1. Culture and Language Module (CLM)

1.1 T5.1: Culture and Language Module (CLM)

Owner(s): INTU
DOA Task: T5.1
Tier: 0
Nature: System
Result: K5.1

This has the purpose of information normalization from multiple languages, with the goals of enabling a search in one language to retrieve information that originated in another language; and to present key elements of such information. To achieve this, it will be based upon Natural Language Processing including POS, morphological analysis, NER, and Sentiment Analysis. This background will be extended via machine learning models of the topic/domain/language register of the source text to support disambiguation and mapping of all the polysemic information in the raw texts to monosemic ontological values (ontologization of the text information); plus, bespoke domain ontologies to convert source texts to ontological graphs that turn the polysemic textual data into monosomic structured knowledge. The output will be an ontology and an independent software module for multi-cultural and multi-lingual data searches, which can also be used in conjunction with other human centric modules such as those in the DARC module.

1.1.1 Introduction

Purpose: To give a data consumer a better understanding of what data offers exist both within their own data space and in other data spaces, which may come from different sectors or countries and in different languages. This will increase the potential for use of a data offer. This is achieved through transformation of human language in shared information and queries into a rich, searchable hierarchical ontological description of the offered data set.

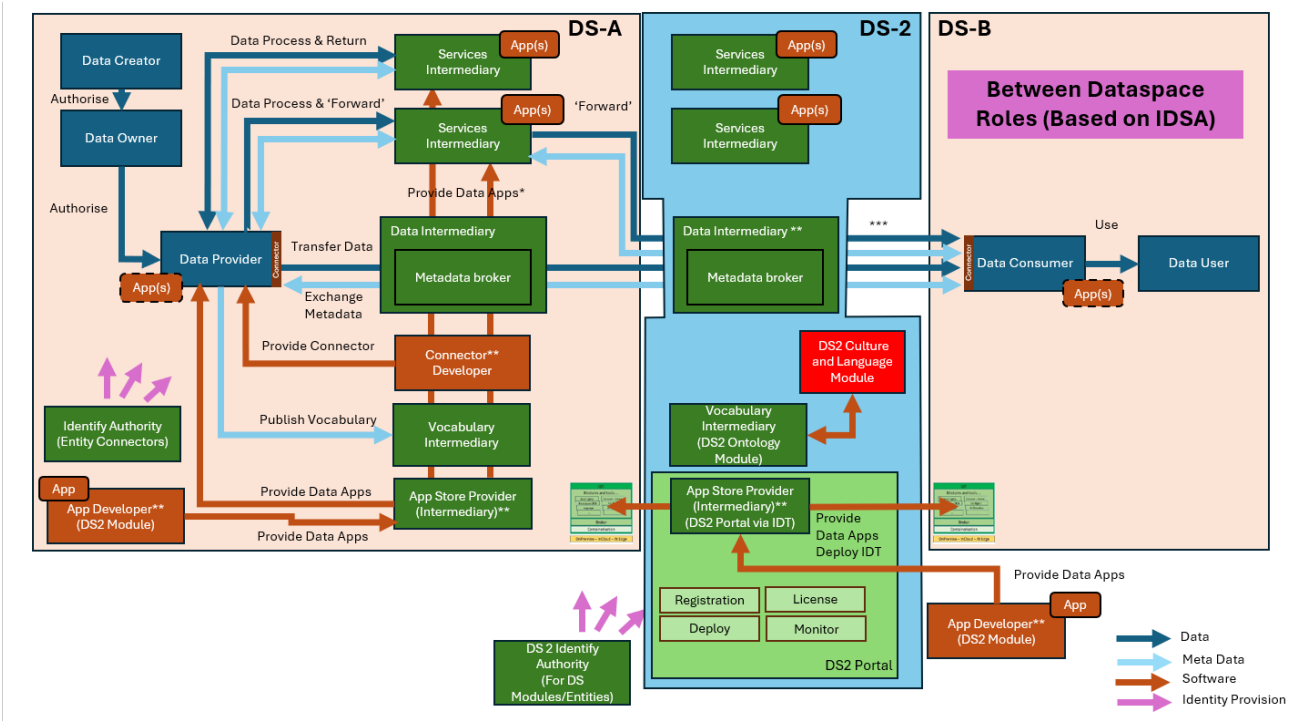
Description: This module will enable the participant end user to search for information in human language and to receive relevant results that would not have been returned by simple queries based on the face value of the search parameters. The module ontologizes the information, creating a representation of the information in an internal “lingua franca” that serves all participants. This “lingua franca” is not “flat” but integrates hierarchy of concepts and human expert knowledge that enables queries on a concept to return its sub-concepts, thus broadening the scope of possible responses. Such responses may represent search on all equivalent terms for the search terms (e.g. the search was for Britain or UK which are ontological equivalents) and/or search within “subsets” of the search terms (e.g. the search was on Britain and the information referred to Manchester which is in England which is in Britain). This is based on converting the human language unstructured data or semi-structured human-language based metadata that is submitted for sharing by any participating dataspace into an ontological digest which contains relevant hierarchal information and attributes of each element. The same process is performed on the fly on queries. Ontologized information will be stored in the Catalogue and used by other modules as required. The module will also dynamically provide the same service to the Chatbot in facilitating queries across languages and focusing likely

responses. The orchestration module regarding possibly disparate expressions of policy, procedures etc. may also make use of this module. The module will be accessible to all other modules with a bespoke REST API.

