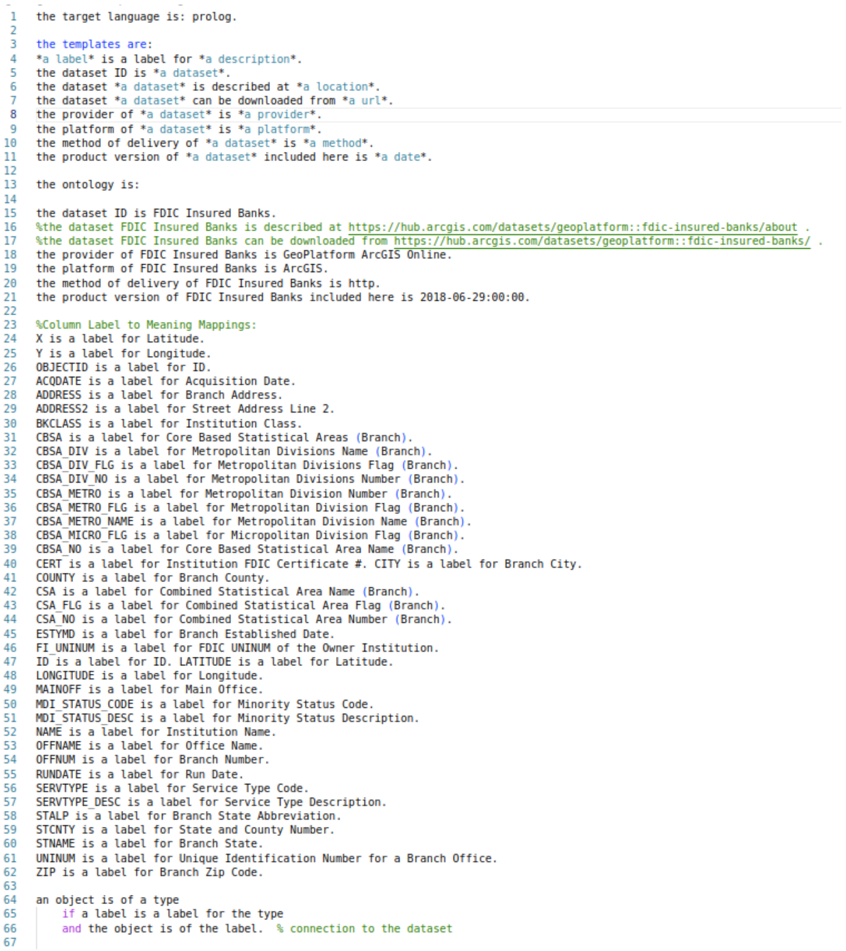
Logical English to annotate metadata and augment semantics descriptions

Jacinto Dávila[1], Miguel Calejo[2] and Andrew Noble[3]

Introduction

We present Logical English (Kowalski et al, 2023), LE, as a suitable alternative to annotate metadata of CSV datasets and augment semantics descriptions. Logical English is a controlled natural language, designed as syntactic sugar for logic programming languages such as Prolog and Datalog. It can used to write documents with a formal structure that can be mapped to computer code, while preserving the readability of a text in English.

