









# Lost in a *sea* of ML models and resources?



#### Table of content

What is MLentory?

- The problem
- Introduction to MLentory
- What are we doing differently

How does MLentory work?

- Architecture Overview
- Docker
- ETL pipeline
- Backend
- LLMs
- Frontend

How are we building MLentory?

- Development Framework
- Version Control
- Code Quality
- Documentation
- Demo
- Lessons Learned
- Next steps

#### **Introducing MLentory**

- MLentory is one of the projects of NFDI4DS.
- MLentory gathers ML model information from diverse platforms, harmonizes this data into a common and standardized format and shares this information in a FAIR Digital Object (FDO) registry.



#### What are we doing differently?

- Collecting resources from different sources.
- Curating a schema for ML models called FAIR4ML.
- Enabling natural text queries to search for resources.
- Keeping track of the history of the metadata.



# How does MLentory work?





#### **Overview - Docker**





- Rapid Prototyping and Experimentation
- Reproducibility and Collaboration
- Simplified Deployment and Scalability

#### **ETL** Pipeline

 The Extractor, Transform and Load components are Python packages made for each platform.

 The scheduler is a Python script that uses Apache Airflow to trigger events.



#### Backend

- A Virtuoso database to keep the newest version of the metadata in RDF format.
- A **PostgreSQL** database to keep the history of all the extracted metadata.
- An **Elasticsearch** and **Qdrant** components to index data from the Databases.
- An **Ollama** instance to empower natural language interactions with the MLentory resources.
- An API written with **FastAPI** to serve all MLentory's services.



#### LLMs: Searching

- 1. Receive a query.
- 2. Identify key elements and expand on them.
- 3. Use the new query to extract information.
- 4. Generate questions for the user.
- 5. Recommend filters to the user.
- 6. Serve collected data.



#### LLMs: Chatting

- 1. Receive a question.
- 2. Gather all possible relevant resources.
- 3. Generate embeddings.
- 4. Compare resources and target question.
- 5. Select a subset.
- 6. Compose the query.
- 7. Get a response



#### User interface

- A Web Interface built using Nuxt3.js
- An API endpoint serving the latest version of the graph in a .ttl file





## How are we building MLentory?

#### Development framework

SemTec  ContoClue  T  T  T  T  T  T  T  T  T  T  T  T  T	STELLA	- 🗄 MLentory   🗄 MeSH   🗄 maPlan	🗄 Bioschemas   🖽 BibLabs   🖶 General   🗄	] Q1 Roadmap   🖽 MicrobiomeAnaly	sis + New v	view 25 😵	Discard Save
O Backlog 3		○ Todo ④ …	O In Progress 6 ····	O Done 2		O FullyDone 10	
<ul> <li>private-mlentory-project #37</li> <li>2025.Q1 ETL for OpenML models metadata</li> <li>(ETL OpenML models)</li> </ul>	*	private-mlentary-project #28     project #28     2025.Q1 Testing HF datasets and arXiv     New artifacts	private-mlentory-project #36     2025.Q1 Getting familiar with HF ETL for ML models     ETL OpenML models	<ul> <li>mlentory_frontend #13</li> <li>Link entities between each other.</li> <li>Frontend</li> </ul>	٩	mlentory-eti-pipeline #41     Getting familiar with FAIR4ML wrt OpenML     ETL OpenML models	Î
mlentory-etl-pipeline #44 Test OpenML data integration     ETL OpenML models	4	⊙ mlentory-eti-pipeline ≇49	<ul> <li>○ private-mlentory-project #26</li> <li>② 2025.Q1 ETL for arXiv metadata</li> <li>New artifacts</li> </ul>	⊙ mlentory_frontend #4 Generate colorful labels (Frontend)	۲	⊘ mlentory_frontend ≠10 Bug: Detailed model view does not show up for dummy data Frontend	
mlentory-etl-pipeline #60     Improve KG creation     New artifacts	۲	mientory_frontend #15     Add FDO download button.     Frontend	mlentory-eti-pipeline #47     Add metadata from arXiv papers.     New artifacts			⊘ mientory-eti-pipeline ≢54	
		mientory-ett-pipeline #43     Meteorating OpenML model information     ETL OpenML models	integrate arXiv metadata to the MLentory graph.			⊘ mlentory_frontend ≢2 Implement the Home page page Frontend	