1.1.2 Where this component fits

1.1.2.1 Big Picture

The function of the module is to serve as an Ontological Mediator for human-language or semi-human language information shared between dataspaces or within participating dataspaces. Its two key tasks are: 1) to convert shared information stored in a Catalogue to a normalized ontological format that bridges the language differences between dataspaces; and 2) to support the query system of the Chatbot by ontologizing the dialogue from the side of the partner dataspace end user and retrieving appropriate information.



The ontological representation of the information that is returned by this module, unlike human language that it draws on, is hierarchical and monosemic (an ontological concept has only one unique meaning) and integrates expert knowledge. Therefore, the analysis by the module both resolves human language polysemy and adds a dimension of hierarchy to the semantics of the language. This enables queries on a concept that is high in the hierarchy of the ontology and to return all information that contains concepts that are beneath the higher concept (the “children” of the higher concept).

To enhance its own functioning, this module will also perform a function of ontology comprehensiveness testing. The module will test the information that is flowing through the system to identify if it fits any of the existing ontologies and to send a warning if new types of information are coming through that require adaptation.

This module will also generate runtime ontologized information that will support the dialogue of the Chatbot module. Since queries by human end users will also be

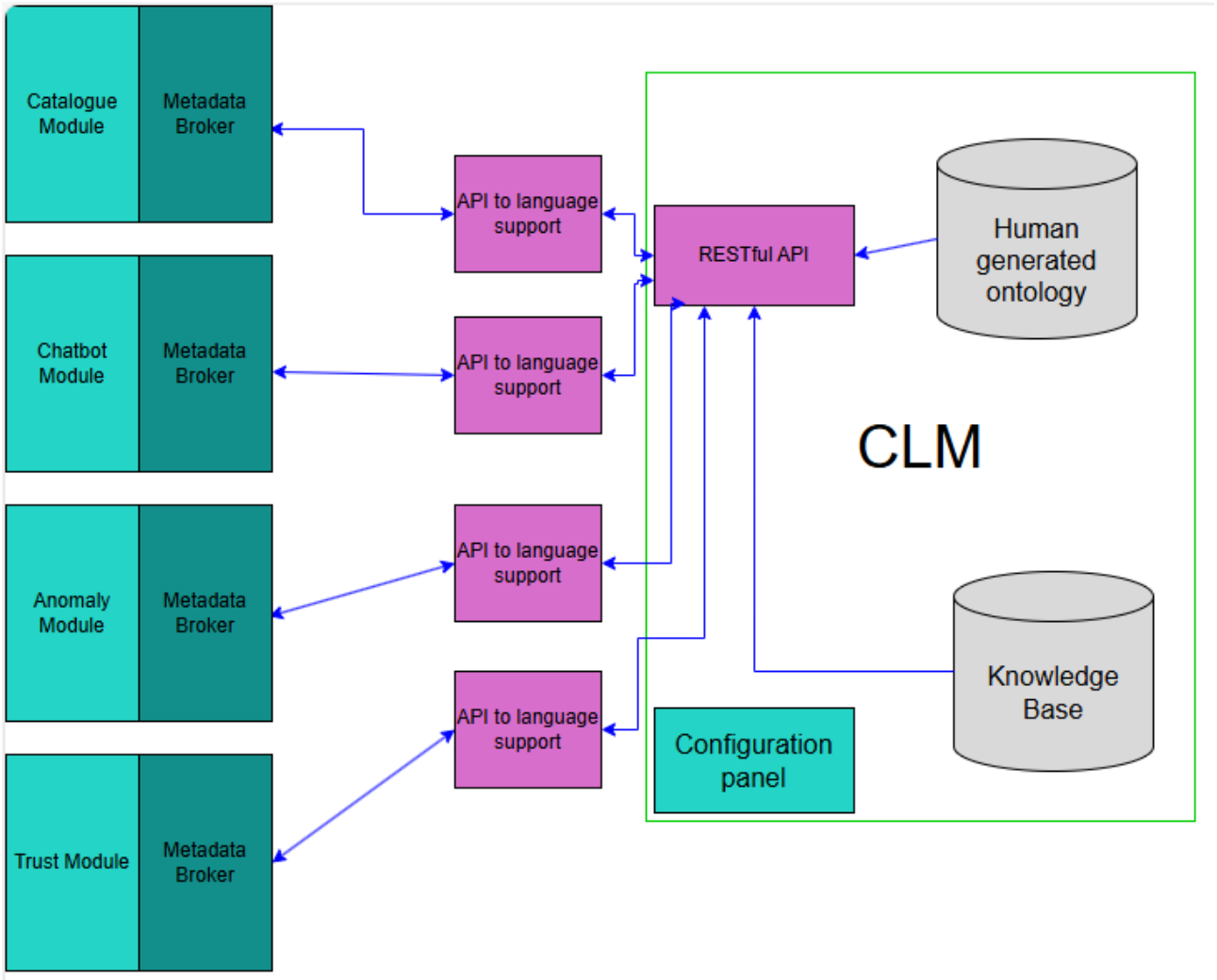
“ontologized”, a user will be able to search in any supported language and domain. The search will be executed as if the user had written the query in the language of the data.