For the purposes of the 2025 OMG Semantic Augmentation Challenge[4], we have prepared an example of a LE for the given dataset FDIC Insured Banks, containing descriptions, in clear English, for each label of the CSV files (taken from the ArcGIS platform) and a collection of statements with the connections between those labels and a number of ontologies, including the suggested (see Figures 1 and 2).



```
the target language is: prolog.

the templates are:
*a label* is a label for *a description*.
the dataset ID is *a dataset*.
the dataset *a dataset* is described at *a location*.
the dataset *a dataset* can be downloaded from *a url*.
the provider of *a dataset* is *a provider*.
the platform of *a dataset* is *a platform*.
the method of delivery of *a dataset* is *a method*.
the product version of *a dataset* included here is *a date*.

the ontology is:

the dataset ID is FDIC Insured Banks.
%the dataset FDIC Insured Banks is described at https://hub.arcgis.com/datasets/geoplatform::fdic-insured-banks/about .
%the dataset FDIC Insured Banks can be downloaded from https://hub.arcgis.com/datasets/geoplatform::fdic-insured-banks/ .
the provider of FDIC Insured Banks is GeoPlatform ArcGIS Online.
the platform of FDIC Insured Banks is ArcGIS.
the method of delivery of FDIC Insured Banks is http.
the product version of FDIC Insured Banks included here is 2018-06-29:00:00.

%Column Label to Meaning Mappings:
X is a label for Latitude.
Y is a label for Longitude.
OBJECTID is a label for ID.
ACQDATE is a label for Acquisition Date.
ADDRESS is a label for Branch Address.
ADDRESS2 is a label for Street Address Line 2.
BKCLASS is a label for Institution Class.
CBSA is a label for Core Based Statistical Areas (Branch).
CBSA_DIV is a label for Metropolitan Divisions Name (Branch).
CBSA_DIV_FLG is a label for Metropolitan Divisions Flag (Branch).
CBSA_DIV_NO is a label for Metropolitan Divisions Number (Branch).
CBSA_METRO is a label for Metropolitan Division Number (Branch).
CBSA_METRO_FLG is a label for Metropolitan Division Flag (Branch).
CBSA_METRO_NAME is a label for Metropolitan Division Name (Branch).
CBSA_MICRO_FLG is a label for Micropolitan Division Flag (Branch).
CBSA_NO is a label for Core Based Statistical Area Name (Branch).
CERT is a label for Institution FDIC Certificate #. CITY is a label for Branch City.
COUNTY is a label for Branch County.
CSA is a label for Combined Statistical Area Name (Branch).
CSA_FLG is a label for Combined Statistical Area Flag (Branch).
CSA_NO is a label for Combined Statistical Area Number (Branch).
ESTYMD is a label for Branch Established Date.
FI_UNINUM is a label for FDIC UNINUM of the Owner Institution.
ID is a label for ID. LATITUDE is a label for Latitude.
LONGITUDE is a label for Longitude.
MAINOFF is a label for Main Office.
MDI_STATUS_CODE is a label for Minority Status Code.
MDI_STATUS_DESC is a label for Minority Status Description.
NAME is a label for Institution Name.
OFFNAME is a label for Office Name.
OFFNUM is a label for Branch Number.
RUNDATE is a label for Run Date.
SERVTYPE is a label for Service Type Code.
SERVTYPE_DESC is a label for Service Type Description.
STALP is a label for Branch State Abbreviation.
STCNTY is a label for State and County Number.
STNAME is a label for Branch State.
UNINUM is a label for Unique Identification Number for a Branch Office.
ZIP is a label for Branch Zip Code.

an object is of a type
    if a label is a label for the type
    and the object is of the label.  % connection to the dataset

