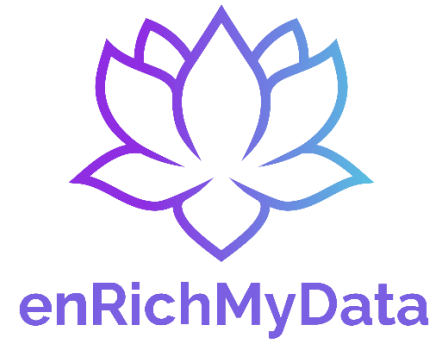

Enabling Data Enrichment Pipelines for AI-driven Business Products and Services

HORIZON-CL4-2021-DATA-01-03

D2.2 enRichMyData Tools v2

Work Package 2

Type of document:	Other
Dissemination level:	Public
Lead beneficiary:	JSI
Authors:	Technical partners contributors
Version:	1.0
Due Date of document:	30.06.2025
Delivery Date of document:	30.06.2025

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

Document History

Version	Date	Contributor	Comments
0.1	01.09.2024	Inna Novalija	Initial draft
0.2	20.09.2024	Inna Novalija, Vincenzo Masucci, Simone Marotta, Nikola Tulechki, Matteo Palmonari, Flavio De Paoli, Roberto Avogadro, Luis Rei, Raghavendra Selvan, Pedram Bakhtiarifard, Cosmin Udroiu, Gøran Svaland, Dumitru Roman	Updated the tools section
0.3	15.10.2024	WP2 technical partners	Added contributions
0.4	15.10.2024	Inna Novalija	Clean up document
0.5	15.12.2024	Inna Novalija	Updated the tools section
0.6	01.06.2025	Inna Novalija, Vincenzo Masucci, Simone Marotta, Nikola Tulechki, Matteo Palmonari, Flavio De Paoli, Roberto Avogadro, Luis Rei, Raghavendra Selvan, Pedram Bakhtiarifard, Cosmin Udroiu, Gøran Svaland, Dumitru Roman	Added information about tools licenses and documentation
0.7	15.06.2025	Inna Novalija, all partners	Adjusted based on internal review comments
0.8	23.06.2025	Inna Novalija, all partners	Clean up document
1.0	30.06.2025	Dumitru Roman	Final editing, version submitted

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

Executive summary

The comprehensive assessment and understanding of the enRichMyData toolbox is best achieved by jointly considering the two key documents, the D2.2 – "enRichMyData Tools" and D3.2 – "Integrated toolbox – Validation plan and testing and evaluation report". These documents have been developed in parallel to ensure a coherent and traceable development lifecycle from requirement definition through to validated deployment.

The D2.2 – "enRichMyData Tools v2" document serves as the foundation as it describes the key improvements and changes in each tool since the previous reporting period, associated KPIs used to monitor tool performance, licensing and openness considerations for reuse and integration. It establishes the initial baseline for functionality and performance expectations based on user needs and system requirements. The deliverable also presents the tool collections in the project that include (i) data discovery (DiscoverR), (ii) data wrapping (WrappR), (iii) data cleaning and transformation (CleanR), (iv) linking and extension (LinkR), (v) data structuring (StructR) and (vi) data classification (ClassifiR), whereas services include (i) interoperable access to datasets (ResourcR), (ii) support for scalable data enrichment pipelines (ScaleR), (iii) support to share and reuse data enrichment pipelines (ReusR), (iv) analysis of streaming data (StreamR) and (v) estimates of energy consumption and CO2 emission of data enrichment pipelines, thus supporting the implementation of "green" (I.e., energy efficient) data enrichment pipelines (GreenR).

Complementing this document, the D3.2 – "Integrated toolbox – Validation plan and testing and evaluation report" captures the iterative development and release cycles of the toolbox, including evolutions, tool updates and added features in each release of the toolbox. It documents the verification and validation processes used to assess compliance with the user requirements and the previously defined KPIs. This includes test cases, validation scenarios, and an evaluation of results against expectations.

Taken together, these documents provide a complete view of the toolbox lifecycle. They define what the toolbox is intended to do, the tool collections and expectations from each tool, while the validation report demonstrates how those intentions were implemented and verified in practice.

Readers are therefore encouraged to consider these documents as complementary and interdependent, especially when evaluating the maturity, quality, and impact of the toolbox.

In this deliverable we specifically provide updates related to tools performed in Y2 and Y3 of the project, as well as information related to tools openness and KPIs.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

Contents

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

List of Tables

List of figures

List of abbreviations

Abbreviation	Description
AI	Artificial intelligence
Apache2.0	Apache License, Version 2.0, see also: https://opensource.org/license/apache-2-0/
API	Application Programming Interface
BSD2Clause	The 2-Clause BSD License, see also: https://opensource.org/license/bsd-2-clause/
CLI	Command Line Interface
CO2	Carbon dioxide
CPA	Column Property Annotation
CTA	Column Type Annotation

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

EMD	enRichMyData
ER	Event Registry
GPL 3.0	GNU General Public License version 3, see also: https://opensource.org/license/gpl-3-0/
HTR1	dataset HardTablesR1
HTR2	dataset HardTablesR2
KPI	Key Performance Indicator
LLM	Large Language Model
MIT	The MIT License, see also: https://opensource.org/license/mit/
MRR	Monthly Recurring Revenue
RDF	Resource Description Framework
REST API	Representational State Transfer API
SSH	Secure Shell
STI	Semantic Table Interpretation
TAO	Tool Augmentation by User Enhancements and Orchestration
TRL	Technology Readiness Level
UI	User interface

List of referenced documents

Abbreviation	Document
D2.1	enRichMyData tools v1
D3.2	enRichMyData integrated toolbox v2

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

1 Deliverable overview

1.1 Deliverable purpose, scope and context

The enRichMyData project provides solutions through a suite of interoperable tools and services, collectively known as the enRichMyData Toolbox. This toolbox comprises various tools, systematically organized into collections, along with comprehensive documentation, accessible at: https://enrichmydata.github.io/toolbox/.

This document provides an overview of the tools and services developed within the enRichMyData Toolbox, with a focus on their openness, interoperability, performance (as measured by key performance indicators, KPIs), and ongoing updates.

In particular, D2.2 highlights:

- Key improvements and changes in each tool since the previous reporting period;
- Associated KPIs used to monitor tool performance;
- Licensing and openness considerations for reuse and integration.

The notion of "openness" in the enRichMyData project refers to the accessibility, transparency, and interoperability of specific tools and services, as well as the enRichMyData Toolbox as a whole. This concept emphasizes the ability of users to access, modify, and integrate various technologies without restrictive barriers, thus fostering innovation and collaboration.

D2.2 is tool-centric, reporting on the design, maturity, openness, and individual progress of each software component. The document discusses the following enRichMyData tools and services (as part of the enRichMyData Toolbox) listed in Table 1 Tools and Services Summary.

Table 1 Tools and Services Summary

Tool Name	Tool Status
SemT-X	Previously SemTUI. Existing tool, updated with new functionalities since M12 and continuously maintained.
ABSTAT	Existing tool, no updates since M12.
OntoText GraphDB	Existing tool, updated with new functionalities since M12 and continuously maintained.
OntoText Refine	Existing tool, updated with new functionalities since M12 and continuously maintained.
Alligator	The new version of s-elBat. Existing tool, updated with new functionalities since M12 and continuously maintained.
Crocodile	Created over the course of enRichMyData, updated with new functionalities since M12 and continuously maintained.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

Lion-Linker	Created over the course of enRichMyData, updated with new functionalities since M12 and continuously maintained.
Koala-UI	Created over the course of enRichMyData, updated with new functionalities since M12 and continuously maintained.
OntoText Reconcile	Since M12 as part of the project OntoText Reconcile was packaged and released a standalone tool.
Wikifier	Existing tool, since M12 the tool has been maintained and updated with new Wikipedia content.
Expert.ai Platform Document Analysis	Customized tool to make it more generic and suitable for different BCs up to month 12. The capabilities of this tool have been utilized in the development of the Expert System Generator tool.
ER Relation Classifier	Created over the course of enRichMyData, updated with new functionalities since M12 and continuously maintained.
Categorizer	Created over the course of enRichMyData, updated with new functionalities since M12 and continuously maintained.
Expert.ai Platform Document Classification	Customized tool to make it more generic and suitable for different BCs up to month 12. The capabilities of this tool have been utilized in the development of the Expert System Generator tool.
Expert.ai Expert System Generator	Created over the course of enRichMyData, built on top of Expert.ai Platform Document Analysis and Expert.ai Platform Document Classification. Regularly maintained.
LamAPI	Existing tool, updated with new functionalities since M12 and continuously maintained.
TAO	Existing tool, updated with new functionalities since M12 and continuously maintained.
SIM-PIPE	Existing tool, updated with new functionalities since M12 and continuously maintained.
Event Registry	Existing tool, since M12 the tool has been maintained and updated with constant flow of news and specific information extracted from the news.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

StreamStory	Existing tool, updated with new functionalities since M12 and continuously maintained.
Time Series Explorer	Created over the course of enRichMyData, updated with new functionalities since M12 and continuously maintained.
CarbonTracker	Existing tool, updated with new functionalities since M12 and continuously maintained.