This task includes the construction of domain ontologies for the domains of the use cases in the project. This task is labour-intensive and requires close collaboration between the use case partners and the partners involved in building the ontology. However, data spaces and data products do not remain constant. They are constantly evolving with new databases and domains. Therefore, this task will include developing automated methods for extending the ontology to new concepts and ideas and for automation of the process of ontology creation process by extending the common language as new things emerge. Such a method requires extensive research.

This task includes the creation of bespoke APIs for the connection between the module and all other relevant modules of the DS2 platform.

The module will be connected to the Catalogue module which will host the ontologized information produced by the analysis of the CLM and to the Chatbot module to assist in drilling down in the dialogue with users to identify the requested information. It may also be connected to the Orchestration Module (tentative and pending decisions).

Schema for the DS2 CLM module



Where	Status

Within a single Dataspace for use between participants in that Dataspace only	Yes. Within a dataspace there will be participants with disparate databases that do not necessarily use the same formats, languages or query systems. The use of the module within a single Dataspace can serve to normalize the data of the Dataspace before sharing it with others.
Deployed and used by a single participant to enable the participant in either an In-Data space or Inter- Data space scenario	N/A
Across Dataspaces without Service Intermediary	Yes: The module can be deployed by a single participant to facilitate data normalization.
Across Dataspace with Intermediary	Yes. The module is a Dataspace Intermediary as it is providing the function of normalization of information between dataspaces.
Other Comments	REST and SOAP calls are supported

1.1.2.2 Within a single Dataspace (where applicable)

Before data is shared between participants of different dataspaces, it should *ideally* reflect a normalized standard of the Dataspace it originates in. Otherwise, DS2 will not be a bridge between dataspaces but rather between databases that exist within each of the participant dataspaces. Therefore, integration of the DS2 CLM module or certain functionalities of it can help normalize the data of a given Dataspace before it is shared across dataspaces.

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

1.1.2.4 Across Dataspaces without Intermediary (where applicable)

The key function of the module is to create a common ontological representation of information from all the domains and in all the supported languages across dataspaces. The output of the module is stored in the Catalogue and is also used by the Chatbot for queries and searches.

1.1.2.5 Acrss Dataspace with Intermediary (where applicable)

The CML module is a Dataspace Intermediary as it provides the function of normalization of information between dataspaces.

