information required to orchestrate FAIR operations on arbitrary data?
- Given the vast training and validation data sets used to create AI, has the need for **explicitly** making FAIR data been obviated?
- If the machine is acting (reliably) on data, is the data by definition machine-actionable?
- Do we invest in making data more FAIR, or do we invest in making machines more intelligent (more discerning about data).
- Who better to make data machine-actionable than a machine?

Might we conclude that AI has obviated the need for FAIR?

Key Considerations:

- **Filtering:** The data collected from these sources are heavily filtered to remove noise, low-quality text, and duplicates.
- **Bias and Ethical Concerns:** Training on public data can introduce biases present in the original content (e.g., social biases, regional biases), which is why researchers try to apply techniques to mitigate these issues.
- **Copyrighted Material:** While efforts are made to use publicly available content, there are ongoing debates and concerns about training on data that may come from copyrighted works, especially when it is scraped without explicit permission.

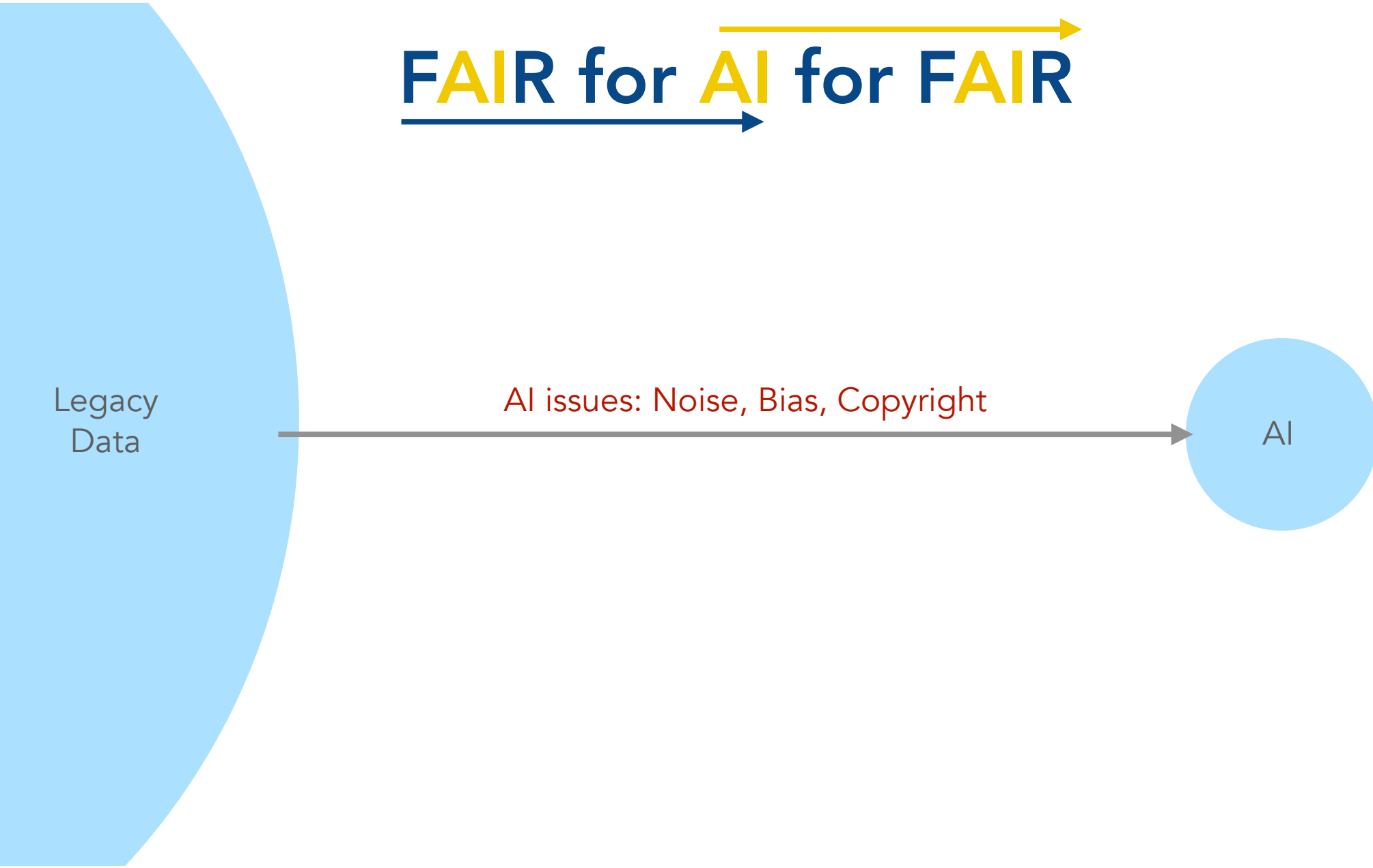
Content generated using OpenAI's ChatGPT-4

