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

HORIZON-CL4-2021-DATA-01-03

D2.1 enRichMyData tools v1

Work Package 2

Type of document:	Other
Dissemination level:	Public
Lead beneficiary:	JSI
Authors:	Krisztian Buza (editor), all other technical partners contributors
Version:	1
Due Date of document:	30.09.2023
Delivery Date of document:	29.09.2023

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	02.08.2023	Krisztian Buza	Initial draft.
0.2	24.08.2023	Roberto Avogadro, Brian Elvesæter, other tech partners	Updated the tools section
0.3	30.08.2023	Krisztian Buza	Added conclusions
0.4	18.09.2023	Krisztian Buza	Add Tab. 2
0.5	25.09.2023	Krisztian Buza	Clean up document
1.0	29.09.2023	Dumitru Roman	Final QA and final 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

enRichMyData will deliver its capabilities as a set of interoperable tools and services that will form the enRichMyData Toolbox. Collections of tools 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 datasets and data enrichment services (ResourcR), (ii) scalable data enrichment pipelines (ScaleR), (iii) shareable, reusable data enrichment pipelines (ReusR), (iv) streaming data enrichment pipelines (StreamR) and (v) services to estimate energy consumption and CO2 emission of data enrichment pipelines, thus supporting the implementation of "green" (I.e., energy efficient) data enrichment pipelines (GreenR). These tools, together with their documentations, are available at https://enrichmydata.github.io/toolbox/. By describing the aforementioned web page, this deliverable gives a short overview of the tools and services of the enRichMyData Toolbox.

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

Contents

1 Project introduction 7
2 Deliverable overview 8
2.1 Deliverable purpose, scope and context 8
2.2 Target audience 8
2.3 Deliverable structure 8
3 The Data Enrichment Pipeline 9
4 Tools 10
4.1 DiscoverR 10
4.1.1 ABSTAT 10
4.1.2 SemTUI 10
4.2 WrappR 11
4.2.1 Ontotext GraphDB 11
4.2.2 Ontotext Semantic Objects 12
4.2.3 Ontotext Semantic Search 12
4.3 CleanR 12
4.3.1 Ontotext Refine 12
4.3.2 RMLMapper 13
4.4 LinkR 13
4.4.1 s-elBat 14
4.4.2 Ontotext Reconciliation 14
4.5 StructR 14
4.5.1 Wikifier 15
4.5.2 Expert AI Platform Document Analyser 15
4.5.3 Event Registry Relation Classifier 15
4.6 ClassifieR 15
4.6.1 InfoMiner 16
4.6.2 Expert AI Platform Document Classification 17
5 Services 18
5.1 ResourcR 18
5.1.1 LamAPI 18
5.2 ScalR and ResuR 18

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

5.2.1 TAO 19
5.3 StreamR 19
5.3.1 Event Registry 20
5.3.2 StreamStory 20
5.4 GreenR 21
5.4.1 Carbontracker 21
6 New Features and Updates 22
7 Conclusion 25

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

List of figures

Figure 1: Overview of the data enrichment pipeline 9
Figure 2: SemTUI 11
Figure 3: OntoText Refine 13
Figure 4: Event Registry 20
Figure 5: StreamStory 21

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/
CO2	Carbone dioxide
GPL 3.0	GNU General Public License version 3, see also: https://opensource.org/license/gpl-3-0/
MIT	The MIT License, see also: https://opensource.org/license/mit/
RDF	Resource Description Framework
TAO	Tool Augmentation by user enhancements and Orchestration
TRL	Technology Readiness Level
UI	User interface

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

1 Project introduction

enRichMyData provides a novel paradigm for building rich, high-quality and valuable datasets to feed Big Data Analytics and AI applications. It aims at facilitating the specification and scalable execution of data enrichment pipelines, with a focus on supporting various data enrichment operations such as discovery, understanding, selection, cleaning, transformation, integration of Big Data from a variety of sources. **enRichMyData** makes this paradigm easily accessible to a wide range of large and small organizations that encounter difficulties in delivering suitable data to feed their data analytics solutions, due to the lack of specific tools and expertise to support cost-effective and energy-efficient management of data enrichment pipelines.

enRichMyData aims to deliver a toolbox consisting of software tools and infrastructure services for setting up, deploying, executing and managing data enrichment pipelines with the following objectives:

- **Objective O1**: Improve data discovery and profiling featuring search on data, ontologies, and semantic data profiles, to identify data that are potentially valuable for data enrichment.
- **Objective O2:** Improve wrapping of data sources in different formats so they can be securely accessed as virtual semantic graphs are used more easily for data enrichment.
- **Objective O3**: Simplify cleaning, linking (to reference resources), and extension of semi-structured data, featuring approaches that enable users to specify such operations visually.
- **Objective O4:** Simplify annotation and classification of textual data, featuring entity and concept extraction, feature extraction (via embeddings), and classification with predefined and custom classifiers.
- **Objective O5**: Support the management of data enrichment pipelines, including creation and operation of data, linking and extension of services, a framework for deployment and execution of pipelines at large scale, and reuse and extension of existing pipelines to deliver a hub of data and services for data enrichment.
- **Objective O6:** Support data streaming in data enrichment pipelines, featuring support for setting up appropriate endpoints and ensuring high throughput pipeline execution.
- **Objective O7:** Monitor and reduce energy consumption for executing data enrichment pipelines by using models to estimate and track their carbon footprint.

enRichMyData validates its plan through a strong selection of complementary business cases offered by commercially – focused organizations targeting the development of novel products in a wide range of domains, including *Marketing data Enrichment for smart-bidding optimization, AI-based Welding Analytics, Smart Maintenance of Medical Imaging Systems, European Register of Entities from Known Actions, Global Innovation Ecosystems Knowledge Graph and Mineral Processing Optimization*.