1.1.3 Component Definition

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

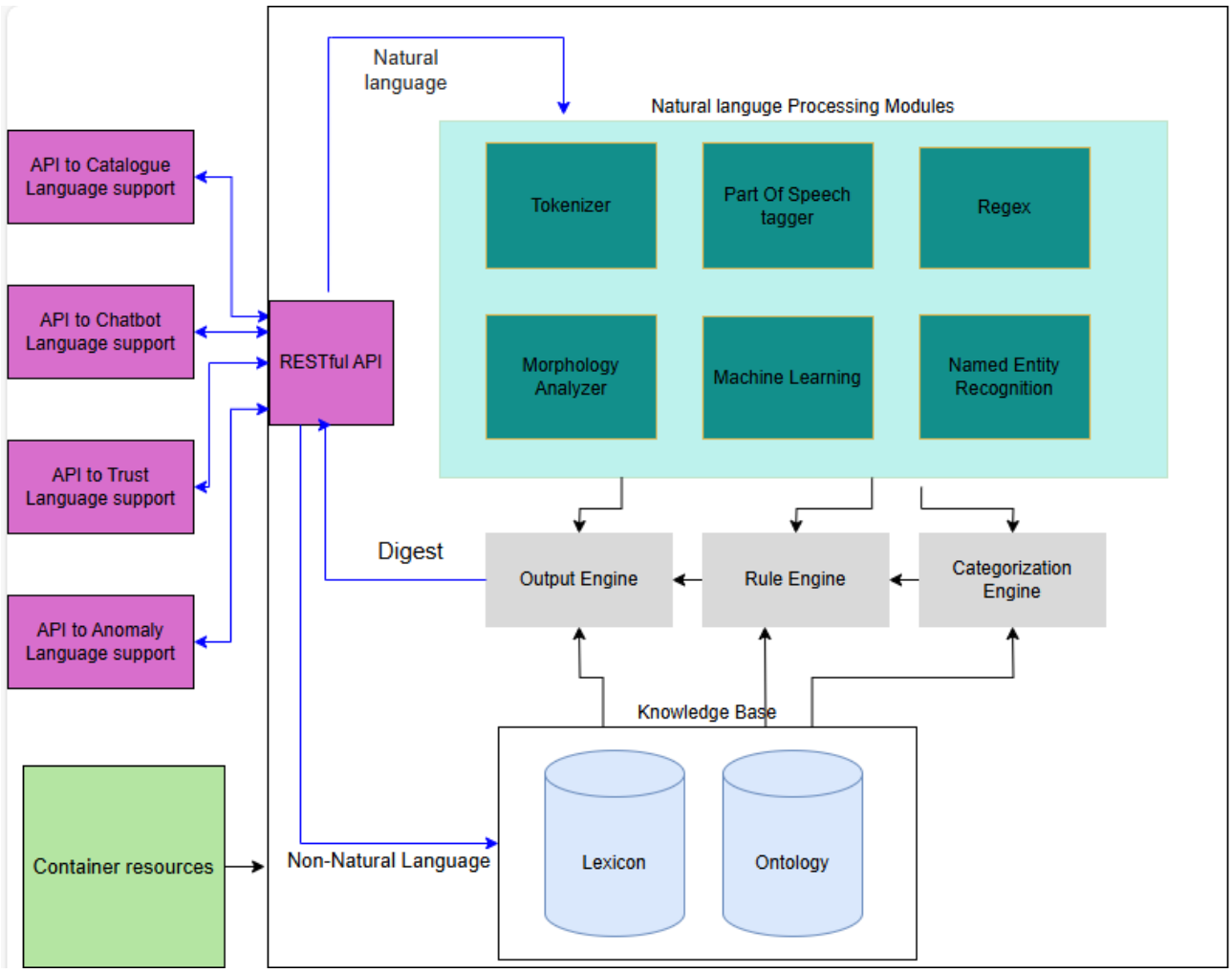


Figure: Internal Architecture of module

This module has the following subcomponent and other functions:

- **Tokenizer**: Fragments a continuous stream of characters into words.
- **Part of Speech (POS) tagger**: Each token is identified as a part-of-speech (e.g. Adjective, noun, etc.) and tagged accordingly.
- **Regex**: Create ontology elements based on the tokens (e.g. telephone numbers, ISBN code)
- **Morphology Analyzer**: Each token is assigned a list of possible stems (each with its current conjugation).
- **Machine Learning**: This block is responsible to provide features needed by subsequent machine language blocks (namely the Named Entity Recognition and the Scoring Engine). Features are mainly patterns of ontological elements such street addresses (composed of city, street and house number) and relationships (two entities related by connection).

- **Named Entity Recognition:** This block is responsible to identify entities even if they were unknown at the time of the ontology creation (places, people, organizations, events and objects).
- **Lexicon**: A glossary of words in each supported language with corresponding grammatical and syntactic constraints that are mapped to the language-independent ontology in way that a lexical entry may correspond to one or more ontological concepts under different syntactic and semantic conditions.
- **Ontology**: A set of concepts and categories organized hierarchically in a subject area that shows their properties and the relations between them.
- **Rule Engine**: This is the component responsible to calculate the theatre of the document (the place where most of the locations mentioned belong to), the topic of the document, and many other document properties.
- **Categorization Engine**: This is the component responsible to predict customised properties of the document using machine learning. For instance, giving a metadata-description a score saying how well it fits the other metadata-descriptions in the same category which the supplier manually specified.
- **Output Engine**: Create a summary and digest of the text.

Ontology Generation and Customization

The current ontology built in as background was meant to apply to raw text and not to the "computerise" that we find in databases.DS2 entails broadening the capability to texts that are not raw human language but have some connection to human language and to disambiguate it. This is necessary for the normalization (or ontologization) of metadata and for building queries.
The generation of domain ontologies and their mapping to lexicons is currently a labour-intensive mainly manual task. This task will include development of a method for off-line semi-automated generation of new domain ontologies and for on-line warning of gaps in the existing domain ontologies (or language mappings) and for content-based anomaly detection.