```

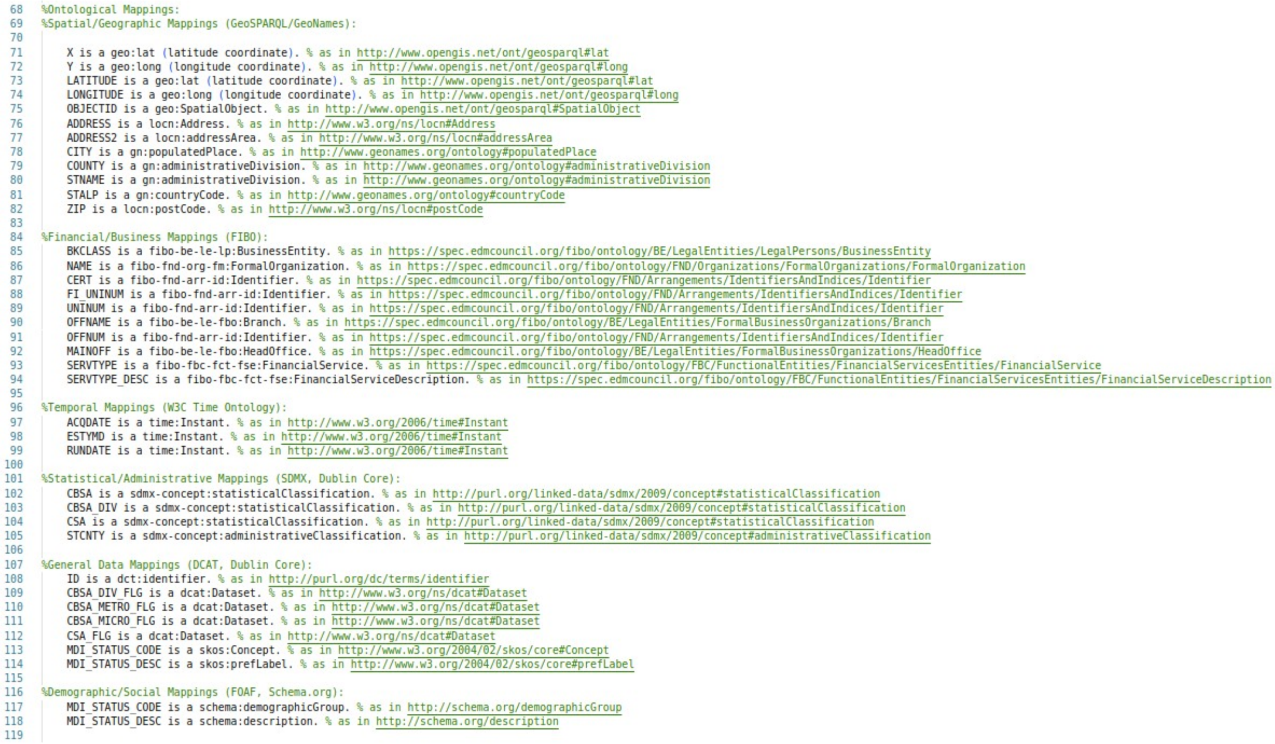
Figura 1: A Logical English Document as metadata

