Future-Proofing Research Data Repositories: Keeping up with the ML/AI Revolution

Stephanie Labou
Data Science Librarian
UC San Diego Library

Before we start – AI vs ML

Artificial intelligence (AI) and machine learning (ML) are often used interchangeably, but there are differences

To summarize from Google Cloud and Amazon Web services pages some relevant differences:

- ML is an application of AI where models use statistical methods to identify patterns and extract meaning from very large amounts of data
- ML relies on structured and semi-structured data whereas AI can use unstructured data

More about AI vs ML, from AWS (https://aws.amazon.com/compare/the-difference-between-artificial-intelligence-and-machine-learning/) and Google Cloud (https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning)

There is no escaping ML

- ML is rapidly becoming popular across research domains
- Social interest in ML (and AI) and questions about transparency
- Various gov memos/directives about sharing data and research outputs

ML-related repository deposits are inevitable, if not already happening - and will only increase

First things first

We need to better understand how ML objects are being shared by practitioners, so we can identify common practices as well as gaps and barriers to findability and reusability.

Machine learning objects are complex

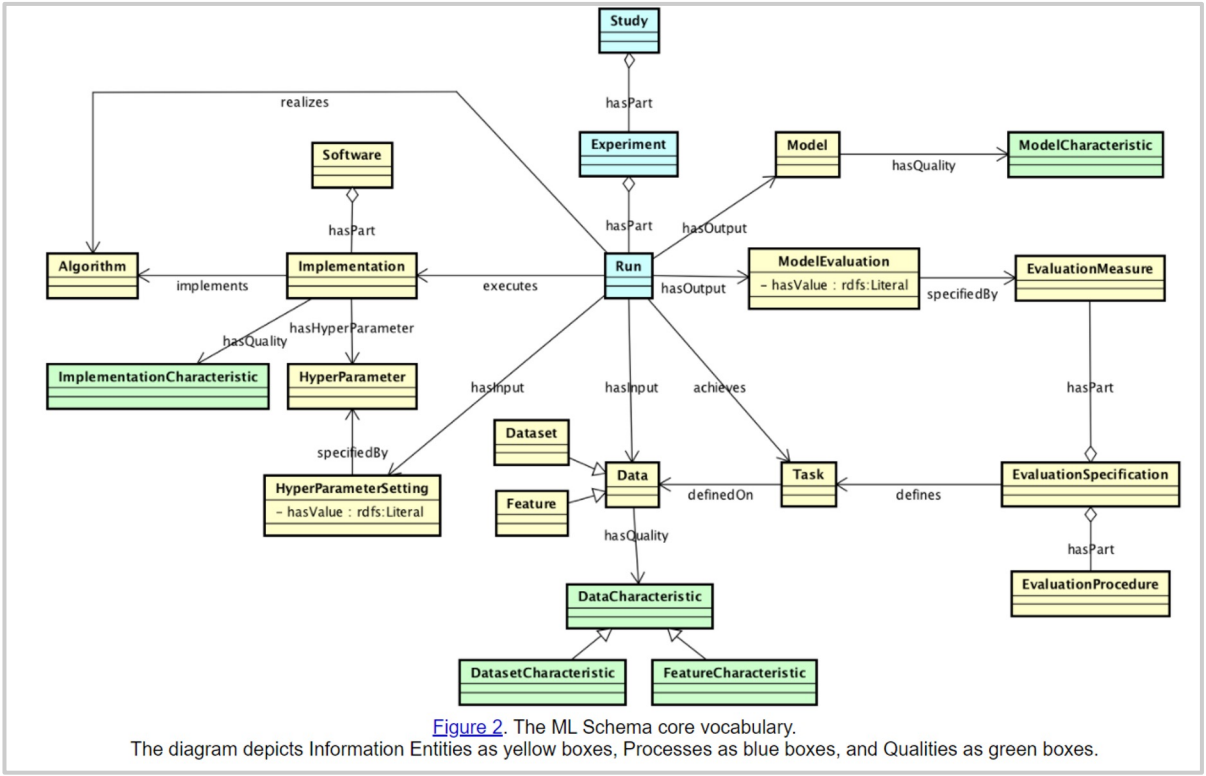


Figure 2. The ML Schema core vocabulary.
The diagram depicts Information Entities as yellow boxes, Processes as blue boxes, and Qualities as green boxes.