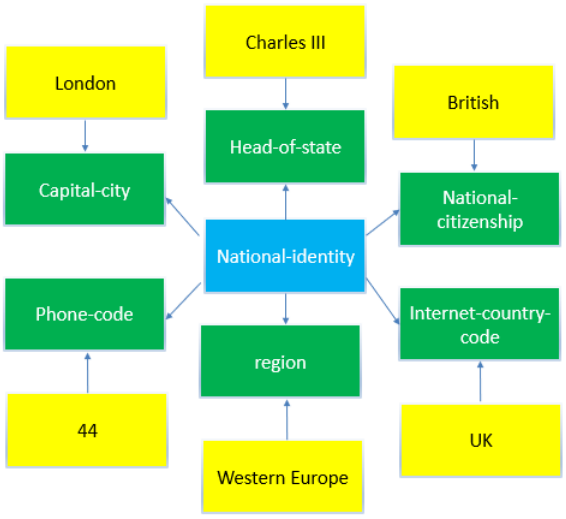
The proposed method for semi-automated domain ontology generation is as follows:

- Create a training set of documents relating to the chosen domain and use unsupervised learning to distribute the documents into categories. The process will use the NLP engine to identify morphological variants of different lexical occurrences to normalize them to their base stem.
- Extract key monograms/bigrams/trigrams using TF/IDF methods.
- Match the lexical instances in the documents with entries in a thesaurus of the same language and tag those, which have multiple "parents" in the concepts of the thesaurus. The key words now will be analysed according to thesauruses to find relationships between them. This will create a set of words, which can be tagged as "synonyms" (according to the thesaurus), antonyms or related words.
- Link the lexical instances with their parents in the thesaurus (two to three levels above) and identify the key "parents" of the lexical instances in the training set as opposed to the non-domain training set. Place the parents in the ontology table as "concepts" and sub-concepts and the actual words or phrases as instances.
- Once we have a statistical "bag of key words" of the training set, we can look for the main features of each category in each cluster of documents. These will be the most frequent lexical features normalized to their "root" word. The synonyms now

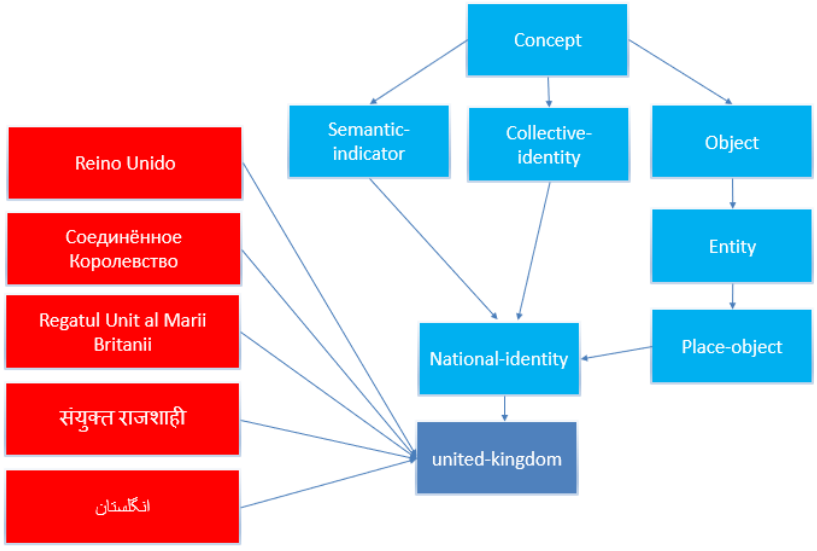
will be linked to each other as separate lexical instances, which lead to an ontological instance. The ontological instance will be the most frequent Key Word with additional information, which leads to its place in the thesaurus.

- A lexicon, which comprises of the key words or phrases, which appear in our domain specific training set as distinct from the frequency of their occurrence in the non-domain, set. If a lexical entry appears in the same frequency (more or less) in the non-domain set, this may indicate ambiguity or, in any case, less relevance to the domain.
- Clusters of entries linked together as synonyms that identify the ontological instance and clusters which are antonyms of those words (deposit will appear in the thesaurus as an antonym of withdraw under the meaning of banking action).
- At this stage, training and testing is performed with manual checks to see if the features of the lexicon and ontology really are typical of the training set.