[1] jd@logicalcontracts.com, jacinto@ula.ve

[2] mc@logicalcontracts.com

[3] andrew@nobleaccounting.com.au

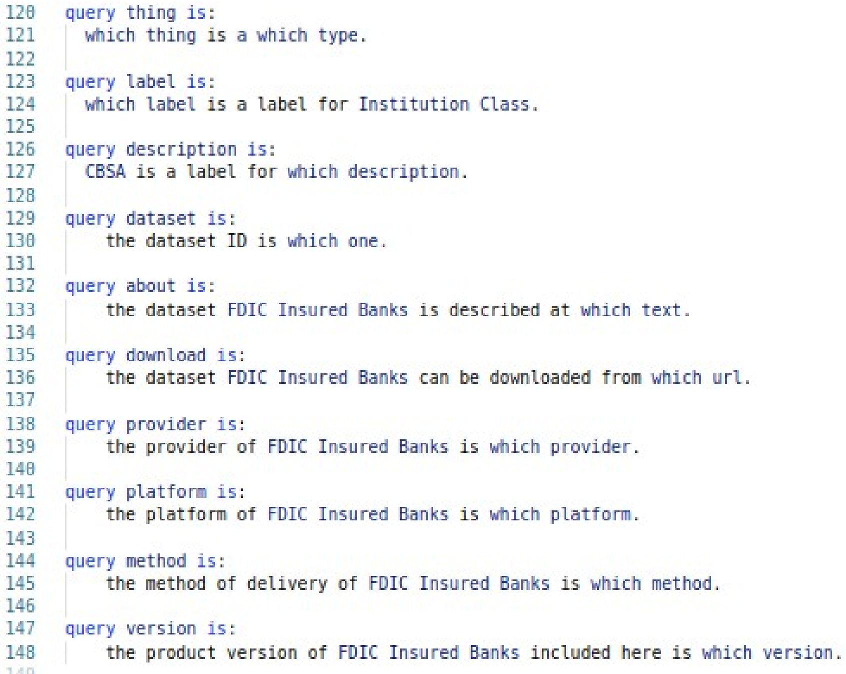
[4] https://www.omg.org/events/omg-semantic-augmentation-challenge/



```
%Ontological Mappings:
%Spatial/Geographic Mappings (GeoSPARQL/GeoNames):

    X is a geo:lat (latitude coordinate). % as in http://www.opengis.net/ont/geosparql#lat
    Y is a geo:long (longitude coordinate). % as in http://www.opengis.net/ont/geosparql#long
    LATITUDE is a geo:lat (latitude coordinate). % as in http://www.opengis.net/ont/geosparql#lat
    LONGITUDE is a geo:long (longitude coordinate). % as in http://www.opengis.net/ont/geosparql#long
    OBJECTID is a geo:SpatialObject. % as in http://www.opengis.net/ont/geosparql#SpatialObject
    ADDRESS is a locn:Address. % as in http://www.w3.org/ns/locn#Address
    ADDRESS2 is a locn:addressArea. % as in http://www.w3.org/ns/locn#addressArea
    CITY is a gn:populatedPlace. % as in http://www.geonames.org/ontology#populatedPlace
    COUNTY is a gn:administrativeDivision. % as in http://www.geonames.org/ontology#administrativeDivision
    STNAME is a gn:administrativeDivision. % as in http://www.geonames.org/ontology#administrativeDivision
    STALP is a gn:countryCode. % as in http://www.geonames.org/ontology#countryCode
    ZIP is a locn:postCode. % as in http://www.w3.org/ns/locn#postCode

%Financial/Business Mappings (FIBO):
    BKCLASS is a fibo-be-le-lp:BusinessEntity. % as in https://spec.edmcouncil.org/fibo/ontology/BE/LegalEntities/LegalPersons/BusinessEntity
    NAME is a fibo-fnd-org-fm:FormalOrganization. % as in https://spec.edmcouncil.org/fibo/ontology/FND/Organizations/FormalOrganizations/FormalOrganization
    CERT is a fibo-fnd-arr-id:Identifier. % as in https://spec.edmcouncil.org/fibo/ontology/FND/Arrangements/IdentifiersAndIndices/Identifier
    FI_UNINUM is a fibo-fnd-arr-id:Identifier. % as in https://spec.edmcouncil.org/fibo/ontology/FND/Arrangements/IdentifiersAndIndices/Identifier
    UNINUM is a fibo-fnd-arr-id:Identifier. % as in https://spec.edmcouncil.org/fibo/ontology/FND/Arrangements/IdentifiersAndIndices/Identifier
    OFFNAME is a fibo-be-le-fbo:Branch. % as in https://spec.edmcouncil.org/fibo/ontology/BE/LegalEntities/FormalBusinessOrganizations/Branch
    OFFNUM is a fibo-fnd-arr-id:Identifier. % as in https://spec.edmcouncil.org/fibo/ontology/FND/Arrangements/IdentifiersAndIndices/Identifier
    MAINOFF is a fibo-be-le-fbo:HeadOffice. % as in https://spec.edmcouncil.org/fibo/ontology/BE/LegalEntities/FormalBusinessOrganizations/HeadOffice
    SERVTYPE is a fibo-fbc-fct-fse:FinancialService. % as in https://spec.edmcouncil.org/fibo/ontology/FBC/FunctionalEntities/FinancialServicesEntities/FinancialService
    SERVTYPE_DESC is a fibo-fbc-fct-fse:FinancialServiceDescription. % as in https://spec.edmcouncil.org/fibo/ontology/FBC/FunctionalEntities/FinancialServicesEntities/FinancialServiceDescription

%Temporal Mappings (W3C Time Ontology):
    ACQDATE is a time:Instant. % as in http://www.w3.org/2006/time#Instant
    ESTYMD is a time:Instant. % as in http://www.w3.org/2006/time#Instant
    RUNDATE is a time:Instant. % as in http://www.w3.org/2006/time#Instant

%Statistical/Administrative Mappings (SDMX, Dublin Core):
    CBSA is a sdmx-concept:statisticalClassification. % as in http://purl.org/linked-data/sdmx/2009/concept#statisticalClassification
    CBSA_DIV is a sdmx-concept:statisticalClassification. % as in http://purl.org/linked-data/sdmx/2009/concept#statisticalClassification
    CSA is a sdmx-concept:statisticalClassification. % as in http://purl.org/linked-data/sdmx/2009/concept#statisticalClassification
    STCNTY is a sdmx-concept:administrativeClassification. % as in http://purl.org/linked-data/sdmx/2009/concept#administrativeClassification

%General Data Mappings (DCAT, Dublin Core):
    ID is a dct:identifier. % as in http://purl.org/dc/terms/identifier
    CBSA_DIV_FLG is a dcat:Dataset. % as in http://www.w3.org/ns/dcat#Dataset
    CBSA_METRO_FLG is a dcat:Dataset. % as in http://www.w3.org/ns/dcat#Dataset
    CBSA_MICRO_FLG is a dcat:Dataset. % as in http://www.w3.org/ns/dcat#Dataset
    CSA_FLG is a dcat:Dataset. % as in http://www.w3.org/ns/dcat#Dataset
    MDI_STATUS_CODE is a skos:Concept. % as in http://www.w3.org/2004/02/skos/core#Concept
    MDI_STATUS_DESC is a skos:prefLabel. % as in http://www.w3.org/2004/02/skos/core#prefLabel

%Demographic/Social Mappings (FOAF, Schema.org):
    MDI_STATUS_CODE is a schema:demographicGroup. % as in http://schema.org/demographicGroup
    MDI_STATUS_DESC is a schema:description. % as in http://schema.org/description