From Publio et al. (2018). ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies. https://arxiv.org/abs/1807.05351

See also: https://ml-schema.github.io/documentation/ML%20Schema.html

A sampling of repositories with ML content

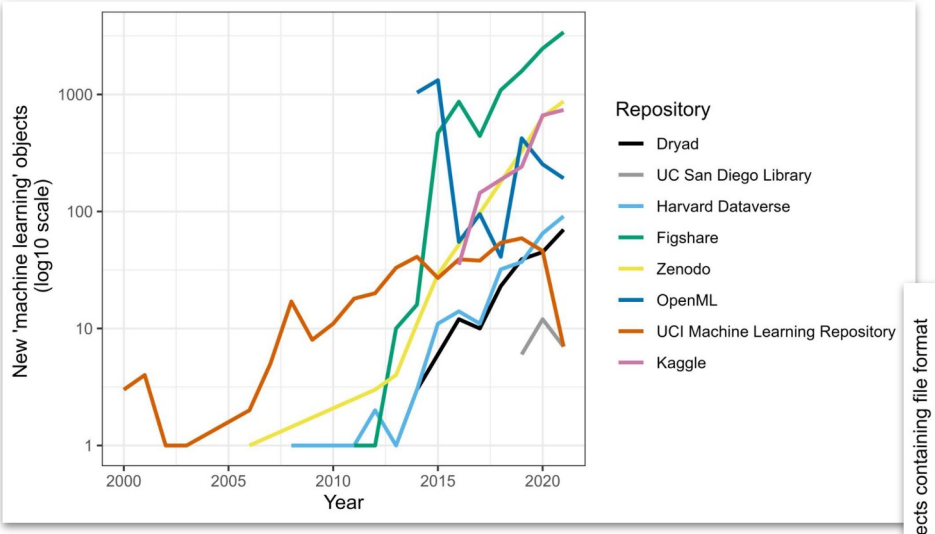
Specialist

- UC Irvine Machine Learning Repository
- OpenML
- Kaggle

Generalist

- Figshare
- Zenodo
- Dryad
- Harvard Dataverse
- UC San Diego Library Digital Collections

There's a lot of ML content in repositories, and it comes in a lot of different formats



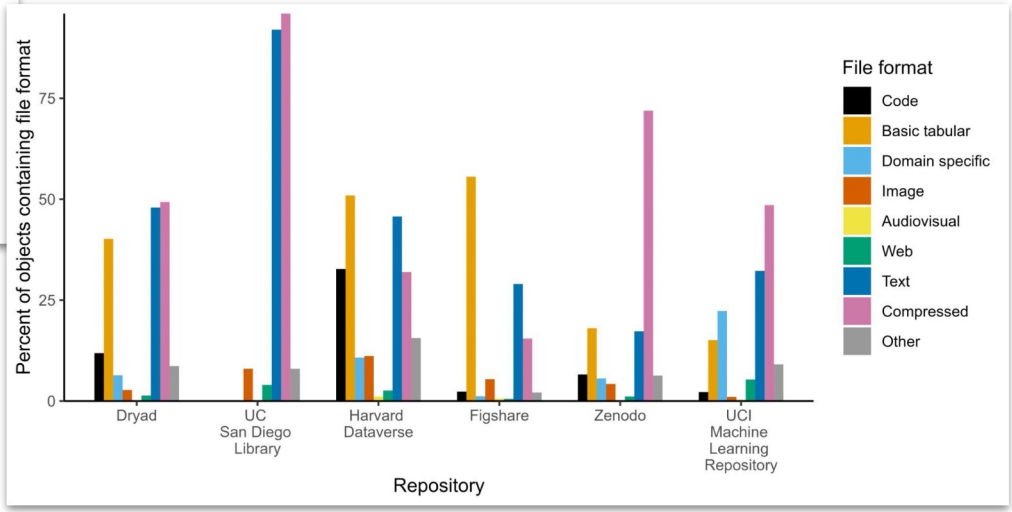


Figure 1 and Figure 2 from Labou et al. 2024 (forthcoming)