#### **Development framework**

C C IC	sed 🛛 🕐 🖸 🖉 🖉 🖉 💭 💭 💭 zbmed-semtec/private-mlentory-project 🛛 Privat
	licarcia opened 3 weeks ago
	Deliverables and milestones
	*A new module in the extraction package.
	*A pull request detailing the changes.

Deliverables and milestones			bit2424
*A new module in the extraction package. *A puil request detailing the changes.	Labels Done enhancement		
			Туре
✓ Sub-issues (∅ 2 of 2)	Pr	eview	
Add HF dataset extraction capabilities. mlentory-etl-pipeline#45	۲		Projects
<ul> <li>O (1) Add integration of the dataset metadata to the models metadata. mlentory- ett-pipeline#46 (0 2 of 2)</li> </ul>	•		SemTec Status FullyDone -
Create sub-issue			Project MLentory
R 🚱 Ijgarcia assigned bit2424 3 weeks ago			Topic New artifacts Start date Jan 1, 2025
Ijgarcia added this to E SemTec 3 weeks ago			End date Jan 27, 2025 Actual start Jan 6, 2025
G github-project-automation moved this to Backlog in <u>SemTec</u> 3 weeks ago			Actual end Feb 3, 2025
Started 3 weeks ago			Milestone
Igarcia moved this from Backlog to In Progress in T SemTec 3 weeks ago			No milestone
Igarcia changed the title 2025.Q1 Extracting HF dataset information 2025.Q1 Add HF datasets 3			Relationships None yet
Igarcia changed the title 2025.Q1 Add HF datasets 2025.Q1 ETL for HF datasets 3 weeks ago			Development

Edit

Assiances

	veeks ago				
Closed wit	1:				
	ntec/mlentory-etl-pipeline				
	ntec/mlentory-etl-pipeline				
	ntec/mlentory-etl-pipeline				
-					
<b>O</b> 🖗 b	t2424 closed this as comp	<u>eted</u> 2 weeks ago			
<b>O</b>	t2424 closed this as <u>comp</u>	<u>eted</u> 2 weeks ago			
<ul> <li>Image: Image: Ima</li></ul>	t2424 closed this as <u>comp</u>	<u>eted</u> 2 weeks ago moved this from In Proc	arress to Done in 🖽 S	emTec 2 weeks and	
<u>о</u> р С 9	t2424 closed this as <u>comp</u> thub-project-automation	<u>eted</u> 2 weeks ago moved this from In Prog	gress to Done in 🖽 🕻	emTec 2 weeks ago	
0 g	t2424 closed this as <u>comp</u> thub-project-automation	eted 2 weeks ago	gress to Done in 🖽 🕻	iemTec 2 weeks ago	

#### Version Control



mlentory-etl-pipeline Public	☆ Edit Pins -	• Watch 3	* 😵 Fork 1 👻 📌 Starred 3 👻
ి main 👻 ిి 18 Branches 🛇 0 Tags	Q Go to file t Add file +	<> Code -	About ®
bit2424 Merge pull request #59 from zbme	xd-semtec/bug/fix_header_extraction 🚥 🗙 627f7e9+2 weeks ago	(1) 327 Commits	This repository aims at exploring APIs from different ML-related platforms so
circleci	👠 💄 Fixing CI/CD database conection errors with script ex		we can harvest and harmonize data from them
.github/ISSUE_TEMPLATE	✤ Update GitHub issue templates to YAML format		
vscode	▲ Fixed gpu not being used		কাষ্ট GPL-3.0 license
code	Improve section extraction and QA parsing with hierarchi		√ Activity
data/configuration	✤ Updated the M4ML schema and formats		
deployment	Refactor model indexing and improve code formatting		☆ 3 stars
docs	Update README.md to reflect changes in project archite		
	Undeted tests to bandle the new package names	last month	