```

Figura 2: The Logical English Metadata includes ontological, (is a), information

The LE document also contain a set of queries which represent typical questions that can be posted to the document, considered as a knowledge base. In this case, questions about the structure and content of the described dataset. These are only referential questions and can be modified or extended at will by any competent writer (Figure 3):

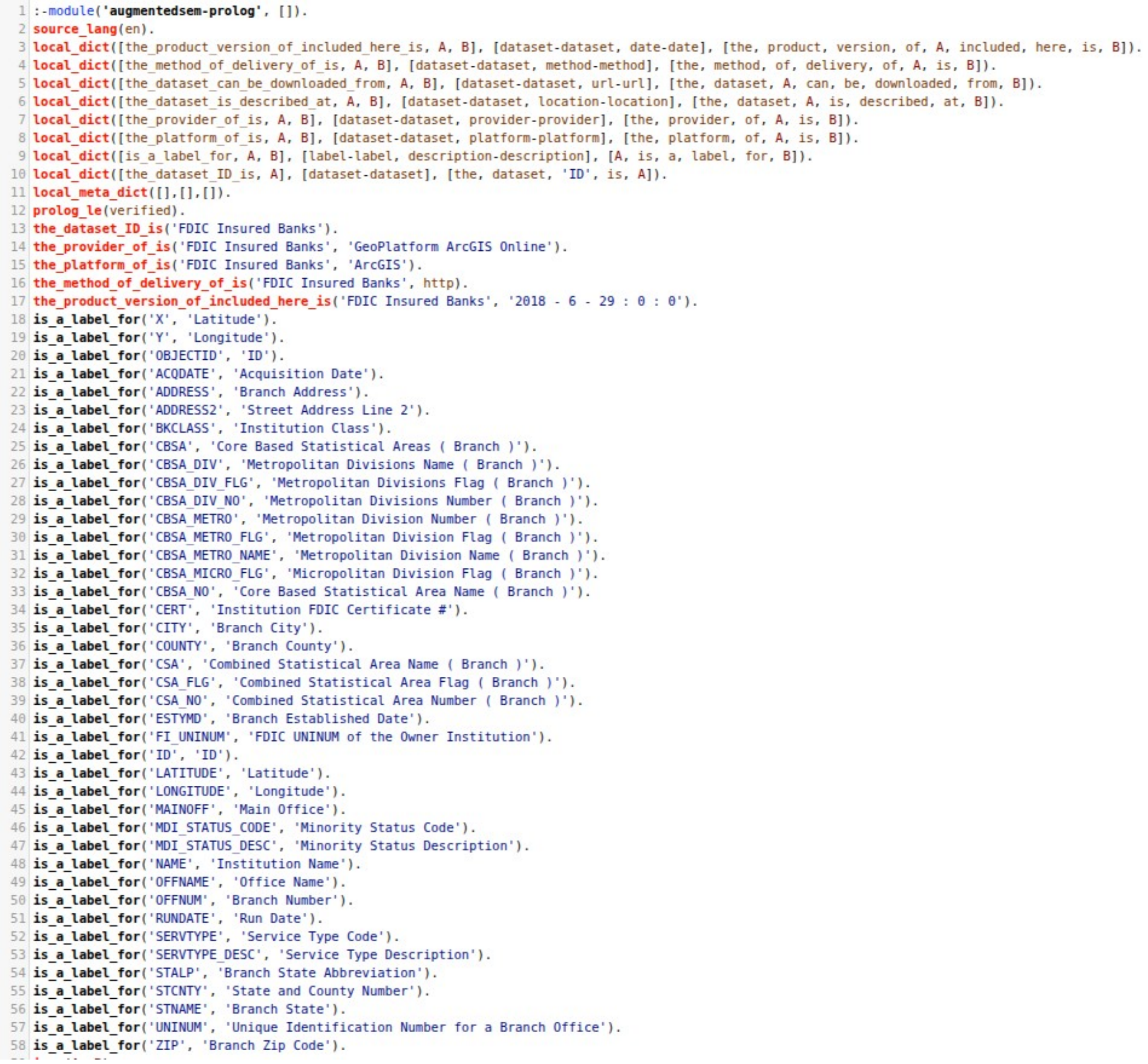


```
query thing is:
  which thing is a which type.

query label is:
  which label is a label for Institution Class.

query description is:
  CBSA is a label for which description.

query dataset is:
    the dataset ID is which one.

query about is:
    the dataset FDIC Insured Banks is described at which text.

query download is:
    the dataset FDIC Insured Banks can be downloaded from which url.

query provider is:
    the provider of FDIC Insured Banks is which provider.

query platform is:
    the platform of FDIC Insured Banks is which platform.

query method is:
    the method of delivery of FDIC Insured Banks is which method.

query version is:
    the product version of FDIC Insured Banks included here is which version.
```