How to make repositories more amenable to ML (and eventually AI) content

Unambiguous citations

Emphasize related resources

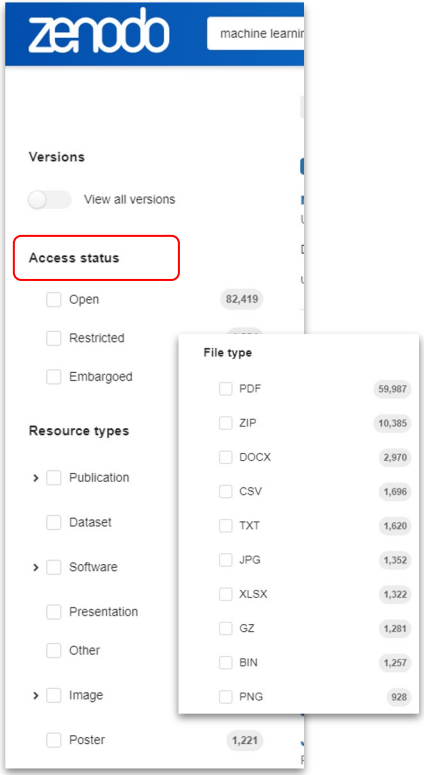
Clear licensing

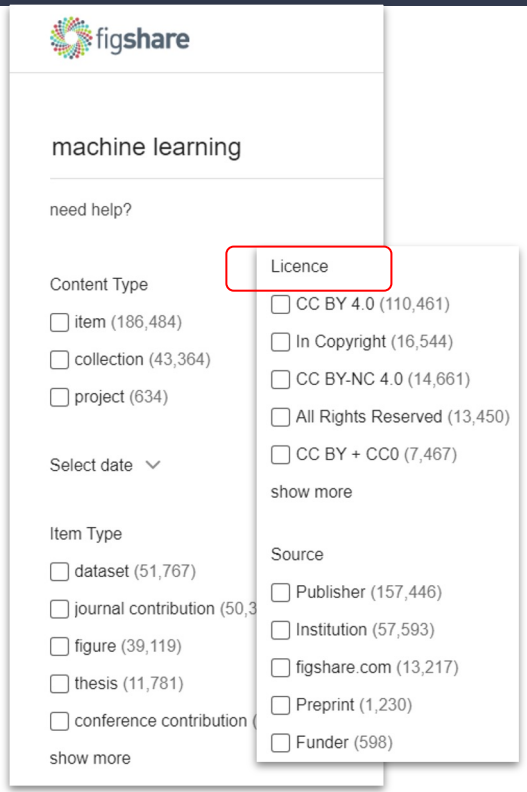
Rich(er) metadata

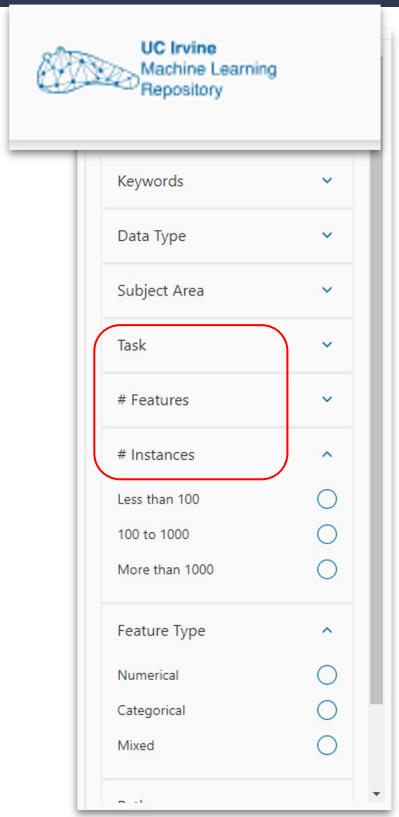
Labels for compressed files

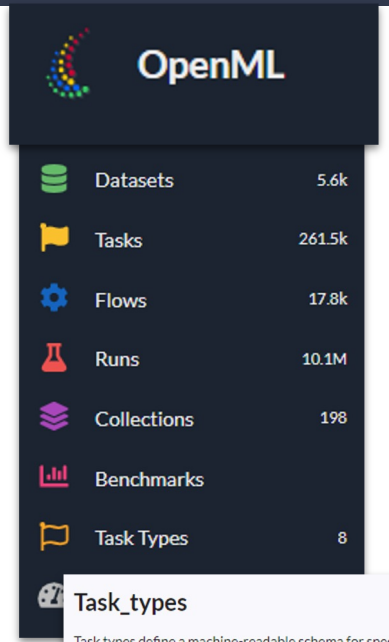
Access at scale

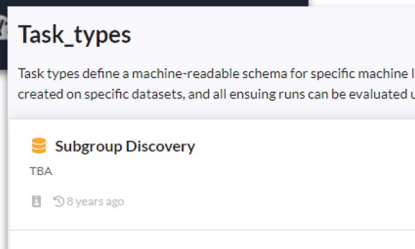
Findability: search strategy & use case are different



