12 DS2 Curation Module (CUR)

12.1 DS2 Curation Module (CUR)

Owner(s):	IBM
DOA Task:	T4.3
Tier:	T2
Nature:	Optional
Result:	K4.3

This task will prepare data to be exchanged between data requestors and data providers in disparate sectors and enterprises. Amongst the topics to achieve this goal, this task will cover the curation of generated/collected data, portability, and transformation of data, covering security data transport protocols and data interoperability issues, include transformations due to data sovereignty requirements, ontological reconciliation, and schema differences. Novel challenges to be addressed by this task include reconciling semantic differences in data classifications between sectors.

12.1.1 Introduction

Purpose: Data obtained from disparate sources runs the risk of remaining siloed unless it is curated to match the format of similar data from other data sets. Manual curation of datasets, however, can be a labour-intensive task, and not suited to DS2's dynamic nature of federating dataspaces. The aim of this task is the automatic creation of pipelines to curate data through machine learning

The Data Curation module is invoked on two or more data sets and aims to identify through machine learn data transformations which need to be performed on fields in order to allow interoperability between the data sets – for example, the conversion of time-data formatting. The required transformation(s) will be automatically selected from a transformation library, and a processing pipeline will be created and executed to curate the data.

Description: In a data space environment, data sources will be provided by different data providers from different environments and different sectors. Involvement of a wider range of data spaces may further exacerbate the situation, as exposes even more data sources and heterogeneous environments.

There may be several issues that occur when a data consumer collects disparate data sets which may come from an assortment of different data providers – for example, data providers from different sectors or countries may provide data in different terminologies or languages. Additionally, mixed data sets may be incompatible in number of ways – data formats may be different, collected data values may have been measured on different scales, etc. The curation of such a mix of data may not only require the programming effort of a data engineer but can be a long and costly process. The Data Curation module aims to use machine learning (ML) to automatically recommend transformations between columns of structured data, and then automatically apply the discovered transformations on the relevant data columns.

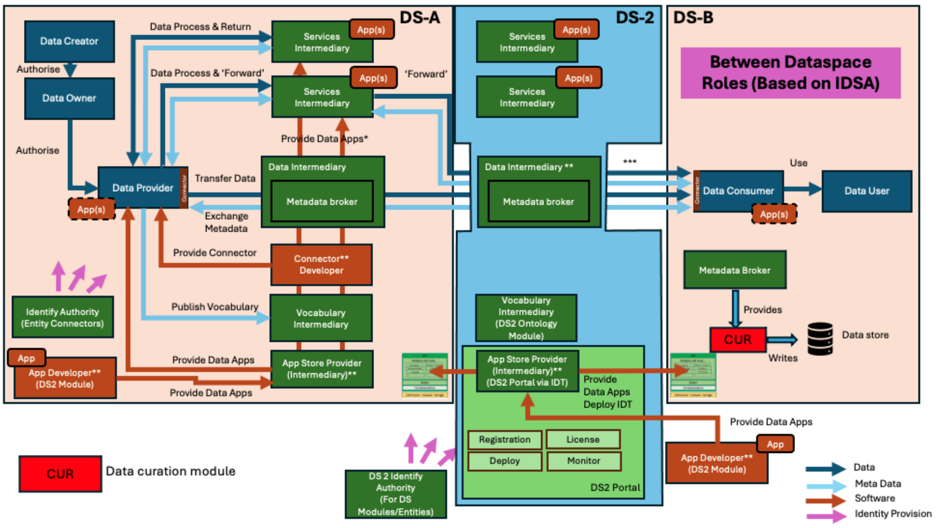
This module can either be triggered as part of a data acquisition workflow or invoked directly by a data consumer on explicit data sets. As opposed to the Transformation functionality of the Data Detection and Data Transformation module which deals with transformations *within* a given data set to correct quality errors, this module deals with the detection of differences in data columns *across* data sets, which may be due to formatting or even column headers.