Figura 3: Prototypical Questions for the LE Metadata

The output generated by the format when used with the source dataset.

Whenever is queried, an LE document is translated into a target language. In this case, we selected Prolog in order to combine it with a transformation of the dataset into Datalog. This could, of course, be done in many other ways, including consulting the dataset on the flight with just in time translations.

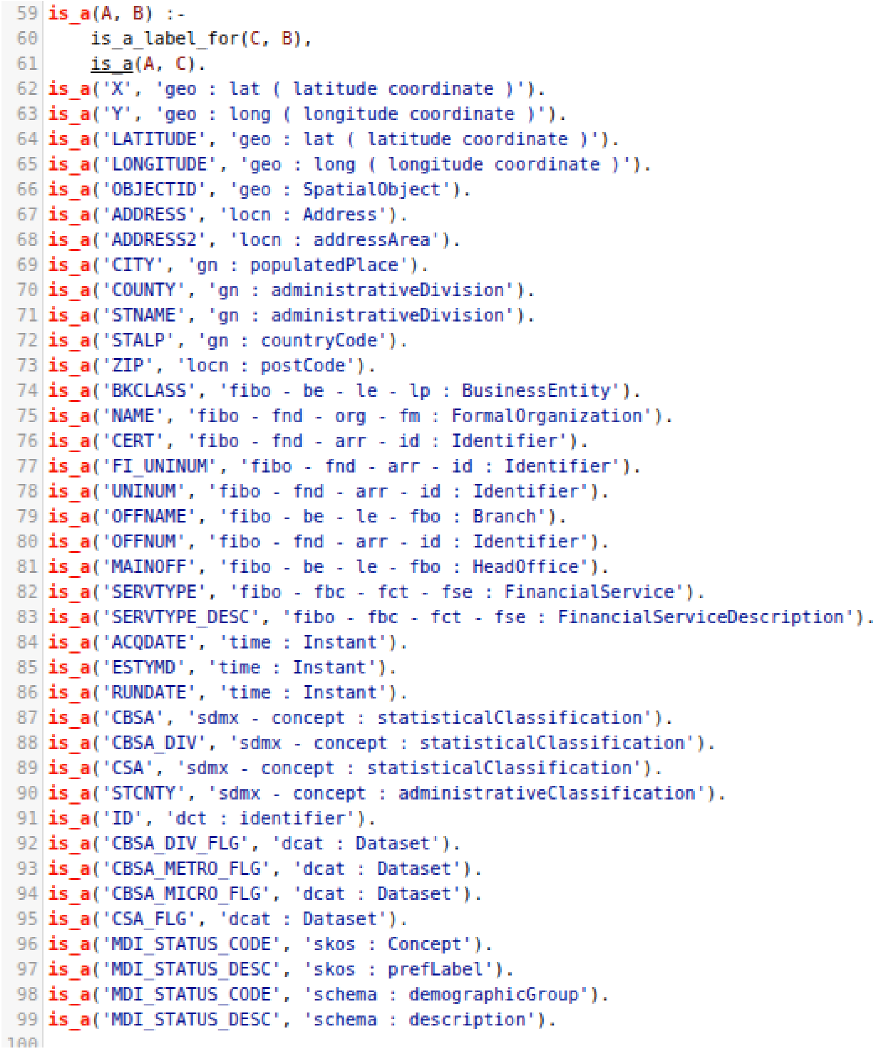


```prolog
:-module('augmentedsem-prolog', []).
source_lang(en).
local_dict([the_product_version_of_included_here_is, A, B], [dataset-dataset, date-date], [the, product, version, of, A, included, here, is, B]).
local_dict([the_method_of_delivery_of_is, A, B], [dataset-dataset, method-method], [the, method, of, delivery, of, A, is, B]).
local_dict([the_dataset_can_be_downloaded_from, A, B], [dataset-dataset, url-url], [the, dataset, A, can, be, downloaded, from, B]).
local_dict([the_dataset_is_described_at, A, B], [dataset-dataset, location-location], [the, dataset, A, is, described, at, B]).
local_dict([the_provider_of_is, A, B], [dataset-dataset, provider-provider], [the, provider, of, A, is, B]).
local_dict([the_platform_of_is, A, B], [dataset-dataset, platform-platform], [the, platform, of, A, is, B]).
local_dict([is_a_label_for, A, B], [label-label, description-description], [A, is, a, label, for, B]).
local_dict([the_dataset_ID_is, A], [dataset-dataset], [the, dataset, 'ID', is, A]).
local_meta_dict([],[],[]).
prolog_le(verified).
the_dataset_ID_is('FDIC Insured Banks').
the_provider_of_is('FDIC Insured Banks', 'GeoPlatform ArcGIS Online').
the_platform_of_is('FDIC Insured Banks', 'ArcGIS').
the_method_of_delivery_of_is('FDIC Insured Banks', http).
the_product_version_of_included_here_is('FDIC Insured Banks', '2018 - 6 - 29 : 0 : 0').
is_a_label_for('X', 'Latitude').
is_a_label_for('Y', 'Longitude').
is_a_label_for('OBJECTID', 'ID').
is_a_label_for('ACQDATE', 'Acquisition Date').
is_a_label_for('ADDRESS', 'Branch Address').