The consortium consists of 13 partners from 11 countries. With a mixture of R&D/technology and business case providers, enRichMyData gathers five large companies, three SMEs, two research institutes, and three universities.

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

2 Deliverable overview

2.1 Deliverable purpose, scope and context

enRichMyData will deliver its capabilities as a set of interoperable tools and services that will form the **enRichMyData** Toolbox. These tools, organized into collections of tools, together with their documentations, are available at

- https://enrichmydata.github.io/toolbox

By describing the aforementioned web page, this deliverable gives a short overview of the tools and services of the **enRichMyData** Toolbox.

2.2 Target audience

This deliverable primarily targets users of the enRichMyData toolbox, that is: software developers, researchers, knowledge engineers and other technical staff.

2.3 Deliverable structure

Section 3 provides an overview of the data enrichment pipeline; this is followed by a short description of tools and infrastructure services associated with individual steps of this pipeline in subsequent sections.

This deliverable is closely related to deliverables D1.2 and D3.1. The enRichMyData architecture, described in Deliverable D1.2, is based on the tool and services presented in this deliverable, whereas the integration platform is detailed in D3.1.

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

3 The Data Enrichment Pipeline

enRichMyData will deliver its capabilities as a set of interoperable tools and services that will form the **enRichMyData** Toolbox. Figure 1 provides the conceptual architecture of the toolbox. At the center lays the notion of a data enrichment pipeline that receives input data to be enriched and data to enrich with (left hand side), and generates enriched data (right hand side). The enrichment process is supported by:

- a set of tools, organized into collections of tools (top of the figure), that provide functional capabilities needed to support the design of pipelines; and
- a set of infrastructure services (bottom of the figure) that support the effective and efficient deployment and execution of pipelines.

enRichMyData, being a toolbox with **loosely coupled but interoperable tools and services** is meant to handle complex data enrichment scenarios, where **tools and services can be combined and customized as needed**.

Collections of tools include tools for: (i) data discovery (DiscoverR), (ii) data wrapping (WrappR), (iii) data cleaning and transformation (CleanR), (iv) data linking and extension (LinkR), (v) data structuring (StructR) and (vi) classification (ClassifiR), whereas services include (i) interoperable datasets and data enrichment services (ResourcR), (ii) scalable data enrichment pipelines (ScaleR), (iii) shareable, reusable data enrichment pipelines (ReusR), (iv) streaming data enrichment pipelines (StreamR) and (v) a library to estimate energy consumption and related CO2 emissions that allows to optimize data enrichment pipelines w.r.t. their environmental footprint (GreenR).

Each collection of tools or infrastructure services is implemented by at least one tool or service. In the subsequent sections, we give a short description of the available tools and services, including references to their GitHub repositories. Note that some of the tools or services are listed in several collections.

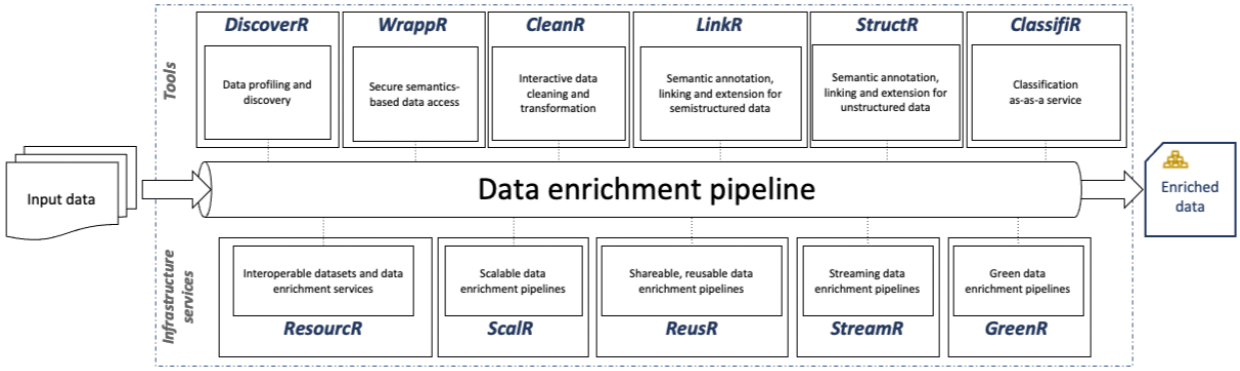

Figure 1: Overview of the data enrichment pipeline

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

4 Tools

4.1 DiscoverR

DiscoverR tools are the components of the enRichMyData toolbox that help users find and understand data that they can use in their data enrichment processes. The tools cover two main families of functionalities not covered well by existing data discovery technology. The first family is about providing deep insights into the content of, possibly large and complex, knowledge graphs represented in RDF, i.e., vocabulary patterns and their distribution; it includes search over and browsing of knowledge graph vocabulary patterns for humans, and API-based access for algorithms. The second family is about providing means to discover content in third-party sources to assist in the enrichment of tabular data; it includes accessing sources available as REST services from a table to enrich and mechanisms to fetch data contained therein and integrate it into the table.