Might we conclude that AI has obviated the need for FAIR?

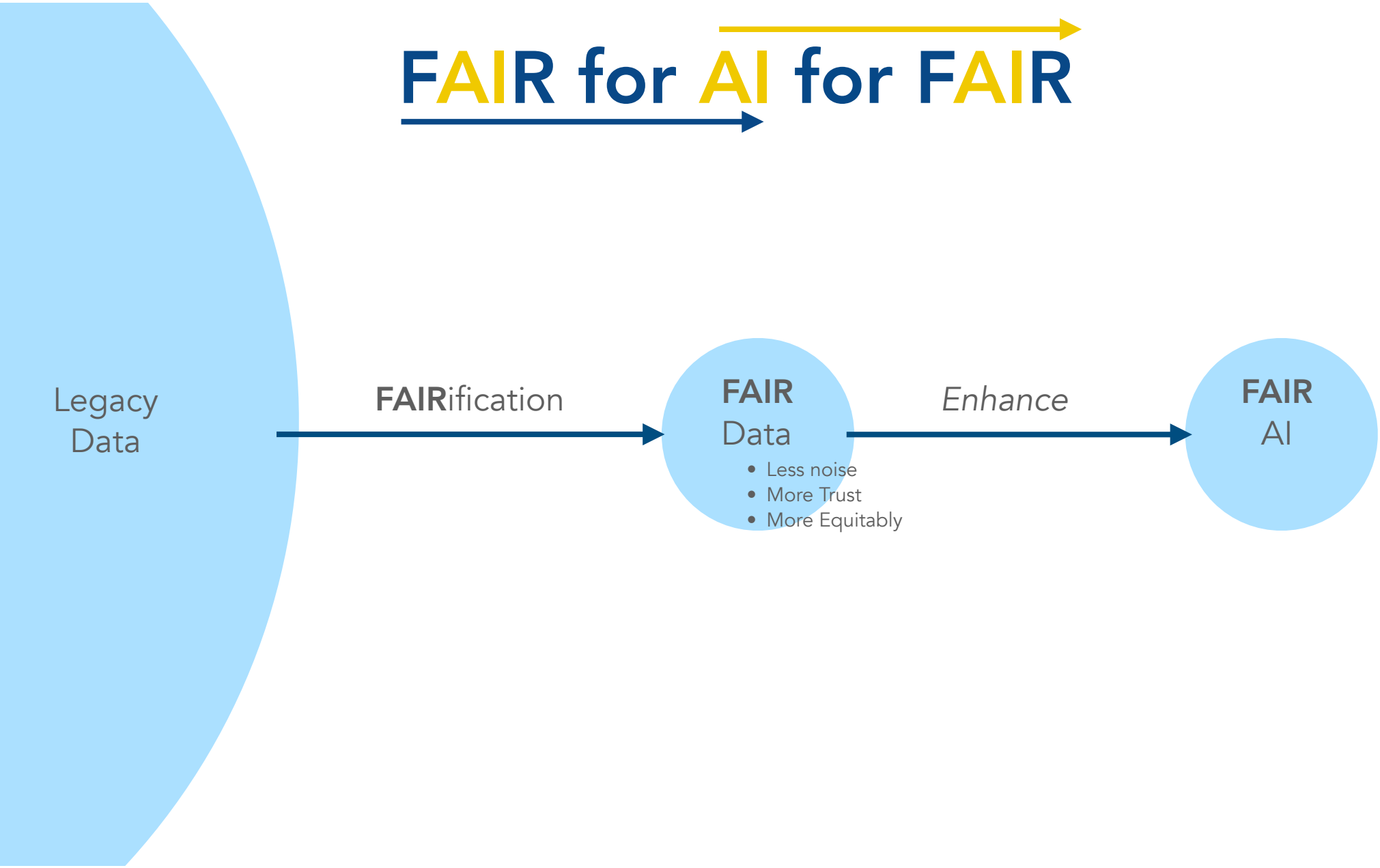
FAIR Considerations:	Key Considerations:
Computability	**Filtering:** The data collected from these sources are heavily filtered to remove noise, low-quality text, and duplicates.
Trustworthiness	**Bias and Ethical Concerns:** Training on public data can introduce biases present in the original content (e.g., social biases, regional biases), which is why researchers try to apply techniques to mitigate these issues.
Equitability	**Copyrighted Material:** While efforts are made to use publicly available content, there are ongoing debates and concerns about training on data that may come from copyrighted works, especially when it is scraped without explicit permission.