1.2 Target audience

Deliverable D2.2 is serving as a resource for technical users - including software developers, researchers, and knowledge engineers - who actively engage with the enRichMyData tools.

1.3 Deliverable structure

This deliverable contains the following sections:

- Section 2 provides the details of the implementation for tools and services developed in enRichMyData, with subsections focused on: short introduction of specific tool/service; summary of updates from last reporting document; tool/service KPIs; tool/service openness.
- Section 3 concludes the document.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2 Implementation of tools and services

2.1 Tool collection: DiscoverR

2.1.1 SemT-X

2.1.1.1 Short introduction

SemT-X is a framework that provides a User Interface (SemT-UI) and a Python library (SemT-Py) to let users enrich tables by combining data linking and data extension services. In this case, the discovery process happens while the user is enriching a data sample with the user interface: the user can discover linking algorithms that are available and use them to bridge to existing data sources (e.g., linking cities described in a column to their id in the Wikidata KG); once the links are found using the selected linking service, additional data can be fetched from the reference data source: the user can explore data available in the data source and specify the data she/he wants to add to the table (e.g., fetching the population of each city from Wikidata). Although this "link and extend" mechanism is inspired by the principles of web-based data exploration of linked open data, SemT-X is not limited to exploiting linked data sources. For example, we tested linking and extension services from private company KGs and included the HERE geocoding service for linking addresses to coordinates; once two columns have geocoordinates, data can be enriched with the shortest route distance (as calculated by a route planning service). In other words, explorative functionalities offered by SemT-X support the discovery of (1) linking and extension services, (2) data fetched from external sources, and (2) possible flaws in the data enrichment process (e.g., wrong links). SemT-X implements, extends, and improves functionalities previously available in a similar tool named ASIA, providing a more complete user experience and the possibility of translating the enrichment operations into code via the definition of Python pipelines.

Tool documentation can be found on the following link: https://i2tunimib.github.io/I2T-docs

2.1.1.2 Summary of updates

Since M12, updates have focused on the development of the SemT-Py Python library, which enables users to access enrichment functionalities directly from a Jupyter notebook1. This allows users to either manually implement or automatically generate the pipelines designed with SemT-UI as Python scripts, which can then be executed in batch mode on new tables. Moreover, the steps in the pipelines can be implemented as Docker containers to generate pipelines executed in workflow management environments such as TAO.

2.1.1.3 KPIs

Usability is evaluated through the user experience questionnaire (UEQ), which is a widely used questionnaire to measure the subjective impression of users towards the user experience of products (https://www.ueq-online.org).

The UEQ questionnaire measures the usability and user experience of a tool using six different scales:

• Attractiveness: the overall impression of a product.

• Perspicuity: how easy is to learn and get familiar with the product?

[1] https://jupyter.org

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

• Efficiency: how much effort is required by the users to complete their tasks?

• Dependability: is the user free and in control of his actions?

• Stimulation: does product arouse excitement?

• Novelty: how innovative and creative is the product?

The UEQ does not produce an overall score for the user experience. Because of the construction of the questionnaire, it does make no sense to build such an overall score (for example by calculating the mean over all scales), since this value cannot be interpreted properly.

The scales of the UEQ are grouped into pragmatic quality (Perspicuity, Efficiency, Dependability) and hedonic quality (Stimulation, Originality). Pragmatic quality describes task related quality aspects, and hedonic quality the non-task related quality aspects.

Values between -0.8 and 0.8 represent a neutral evaluation of the corresponding scale, values > 0,8 represent a positive evaluation and values < -0,8 represent a negative evaluation.

The range of the scales is between -3 (horribly bad) and +3 (extremely good). But in real applications, in general, only values in a restricted range will be observed. It is due to the calculation of means over a range of different persons with different opinions and answer tendencies (for example the avoidance of extreme answer categories) extremely unlikely to observe values above +2 or below -2.

Table 2 SemT-X KPIs Description

What (types)	How (process)	Values (concrete values)
Usability	Usability User-study with UEQ after task completion	Usability User Experience Questionnaire (UEQ) – results above the 0.8 positive threshold for Attractiveness, Pragmatic Quality, and Hedonic Quality

2.1.1.4 Tool openness

Open-source code. Users can download the code, host it, and run it. **Tool is currently licensed under: Apache 2.0.**

2.1.2 ABSTAT

2.1.2.1 Short introduction

ABSTAT supports pattern-based profiling of very large KGs, as well as explorative searches on top of these profiles. In the backend, profiles list all the schema-level connections existing in a graph as well as several statistics, thus providing a schema-level complete summary of the data stored in the KG. In the frontend, explorative queries support humans and machines in searching relevant data ("Which connections do the graph represent between cities and sports teams?"), filter and browse all available connections (e.g., finding all the properties used to describe entities of the class dbo:City, or finding that the unique property connecting dbo:City and dbo:SportTeam in DBpedia is dbo:wikiPageWikiLink

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

and that about 9000 of these connections exist). These profiles have been proved useful to help users formulate queries over KGs with complex schemas, annotate tables, detect quality problems, and to help machines select the most relevant features.

Tool documentation can be found on the following link: https://bitbucket.org/disco_unimib/abstat/src/master

2.1.2.2 Summary of updates

The tool, part of the assets of the project, has not been updated from M12.

2.1.2.3 KPIs

ABSTAT-HD is the distributed version of ABSTAT. Also, "full" (profiling mode) refers to profiling using pattern minimalization mechanisms based on the ontology (a key distinguishing feature of ABSTAT); "core" refers to profiling without these mechanisms. ABSTAT KPIs are defined based on a more comprehensive evaluation campaigns presented in two published papers[2],[3].

Table 3 ABSTAT KPIs Description

What (types)	How (process)	Values (concrete values)
Efficiency **Focus on ABSTACT-HD, the distributed version of ABSTAT**	Efficiency Measuring execution times and controlling the impact of different variables on the execution times (e.g., size of the data, with or w/o minimalization, etc.); however, for conciseness reasons, comparative measures related to execution times and size of processed graphs are more informative (see below). Comparing execution times of ABSTAT vs ABSTAT-HD. Speed measure: speed-up factor, e.g., "*n* x = *n* times faster than". Comparing the size of the graphs that could be processed by ABSTAT to size of graphs processed with other profiling	Efficiency Speed: On a 1 node distributed infrastructure, ABSTAT-HD is on average: 3x faster than ABSTAT in core profiling mode; 42x faster than ABSTAT in full profiling mode. Graph size: experiments with ABSTAT were run, with a 5-node distributed infrastructure, on an 8.18B triple graphs, 6x larger than graphs profiled with DistLODStat (efficient profiling method computing statistics but not patterns)

[2] Alva Principe, R. A., Maurino, A., Palmonari, M., Ciavotta, M., & Spahiu, B. (2021). ABSTAT-HD: a scalable tool for profiling very large knowledge graphs. The VLDB Journal, 1-26.

[3] Spahiu, B., Palmonari, M., Alva Principe, R. A., & Rula, A. (2023). Understanding the structure of knowledge graphs with ABSTAT profiles. Semantic Web, (Preprint), 1-27.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

	techniques for which experiments are reported. Graph size measure: increased size factor, e.g., "*n* x = *n* times bigger than".	
Conciseness **Compression rate obtained with pattern minimalization vs. trivial pattern extraction (more relevant for graphs with rich ontologies)**	Conciseness Comparing conciseness of profiles when minimalization is used vs. when it is not used. Measure: compression rate = patterns with minimalization/patterns without minimalization	Conciseness Compression rate = 0,56 on DBpedia 2014 (566M triples).
Usefulness **Impact on downstream tasks**	Usefulness Previous versions of ABSTAT were evaluated on other downstream tasks. Here we focus on usefulness for end users using a flagship task: query writing, evaluated with the proxy, experiment-friendly task of SPARQL query completion. The study is based on A/B test and measures the performance of query completion with ABSTAT vs Protegè to demonstrate the usefulness of the profiles wrt the baseline approach based on ontology exploration. User sample: 117; # queries = 3. Measures: • macro-averaged increase in query completion accuracy; • macro-averaged decrease in time required for query completion.	Usefulness Macro-averaged increase in query completion accuracy vs Protege = +10% Macro-averaged decrease in time required for query completion = -49,29%

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.1.2.4 Tool openness

Open-source code. Users can download the code, host it, and run it. There is an online version that can be freely accessed. **Tool is currently licensed under: GPL 3.0.**

2.2 Tool collection: WrappR

2.2.1 OntoText GraphDB

2.2.1.1 Short introduction

GraphDB is a tool, commercially developed by Ontotext. Part of the features of the tool are co-developed in the enRichMyData project. These are the "Talk To Your Graph" LLM interface and the GraphQL[4] abstraction layer, previously a separate tool but finally fused with GraphDB.