is_a_label_for('ADDRESS2', 'Street Address Line 2').
is_a_label_for('BKCLASS', 'Institution Class').
is_a_label_for('CBSA', 'Core Based Statistical Areas ( Branch )').
is_a_label_for('CBSA_DIV', 'Metropolitan Divisions Name ( Branch )').
is_a_label_for('CBSA_DIV_FLG', 'Metropolitan Divisions Flag ( Branch )').
is_a_label_for('CBSA_DIV_NO', 'Metropolitan Divisions Number ( Branch )').
is_a_label_for('CBSA_METRO', 'Metropolitan Division Number ( Branch )').
is_a_label_for('CBSA_METRO_FLG', 'Metropolitan Division Flag ( Branch )').
is_a_label_for('CBSA_METRO_NAME', 'Metropolitan Division Name ( Branch )').
is_a_label_for('CBSA_MICRO_FLG', 'Micropolitan Division Flag ( Branch )').
is_a_label_for('CBSA_NO', 'Core Based Statistical Area Name ( Branch )').
is_a_label_for('CERT', 'Institution FDIC Certificate #').
is_a_label_for('CITY', 'Branch City').
is_a_label_for('COUNTY', 'Branch County').
is_a_label_for('CSA', 'Combined Statistical Area Name ( Branch )').
is_a_label_for('CSA_FLG', 'Combined Statistical Area Flag ( Branch )').
is_a_label_for('CSA_NO', 'Combined Statistical Area Number ( Branch )').
is_a_label_for('ESTYMD', 'Branch Established Date').
is_a_label_for('FI_UNINUM', 'FDIC UNINUM of the Owner Institution').
is_a_label_for('ID', 'ID').
is_a_label_for('LATITUDE', 'Latitude').
is_a_label_for('LONGITUDE', 'Longitude').
is_a_label_for('MAINOFF', 'Main Office').
is_a_label_for('MDI_STATUS_CODE', 'Minority Status Code').
is_a_label_for('MDI_STATUS_DESC', 'Minority Status Description').
is_a_label_for('NAME', 'Institution Name').
is_a_label_for('OFFNAME', 'Office Name').
is_a_label_for('OFFNUM', 'Branch Number').
is_a_label_for('RUNDATE', 'Run Date').
is_a_label_for('SERVTYPE', 'Service Type Code').
is_a_label_for('SERVTYPE_DESC', 'Service Type Description').
is_a_label_for('STALP', 'Branch State Abbreviation').
is_a_label_for('STCNTY', 'State and County Number').
is_a_label_for('STNAME', 'Branch State').
is_a_label_for('UNINUM', 'Unique Identification Number for a Branch Office').
is_a_label_for('ZIP', 'Branch Zip Code').
```