4.1.1 ABSTAT

Since knowledge graphs (KGs) play a crucial role in data enrichment, either as target data sources of interest or as bridges to reach additional sources, ABSTAT supports pattern-based profiling of even 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?"), and 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 and that about 9000 of these connections exist). These profiles have been proved useful to help humans formulate queries over KGs with complex schemas, annotate tables, detect quality problems, and to help machines select the most relevant features. In enRichMyData, profiles of well-established KGs like WikiData and DBpedia will be considered, but also of KGs produced and used in the project.Home page: http://abstat.disco.unimib.it/

Code: https://bitbucket.org/disco_unimib/abstat/src/master/

License: GPL3.0

TRL: 6

4.1.2 SemTUI

SemTUI is a framework that provides a User Interface (UI) 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 DBpedia 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 DBpedia). Although this "link and extend" mechanism is inspired by the principles of web-based data exploration of linked open data, SemTUI 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 and; once two columns have geocoordinates., data can be enriched with the shortest route

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

distance (as calculated by a route planning service). In other words, explorative functionalities offered by SemTUI 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). SemTUI implements, extends, and improves functionalities previously available in a similar tool named ASIA, providing a better user experience and the possibility of translating the enrichment operations into code that can also be manipulated from developer-friendly interfaces like a notebook.

Home page: https://i2tunimib.github.io/I2T-docs/

Code: https://github.com/I2Tunimib

License: Apache2.0

TRL: 4

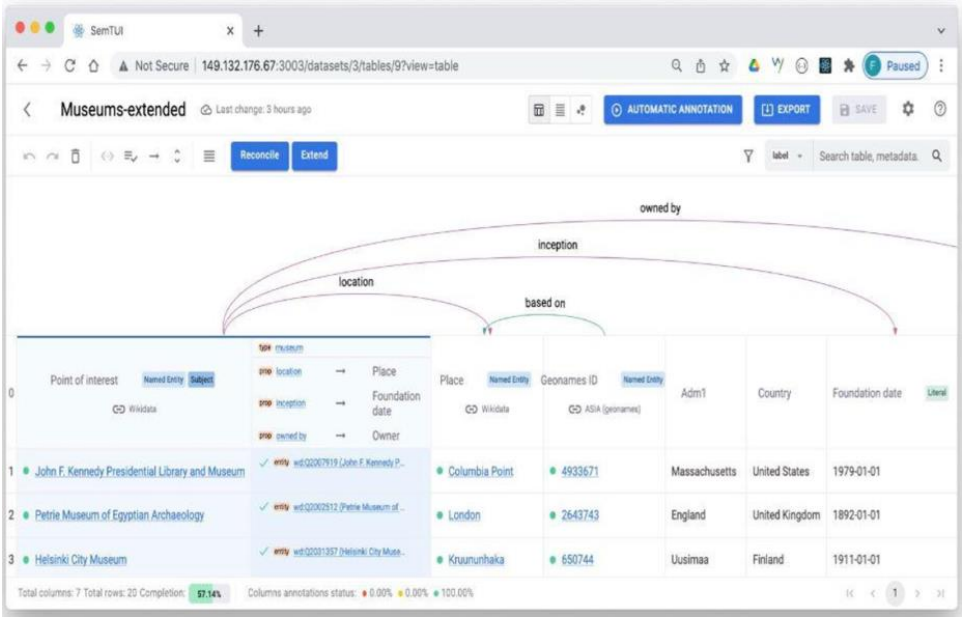

Figure 2: SemTUI

4.2 WrappR

WrappR provides data access using a virtual semantic layer and ensures secure access. WrappR is delivered as a semantic graph database with efficient reasoning, cluster and external index synchronization support. It provides a variety of different type of APIs and access methods as well as different types of data federation and virtualization. Through semantic data access and integration, WrappR provides a practical, robust and versatile tool to improve access to data.

4.2.1 Ontotext GraphDB

Ontotext GraphDB is a highly efficient and robust graph database with RDF and SPARQL support. It supports a number of plugins and connectors such as MongoDB connector for JSON store access, JDBC for exposing RDF as a virtual relational DB, ONTOP for virtual SPARQL access.

Home page: https://graphdb.ontotext.com/

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

Doc: https://graphdb.ontotext.com/documentation/10.1/
License: proprietary
TRL: 8

4.2.2 Ontotext Semantic Objects

The service named "Semantic Objects" is a declaratively configurable service for querying and mutating knowledge graphs which automatically transpiles GraphQL queries and mutations into optimized SPARQL queries.