GraphDB itself is an RDF database that is available in three versions: Lite, Standard, and Enterprise, each offering different capabilities such as managing up to 100 million RDF statements in memory, handling tens of billions of RDF statements on a single server and adding clustering capabilities as well as full-text search connectors. It is a leading database for managing semantic information with numerous large production installations in big enterprises, aiming to simplify the development and continuous operation of knowledge graphs through various integration extension points and plugins. GraphDB supports W3C standards, all major syntaxes, and query languages related to semantic repositories, and implements efficient operational and enterprise features such as inferencing, visual data exploration, high availability, distribution, and consistency checks. At a strategic level, GraphDB enables better discoverability of content assets, rich semantic context, and easier knowledge exploration and navigation, while at an operational level, it helps develop data-centric infrastructure, unlock knowledge from enterprise data silos, and establish organizational information standards compliant with W3C specifications.

2.2.1.2 Summary of updates

Since Version 11 of GraphDB, GraphQL support has been added natively to the tool. This was the functionality co-developed in the project, provided by Sematic Objects, which has been discontinued as a standalone tool.

This major evolution co-developed in the project adds a comprehensive suite of features giving a high-level of abstraction to both developers and business users, simplifying the process of building knowledge graphs and encoding business knowledge. They are different from object-relational mappings as they are dedicated cloud-native microservices that optimize query/mutation invocation and SPARQL generation for efficient data retrieval. The GraphQL endpoint is described using the Semantic Objects Modeling Language (SOML), which is a simple way to define business objects or domain objects within the platform, offering abstractions comfortable for business users.

A second major evolution co-developed in the project is the development of the "Talk To Your Graph" features in v 10.8 of GraphDB. These were a direct result of the release and success of LLM's in 2022 and provide an interface for a LLM to enriched data, structured in RDF.

Until M24 the tool has evolved with the following major features:

[4] https://graphql.org/

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

- Updated Elastic Search and RDF4J[5] (version. 10.7)
- New comprehensive documentation (version. 10.7)
- Upgraded YASGUI component in GraphDB Workbench (version. 10.6)
- Support for JSON-LD[6] 1.1 (version. 10.5)
- ChatGPT Retrieval Plugin (version. 10.4)
- Talk To Your Graph v1.0 (version. 10.3)
- Talk to Your Graph 2.0 (version. 10.8)

Tool documentation can be found on the following link: https://graphdb.ontotext.com/documentation.

2.2.1.3 KPIs

Built-in support for GraphQL (Prevously Ontotext Semantic Objects). As the focus in EMD is the LLM integration, we have evaluated the development by measuring the following KPIs:

Table 4 OntoText GraphDB KPIs Description

What (types)	How (process)	Values (concrete values)
Performance **We are measuring the LLM / RDF integration developed in the project.** **(TalkToYour Graph feature, GDB v 10.8)**	Performance Using internal question Answering Benchmarks compiled from existing commercial use-cases	Performance Accuracy values compared to the baseline of the benchmark. Target value >85% accurate queries corresponding to natural language questions. Evaluated by hand.
Integration **We are evaluating the successful integration of Sematic Objects (perv TRL7) in GraphDB (TRL9)**	Integration Suite of integration and regression tests	Integration Pass/fail No regression wrt existing functionalities of GraphDB

[5] https://rdf4j.org/

[6] https://json-ld.org/

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.2.1.4 Tool openness

GraphDB is a commercial product, but has a free-to-use version. The tool supports Open W3C Standards (RDF, SPARQL, GraphQL, SHACL) and GraphDB Free is widely used in hundreds of researchers[78]. Tool is developed as a commercial tool – proprietary with free use options as specified in https://raw.githubusercontent.com/Ontotext-AD/reconciliation/refs/heads/main/LICENSE.

2.3 Tool collection: CleanR

2.3.1 OntoText Refine

2.3.1.1 Short introduction

Ontotext Refine is a data transformation tool for converting structured data into RDF using a visual mapping interface and a CLI for automation. It allows users to clean, transform, and map tabular and lightly nested JSON or XML data into a knowledge graph, supporting various data cleaning and transformation algorithms. Users can also integrate their own custom OpenRefine extensions and automate transformations using Refine CLI. Additionally, Ontotext Refine can be used with or without backing from a semantic repository and accessed through a REST API for task automation.

Tool documentation can be found on the following link: https://platform.ontotext.com/ontorefine.

2.3.1.2 Summary of updates

Version 1.2.1, co-developed in the project was released on 19 June 2023 with an overall update of all external dependencies The update includes: libraries, docker base images, Java versions, etc.

2.3.1.3 KPIs

Table 5 OntoText Refine KPIs Description

What (types)	How (process)	Values (concrete values)
Usability **The improvements in Ontotext Refine involve the development of new features, allowing the integration of the tool in a production pipeline handling data in batch mode and at scale.**	Usability Ability to use the tool in EMD use cases and handle the large amounts of data supplied by the use cases.	Usability Ability to complete the task. (target: yes) Integration in the EMD pipeline (TAO) (yes/no) Time to complete the task

[8] **Diverse Uses of a Semantic Graph Database for Knowledge Organization and Research.** Alexiev, V. In *European Data Conference on Reference Data and Semantics (ENDORSE 2021)*, pages 47, July 2021. European Commission: Directorate-General for Informatics, Publications Office of the European Union, ISA2 Programme. Workbench is Open Source; builds upon and financially contributes to Open Source software (RDF4J). https://bibbase.org/network/publication/alexiev-diverseusesofasemanticgraphdatabaseforknowledgeorganizationandresearch-2021?_gl=1*lbn7xn*_ga*MTYxNzEyNTg0Ni4xNzQyNTY1MDk1

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.3.1.4 Tool openness

Commercial product. Free to use without limitations. Compatible with OpenRefine interfaces and the W3C Reconciliation protocol.

Tool is developed as **commercial** - proprietary with free use options as specified in https://raw.githubusercontent.com/Ontotext-AD/reconciliation/refs/heads/main/LICENSE.

2.4 Tool collection: LinkR

2.4.1 SemT-X

2.4.1.1 Short introduction

SemT-X is also considered to be part of the LinkR collection because it (1) offers access to linking services by implementing the W3C Reconciliation API specifications; (2) supports revision of linking annotations at either schema or entity level; and (3) supports the creation of enrichment pipelines. A description of its features is available in the previous section about the DiscoverR collection.

2.4.2 Alligator

2.4.2.1 Short introduction

The new version of s-elBat developed within the project is now referred to as Alligator. It provides algorithms for all Semantic Table Interpretation (STI) tasks, which altogether deliver complete automatic annotations using reference KGs such as Wikidata, DBpedia, covering entity linking, also referred to as cell-entity annotations, column-type annotations and column property annotations. The tool can also be used only for entity linking, specifying a set of columns to consider, behaving as a reconciliator, in compliance with the W3C Reconcilation Service API. Since Wikidata is becoming the reference encyclopedic knowledge base, the maintenance of the tool focuses on Wikidata. After the annotation, the data in the table can be transformed into a graph format, the structure is compliant to the desired ontology and the entities are interlinked with the KG. The tool has been competing in several international competitions established for STI, scoring frequently among top performing systems.

2.4.2.2 Summary of updates

No major functional update from M12. Most of work has been dedicated to the extensive evaluation and comparison with LM-based and LLM-based algorithms on the entity linking task (respectively, TURL, and TableLlama and GPT), which also resulted in the acquisition of novel, mostly open-source, alternative algorithms for entity linking. Additionally, a complete refactor of the codebase has been conducted, with the development of a full-fledge python library, with an increased usability when integrated with data enrichment applications, e.g., SemT-X.

Moreover, a paper has been published to evaluate it against major LLMs counterparts and can be found at: https://arxiv.org/pdf/2408.06423

Tool documentation can be found on the following link: https://github.com/enRichMyData/alligator.

