Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization
TAILOR
Grant Agreement Number 952215

Integration of OpenML with AI4EU v.1
Progress Report

Document type (nature)	Report
Deliverable No	11.3
Work package number(s)	11
Date	16-05-2022
Responsible Beneficiary	#12, TU/e
Author(s)	Joaquin Vanschoren
Publicity level	Public
Short description	OpenML is an open source platform for sharing machine learning datasets, algorithms, and models. We plan to interface OpenML with the AI4EU platform, so that AI4EU resources can be accessed via OpenML interfaces already used by many AI researchers, and OpenML resources can be viewed via AI4EU. (See also Appendix 1)

History			
Revision	**Date**	**Modification**	**Author**
version 1	2022-05-17	-	Joaquin Vanschoren

Document Review		
Reviewer	**Partner ID / Acronym**	**Date of report approval**
Peter Flach	UNIVBRIS, #16	17 May, 2022
Marc Schoenauer	INRIA, #3	17 May, 2022

Table of Contents

Summary of the report

A one-way interactivity between OpenML and the AI4EU Experiments Platform has been implemented in a collaboration between TU/e and IAIS Fraunhofer (from AI4EU). This allows users to import any OpenML dataset into the experiment platform and run experiments on this. This in turn allows AI4EU users to test models on thousands of machine learning datasets. We had planned further integration, but this hinges on AI4EU interoperability which is currently not yet available:

- Importing thousands of OpenML datasets (and other resources) into the AI4EU search engine. This is not yet possible since AI4EU currently has no way to add datasets programmatically (e.g. via an API). OpenML has 1000s of datasets and more are added every day, so doing this manually through a web form is not feasible. People can still search datasets via the OpenML website or OpenML API and import the ones they need into the AI4EU Experiments platform.
- A two-way integration that also allows sharing AI4EU resources with the OpenML ecosystem. This would allow using AI4EU resources (e.g. datasets and models) within all the AI tools connected to OpenML and thus give a much larger section of the AI community convenient access. Sadly, AI4EU still doesn't have any form of interoperability to either pull in new resources or push new resources out.

We hope that these valuable capabilities will become possible in the future development of the platform, as foreseen in the AI4Europe proposal.

1. Original goals

Our goal in this task was to fully Interface AI4EU with OpenML, one of the most widely used machine learning collaboration frameworks, which predates AI4EU and shares many of the same goals.

One of Europe's greatest strengths is that it harbors a vibrant AI community. This includes world-leading AI scientists, strong entrepreneurial hubs with hundreds of AI startups (e.g. Paris, Berlin, London, Barcelona, Amsterdam,...), world-leading machine learning tools such as scikit-learn (developed originally at INRIA, Paris) used by thousands of companies and a vital part of many other AI frameworks, a huge community of practice consisting with hundreds of thousands of data scientists who know how to use existing AI libraries to build new AI systems, and well-organized AI organizations such as CLAIRE, ELLIS, EURAI, and the European AI Networks of Excellence. They represent a large pool of AI experts which are indispensable to realize large-scale adoption and leadership in AI.