Content generated using OpenAI's ChatGPT-4

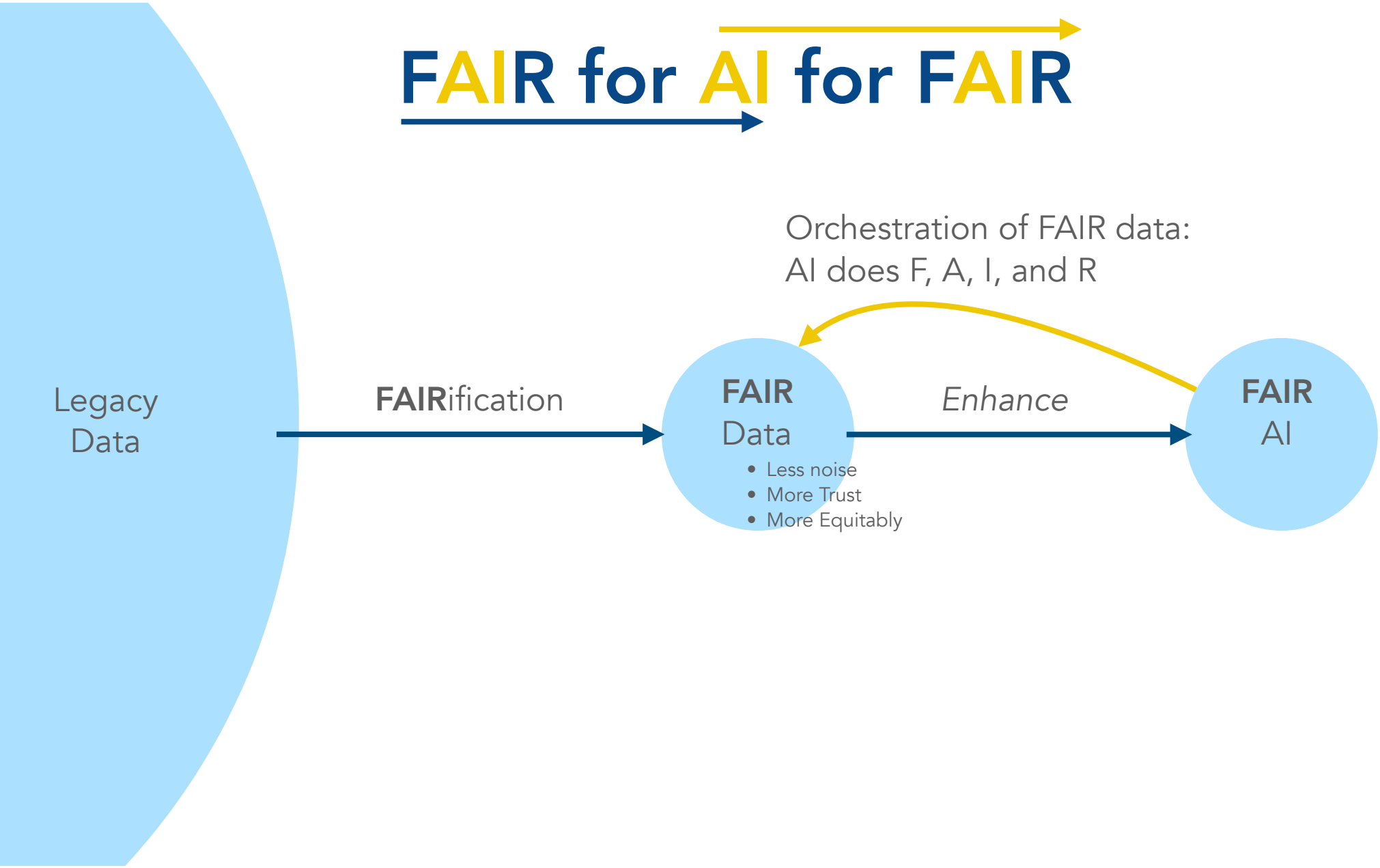
FAIR for AI for FAIR



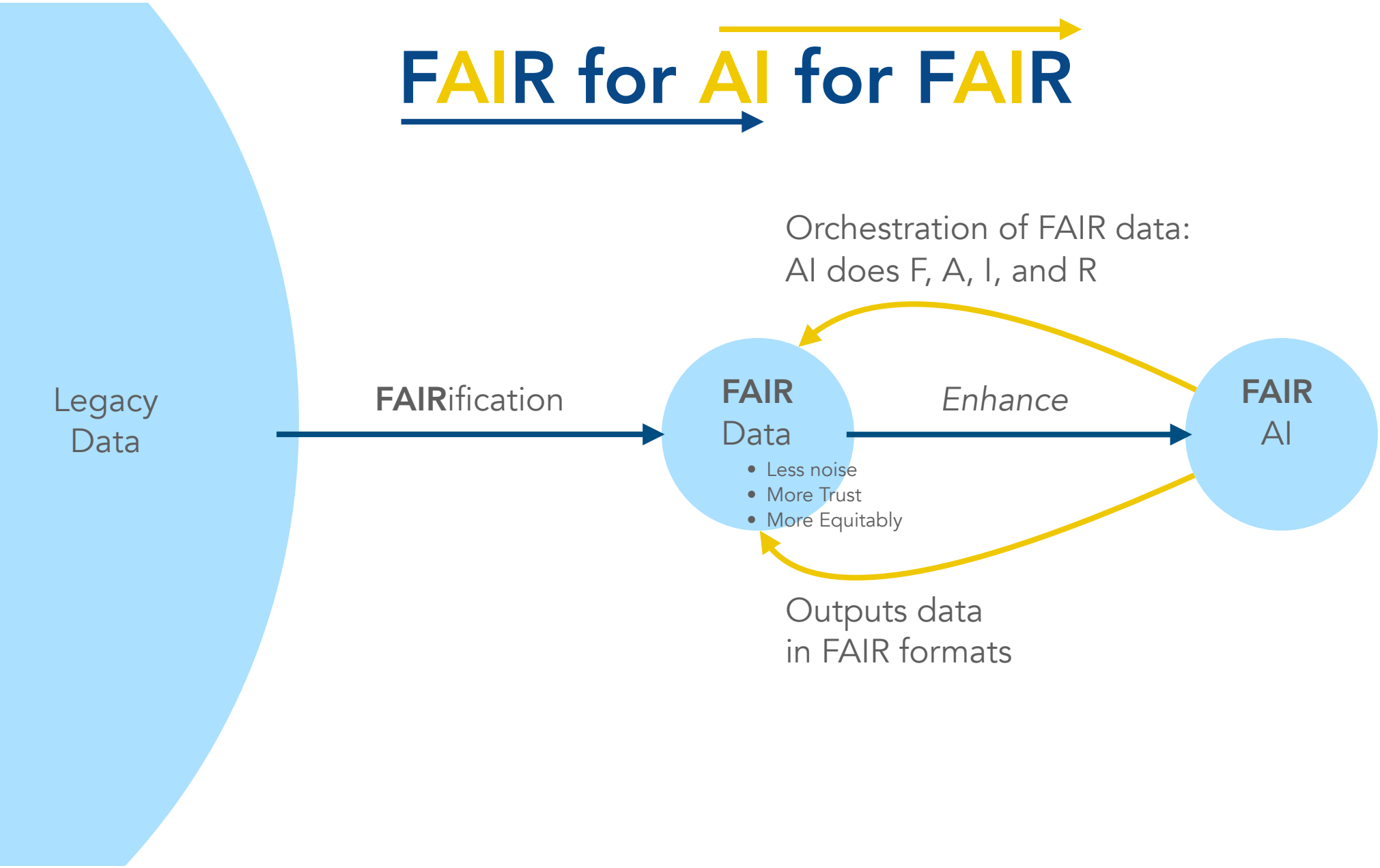
FAIR for AI for FAIR



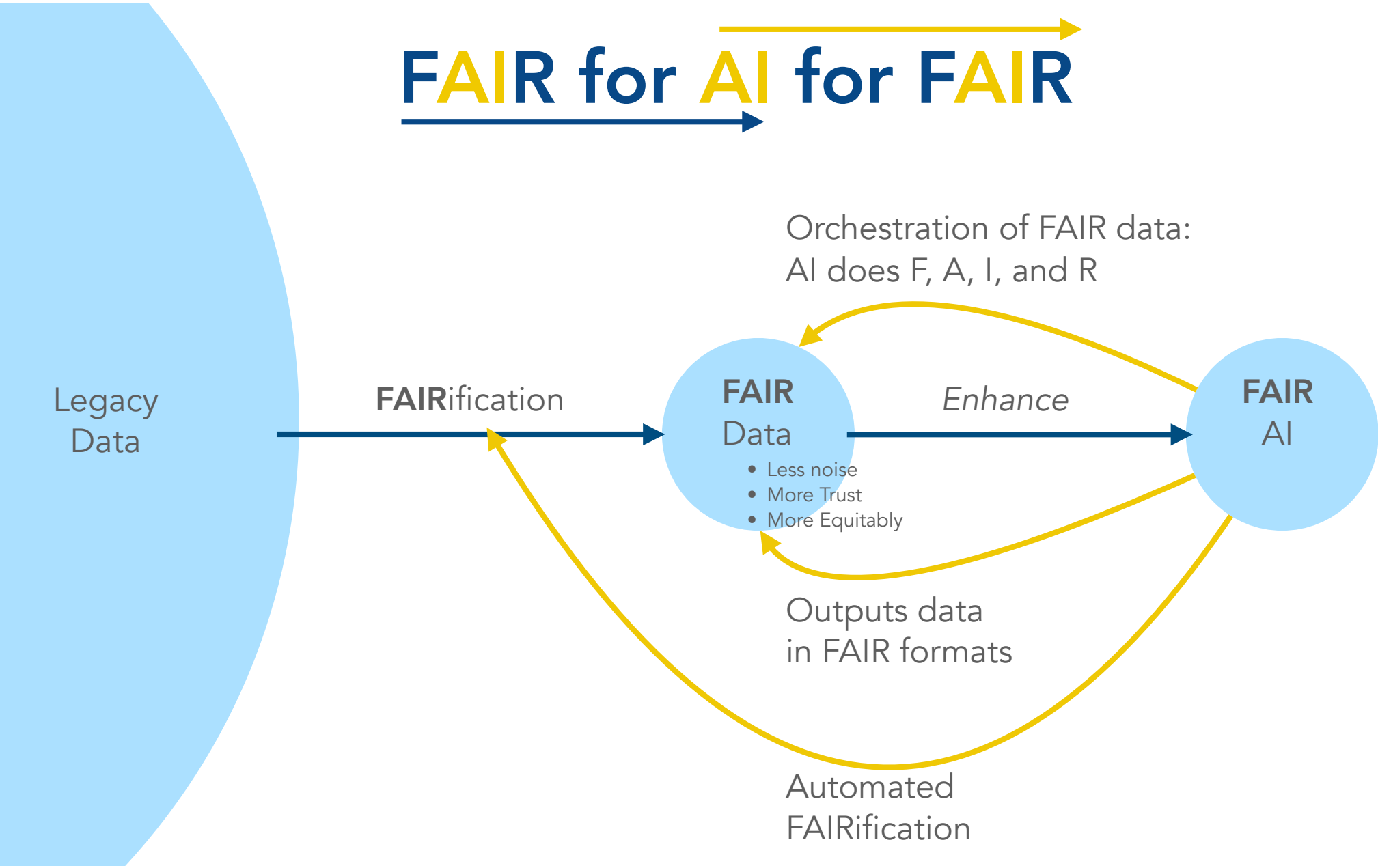