2.4.2.3 KPIs

Table 6 Alligator KPIs Description

What (types)	How (process)	Values (concrete values)

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

Quality of annotations **To evaluate the quality of annotations we focus on the accuracy of the disambiguation process as done in most impactful related work. Observe that the quality of the list of retrieved candidate entities is measured to evaluate LamAPI (ResourcR)**	Quality of annotations We compare the approach with state-of-the-art approaches on a comprehensive benchmark for semantic table annotation. The quality can be measured in terms of accuracy with a set of candidates retrieved by an oracle, which is equivalent for all the approaches evaluated on the same task. As a result, we consider the following measure Accuracy (ED): number of correctly disambiguated mentions/ total number of mentions	Quality of annotations Linking (CEA) Accuracy >= 0.8
Full Semantic Table Interpretation (STI)	As Alligator provides full STI annotations, we report the F1 score for Column Type Annotation (CTA) and Column Property Annotation (CPA) tasks	CTA F1 >= 0.8 CPA F1 >= 0.8
Efficiency	Efficiency Efficiency is measured as average processing time, i.e., average time in seconds required to return a link for a mention considering a set of a fixed number of candidates.	Efficiency Average Processing Time (APT) s/mention <= 1 s/mention

2.4.2.4 Tool openness

Open-source code. Users can download the code, host it, and run it. The code for replicating the experiments is also open-source. **Tool is currently licensed under: Apache 2.0.**

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.4.3 Crocodile

2.4.3.1 Short introduction

Crocodile is a Python library for performing entity linking over large-scale tabular data. It links entity mentions in table cells to knowledge graph entities, such as those in Wikidata, by retrieving candidates from an external service and selecting the most likely match based on table context. Crocodile is designed for scalability and robustness, supporting multiprocessing and batch processing. It has been evaluated on standard benchmarks and applied to large business case tables in real-world settings.

Tool documentation can be found on the following link: https://github.com/enRichMyData/crocodile.

2.4.3.2 Summary of updates

Since M12, Crocodile has been actively extended to enhance its scalability, flexibility, and integration capabilities. It supports efficient batch processing of large tables with full multiprocessing, enabling parallel disambiguation of rows across multiple workers. Functionally, Crocodile can be seen as a simplified version of Alligator: while Alligator combines two ML models in a pipeline, Crocodile uses a single ML model for final prediction, preceded by a heuristic-based initial ranking. This design choice allows for improved performance and reduced complexity in large-scale applications.

Additional updates include:

- Implementation of a Bag-of-Words (BoW) similarity method
- Improved integration with external components such as Koala-UI and LamAPI

These updates have been evaluated on benchmark datasets (e.g., HTR1, HTR2) and applied in real-world business cases, contributing to the tool's stability and usability in production scenarios.

2.4.3.3 KPIs

Table 7 Crocodile KPIs Description

What (types)	How (process)	Values (concrete values)
Disambiguation Accuracy	Entity linking accuracy on benchmark datasets comparable to Alligator. Accuracy is measured by assuming the correct candidate is always present in the list of candidates to choose from. Specifically used: HTR1 HTR2	HTR1 accuracy > 0.80 HTR2 accuracy > 0.70
Efficiency	Efficiency is measured as average processing time, i.e., average time in seconds required to return a link for a mention considering a set of a fixed number of candidates.	Average Processing Time (APT) s/mention <= 1 s/mention

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.4.3.4 Tool openness

Crocodile is open-source, with code available at https://github.com/enRichMyData/crocodile. **Tool is currently licensed under: Apache License 2.0.**

2.4.4 Lion-Linker

2.4.4.1 Short introduction

Lion-Linker is a Python library and CLI tool for LLM-based tabular enrichment, thus performing the full Semantic Table Interpretation (STI) pipeline. It combines candidate retrieval with prompt-based disambiguation using local or remote language models (e.g., Mistral 7B via Ollama and/or OpenRouter). It supports integration with multiple retrievers (LamAPI, Wikidata Lookup, OpenRefine) and allows fine-grained control over table context, candidate formatting, and prompt tuning.

Tool documentation can be found on the following link: https://github.com/enRichMyData/lion_linker.

2.4.4.2 Summary of updates

Since M12, Lion-Linker has evolved into a practical tool for LLM-based Semantic Table Interpretation (STI). Development has focused on stabilizing the core pipeline and making it usable across different environments. The tool now supports multiple language models through Ollama (for local models) and OpenRouter (for remote APIs), allowing flexibility in model choice.

Key improvements include:

- Integration with multiple candidate retrievers: LamAPI, Wikidata Lookup, and OpenRefine;
- A configurable prompting system to guide LLM-based disambiguation
- A unified interface via Python API and CLI
- Support for Docker-based deployment for easier reproducibility

These updates improve the usability and adaptability of Lion-Linker across different linking scenarios and backends.

2.4.4.3 KPIs

Table 8 Lion-linker KPIs Description

What (types)	How (process)	Values (concrete values)
LLMs model interoperability	Employment of 2+ Large Language models	• Possibility to use all Ollama models • Possibility to use all OpenAI models • Possibility to use Grok models
Retriever Flexibility	3+ candidate retrievers	• Wikidata Lookup Service • LamAPI

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

		• OpenRefine

2.4.4.4 Tool openness

Lion Linker is open-source and available at https://github.com/enRichMyData/lion_linker. It offers a Python API, CLI interface, and Docker deployment, and is compatible with any LLM that supports local or API-based access. **Tool is currently licensed under: MIT license.**

2.4.5 Koala-UI

2.4.5.1 Short introduction

Koala-UI is a web-based user interface for inspecting and verifying entity linking results over tabular data. It connects to the backend of a linking algorithm and visualizes the results, as long as the backend supports the required W3C format, allowing users to manually validate or correct the predictions. It supports exporting validated results for use in downstream applications or as training data and is optimized for usability in expert annotation workflows.

Tool documentation can be found on the following link: https://github.com/enRichMyData/koala_ui

2.4.5.2 Summary of updates

Since M12, Koala-UI has been extended with features aimed at improving usability and better integration with the entity linking tools. The UI now provides enhanced support for visualizing and validating entity annotations at the table and cell level, including improved handling of large tables and better feedback when interacting with external services.

Key updates include:

- Improved compatibility with Crocodile, supporting real-time inspection of entity linking results
- Support for manual correction of predicted entities and exporting validated results
- Enhanced feedback messages and loading indicators to improve the user experience during annotation and correction
- General UI improvements to streamline the annotation workflow and reduce friction for expert users

These improvements were guided by user feedback and have been used in annotation workflows to support the validation of results produced by tools such as Crocodile, Alligator and Lion-Linker.

2.4.5.3 KPIs

Table 9 Koala-UI KPIs Description

What (types)	How (process)	Values (concrete values)
Usability	Tool usability was evaluated through a user questionnaire	The feedback was positive, and most users described the interface as intuitive and easy to use

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

Integration	Koala-UI enables easy revision of annotations via communication with Crocodile endpoints	All tested endpoints functioned as expected

2.4.5.4 Tool openness

Koala-UI is open-source, with code available at https://github.com/enRichMyData/koala_ui. **Tool is currently licensed under: MIT License.**

2.4.6 OntoText Refine

Ontotext Refine is part of both LinkR and CleanR tool collections was described in Section 2.3.1.

2.4.7 OntoText Reconcile

2.4.7.1 Short introduction

Ontotext Reconcile is a service developed by Ontotext that allows users to leverage the GraphDB Elasticsearch connector to reconcile textual references against their own data. The service enables users to create Elasticsearch indexes from their RDF repository, feed data, and provide an Elasticsearch query template using the Mustache templating system[9]. Additionally, Ontotext Reconcile can automatically generate default queries if customization is not desired. However, it is primarily intended for reconciliation against textual references and may require additional logic for reconciling identifiers.

Tool documentation can be found on the following link: https://www.ontotext.com/company/news/graphwise-reconciliation-build-custom-reconciliation-services-over-your-knowledge-graphs.

2.4.7.2 Summary of updates

Since M12 as part of the project Ontotext Reconcile was packaged and released a standalone tool.

2.4.7.3 KPIs

Table 10 OntoText Reconcile KPIs Description

What (types)	How (process)	Values (concrete values)
Performance **The ability to perform reconciliation task over an arbitrary RDF graph at speed and scale. We compare the service to similar reconciliation services (such as Wikidata), which are very slow and are looking for a much faster realization.**	Performance Measuring the time to execute a reconciliation candidate selection task for large input files	Performance Benchmark. Ability to process 10K lines of input data in less than 1 minute

[9] https://github.com/mustache/mustache

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.4.7.4 Tool openness

It depends on the commercial version of Ontotext GraphDB. Free to use with own Elastic Search Implementation. The **Tool is developed as a commercial product.**

2.5 Tool collection: StructR

2.5.1 Wikifier

2.5.1.1 Short introduction

The JSI Wikifier is a web service that takes a text document as input and annotates it with links to relevant Wikipedia concepts. JSI Wikifier is regularly maintained, so the most recent Wikipedia dump is available for semantic annotation.

Tool documentation can be found on the following link: https://wikifier.org/info.html.

2.5.1.2 Summary of updates

Wikifier has been regularly updated with the latest data from Wikipedia.