Home page: https://platform.ontotext.com/semantic-objects/
License: proprietary
TRL: 7

4.2.3 Ontotext Semantic Search

The Semantic Search provides a way to index the data from GraphDB in Elasticsearch and run queries against it.

Home page: https://platform.ontotext.com/semantic-search/
License: proprietary
TRL: 7

4.3 CleanR

CleanR supports the specification of data manipulation transformations, including data cleaning operations and the generation of knowledge graphs from various data formats. Users specify transformations interactively from a user interface, while specifications will be stored in a machine-readable format to be replicated and reused. CleanR provides a broad set of AI-enabled data transformations (e.g., ML-based recommendations) and integrates them with generic linking and extension functionalities provided by the ResourcR. CleanR enables data cleaning and enrichment operations to be shared (as asset, text or executable), managed and, if needed, incorporated as steps in the data pipelines in the ScalR component.

4.3.1 Ontotext Refine

Ontotext Refine (OntoRefine) is a free application for automating the conversion of messy string data into a knowledge graph.

Home page: https://www.ontotext.com/products/ontotext-refine/
Doc: https://platform.ontotext.com/ontorefine/
License: proprietary
TRL: 8

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

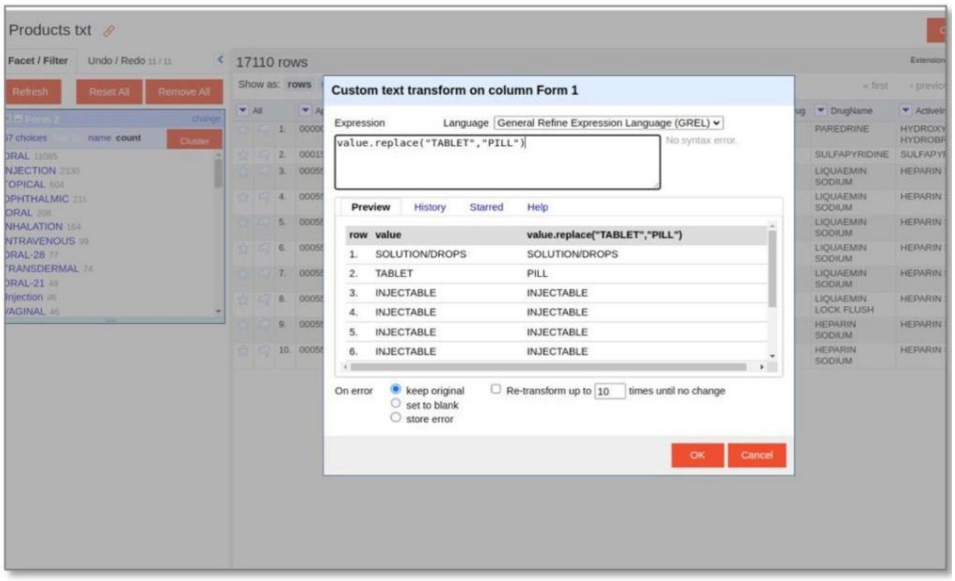

Figure 3: OntoText Refine

4.3.2 RMLMapper

RMLMapper executes RDF Mapping Language (RML) rules (https://rml.io/specs/rml/) to generate Linked Data from multiple originally (semi-)structured data sources.

Home page: https://github.com/RMLio/rmlmapper-java

Doc: https://github.com/RMLio/rmlmapper-java

License: MIT

TRL: 7

4.4 LinkR

When users need to enrich their dataset with an external data source, the task of linking values from both sources becomes a critical hurdle to overcome. It's the key to unlocking the external data and seamlessly blending it with the original dataset. In addition, Knowledge Graphs (KGs) provide a powerful abstraction to support AI applications and interoperability. As most data arrives in loosely structured formats like tables and JSON files, one invaluable asset for data enrichment lies in simplifying the transformation of tables into graphs. This transformation necessitates linking—specifically, connecting elements of the input dataset (columns, rows, and more) to the ontology's classes and properties employed in graph modeling.

Tools that are part of the "LinkR" collection provide support for a variety of linking tasks that solve the above problems. Tools in the LinkR collection include: SemTUI, e-elBat, Ontotext Refine and Ontotext Reconciliation. We refer to Section 4.1.2 for the description of SemTUI and to Section 4.3.1 for the description of Ontotext Refine.

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

4.4.1 s-elBat

s-elBat provides algorithms for all Semantic Table Interpretation (STI) tasks, which altogether deliver complete automatic annotations using reference KGs such as WikiData, DBpedia. After the annotation, the data in the table can be transformed into a graph format, the structure will be compliant to the desired ontology and the entities will be interlinked with the KG. The tool has been competing in several international competitions established for STI, scoring frequently among top performing systems.