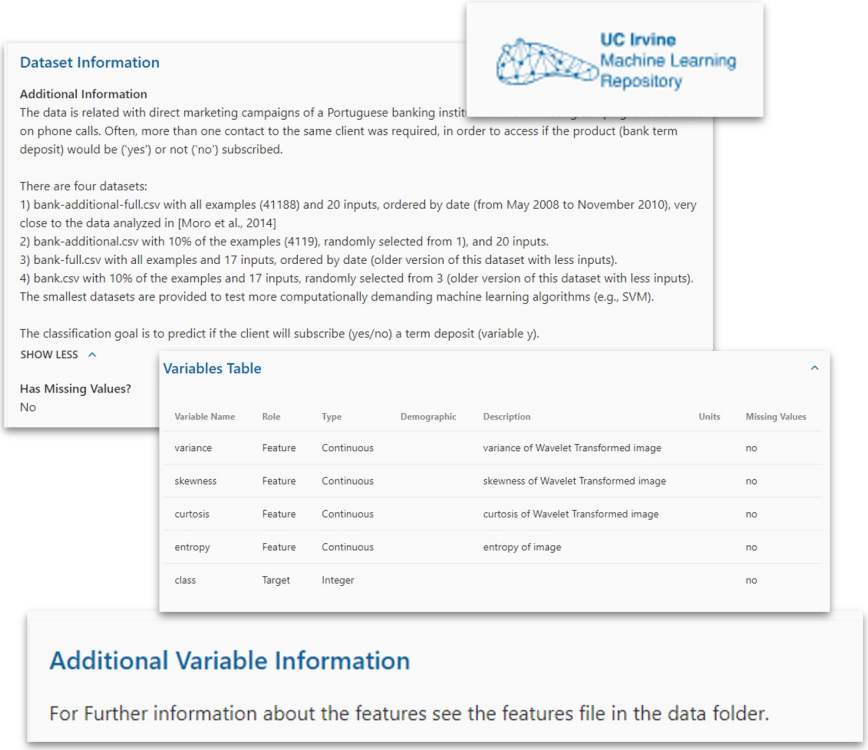








ML-useful metadata



Additional metadata fields: number of instances, task type

Even just a structured README template or features file (since adding additional metadata fields can be a big ask)

Your system may vary, but any way to enable faceting/findability by *structure* of data (or code, or other)

Make repository content ML/AI-ready

(or at least, more ready than now)

First use case was how do we make ML-related objects in repositories more suited to how ML research is conducted in practice

Second use case is making content in repositories (which may or may not be explicitly ML) amenable to ML or AI methods

Libraries especially collect good data!