2.5.1.3 KPIs

Table 11 Wikifier KPIs Description

What (types)	How (process)	Values (concrete values)
Scalability **Number of documents annotated**	Scalability The semantic annotation process is based on regular usage of Wikifier API and additional annotation when a new set of documents becomes available.	Scalability Values: minimum 1.000.000 documents
Annotation precision	Annotation precision Measuring Precision Top N for semantic annotation/wikification task based on developed gold standard.	Annotation precision Annotation Precision Top 1 > 0.8

2.5.1.4 Tool openness

Free service for common usage and partners. Free registration by email. Commercial use negotiable. **The tool is developed as proprietary product.**

2.5.2 Expert.AI Platform Document Analysis

2.5.2.1 Short introduction

Expert.AI Platform's Document Analysis APIs are a solution in Natural Language Processing and Understanding. This technology provides the power of deep language intelligence, allowing for comprehensive document analysis, which features text subdivision, part-of-speech tagging,

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

morphological analysis, lemmatization, syntactic analysis, and semantic analysis. In essence, it comprehensively analyses and understands text, providing valuable insights to make sense of unstructured data. This capability extends across multiple languages, including English, Spanish, French, German, and Italian. One of the main draws of this technology is that it allows for a fast and easy understanding of the essence of textual data – the Key Elements feature extracts key lemmas, phrases, Knowledge Graph concepts, sentences, and Knowledge Graph topics from text, each accompanied by its relevance score. This empowers users to grasp the crucial elements within their texts.

With the Document Analysis capabilities, users can enrich their textual data by identifying named entities, including people, organizations, places, and values like currency amounts and percentages. These entities are linked to open data sources such as Wikidata and GeoNames, adding an additional layer of knowledge to the data. The Relations feature provides insights into the semantic relationships within text, going beyond keywords to answer complex questions like "who did what when?" and uncovering intricate information. All results are conveniently accessible in JSON format, facilitating easy integration into composite pipelines for solving complex use cases. The tool's versatility and detailed insights make it a valuable tool for a wide range of applications, providing a solid foundation for innovative projects.

Tool documentation can be found on the following link:

https://docs.expert.ai/platform/latest/nl-flow-api/

2.5.2.2 Summary of updates

As of this reporting period, there have been no updates specific to the Expert.AI Platform Document Analysis tool since D2.1. However, it is important to note that the capabilities of this tool have been utilized in the development of the Expert System Generator tool.

2.5.2.3 KPIs

Table 12 Expert.ai Platform Document Analysis KPIs Description

What (types)	How (process)	Values (concrete values)
Quality of extracted information **What you plan to measure (types): The quality of extraction in terms of entities extracted**	Quality of extracted information Over the last few years, the application has also been tested in other contexts with manual verification and by customers who have provided us with the results of their experiments.	Quality on NER Values for the measure (concrete values): Precision, Recall and F-score > 80%

2.5.2.4 Tool openness

Usable for the project period and 6 months after without credentials. Freemium License (1 month or 1000 text analyzed). **The tool is developed under: Freemium License.**

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.5.3 ER Relation Classifier

2.5.3.1 Short introduction

The *Event Registry (ER) Relation Classifier* identifies and categorizes relations and events in text using a predefined taxonomy. As part of its core functionality, it also performs Named Entity Recognition and enriches text using Wikifier. The tool's internal pipeline is very efficient, with hundreds of news articles to be classified per second, effectively enabling real-time monitoring of relations/events. It is an integral part of the Event Registry service, with outputs accessible via both the web platform and the ER API.

For the enRichMyData project, we leveraged the ER data management interfaces to create and annotate examples for newly defined relation classes. These examples were integrated into the supervised training and test sets, after which the classifier was retrained. As a result, the system can now detect and track these new event classes in real-time news streams via the ER API. Additionally, we applied the updated classifier to a two-year archive of news articles, selected via Wikifier concepts related to AI, to conduct retrospective analyses for the Innograph use case.

Documentation for the tool is available at:

- Event Registry Platform: https://eventregistry.org
- GitHub Repository: https://github.com/EventRegistry

2.5.3.2 Summary of updates

Initially, we introduced a broad relation class called *"technology use"*, designed to capture any instance of an organization (e.g., a company or government agency) utilizing a technology (e.g., "3D printing", "magnetic resonance imaging"). To enhance analytical precision and better focus on specific innovation areas, we subdivided this general class into two more specific relation types: *"technology use AI"* and *"technology use blockchain"*. These finer-grained classes enable us to better identify and evaluate the adoption of particular technologies.

The primary objective was to track the adoption of Artificial Intelligence (AI), which is the central focus of the InnoGraph use case as it identifies instances of an organization using an AI technology. The inclusion of blockchain served both as a demonstration of the model's flexibility across diverse technologies and as a comparative benchmark.

The work throughout the project progressed as follows:

- **Year 1:** Developed the initial *"technology use"* relation class
- **Year 2**: Introduced the refined classes "technology use AI" and "technology use blockchain"
- **Year 3**: Enhanced the classifier through the addition of new annotated training and test data. Applied the retrained classifier to extract instances of *"technology use AI"* from two years of archived news articles

2.5.3.3 KPIs

Table 13 ER Relation Classifier KPIs Description

What (types)	How (process)	Values (concrete values)

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

Scalability	Scalability Number of articles processed	Scalability Number of instances processed over 10.000.000 Relevant news articles identified > 1.000.000c
Effectiveness: Classifier Performance	Effectiveness: Classifier Performance F1 measure (per class) Technology-use Technology-use + AI Technology-use + Blockchain	Effectiveness: Classifier Performance F1 measure (per class) > 0.7

2.5.3.4 Tool openness

Subscription-based commercial service. **The tool is developed as a commercial product by Event Registry.** JSI used this tool and associated support tools to add new classes of relations and analyze historical news data.

2.6 Tool collection: ClassifiR

2.6.1 Categorizer

2.6.1.1 Short introduction

Categorizer is a multi-faceted research effort encompassing the semi-automatic adaptation of existing taxonomies to new use cases, automatic taxonomy alignment, and the development of corresponding text classifiers. While the methods employed are broadly applicable, we focused specifically on the needs of InnoGraph business case as the primary use case throughout the conception, planning, and development phases.

For InnoGraph, we developed a custom knowledge taxonomy derived from the Cooperative Patent Classification (CPC) system. This taxonomy enables the classification of diverse textual sources—such as patents, code documentation, news articles, and scientific publications—into more than 80 distinct innovation domains. Using a multi-label classification approach, it captures the interdisciplinary nature of modern innovation.

The resulting taxonomy and the text classifier, which was developed by fine-tuning pre-trained transformer language models, are publicly available and open access. Categorizer remains under active research and development.

Tool documentation can be found on the following link: https://github.com/elmotamedi/TaxoRefine.

2.6.1.2 Summary of updates

All components of the tool—including the adaptation of the taxonomy, algorithmic approaches to taxonomy alignment, classifier development, evaluation, and dissemination—were carried out as part of the InnoGraph business case.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

Key updates and milestones include:

- **Taxonomy Design and Adaptation**: We created a domain-specific innovation taxonomy for InnoGraph, adapted from the Cooperative Patent Classification (CPC) system. This required the identification, restructuring, and extension of relevant CPC codes to better reflect interdisciplinary innovation areas.
- **Taxonomy Matching and Alignment**: We implemented semi-automatic and automatic methods to map external and internal taxonomies, enabling reuse of structured knowledge across different datasets and text domains.
- **Classifier Development**: We developed a multi-label text classifier using fine-tuned transformer-based language models (e.g., BERT variants). The classifier is capable of labeling texts from varied sources (patents, scientific literature, technical documentation, news) according to the custom innovation taxonomy.
- **Supervised Dataset**: A supervised dataset was generated by subsampling the Google Patents dataset.
- **Evaluation and Refinement**: Classifier performance was evaluated through standard metrics (e.g., F1 score per class).
- **Deployment and Reusability**: The resulting taxonomy and classifier are open access, well-documented, and designed to be reused in future projects.
- **Research Outputs**: Several scientific papers describing the methodology and findings have been produced or are in preparation. These contributions ensure the visibility and reproducibility of the work within the research community.

In summary, *Categorizer* has matured from a conceptual research idea into a fully operational and extensible system. Its components—taxonomy, methods, and classifiers—are positioned to serve both current and future innovation-tracking applications beyond InnoGraph business case.

2.6.1.3 KPIs

Table 14 Categorizer KPIs Description

What (types)	How (process)	Values (concrete values)
Effectiveness: Classifier Performance	Effectiveness: Classifier Performance Identifying and improving the categorized performance metrics. Enhancing the classification task by balancing the dataset using balancing techniques for multi-label problems and leveraging larger pretrained models. Examining the different knowledge fields	Effectiveness: Classifier Performance F1 measure > 0.7

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

	to better understand the variations in classifier performance across them.	

2.6.1.4 Tool openness

Open model available for download. The tool is currently licensed under: **MIT license**.

2.6.2 Expert.AI Platform Document Classification

2.6.2.1 Short introduction

Expert.AI Platform's Document Classification APIs are a solution in Natural Language Processing and Understanding. This technology offers a range of taxonomies along which it's possible to categorize textual data, including I.P.T.C. Media Topics, GeoTax, Emotional Traits, and Behavioral Traits. Our supported languages include English, Spanish, French, German, and Italian. The outcome of our classification results is provided in JSON format. For example, we submitted a document about the European Union's carbon emission reduction efforts, and it was categorized under "Environmental Policy" and "Air Pollution" within their respective taxonomies. Similar results can be expected uniformly across all our available taxonomies.