Challenges include defining a standard format for describing transformations which can be interpreted by an LLM, machine learning determination of required transformation, automatic selection of correct transformation, machine learning reflection to evaluate the selection without actual execution and finding a correct use case to highlight this technology.

This task will use the Orchestrator's Service Registry for transformation routine options.

12.1.2 Where this component fits

12.1.2.1 Big Picture



12.1.3 Where this component fits

Where	Status
Within a single Dataspace for use between participants in that Dataspace only	No: This module runs on local collected data
Deployed and used by a single participant to enable the participant in either an In-Data space or	Yes: This is the primary mode of use DS2 Service Registry

Inter- Data space scenario	
Across Dataspaces without Service Intermediary	No: This module runs on local collected data (although of course the data may have been obtained from different dataspaces)
Across Dataspace with Intermediary	No: This module runs on local collected data (although of course the data may have been obtained from different dataspaces)
Other Comments	N/A

12.1.3.1 Within a single Dataspace (where applicable)

N/A – runs on a single data consumer/producer site.

12.1.3.2 Deployed and used by a single participant (where applicable)

The Data Curator module will analyze datasets to determine relationships between data fields, applying transformations if required to bring the representation of a data field in one data set to the representation of an identified field in another data set.

This module will be invoked on explicit data sets and will require as input schema descriptions of the data sets, sample data frames from the data sets, and a metadata description of the transformation modules available. Using machine learning, this module will return the discovered relationships between the data sets, and if a transformation is required to resolve differences between the matched fields (e.g. formatting, units…) then the module selects the appropriate transformation module and apply it, creating a new version of one of the input datasets.

12.1.3.3 Across Dataspaces without a Service Intermediary (where applicable)

No, this runs on local data that has already been downloaded.

12.1.3.4 Across Dataspaces with a Service Intermediary (where applicable)

No, this runs on local data that has already been downloaded.

12.1.4 Component Definition

The figure below represents the actors, internal structure, primary sub-components, primary DS2 module interfaces, and primary other interfaces of the module.

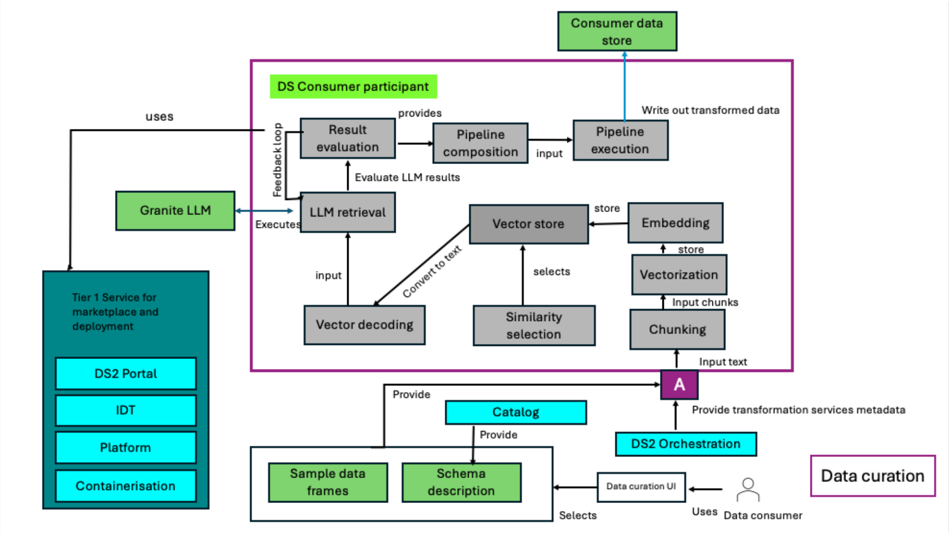


Figure 1: Schema for the Module

This module has the following subcomponent and other functions. Note there is no significant UI involved since the component is essentially a series of back-end operations:

Data Curation Module:

- **Vector store:** To implement RAG (Retrieval Augmented Generation) for the Machine Learning retrieval from the LLM, a vector store needs to be created from the Embedding Module
- **Chunking:** Breaks large text segments into smaller ones. The text to be chunked for addition to the vector store will come from metadata descriptions of the data sets from a catalogue (TBD), a description of the available transformation modules (currently envisioned to be held in the DS2 Service Registry), and sample data frames.
- **Vectorization:** Encodes textual inputs from the chunking into a vector for storage in the vector store
- **Embedding:** Enters the generated vector(s) into the vector store.
- **Similarity Selection:** Determines the best match for the vector search.
- **Vector decoding:** Decodes a stored vector into text for input to the LLM
- **LLM retrieval:** Based on the metadata descriptions of the available transformation modules from the Orchestration module and the result evaluation for correlation results, determines the most suitable transformation module to use.

- **Result evaluation:** Compares the answer returned from the LLM to original input specifications (schema descriptions, sample data frames) to validate results. There is a feedback loop that can reevaluate the query if the validation step fails.
- **Pipeline composition:** Creates an executable pipeline for data transformation. This stage will configure a pipeline to use the identified data transformation module on the dataset.
- **Pipeline execution:** Executes the pipeline and stores the transformed data set(s) back to the original Consumer data store.

Interacting, external modules to the Data Retrieval Module include:

- An LLM (expected to be IBM Granite) will be required. The trained LLM will be supplemented as part of the RAG process to acquire knowledge of the available transformation modules as described.
- Data schema description and sample data frames for the data sources are required
- Tier 1 Service Stack containerization and deployment.

12.1.5 Technical Foundations and Background

The Data Curator component will use the Open-Source IBM Granite LLM to determine relationships between data sources. As such, this component will play an essential role in the curation of data.

The DS2 Service Registry is expected to store the metadata for the transformation modules.

Subcomponent/Component	Owner	License
IBM Granite LLM foundation model	IBM	Apache 2.0

12.1.6 Interaction of the Component

The following table specifies the primary input/output controls/data to blocks which are not part of the module

With Module/Feature	Receives From/Gives To	What
Catalog Meta Data Broker	Receives from	Metadata describing the data sets. In particular, a schema description of the data sets is expected.
DS2 Orchestration Service Registry	Receives from	Storage for the metadata description of the available data transformations and a query interface.

12.1.7 Technical Risks

Risk	Description	Contingency Plan
Data sets are intrinsically disjoint one from another	The data sets for a given use case do not contain data which can be curated – e.g. no overlapping fields.	Look for additional data sources
Too few and too simple data sets to require this	There will only be a very limited number of data sets with limited fields, which consequently do not require a sophisticated tool for curation.	Look for additional data sources

Cannot find an existing transformation routine	The LLM will attempt to determine required mappings between columns in different data steps and match an appropriate transformation routine to a transformation from a catalogue of transformations. Research is required to see if this can really work for the general case.	Develop additional transformation routines to match use cases.
Cannot adequately define transformations in metadata	A generic format is required for defining transformations that an LLM can understand.	If required, this format will be iteratively developed, increasing the complexity of the functions as it develops

12.1.8 Security

Security Issue	Description	Need
Secure communication for metadata	The Data Curator will run locally within a Kubernetes cluster	This might not require special attention but the fit of this to the containerisation of IDT needs to be examined

12.1.9 Data Governance

Data Governance Issue	Description	Need
None	All data used is data being made available by other DS2 components	N/A

12.1.10 Requirements and Functionality

This module will be used in the following use cases:

City Scape	TBD
Green Deal	TBD
Agriculture	TBD
Inter-Sector	TBD

Commented [ES1]: Currently no commitment from the use case partners to use this module.

Their requirements and functions/extensions to achieve them relative to this module, specifically extracted from the use case are as per the table below noting that in many cases further discussion might need to take placed between pilot partner and module partner to determine if a fit or the scope of the precise fit.

Note for this module due to a late directional change, it has not been possible to finalise in which cases/scenarios will be used. This will require a physical discussion planned for a Plenary in Cluj in October and it is anticipated that at least one scenario from each case would utilise this module.