→ impactful, often publicly funded, well documented

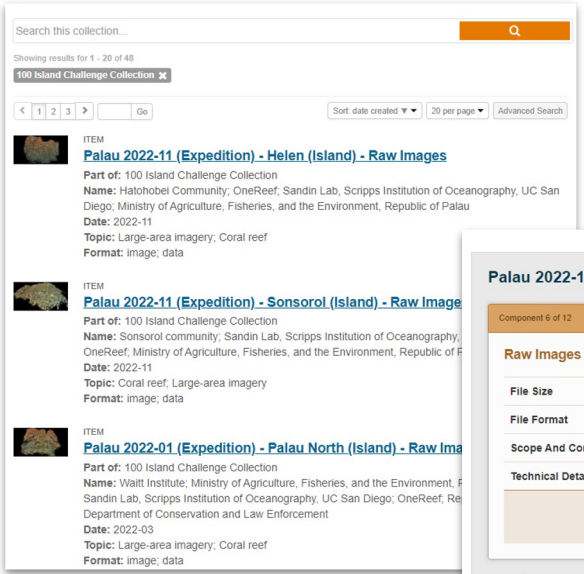
Ease of (manual) bulk downloads

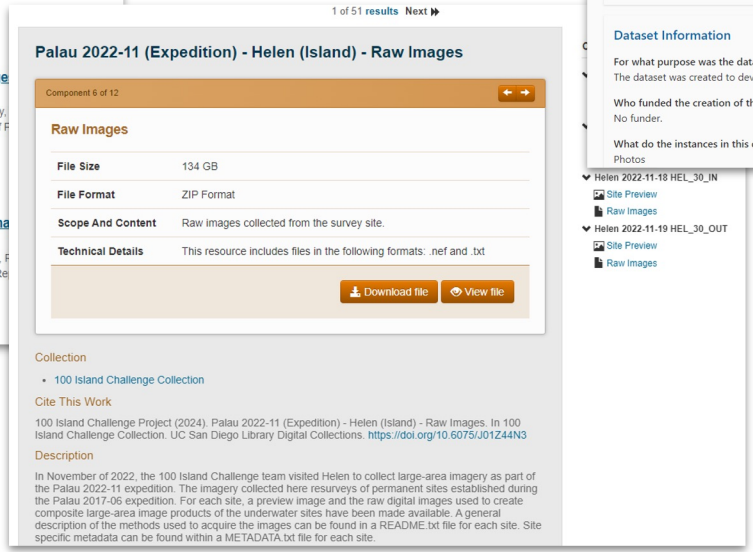
Ease of access at the scale needed for ML/AI is paramount to surface and make content visible and high value

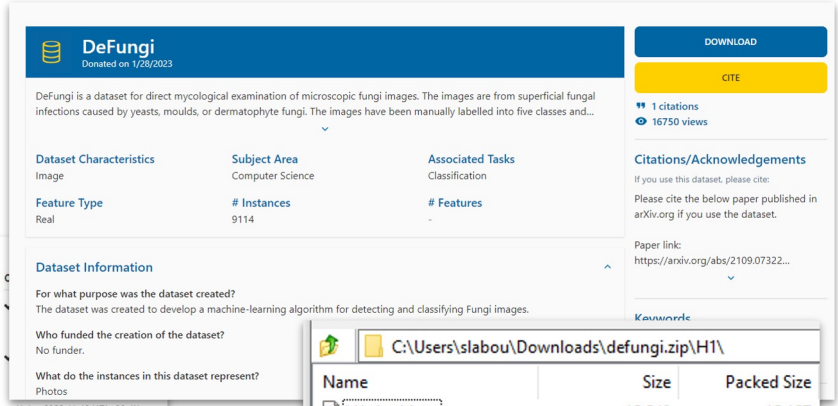
If it’s too challenging to download manually, or web scrape, or otherwise access programmatically, end users may look elsewhere

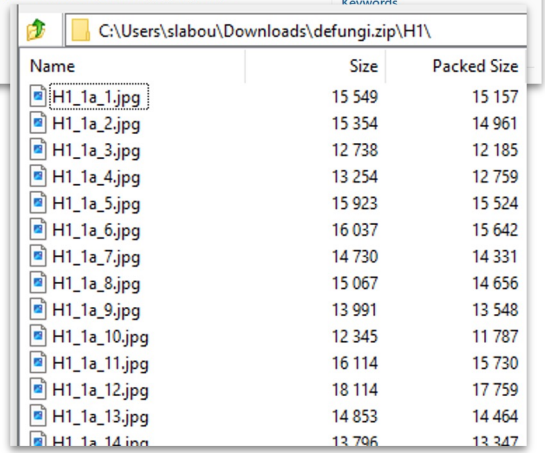
Also worth considering as an explicit aspect of accessibility in a FAIR context

Ease of (manual) bulk downloads









API (Application Programming Interface)

Changing unit of bundling objects, so users can more easily bulk download content of interest, is likely a big ask, requiring big back end changes

An API (in theory) would allow users to programmatically download the metadata & data for their items of interest

Caveat: an API is *also* a big ask, in terms of up front development, potential security concerns, and sustainability

Decisions, decisions…

To be clear: Not every repository or each piece of repository content may be suitable! Repositories don’t have to be all things to all end users

However, consider:

- Are collection items likely to be of interest to ML (or AI) practitioners?
- Is there a benefit to having repository items higher visibility/use in the area of ML/AI?
- Where and how (and if) to invest additional resources to change ingest process, metadata fields, and/or access methods?

Acknowledgements

This project was funded by a Librarians Association of the University of California (LAUC) Research Grant and the UC San Diego Library Research Data Curation Program.

Forthcoming paper in *Journal of eScience Librarianship*

Co-authors:

Abby Pennington (Library)

Ho Jung Yoo (Library)

Michael Baluja (UC San Diego grad student)