Our geographical taxonomy, GeoTax, a feature within our suite, adeptly classifies text based on geographical references, covering a spectrum from countries to cities.

Our Emotional and Behavioral Traits taxonomies offer a nuanced understanding of human emotions and personality traits in both English and German texts. For instance, the analysis of a text detailing a farmer's association with the Sustainable Australia Party reveals emotions such as 96% disappointment and 3% happiness. Similarly, a news piece on Greta Thunberg unveils behavioral traits of 71% initiative and 29% competence.

Our document classification technology is just the starting point. It enables its users to gather actionable insights which can then be integrated into diverse applications, fostering innovation in subsequent use cases across industries.

Tool documentation can be found on the following link: https://docs.expert.ai/platform/latest/nl-flow-api/.

2.6.2.2 Summary of updates

As of this reporting period, there have been no updates specific to the Expert.AI Platform Document Classification tool since D2.1. However, the capabilities of this tool have been utilized in the development of the Expert System Generator tool.

2.6.2.3 KPIs

Table 15 Expert.ai Platform Document Classification KPIs Description

What (types)	How (process)	Values (concrete values)
Quality of document classification	Quality of document classification Comparing the context identified by the tool with the	Quality of document classification

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

The quality of document classification in terms of F1, precision, recall	manually reported category using Spend Network Data	Values for the measure (concrete values): F1,Precision,Recall > 80%

2.6.2.4 Tool openness

Usable for the project period and 6 months after without credentials. Freemium License (1 month or 1000 text analyzed). **The tool is currently licensed under: Freemium License.**

2.6.3 Expert.AI Expert System generator (Q&A)

2.6.3.1 Short introduction

Expert.AI Answers allows for the generation of expert systems, these expert systems capture the knowledge contained in one or more PDF documents through hybrid content indexing that is based on symbolic/semantic analysis and the use of neural networks the information is organized into a RAG structure, the knowledge thus perfectly organized can then be queried with pointed questions posed in natural language, the answer can be provided either through ranking sorted by confidence level, also reporting the reference to the part of the pdf that contains the answer, and this allows for traceability and explainability of the answer, but the answer can also be generated through the use of LLMs but which are very controlled to avoid hallucinations.

Tool documentation can be found on the following link: https://drive.google.com/file/d/1pXRuh5xSF0vn_rcSGysABo_2AZt7EPnH/view?usp=sharing.

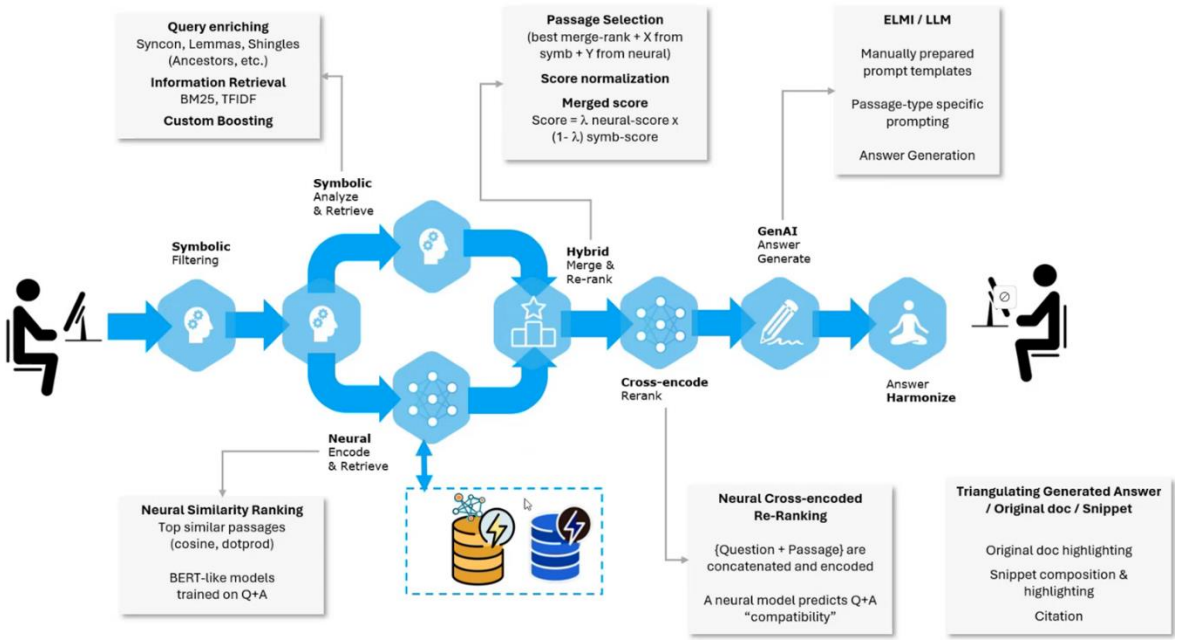

Figure 1 Expert.ai Expert System generator (Q&A) pipeline

2.6.3.2 Summary of updates

The updates for the tool included:

- Restructuration of tool APIs and addition of new endpoints for easier maintenance and usage;

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

- Improvement in the parsing of tables in PDF files;

2.6.3.3 KPIs

Table 16 Expert.ai Expert System generator (Q&A) KPIs Description

What (types)	How (process)	Values (concrete values)
Quality if context identification **We are measuring the correctness of the answers provided both by analyzing the generated answer and by analyzing the position in the ranking of the correct answer relative to all those provided**	Quality if context identification How to measure (process): Comparing the context identified by the tool with the manually reported by BOSCH and PHILIPS	Quality if context identification Values for the measure (concrete values): Number of correct and acceptable answers > 80% MRR> 85%

2.6.3.4 Tool openness

Usable for the project period and 6 months after without credentials. Freemium License (1 month or 1000 text analyzed). **The tool is currently licensed under: Freemium License.**

2.7 Service collection: ResourcR

2.7.1 LamAPI

2.7.1.1 Short introduction

LamAPI, has the goal of digesting a data source that we want to link to, in our case a knowledge graph, and preparing data structures that support linking and reconciliation. The primary feature is obviously to index the knowledge graph to support fast entity lookup and follow-up matching operations. However, LamAPI also implements a variety of other features that all help to back the linking task. It supports the retrieval of useful data from the indexed graph, including entity relations and literal values. It stores links across graphs (e.g., sameAs), so that they can be exploited for cross-graph lookup. More recent features support entity embeddings, such as RDF2vec, which can be used to estimate the relatedness between candidate matches for different values in the table. Finally, it also provides services to analyze the input data, such as a data type identification service, which is crucial for identifying specific data types such as numbers, dates, emails, telephone numbers, street addresses and URLs. LamAPI has been extensively tested and engineered with large knowledge graphs such as Wikidata and DBpedia, which helps to make its configuration easier when a user wants to make a new knowledge graph, e.g., a proprietary graph, available for linking and reconciliation.

Tool documentation can be found on the following link: https://github.com/unimib-datAI/lamAPI.

2.7.1.2 Summary of updates

Updates from M12 include two main types of changes to LamAPI. We introduced a first approach to filter candidates by coarse types, using the four classes PERSON, LOCATION, ORGANIZATION and OTHER; this approach is based on a method that clusters Wikidata types into these coarse types, by re-implementing an approach proposed in the state-of-the-art. This functionality is targeted to improve the capability to filter candidates that have specific types.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

We refactored the code to improve the usability of LamAPI (specifically for the lookup endpoint) by adding: the possibility to specify the language; the possibility to specify the NER type (see above); the possibility to specify that QIDs must be in response. Moreover, a soft filtering strategy over types, given the Transitive Closure (TC) of the underlying Knowledge Graph (KG), has been implemented.

2.7.1.3 KPIs

Table 17 LamAPI KPIs Description

What (types)	How (process)	Values (concrete values)
Efficiency **Capability to retrieve candidates upon a given query efficiently.**	Efficiency Efficiency is measured as average retrieval time (seconds) per entity	Efficiency Average Retrieval Time (ART) s/mention <= 0.5 s/mention
Quality of entity retrieval **We evaluate the quality of the ranked list of candidates retrieved using LamAPI.**	Quality of entity retrieval Measures: Coverage of retrieved entities is measured as the number of correct candidates found in the top-k candidate list for a dataset. Comparison of coverage of LamAPI and the Wikidata lookup service (used in several related work)	Quality of entity retrieval Coverage >= 0.8
Impact of types during entity retrieval	Candidate retrieval with enhanced type search (given by the transitive closure) improves coverage on domain specific data	Delta% >= 0.1

2.7.1.4 Service openness

Open-source code. Users can download the code, host it, and run it. Code for replicating the experiments is also open source. **The tool is currently licensed under: Apache 2.0.**

2.8 Service collection: ScalR and ReusR

2.8.1 TAO

2.8.1.1 Short introduction

TAO is an open-source, lightweight, generic, extensible, and distributed orchestration framework and implements both ScalR and ReusR functionalities. It allows to reuse (i.e., integrate) enRichMyData tool

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

collections as well as specify custom user tools. This framework enables processing composition and distribution in such a way that end users could define processing workflows by themselves and easily integrate additional processing modules, without extensive programming knowledge.