WHERE	WHAT	WHY	Run/Design Time	Priority
	Use Case 1: City Scape			

TBD				
	Use Case 2: Green Deal			
TBD				
	Use Case 3: Agriculture			
TBD				

12.1.11 Workflows

The following sub-sections describe the sequence diagrams of the Module

12.1.11.1 Read Data Sources

This feature provides the capability for a data owner to initiate data curation amongst the data sources that it has access to. The sequence diagram is shown in Figure 2

The main steps/functionalities are as follows:

- Data owner initiates a request for data curation from the Data Curation Module on given data sources, passing the data schemas and sample data frames of the data sources.
- A metadata description of all available data transformation modules is obtained from the DS2 Orchestration Service Registry provided that a party has added a service there
- Input data is vectorized and embedded in the vector store
- A vector search for the query to the LLM is executed
- The returned vector is decoded to text
- The LLM is queried and evaluates data column matches and required transformations
- LLM results are checked again input, and re-executed if necessary
- If transformation is required, a transformation pipeline is automatically created and executed
- Transformed data is written out to the local store.

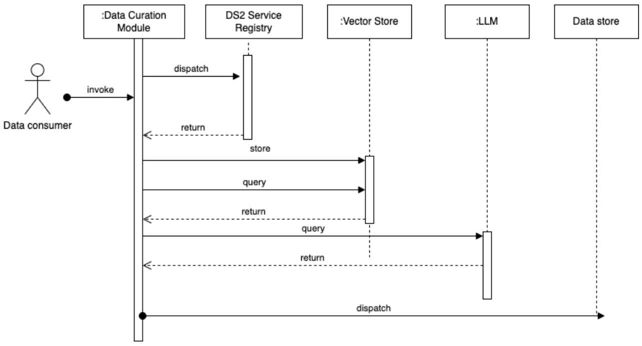


Figure 2: Data curation request

12.1.12 Role, Resourcing, and Milestones

Sub-component	Main Activity	M18	M24	M30	M36
Catalog	- Definition of transformation description metadata understandable by an LLM which will be stored in the Catalog. This may have implications on the API of the transformations to be developed	■			
LLM retrieval/result evaluation	- Obtaining correct transformation recommendations from the LLM	■			
Chunking	- LLM result accuracy will depend on the proper selection of chunks from the supplementation RAG information	■			
Data Curator	- Working with other use case to add synthetic data sets/extra columns to show case the abilities of the Data Curator		■		
Data Curator	- Testing and refinements to LLM RAG process, transformation descriptions, potentially new transformation, etc		■		
Data Curator	- Working with use case partners with their data and requirements		■		
Data Curator	Containerization, DS2 integration…			■	
Data Curator	Final version			■	
Table Total/DOA Task Total/Resilience	**Comments:**			**20**	

12.1.13 Open Issues

The following table summarise open issues/uncertainties that need to be resolved during the next stages or implementation.

Issue	Description	Next Steps	Lead or Related Component
Catalog Meta Data Broker	Metadata describing the data sets. It is TBD where this information will come from, since it refers to data sets that may have already been downloaded from other sources or are legacy to the organization.	Continue researching what metadata is required and what metadata can be automatically determined	VTT and IBM
Definition of transformations	Define a standard interface for defining transformations that the LLM will be able to understand	Ongoing research	IBM
Obtain the metadata description of the available transformations	Need to register the available transformations and their metadata in the DS2 Service Registry and provide a search mechanism	Definition of the Service Registry and the metadata requirements for describing transformations	IBM, SAG, ICE
Determine how the data curation process will be initiated	Determine whether the data curation process will be manually initiated by the data owner, or if, for example, it will part of the data acquisition pipeline process.	Start with the simpler manually initiated process and then evaluate.	Green Deal datasets
Implementation of transformations	Create a set of transformations that are relevant to our use cases	Obtain data sets from use case partners	Use case partners, SAG