Home page: https://bitbucket.org/disco_unimib/selbat/

License: Apache2.0

TRL: 5

4.4.2 Ontotext Reconciliation

Ontotext Reconciliation is a tool for exposing Reconciliation API over RDF data indexed in GraphDB. It simplifies the creation of a flexible reconciler service for OpenRefine-compliant APIs. By leveraging RDF graph content and ontology, users can build a robust reconciliation service without modifying code. The tool supports entity type specification for filtering, optional feature extraction, and efficient searches using GraphDB connectors. These connectors ensure synchronization, provide custom mapping for RDF types, and offer scored matches and custom processing options. Overall, the tool enables the creation of reconciliation services with ease and flexibility, enhancing data matching capabilities.

Home page: https://www.ontotext.com/knowledgehub/videos/kgf21-talks-reconciliation-server-demonstration-against-wikidata/

License: proprietary

TRL: 5

4.5 StructR

Data extraction and structuring from unstructured text sources have always been a challenging task in the field of data analytics. To tackle this challenge, we introduce StructR, a powerful component within the enRichMyData toolbox that specializes in extracting structured data from textual content. StructR offers a range of advanced techniques, including entity recognition and linking, relation extraction, event extraction, and temporal information extraction, to unlock valuable insights from unstructured text.

Entity recognition and linking are vital components of StructR, enabling the identification and linking of entities mentioned in the text to relevant knowledge bases or resources. The component incorporates two powerful entity linking tools: Wikifier and the Expert AI Platform for Document Analysis. These tools leverage advanced algorithms to recognize and link entities, providing enriched context and enhancing the understanding of the text.

Relation extraction is another key functionality of StructR, enabling the identification and extraction of relationships between entities mentioned in the text. This helps uncover connections, associations, and dependencies between different entities, facilitating a deeper understanding of the underlying information.

Event extraction, yet another essential capability of StructR, focuses on identifying events or occurrences mentioned in the text. By automatically detecting and extracting event information, researchers can gain insights into various activities, incidents, or developments described in textual data.

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

The Expert.ai Platform for Document Analysis and the Event Registry Event Types tool empower researchers and businesses alike to efficiently extract precise and valuable relation and event information from diverse textual data sources spanning various domains.

StructR also incorporates temporal information extraction, enabling the extraction of time-related details from the text. This includes identifying dates, time expressions, temporal relations, and durations, enabling researchers to analyze and understand the temporal aspects associated with the extracted data.

By leveraging the power of StructR, researchers can harness the potential of unstructured text data for data-driven insights and decision-making. StructR offers efficient and accurate extraction of structured information, enabling a deeper understanding of text data and facilitating downstream analysis and applications.

The enRichMyData project has carefully curated and integrated cutting-edge tools within StructR. Through collaborative efforts and expertise, the component seamlessly integrates entity linking tools like Wikifier and the Expert AI Platform for Document Analysis, ensuring high-quality entity recognition and linking capabilities.

4.5.1 Wikifier

The JSI Wikifier is a web service that takes a text document as input and annotates it with links to relevant Wikipedia concepts (entities).

Home page: https://wikifier.org/info.html

License: proprietary

TRL: 9

4.5.2 Expert AI Platform Document Analyser

With the Natural Language API's document analysis capabilities, you can perform deep linguistic analysis, keyphrase extraction, named entity recognition, relation extraction and sentiment analysis.

Home page: https://try.expert.ai/document-analysis

License: proprietary

TRL: 9

4.5.3 Event Registry Relation Classifier

Event Registry Relation Classifier extracts relations between entities from news data, such as a company bought another company.

Home page: https://eventregistry.org/

License: proprietary

TRL: 9

4.6 ClassifieR

ClassifiR simplifies the task of labeling and categorizing entire documents based on predefined taxonomies, industry classifications, or customized label sets. It works seamlessly with StructR, which identifies text segment properties, providing a comprehensive data analysis solution.

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

With a user-friendly graphical interface, ClassifiR facilitates the creation and exploration of custom ontologies through clustering, labeling, and querying. Its interactive capabilities empower users to effortlessly develop and train personalized classifiers that automate the classification process. The results are conveniently accessible through a unified endpoint, regardless of the chosen classification method.

Document classification, a fundamental task in NLP, involves categorizing text documents into predefined classes or categories based on their content. The goal is to automatically assign the most appropriate category or label to each document, enabling efficient organization, retrieval, and analysis of large text corpora. The most relevant types of document classification, to our context, include topic classification, news categorization, sentiment analysis, emotion classification, and intent prediction.

Document classification can be divided into the following:

1. Multiclass Classification: In multiclass classification, each document is assigned to one and only one class or category. The goal is to accurately assign each document to a single predefined class label from a set of multiple mutually exclusive classes. For example, if there are three classes (A, B, C), a multiclass classifier would assign each document to one of these three classes.
2. Multilabel Classification: In multilabel classification, each document can be assigned to multiple class labels simultaneously. Instead of being limited to a single class label, a document may belong to multiple categories or have multiple attributes. The classifier assigns a binary label to each class, indicating whether the document belongs to that class or not. This allows for more flexibility and captures the possibility of documents having multiple topics or attributes. For instance, a document might be labeled as belonging to both “Sports” and “Entertainment” categories.
3. Hierarchical Classification: Hierarchical classification involves organizing classes or categories in a hierarchical or tree-like structure. Instead of directly assigning documents to specific classes, the classifier operates in a hierarchical manner, making decisions at different levels of the hierarchy. Each class is organized into parent and child relationships, where the child classes represent specific subcategories or attributes of the parent classes. This approach allows for a more structured and granular classification scheme. For example, in a hierarchical classification system for news articles, the top-level classes could be “Sports,” “Politics,” and “Entertainment,” with further subcategories such as “Football,” “Basketball,” “Elections,” “Legislation,” “Movies,” and so on.