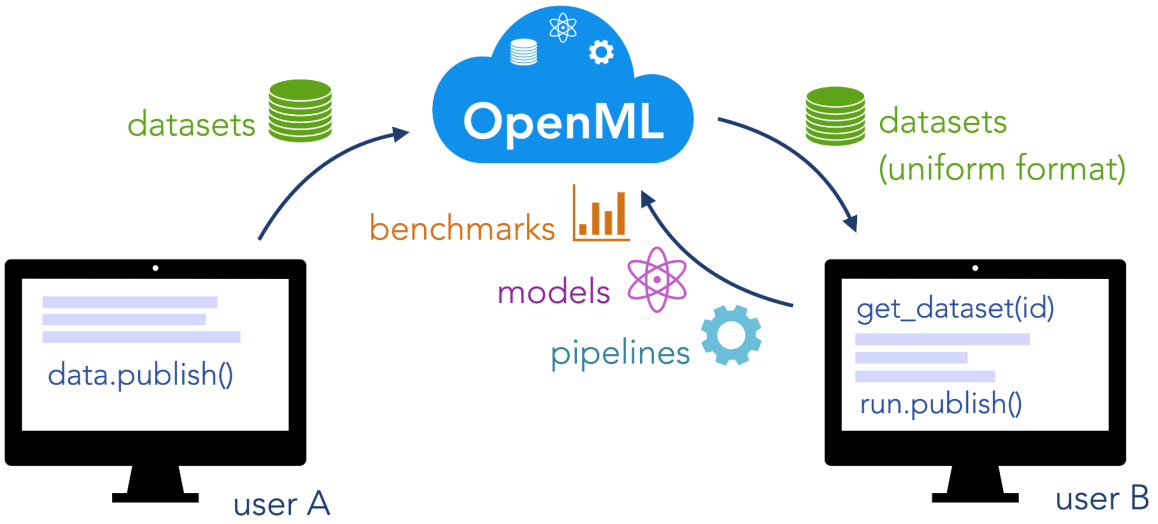
We want to give this entire community easy access to the AI4EU platform. We planned to do this by connecting the AI tools that people already use every day to the AI4EU platform through simple helper libraries or plugins. Hence, these users can continue to build AI systems in their own familiar way (which may in itself be a process refined over decades), but get additional access to AI4EU resources such as datasets, algorithms, benchmark results, as well as data storage and compute. The integration of AI4EU and OpenML will also make it drastically easier for people to share new datasets, algorithms, and benchmarks directly from these existing tools to the AI4EU platform.

Connecting these existing machine learning tools to AI4EU platform can be done efficiently through OpenML, a Europe-based open-source platform for machine learning that allows frictionless sharing of machine learning datasets, pipelines, models, and benchmarks, and that is already integrated with many popular machine learning libraries. OpenML currently contains over 20000 datasets, 8000 machine learning pipelines, and 10 million machine learning benchmarks, contributed by 13000 registered users. OpenML is used by over 250,000 users yearly, and is part of over 800 scientific studies.