🛉 🊧 Add metadata retrieval endpoints for model details and history... 🗅 origin/main

- Consolidate ModelHistoryController into ModelController bit2424
- Add .cursorrules to gitignore bit2424
- Participation of the search and model history controllers bit2424
- Add RO-Crate generation endpoint and controller bit2424
- Add Docker deployment configuration for GPU and non-GPU environments bit...
- Add project structure section to README bit2424

P main - P 7 Branches 🛇 0 Tag	ps Q. Go to file • Add file • Code	About
bit2424 Merge pull request #12 from	n zbmed-semtec/feature/adding_extraction_me 🚥 0143/11-last week 🕥 <b>76 Commit</b>	A web-interface to search for machine learning model metadata with natural
🖿 .nuxt		language.
🖿 .vscode	Added the vuetify library 3 months ag	10 Ar Activity
components	Add method filtering and interactive method selection in last week	
layouts		o 0 stars 0 3 watching
mock		
🖿 pages	Add extraction metadata visualization in ModelDetails c 2 weeks ag	0 Poleacos
plugins		
🖿 public		

#### **Code Quality**











#### Documentation

Inline Code Documentation	README.md in relevant folders	TDD with the big picture.
<pre>def query(self, sql: str, params: Dict[str, Any] = None) -&gt; pd.DataFrame:     """     Execute a SQL query and return results as DataFrame.     Args:         sql (str): SQL query to execute         params (Dict[str, Any], optional): Query parameters</pre>	MLentory Extraction/Transformation/Loader (ETL) Pipeline MLentory is centered around information on ML models, how to harmonize that data, and how to make it available and searchable on an FDO (FAIR Digital Object) registry. Purpose	MLentory Technical Design Document Created by Nelson David Quiñones Virgen and Leyla Jael Castro.
Returns: pd.DataFrame: Query results as DataFrame """ def get_recent_models_metadata( self, limit: int, latest_modification: datetime, threads: int = 4	To build a system that extracts ML (Machine Learning) model information from different platforms, normalizes that data in a common format, stores it, and shares it in a FDO registry to facilitate IR (Information Retrieval) and comparison/recommendation systems. This Into (Technical Design Document) will help new contributors understand and old ones remember what decisions were made on the system's design, the motivation behind them, and their impact. The document focuses on the design of the ETL pipeline to collect, transform, and store extracted information.	Table of Contents 1 Acronyms 3 Version 3 Purpose 4 Background 4
<pre>) -&gt; pu.bd(arrame: Retrieve recent models metadata from HuggingFace API. Args:</pre>	Run The Project         There are different things you can execute in this project.         • The first one is the whole ETL pipeline, which is the main component of the project. See instructions here: ETL pipeline         • The second one is the test component, which is the component that tests the ETL pipeline. See instructions here: Test Component         Background	Requirements     5       Functional Requirements     5       Non-Functional Requirements     5       Detailed design     6       ETL pipeline design:     6       Extract stage     6       Data extraction for the HF Platform     10       Transform stage     12       Load stage     15       General Considerations     17
	This project is part of the NFDI4DataScience initiative, a German Consortium whose vision is to support all steps of the complex and interdisciplinary research data lifecycle, including collecting/creating, processing, analyzing, publishing, archiving, and reusing resources in Data Science and Artificial Intelligence.	Tools     19       Frogramming Language     19       Version Control Tools     20       Database Tools     21       Deployment Tools     22       Continuous Integration and Deployment     22       Orchestration Tools     24       Implementation plan     27       Month 1: Define ETL architecture and Tool selection     27       Month 2: Access analysis and Technology familiarization.     28



#### Lessons Learned:

- You wont need it
- KISS
- Create an MVP



#### Next steps

- Make the chat affect GUI elements.
- Improve our unstructured text extraction:
  - Test more LLMs.
  - Improve context selection.
  - Explore better prompts.
- Add more resources and platforms.
- Deploying LLMs in production.
- Share our tool and get more feedback!



### Thanks for listening!