FAIR for AI for FAIR

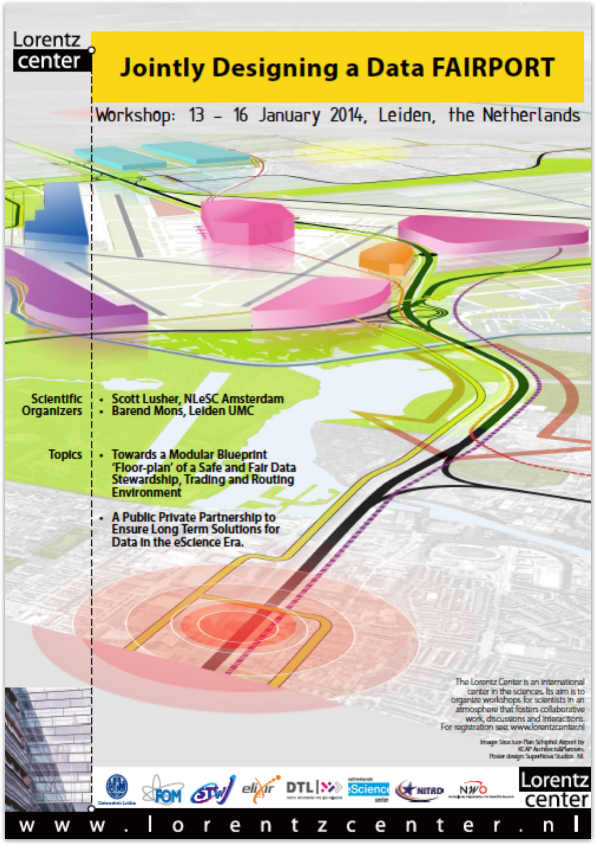


FAIR for AI for FAIR



FAIR for AI for FAIR





2014

FAIR as in "fair"

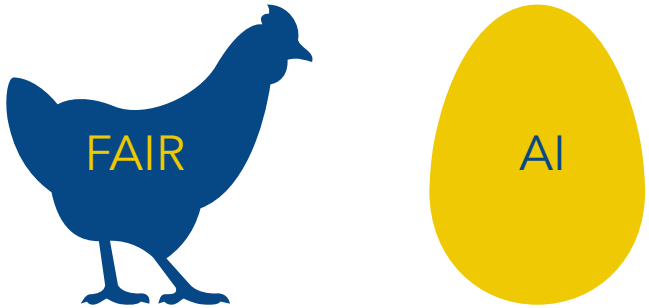
2024

FAIR as in "AI-Ready"

OpenAI
Microsoft
AI
Meta
Amazon Q
Google AI

EOSC Symposium 2024
21 October (15:30 to 16:30)
Examples of EOSC scientific, societal and economic impact session (Part 1)

FAIR well



Erik Schultes
http://orcid.org/0000-0001-8888-635X
eriks@gofair.foundation
These slides: https://osf.io/cr9sv

GO FAIR foundation

LACDR

LIFES Leiden Institute for FAIR & Equitable Science