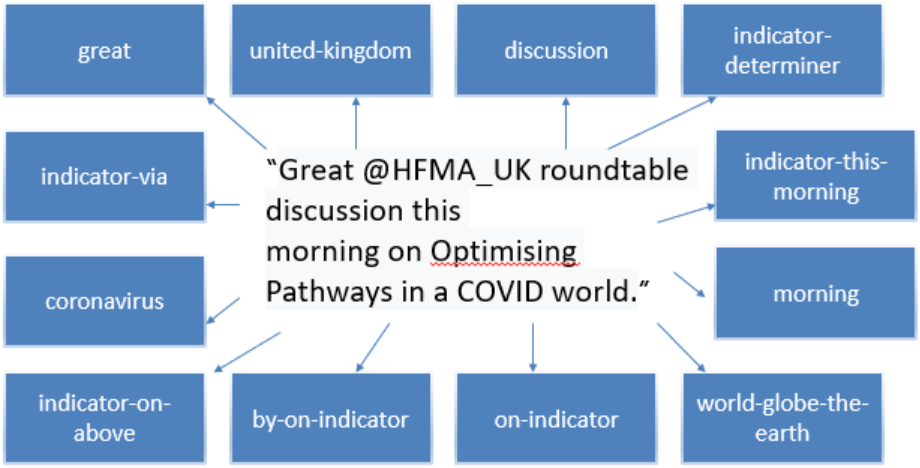
Example of ontology element attributes



Lexicon-Ontology Relationship



Example of a digest of text



1.1.4 Technical Foundations and Background

The module will use Propriety software from INTU as well as specifically developed components for DS2. This includes INTU's patented models for Natural Language Processing including POS, morphological analysis, NER, and Sentiment Analysis

Subcomponent/Component	Owner	License
ntuScan	INTU	Propriety

1.1.5 Interaction of the Component

The interactions of the NLP/Ontological module as seen in this stage of the project are as follows:

- DS2 DARC Module: The chatbot part of the DARC is the first and main component that can benefit from the CLM module as it interacts directly with human natural language. The CLM module will convert a natural language phrase (in any supported language) that is input from a human user into an ontological digest and categorization score and pass the results on to the Chatbot. This digest will be matched to data in the catalogue, enabling the Chatbot component to reply to the user with suggestions for responses to their query. Once the dialogue of the chatbot with the user decides the correct domain specific meaning of a term it will be able to propagate the information back to the module to simplify future interactions. A user in one dataspace may write that they are looking for information on "courts". This could mean a court of law, a tennis/basketball court, a royal court or a schoolyard. Since the lexeme "court" is mapped in the lexicon to all three - the user will then be asked which they want. This will enable the CLM to disambiguates and send a query for information (e.g. on court-of-law). Since the same normalization has already been applied to the information from dataspaces in the catalogue - the query will be exact.
- Catalogue Module: To enable normalized access to the catalogue, it will include, in addition to the input data, a representation of the same data in a language independent, monosemic, and searchable form. This representation will be an

ontological digest graph of the input data that is generated by the module and enables drill-down in the ontological hierarchy (from general concepts to specific instances) and discovery of implicit affinities through attributes of instances. This format will be generated for the catalogue both for metadata and natural language text.

- DS2 Data Detection and Data Transformation Module: the module may participate in anomaly detection of non-natural language data that is influenced by the source languages. The values of metadata may differ from one DS to another. The Metadata Broker sends the data to the module, which converts it to normalized ontological format and identifies possible anomalies in the data. Such anomalies may be lower and upper bounds of physical measurements, dates, et alia. present incumbent of selected office positions, etc. This information should be further enhanced in the configuration phase and manually verified by the data supplier. This may augment the functionality provided by the INDRA and SAG modules.

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
Chatbot	Gives and receives	The Chatbot passes on the raw dialogue with the participant to the DS2 CLM module which returns an ontological representation of the query (ontological digest and classification) which allows the Chatbot to access the catalogue and offer the participant relevant answers to his query.
Catalogue	Receive From	The human language and associated metadata are uploaded to the Catalogue and sent to the DS2 CLM module for analysis which then returns the ontological digest and classification of the information.
Catalog Manager Metadata Broker	Give To	Matches metadata that is not human language but may reflect use of different wording for the same meaning.

1.1.6 Technical Risks

Risk	Description	Contingency Plan
No cross-language domains	No case study on same domain across languages or data types	Dialogue with the Use Case partners to define the domains and the data types.
No training data for models	Privacy or other constraints in providing data	Pre-definition of the concepts to be passed on after analysis
Synchronization of the NLP with the chatbot features	The chatbot must be built based on a search mechanism and only then implementation of the NLP	A template for Q&A will be prepared which can integrate into a chatbot
Short text NLP across languages	To determine the languages of queries and mapping from the natural language to query language	Adopt the proper Multilingual Language Model.

Conversational Agent fails human-centric aspects	Failing to understand the human-centric aspects of complex DLCs is critical for the successful implementation of the Conversational Agent.	Outreach to the users in the project to receive the query languages and parameters of their DB
Chatbot method turns out unreliable after effort	Many of the queries can be dealt with simple ontological wizards - this may turn out too late	Specify and request training data from uses cases for improving the accuracy of the AI-powered chatbot.
Execution time exceed limits	The time required to process a text is not bound.	Set reasonable limits.
Very large texts	Extremely large amounts of data	Set reasonable limits.