In summary, multiclass classification assigns each document to a single class label, multilabel classification allows for multiple class labels per document, and hierarchical classification organizes classes in a hierarchical structure to provide a more structured and granular classification scheme.

4.6.1 InfoMiner

InfoMiner, an offshoot of the original Ontogen, provides a modern web user interface with useful visualizations underpinned by data analysis and machine learning algorithms with the objective of rapidly constructing labeled datasets, their taxonomies, and classifiers.

- Data Grouping features allow the user to quickly identify similar documents with user defined metrics. This is supported by automatic methods such as clustering.
- Smart Visualization techniques allow the user to understand the data quickly. InfoMiner uses centroid based methods to summarize each cluster. It also automatically creates visualizations such as word clouds, treemaps, and timelines.
- Data Filtering allows the user to query data over its set of properties, the most important of which typically being its textual content or metadata.

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

- Taxonomy creation allows the user to use all the previously mentioned methods (analysis, grouping, filtering) to create a taxonomy and navigate.

Home page: https://github.com/Infominer-JSI
Code: https://github.com/Infominer-JSI/infominer-js
License: BSD2Clause
TRL: 7

4.6.2 Expert AI Platform Document Classification

Document Classification by Expert AI is meant to analyze text to label and identify media topics, emotional traits, geographical references, and more.

Document classification determines what a text is about in terms of categories of a taxonomy.

Available taxonomies are shown in Table 1.

Table 1: Available taxonomies in Expert AI Platform Document Classification

Taxonomy	English	Spanish	French	German	Italian
iptc	✓	✓	✓	✓	✓
geotax	✓	✓	✓	✓	✓
Emotional-traits	✓			✓	
Behavioral-traits	✓			✓	

Home page: https://try.expert.ai/document-classification
License: proprietary
TRL: 9

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

5 Services

5.1 ResourcR

ResourcR provides infrastructure components to support the creation of linking services for a given dataset from a data provider as well as access mechanisms such as search and query. ResourcR enables performant linking and search functionalities with limited effort and exposes them as search and linking APIs. The combination of ResourcR and LinkR makes it possible to turn semantic data produced with the toolbox into resources immediately available for reuse.

5.1.1 LamAPI

Entity search and candidate retrieval. Retrieval of properties associated with entities. Target Knowledge Graph: Wikidata, DBPedia.

Home page: https://bitbucket.org/disco_unimib/lamapi/src/master/

License: Apache2.0

TRL: 5

5.2 ScalR and ResuR

Executing cleaning, transformation, and linking at large scale requires infrastructural components that allow for scalability. As the scalability is the ability of a system to sustain increasing workloads by making use of additional resources, the implementation of a system with this characteristic is an essential step in a big data pipeline to avoid common performance bottlenecks.
There are two main ways of scaling:

Scaling up, or vertical scaling: means using more powerful hardware and more memory. This method offers the best performance, since everything works on the same machine. A possible limitation could be related to the speed of growth of the process; for a fast process, it represents just a short-term solution, and frequent updates became more and more expensive due to hardware limitations.

Scaling out, or horizontal scaling: means adding new power across the infrastructure and not in the same machine. This solution uses parallel computing to increase the performance of the infrastructure and is also valid in the long term. At the same time, moving from a single machine to a distributed system leads to lower speed and higher complexity.

The main goal of the ScalR is to provide horizontal scalability of data enrichment pipelines using software containers and support for management of the different procedures associated with the execution of data enrichment pipelines flexibly on heterogeneous computing infrastructures.

The goal can be achieved by promoting the reuse and modification of existing data enrichment pipelines by exposing them as an integrated deployable unit, as opposed to ad-hoc, non-reusable pieces of code. For this reason, most of the tools that handle scalability can also be considered data orchestrator tools, since, in general, it is usually more convenient to use an automatic tool to orchestrate a pipeline, rather than combine different scripts. This option allows the user to use different languages, like batch or Python files, and use them in the same pipeline with no difficulties. A data orchestrator can receive as input different types of files, handle the scalability, create and schedule the progress of the entire data pipeline. Other advantages of the use of a data orchestrator are the good readability of the data flow, a limited space for error since a great number of activities are managed automatically, and better data quality visibility.