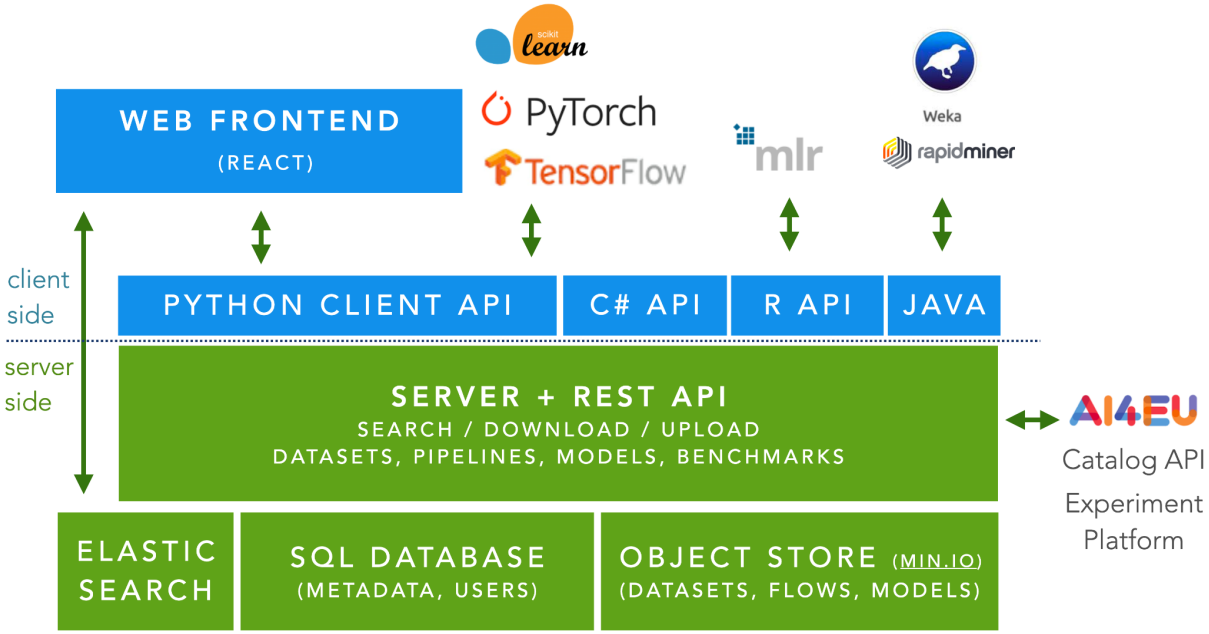
2. Approach

The integration should occur through the OpenML API (programming interface). As illustrated below, this API allows anyone to search, upload, or download AI resources, such as datasets, machine learning models, and benchmarks, with only a few lines of code. This can be done from almost any environment, such as existing machine learning tools, Jupyter notebooks, or custom programs. OpenML has official support for Python, R, and Java, but bindings in other languages (such as Julia, C++, and Rust) were also developed by various users. OpenML also has a website (www.openml.org) that gives access to all resources through the browser. This approach has a number of important benefits:

- People can use the machine learning tools they already know and love. OpenML integrations ensure that all complexity is hidden to end users.
- No scalability issues: people can run models anywhere they want, on any hardware
- People can freely build upon OpenML to offer additional services, e.g. automated model tuning, data cleaning, large-scale benchmarking, and many more. OpenML is fully open source.



A more detailed schema of OpenML is shown below, as well as how we planned to integrate OpenML with AI4EU. In short, we planned to work with the APIs that were expected to appear at the time of proposal writing[1]. More specifically we expected a 'Catalog API' that would allow us to upload and download resources from AI4EU, and access to the 'Experiment Platform' that would allow us to run experiments.



If realized, these APIs would give OpenML users transparent access to all AI4EU resources and allow them to run experiments on AI4EU. Vice versa, it would give AI4EU users access to all OpenML resources, including thousands of datasets and millions of benchmarking results.

[1] A schema and discussion on the proposed AI4EU APIs can be found in https://aclanthology.org/2020.iwltp-1.15.pdf

3. Achieved results

We completed extensions to OpenML that will simplify the interaction with AI4EU, as well as improve OpenML itself. These include:

- Support for compressed datasets formats, allowing the sharing of much larger datasets
- A modernized website giving easier access to all resources through the browser.
- General improvements to the OpenML API, especially the Python API.

In collaboration with AI4EU partners, especially IAIS Fraunhofer, we established a one-way interaction between the AI4EU Experiment platform and OpenML. This allows AI4EU to pull in any OpenML datasets and run experiments on these datasets inside the AI4EU Experiment platform. This allows, for instance, to benchmark AI4EU models across many OpenML datasets, and thus compare them to the state of the art, as results of the best algorithms on all datasets are readily available on OpenML. The elements of this interaction have been implemented in Python, and tested and debugged on 82 datasets.

We documented these extensions, including sample code and full use cases: More information on the 'OpenML Data Integration' of AI4EU is available here:

- Documentation and code
- Additional examples

As such, anyone can search for OpenML datasets on the OpenML website (www.openml.org), remember their IDs, and then download and use them inside the AI4EU experiment platform. The integration takes care of correctly converting OpenML datasets to the internal data representation used in the AI4EU experiment platform (protobuf).

4. Future work

The additional goals, especially the two-way interaction between OpenML and AI4EU, are not (yet) feasible since AI4EU doesn't yet have any APIs to access, search, or download AI4EU resources. We hope that these will become available in the near future, especially since interoperability with other platforms is a key requirement for the AI-on-demand platform. We are especially hopeful that the new AI4Europe project will help us build a full integration, and allow the wider machine learning community to use it more easily.

Appendix 1: Text of the original deliverable

Rationale: OpenML is an established platform for sharing machine learning datasets, algorithms, and models. It contains repositories of over 20000 datasets, 14000 machine learning pipelines, and 10 million machine learning experiments. It guarantees that all shared results are reproducible, and allows easy sharing and downloading of all resources via APIs and integrations into the most commonly used machine learning libraries. OpenML is an open source platform primarily developed by AI researchers in the Netherlands, Germany, and France. It is used yearly by over 150.000 people.
Approach: We plan to interface OpenML with the AI4EU platform, so that AI4EU resources can be accessed via OpenML interfaces already used by many AI researchers, and OpenML resources can be viewed via AI4EU. We will also engage in discussions with AI4EU developers to create a more integrated, vibrant ecosystem. OpenML is leveraged in WP7 and Task 9.1, and this integration will therefore also make this work directly available to the AI4EU platform.