1.1.7 Security

Security Issue	Description	Need
Inter-participant trust	Certain Ontology packages may be restricted to designated partners	Management of permissions
Protection of the Ontology	In case of cyber attacks	Encryption

1.1.8 Data Governance

Data Governance Issue	Description	Need
Metadata Catalogue	This module will access metadata associated with a dataset from the Metadata Catalogue	Published Metadata Catalogue information will be freely available to all entities that are allowed to access the Metadata Catalogue.
Handling of business process	Business processes may represent confidential information	The NLP/Ontological component may be configured to prevent such information from being transferred if it exists or to issue a warning.
Handling of personal data	This component should be GDPR compliant and to include constraints to prevent exposure of personal data.	User/Provider should ensure personal data transferred is transferred according to relevant regulations. The NLP/Ontological component may be configured to prevent such information from being transferred if it exists or to issue a warning.

1.1.9 Requirements and Functionality

This module will be used in the following use cases:

Agriculture	✓
City Scape	✓
Green Deal	✓
Intersector	✓

The 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 place between pilot partner and module partner to determine if a fit or the scope of the precise fit.

In this specific case as a system module, it was not anticipated there would be any specific user requirement issue and so far, this is the case as can be seen from the table below.

WHERE	WHAT	WHY	Run/Design Time	Priority
	Use Case 1: Agriculture			
N/A	N/A	N/A	N/A	N/A
	Use Case 2: Agriculture			
N/A	N/A	N/A	N/A	N/A
	Use Case 1: Green Deal			
N/A	N/A	N/A	N/A	N/A

1.1.10 Workflows

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

1.1.10.1 Disambiguate phrase request

This feature provides the capability to convert a phrase into a list of abstract monosemic terms that will aid the requestor disambiguate the phrase.

The main steps/functionalities are as follows:

- Put textual data as REST request
- Get answer as Json

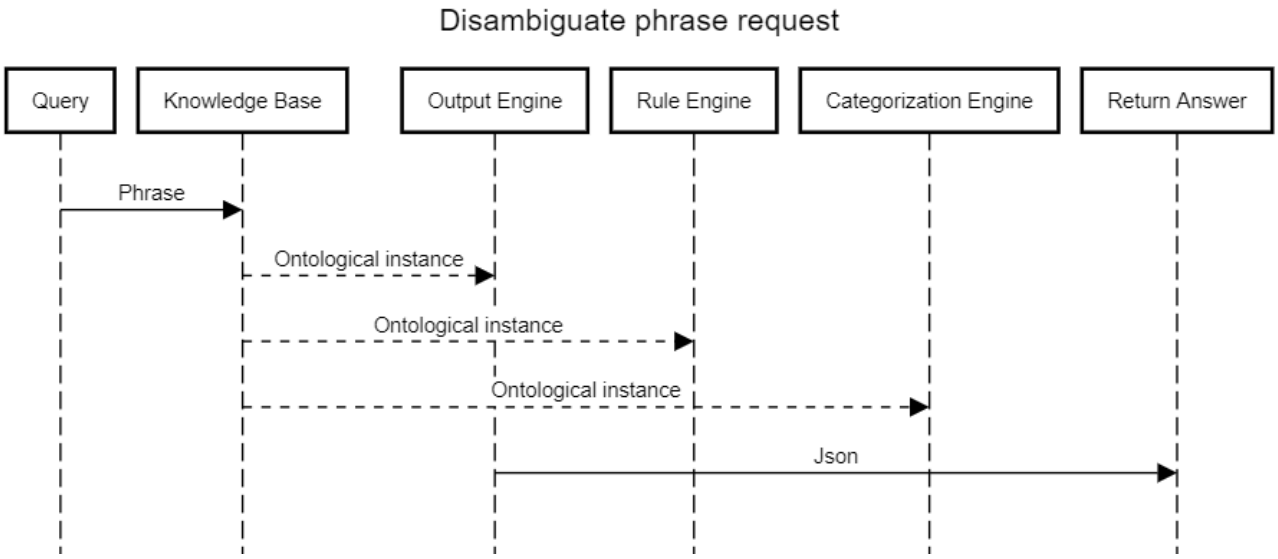


Figure 1: Disambiguation sequence diagram

1.1.10.2 Generate digest request

This feature provides the capability to summarize a text into a graph-like structure of hierarchical monosemic ontological terms that will enable the requestor to normalize compare the text to other texts.