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

5.2.1 TAO

In the enRichMyData project a tool from CS GROUP–ROMANIA called TAO (which stands for Tool Augmentation by user enhancements and Orchestration) is used as for providing the main ScalR functionalities. TAO is an open source, lightweight, generic, extensible, and distributed orchestration framework. It allows to reuse (i.e., integrate) commonly used toolboxes (such as, but not limited to some EarthObservation processing tools like SNAP, Orfeo Toolbox, GDAL, PolSARPro, etc.). This framework allows for processing composition and distribution in such a way that end users could define processing workflows by themselves and easily integrate additional processing modules, without any programming knowledge requirements.

TAO platform provides a means for orchestrating heterogeneous processing components and libraries to process scientific data. This is achieved in the following steps:

1. Preparation of resources (including processing components or tools) and data input.
2. Definition of a workflow as a processing chain.
3. Execution of workflows, scaling up from one to as many nodes as made available.
4. 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).

An important TAO component is the DRMAA (Distributed Resource Management Application API), which provides a standardized access to the DRM 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 and SLURM executions. Support for Kubernetes and CWL (Common Workflow Language) implementation is also in progress. The DRMAA implementations are provided as plugins that allow a high flexibility as the current implementation can be changed easily with another.

Home page: https://github.com/tao-org

License: GPL3.0

TRL: 7

5.3 StreamR

In today's fast-paced, data-driven world, the ability to extract valuable insights from streaming data in real-time is more crucial than ever. That's where StreamR, a powerful component within the enRichMyData toolbox, comes into play. Designed to tackle the challenges of streaming data analysis, StreamR revolutionizes the way organizations uncover real-time insights and drive informed decision-making.

StreamR, powered by Streamstory and EventRegistry, allows organizations to unlock the true potential of streaming data. By harnessing real-time insights from complex multivariate time series data streams and providing real time intelligence derived from news events, businesses can drive innovation, optimize operations, and make data-driven decisions that propel them ahead in a rapidly evolving world.

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

5.3.1 Event Registry

EventRegistry is the world's leading news intelligence platform. EventRegistry tracks over 150,000 news sources in 50+ languages and provides near-real-time article processing. The continuous stream of articles, grouped into events, offers a wealth of text data enriched with metadata, including mentioned entities, topics, and sentiment. Tracking news events with EventRegistry is crucial for businesses and industries, as it offers valuable insights into market dynamics, industry trends, and emerging opportunities. By monitoring news coverage, organizations can make informed decisions, adapt strategies, manage reputation, and gain insights into competitor activities. News event tracking also aids in risk management, business continuity planning, and staying informed about socio-political and economic developments. With EventRegistry, businesses can navigate uncertainties, seize growth prospects, and maintain a competitive edge in today's dynamic business landscape.

Figure 4: Event Registry

Home page: https://newsapi.ai/
Code: https://github.com/EventRegistry
License: proprietary
TRL: 7

5.3.2 StreamStory

Streamstory, a fully integrated solution within StreamR, offers a comprehensive toolkit for analyzing and understanding multivariate time series data. With its advanced features including time series clustering, state modeling, visualization, statistical analysis, and predictive analytics, Streamstory enables businesses to harness the hidden potential within their time series data. It has emerged from prestigious European research projects, where it found diverse applications in predictive failure analysis, transportation optimization, energy demand forecasting, and water management. Streamstory empowers industries across sectors such as supply chain, manufacturing, and resource management by optimizing processes, detecting anomalies, and ensuring seamless operations.

Home page: http://atena.ijs.si:8080/
Code: https://github.com/E3-JSI/StreamStory2
License: proprietary

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

TRL: 4

Figure 5: StreamStory

5.4 GreenR

GreenR provides infrastructure components to support monitoring of data enrichment pipelines in terms of their environmental impact. It monitors the carbon footprint of the various components in the pipeline and provides the results to the use through a dashboard to log and modulate the environmental impact due to the heavy computations within the pipelines.

5.4.1 Carbontracker

Carbontracker is a tool for tracking and predicting the energy consumption and carbon footprint of training deep learning models.

Home page: https://github.com/lfwa/carbontracker/

License: MIT

TRL: 5

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

6 New Features and Updates

Table 2 summarizes new/extended features and updates (adaptations) of tools in the first year of the project, as well as the requirements addressed by these tools and their new features (updates) according to deliverable D1.1.

Table 2: New/extended features and updates of the tools

Name of tool	**New / extended features and updates**	**Requirements addressed**
Abstat	- No development in the first year of the project	RQ-DiscoverR-4
SemTUI	- Improvement of semantic table interpretation algorithms (selBat) - HITL entity linking for tables: Algorithm: normalization of matching score with NIL prediction	RQ-DiscoverR-1
Ontotext GraphDB	- Upgrade to ONTOP 5 (virtualization of RDB) - OpenSearch support - ChatGPT integration: (i) ask GPT (scalar, list, table), (ii) GPT Explain of SPARQL query and/or results	RQ-WrappR-3 RQ-WrappR-1 RQ-WrappR-2
Ontotext Semantic Objects	- Deployment and operations improvements: key components updated to newer versions - Updated components to work with GDB 10.X	RQ-WrappR-2
Ontotext Semantic Search	- Updated components to work with GDB 10.X	RQ-WrappR-2
Ontotext Refine	- No development in the first year of the project	RQ-CLEANR-1 RQ-CLEANR-2 RQ-CLEANR-3 RQ-CLEANR-4
RMLMapper	- Open source alternative to generating linked data, no development in the first year of the project	RQ-CLEANR-1 RQ-CLEANR-2 RQ-CLEANR-3 RQ-CLEANR-4
selBat	Revised linking pipeline: - New ML model for entity linking	RQ-DiscoverR-4