The following operations are possible to be made by TAO:

- User management and login.
- Creation of descriptors for the Docker images that are used in the pipelines.
- Creation of processing component descriptors for the tool application or services that are used in pipelines.
- Creation of workflow definitions by chaining and interconnecting the processing components of the tools (the output of a processing component is linked to the input of the next processing component) in order to create the pipelines.
- Reuse of existing workflow components.
- Version control for components (Docker components, processing components or workflows) to have a history of those components.
- The platform provides thus a means for orchestrating heterogeneous processing components and libraries to process the enRichMyData specific data. This is achieved in the following steps:
- Preparation of resources (including processing components or tools) and data input.
- Definition of a workflow as a processing chain.
- Execution of workflows, scaling up from one to as many nodes as made available.
- Retrieval and visualization of the results, which allows one to see what is executing in the system and where it is executing, as well as the system resource usage (CPU, memory, storage space).

From the TAO web interface, the user has the possibility to add resources (application contained a workflow by adding the desired data source to be used as the initial data provider in the workflow and also to chain the desired processing modules.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

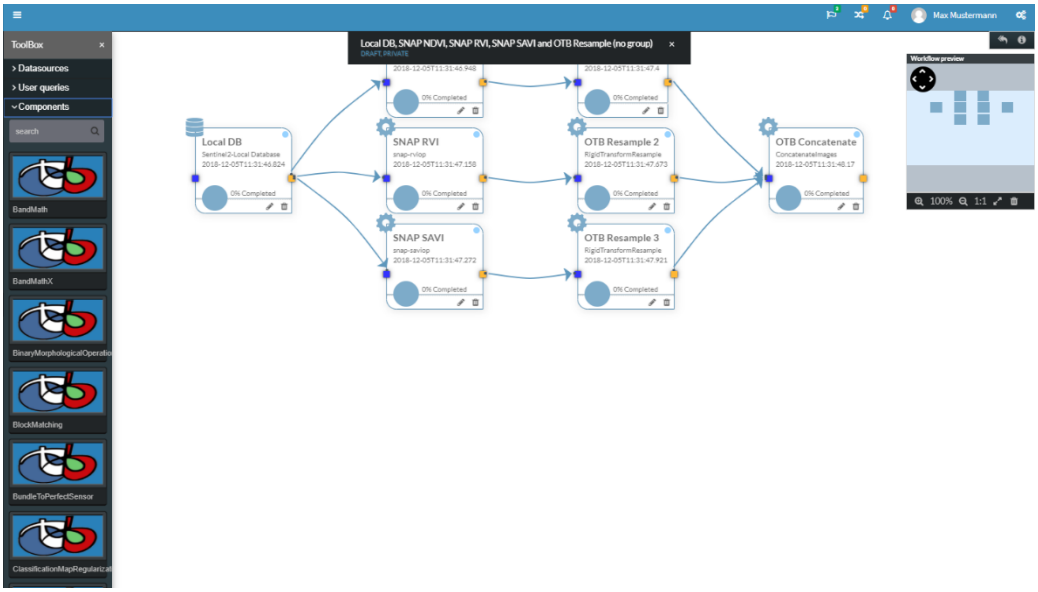

Figure 2 TAO workflow design

The chaining of the modules is performed by the user by drag and dropping the available processing components on the drawing canvas and chain them using directed connectors.

Another relevant component is the Distributed Resource Management Application API (DRMAA)[10], which provides standardized access to the distributed resource management systems for execution resources. It is focused on job submission, job control, reservation management, retrieval of jobs, and machine monitoring information. Currently, there are supported DRMAA implementations for local, remote SSH, Torque[11], SLURM[12] and Kubernetes[13] executions.

Tool documentation can be found on the following link: https://github.com/tao-org.

2.8.1.2 Summary of updates

Since M12, TAO was evolved in order to bring new features that can be used for creating generic pipelines and for execution scaling. Thus, the functionalities concerning the earth observations specificities were decoupled from the core of TAO. The Kubernetes as a resource manager was added to TAO for supporting the execution of pipelines in Kubernetes clusters. The possibility of exporting the TAO pipelines from TAO's internal format into CWL and Argo workflow format and integration with SIM-PIPE was also implemented, allowing now creation of projects, dry runs and performing execution of dry-runs in SIM-PIPE. The CarbonTracker functionalities for estimating CO2 consumption during the execution of the pipelines was integrated into TAO by identifying consumption on each step and performing aggregation at job execution level. Documentation (software user manual, changelogs etc.) was created and published. An installation package for automatic installation was created including enRichMyData tools and services that can be deployed and integrated with TAO. Included enRichMyData tools as application containers, created processing components for the tools

[10] https://www.drmaa.org/

[11] https://github.com/xrg/torque

[12] https://slurm.schedmd.com/documentation.html

[13] https://kubernetes.io/

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

and developed helper tools for different operations in the pipelines (merging files, adding adaptors for different services) and implemented and executed the pipelines for SpendNetwork, JOT, Philips, Bosch and InnoGraph use cases. Pipelines deployed along with the installation package as optional package.

2.8.1.3 KPIs

Table 18 TAO KPIs Description

What (types)	How (process)	Values (concrete values)
Performance **Execution performance**	Performance Having a pipeline with several tools, measure the overhead of TAO in the execution process	Performance Max 2 seconds overhead between scheduling two pipeline steps.
Scalability **Scaling performance**	Scalability Execute workflow with and without scaling and measure improvements	Scalability Improve performances with at least 25% when having an additional node
Usability **How easy to use and build pipelines**	Usability User-study with task completion	Usability Time to complete tasks; scores reported in the questionnaire

2.8.1.4 Tool openness

The TAO framework, which is used for tools integration and creation of pipelines, is available as opensource on GitHub in the following location https://github.com/tao-org. It has an open architecture allowing external tools to be integrated and used in the pipelines. The software user manual (SUM) provides detailed documentation on how new tools can be integrated. For the integration of the tools, there is no limitation, any tool that can be executed (either through command line or through the invocation of an API, no matter if they are internal or 3rd party tools that are necessary for business cases) can be integrated as TAO always uses Docker images encapsulating the tools. All the Docker images and tools are added into TAO through their descriptors that define what applications are available and their invocation and interconnectivity information.

The execution and orchestration framework represented by TAO allows also to be used via the exposed REST API interface, all the functionalities available from the web interface being available also through the REST API. Additionally, TAO is implementing an OGC WPS (Web Processing Service) client and server allowing both to invoke external tools providing this interface but also to be invoked by external application in order to invoke the implemented pipelines. To be able to execute workflows implemented in TAO on other external environments or platforms, the export from TAO internal format into CWL, JSON, Bash or Argo is also supported.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

As TAO has a pluggable architecture, new resource managers, data sources or processing services can be further added. Thus, for example, the current list of resource managers represented by the local invocations, SSH, Slurm, Torque or Kubernetes, can be extended programmatically with new ones.

During each release, an installation package is created containing all the installation scripts and the needed libraries to quickly set up a fully functional toolbox instance. The installation package contains all dependencies needed (database creation, additional packages installation) but also all the docker images for the tools supported in that version. The installation package is published and available for free download, installation and usage on local premises or in cloud environments. The installation of the toolbox is straightforward and very simple, by running a single script all installations and configurations being automatically performed. **The tool is currently licensed under: GNU General Public License version 3 – Open Source Initiative.**

2.8.2 SIM-PIPE

2.8.2.1 Short introduction

SIM-PIPE is a Kubernetes and Argo Workflows based tool to execute dry-runs of data pipelines and provide detailed metrics and predictions of resource consumption for the individual steps of data pipelines.

Tool documentation can be found on the following link: https://github.com/DataCloud-project/SIM-PIPE.

2.8.2.2 Summary of updates

Since M12, SIM-PIPEs controller API has been extended with endpoints to perform and return resource predictions for CPU, memory and duration. Linear regression and power law regression methods for resource predictions has been implemented in the backend.

In order to perform emulation of hardware, we developed scripts for creating virtual machines that can be added to the Kubernetes cluster. By specifying node affinity, data pipeline steps can be scheduled to run on those nodes.

To facilitate prediction of energy and CO2 consumption of data pipelines, a docker image containing the GreenR component CarbonTracker was created. This docker image was extended with a FastAPI server to query CarbonTracker using resource metrics from SIM-PIPE. This enables SIM-PIPE to report CO2 and energy consumption of data pipeline dry-runs.