The main steps/functionalities are as follows:

- Put textual data as REST request
- Get answer as RDF/XML

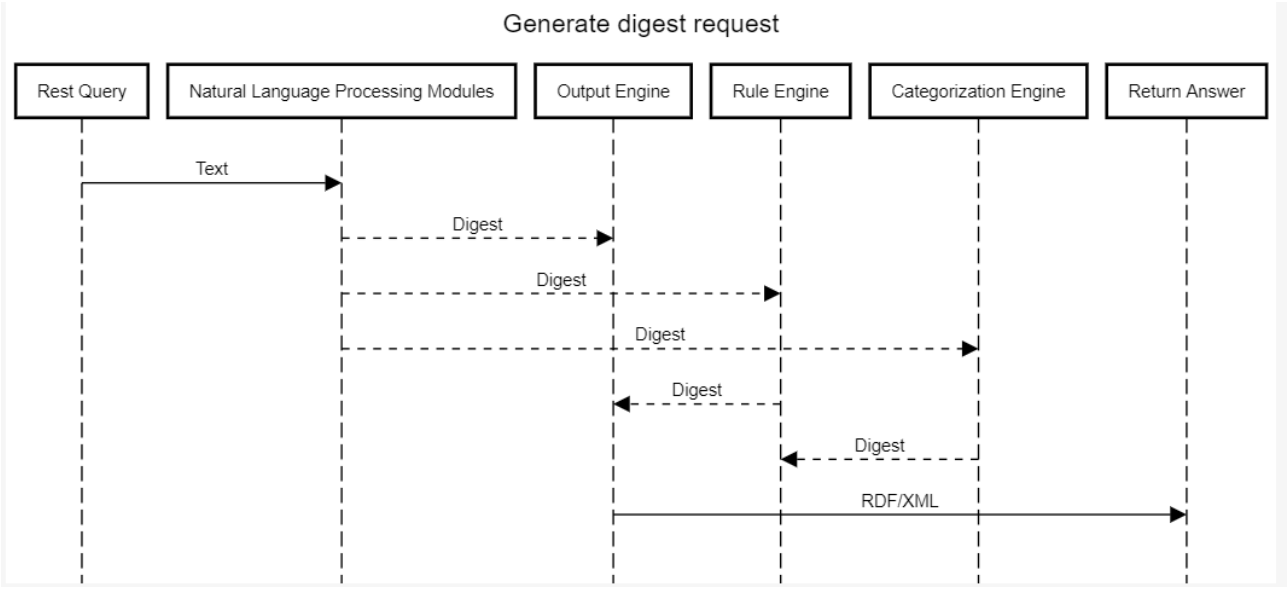


Figure 2: Generate digest sequence diagram

1.1.11 Role, Resourcing, and Milestones

Sub-component	Main Activity	INTU Person Months	M18	M24	M30	M36
Participation in discussions with other partners to find other areas for integration		2				
Romanian NLP	Research into the language	4				
Romanian Tokenizer	Research and implementation of tokenization rules	1				
Romanian Morphology	Research and building of a morphological analyser, including stems, domain-specific terminology, etc.	2				
Romanian POS models	Collection, annotation, and training of models	2				
Annotation of texts in Romanian for SNER	Collection, annotation, and training of models	2				
Integration and Testing	Building the Romanian module	4				
Slovenian NLP	Research into the language	3				
Slovenian Tokenizer	Research and implementation of tokenization rules	1				
Slovenian Morphology	Research and building of a morphological analyser, including stems, domain-specific terminology, etc.	2				
Slovenian POS models	Collection, annotation, and training of models	2				
Annotation of texts in Slovenian for SNER	Collection, annotation, and training of models	2				
Integration and Testing	Building the Slovenian module	4				
Greek NLP	Research into the language	4				
Greek Tokenizer	Research and implementation of tokenization rules	1				
Greek Morphology	Research and building of a morphological analyser, including stems, domain-specific terminology, etc.	2				
Greek POS models	Collection, annotation, and training of models	2				
Annotation of texts in Greek for SNER	Collection, annotation, and training of models	2				
Integration and Testing	Building the Greek module	4				
Ontology Generation per domain	Collection of data, classification and generating manual domain-specific ontologies	2				

Integration into the Catalogue	Database adaptation for ontological data, optimization of the process of exchange of information between the originating dataspace, the CLM module and the Catalogue	2				
Integration into the Chatbot	Rules for the exchange of information between the Chatbot and the CLM module	3				
API	Building bespoke API for all DS2 modules	6				
Automated Ontology R&D	R&D and testing of a method for automated ontology building and anomaly alerts in the existing ontology	4				

1.1.12 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
Use cases with human natural language	Add to each use case elements of human natural language	Wait for decision	Use Cases
APIs	Specifics of the API awaiting decisions by other modules	Further connector research	All modules using the ontology module