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

	- Human-In-The-Loop approach to revise uncertain outcomes	
Ontotext Reconciliation	- No development in the first year of the project	RQ-LinkR-6
Wikifier	- New feature that allows to annotate text with part of speech tags - New feature that allows to use custom vocabulary, such as company names - Annotations based on new wikipedia dump	RQ-StructureR-11
Expert AI Platform Document Analyser	- Design of a component for the parsing and segmentation of PDF documents, often ranging several hundred pages that need to be split into smaller sections - Design of a component for semantic text analysis in in English language, for each passage we keep and index: (i) document, page, geo-coordinates and offset of the passage, (ii) passage full-text index (with symbolic metadata such as lemmas, entity tags, etc.), (iii) passage neural embedding - Design of a component for the semantic analysis of natural language requests in English language, each question must be analyzed by symbolic NLP processor that identifies aspects as: (i) symbolic concepts (related to a given ontology/knowledge-graph), e.g. synonyms <methodology,method,procedure,process> belong to a specific unique concept, (ii) related concepts (e.g. <extinguisher, fire extinguisher> is more specific than <device> but less specific than <carbon dioxide fire extinguisher>) that can be used to expand or restrict the query, (iii) lemmas, entities, etc.	RQ-StructureR-1 RQ-StructureR-3 RQ-StructureR-5 RQ-StructureR-7
Event Registry Relation Classifier	- Extension of the set of relations that can be extracted, in particular: the “technology use” (company uses a technology) relation has been added - Models have been improved with additional annotations	RQ-StructureR-12
InfoMiner	- Online deployment of InfoMiner	RQ-ClassifiR-2
Expert AI Platform Document Classification	- Design of a component able to analyze a Documents that is classified according to	RQ-ClassifiR-1 RQ-ClassifiR-2

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

	keywords/known problem descriptions/AI related topics - The entire pipeline is deployable in a private environment (either on-premises or in a secure cloud environment accessible only by the business partner)	
LamAPI	- Datatype analysis with regex and spacy - Addition of RDF2VEC embeddings - Indexing of Crunchbase (InnoGraph + SN)	RQ-ResourcR-3
TAO	- Decoupled the Earth Observation specific features from the TAO data model and generalization of the TAO data model - Started development of components specific for enRichMyData for tools and input data - Started the updates for enRichMyData TAO web interface - Created enRichMyData installation package - Implementation of the first two enRichMyData pipelines into TAO. - Deployment of TAO to a first enRichMyData integration platform	RQ-ScalR-1 RQ-ScalR-3 RQ-SCALR-9
Event Registry	- Event Registry is being actively developed, nevertheless not within this project	RQ-StreamR-1
StreamStory	- Support business case owners with local installation of StreamStory	RQ-StreamR-2 RQ-StreamR-3
Carbontracker	- Features supporting country specific data, measurement factors and logging labels that are more expressive - Features for command line interface that can be wrapped around any executable code, support for live carbon footprint estimation from 200 regions across the world	RQ-GreenR-3

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

7 Conclusion

The enRichMyData project aims to improve data management and enrichment with the introduction of the enRichMyData Toolbox. Through a comprehensive suite of interoperable tools and services, enRichMyData addresses a wide spectrum of data challenges, offering innovative solutions that span data access, cleaning, transformation, annotation, linking and classification. This toolbox empowers users to seamlessly navigate and manipulate both structured and unstructured data, unlocking the true potential of their datasets.

With collections of tools like DiscoverR, WrappR, CleanR, LinkR, StructR, and ClassifiR, enRichMyData provides a diverse toolkit that caters to a multitude of data processing needs. Additionally, the suite of services, including ResourcR, ScaleR, ReusR, StreamR, and GreenR, ensures that users can enrich their data efficiently, sustainably, and at scale.

The enRichMyData Toolbox isn't just a collection of tools and services; it's a transformative approach to data enrichment that emphasizes interoperability, collaboration, and user empowerment. By centralizing these capabilities and offering them through a unified platform, enRichMyData paves the way for enhanced data-driven decision-making and innovation across industries and domains.

As highlighted in this deliverable, interested individuals and organizations can readily access the enRichMyData Toolbox and its associated documentation at https://enrichmydata.github.io/toolbox/. This web page acts as a gateway to the wealth of resources and functionalities that the toolbox has to offer, serving as a hub for users to explore, learn, and integrate these tools and services seamlessly into their workflows.

In essence, the enRichMyData project presents a holistic solution to the challenges of data manipulation and enrichment. By providing a dynamic and versatile toolbox, enRichMyData empowers users to unleash the full potential of their data, driving innovation, insights, and progress.

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