2.8.2.3 KPIs

Table 19 SIM-PIPE KPIs Description

What (types)	How (process)	Values (concrete values)
Interoperability **The ability for TAO to use the SIM-PIPE API**	Interoperability Integrate TAO to make use of the SIM-PIPE API to execute dry-runs, perform resource predictions and retrieve resource predictions on the SpendNetwork use case.	Interoperability Boolean values. Is TAO able to us SIM-PIPE API to: Execute dry-runs

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

		Perform resource prediction Retrieve resource prediction Retrieve resource prediction without any further processing of the data

2.8.2.4 Service openness

SIM-PIPE is available as opensource software in the following GitHub repository: https://github.com/DataCloud-project/SIM-PIPE. Description of how to install and run the software is provided in the README. The tool can be used through its GraphQL API, which schema documentation can be retrieved using common API query tool such as Insomnia or Apollo GraphQL studio. The tool can also be used through its frontend (GUI). **The tool is licensed under: Apache 2.0.**

2.9 Service collection: StreamR

2.9.1 Event Registry

2.9.1.1 Short introduction

Event Registry is an AI-powered platform that continuously monitors and analyzes global news from over 100,000 sources in real time. It automatically detects, clusters, and categorizes news articles into structured events, enabling users to track emerging topics, trends, and developments across languages and regions. The service provides advanced search, event filtering, and trend detection tools, making it valuable for media monitoring, competitive intelligence, and data-driven decision-making. In enRichMyData Event Registry service is used as news provider for InnoGraph business case.

Tool documentation can be found on the following link: https://eventregistry.org and https://github.com/EventRegistry.

2.9.1.2 Summary of updates

Released before start of project. Constantly updated world's leading news intelligence platform.

2.9.1.3 KPIs

Table 20 Event Registry KPIs Description

What (types)	How (process)	Values (concrete values)
Scalability	Scalability	Scalability
Number of documents processed	Constant document streaming	Number of documents provided > 1.000.000

2.9.1.4 Service openness

Event Registry can be used by anyone given a subscription to the API. Subscription based commercial service available to anyone who register. **The tool is developed as commercial product.**

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.9.2 StreamStory

2.9.2.1 Short introduction

StreamStory is a multi-scale data analysis tool for multivariate continuously time-varying data streams. It represents the data streams in a qualitative manner using states and transitions. Users can upload their own dataset or use one of the pre-loaded datasets.

StreamStory can also be used as a monitoring tool, showing in real-time the state of the monitored process, activity and anomaly detection. During enRichMyData project StreamStory API has been developed that supports the following operations:

- Get a list of models
- Get model by Universally Unique Identifier
- Delete model by Universally Unique Identifier
- Get data classification

Tool documentation can be found on the following link: https://github.com/E3-JSI/StreamStory2.

2.9.2.2 Summary of updates

From M12, StreamStory API has been developed and opened to the general usage (http://streamstory.ijs.si/api/v1/docs/#/). Currently StreamStory user can use the tool via online user interface, download the source code from GitHub or use StreamStory programmatically (via API).

2.9.2.3 KPIs

Table 21 StreamStory KPIs Description

What (types)	How (process)	Values (concrete values)
Usefulness	Usefulness	Usefulness
User evaluation survey	StreamStory tool (user evaluation survey)	User evaluation survey

2.9.2.4 Service openness

Open Source Free Software available for download. Open API. **The tool is currently licensed under: MIT License.**

2.9.3 Time Series Explorer

2.9.3.1 Short introduction

This tool is focused on data cleaning, de-noising, graph methods application. Implementation of descriptive statistics, filters, transition diagrams, visualizations and predictions are done for specific time series data (BGRIMM business case).

Tool documentation is available based on request on the following link: https://github.com/E3-JSI/gm-sequence-mining.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

2.9.3.2 Summary of updates

Created over the course of enRichMyData, updated with new functionalities since M12, that are specific for BGRIMM business case. The anomaly detection is built upon graph-based time series modeling - transformation of multivariate time series into graph representations to predict anomalies (overloads) in a milling plant process. The process incorporates a number of steps, including data simplification and feature engineering; graph construction; embedding and representation learning.

2.9.3.3 KPIs

Table 22 Time Series Explorer KPIs Description

What (types)	How (process)	Values (concrete values)
Performance	Performance Tool is applied in BGRIMM business case for time series data analysis and predicting the power of the SAG mill with machine learning and graph methods, the main metrics is ROC AUC	Performance ROC AUC > 0.7

2.9.3.4 Service openness

The tool is currently developed as proprietary solution for BGRIMM business case.

2.10 Service collection: GreenR

2.10.1 CarbonTracker

2.10.1.1 Short introduction

CarbonTracker is a tool designed to seamlessly measure and predict the carbon footprint of machine learning models. It tracks hardware power consumption and local energy carbon intensity during training, providing accurate, real-time measurements. The tool supports Intel CPUs, NVIDIA GPUs, AMD hardware, and Apple silicon with live updates. It estimates total emissions after the first epoch and can be integrated into existing solutions via CLI or Python bindings.

Tool documentation can be found on the following link: https://carbontracker.info/.

2.10.1.2 Summary of updates

Since M12, CarbonTracker has been extended to support a wider range of deployment environments and use cases. A fallback mechanism was added to handle scenarios where hardware information is missing or inaccessible. This ensures that CarbonTracker continues to provide estimates using predefined baseline values. This makes it fully usable on virtual machines, where hardware metrics are often unavailable.

To support reproducibility and post-hoc analysis, CarbonTracker can now generate reports from log files, tracking and comparing power usage across different components and stages of a compute task. In addition, a new emulation mode was introduced that allows users to define custom hardware

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

profiles by specifying TDP/wattage and utilization levels, making it possible to estimate energy consumption for arbitrary or hypothetical setups.

2.10.1.3 KPIs

Table 23 Carbontracker KPIs Description

What (types)	How (process)	Values (concrete values)
Interoperability **Ability to integrate across heterogeneous hardware platforms and architectures.**	Interoperability Verified support for Intel, AMD, NVIDIA, and Apple silicon through platform-specific testing. Test suite with 99% code coverage.	Interoperability Pass: Cross-platform compatibility test passed (x86, ARM, GPU) Non-pass: Fails one or more platforms.
Task-Agnostic Usage **Usability across different ML and non-ML workflows without task-specific assumptions.**	Task-Agnostic Usage Provides both CLI and Python API for seamless integration into diverse pipelines and scripts.	Task-Agnostic Usage Pass: Integration test across 4 workflow types passed, i.e., cli, python API, and variation specific to ml and non-ml. Non-pass: Can only be used for ML-specific workflows.
Accuracy **Precision of carbon emission estimates based on real-time and location-specific data.**	Accuracy Uses real-time carbon intensity from electricity maps for region specific values.	Accuracy Pass: Uses real-time region-specific values, and when not available fallback values that are still region-specific. Non-pass: Uses fixed carbon intensity values.
Integration & Fallback **Robustness to environments lacking direct hardware power measurement access.**	Integration & Fallback pip-installable; falls back to predefined CPU energy profiles when hardware data is missing (e.g. on VMs).	Integration & Fallback Pass: Fallback mechanism test passed; baseline estimates available Fail: Cannot be used in VMs, cloud services, etc.

2.10.1.4 Service openness

CarbonTracker is open-source and publicly available on GitHub under is currently licensed under the **MIT license** to facilitate broad access to the tool and transparency in results. The real-time use of

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

localized carbon intensity data ensures that CarbonTracker provides accurate and transparent measurements for reducing the carbon footprint of machine learning models. Additional information can be found at carbontracker.info, offering regular roadmap updates, documentation, and concrete examples to support the user community.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.

3 Conclusions

This deliverable presents a focused update on the development and progress of the tools and services within the enRichMyData Toolbox, with a particular emphasis on their openness and measurable performance. It builds upon the foundations laid in deliverables D2.1 and D3.2, by providing a consolidated overview of recent improvements and ongoing enhancements to the toolbox components.

The report includes detailed summaries of each core tool from DiscoverR, WrappR, CleanR, LinkR, StructR, ClassifiR, as well as supporting services such as ResourcR, ScaleR, ReusR, StreamR, and GreenR. For each component, we provided:

- a short description of its functionality
- a summary of updates and improvements since M12
- an evaluation of openness (including licensing and reusability aspects)
- a set of Key Performance Indicators (KPIs)

These KPIs serve both as internal development benchmarks and as practical guidance for users to evaluate the maturity and effectiveness of each tool in applied contexts.

This deliverable serves not only as an implementation update, but also as a progress tracking document that supports transparency and informed adoption of the enRichMyData tools.

THIS PROJECT HAS RECEIVED FUNDING FROM THE EUROPEAN UNION'S HORIZON EUROPE RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO 101070284.