Figura 4: Logical English translated into Prolog/Datalog

Figure 4 shows the first part of the file with the translation from the original LE document.

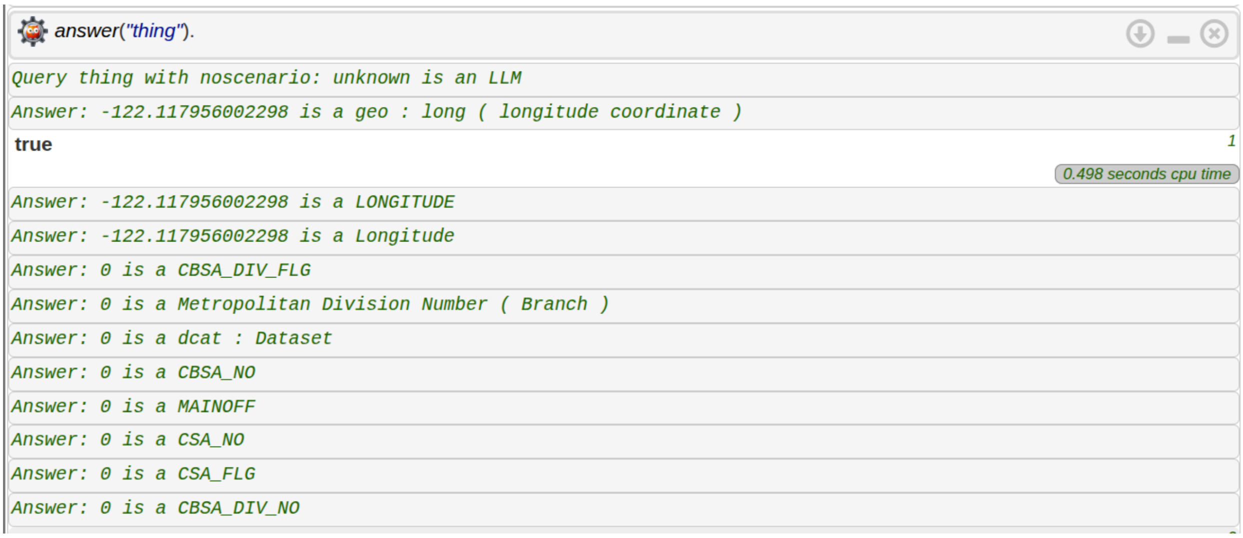
Figure 5 shows the next part which contains a partial rendition of the is_a/2 relation. Those correspond mainly to the ontological information provided by users. Some other essential components of the same relation are depicted in Prolog and are not shown to the final user, like the transitivity rule and the rule that connects the actual data in the dataset with the their types.



```
is_a(A, B) :-
    is_a_label_for(C, B),
    is_a(A, C).
is_a('X', 'geo : lat ( latitude coordinate )').
is_a('Y', 'geo : long ( longitude coordinate )').
is_a('LATITUDE', 'geo : lat ( latitude coordinate )').
is_a('LONGITUDE', 'geo : long ( longitude coordinate )').
is_a('OBJECTID', 'geo : SpatialObject').
is_a('ADDRESS', 'locn : Address').
is_a('ADDRESS2', 'locn : addressArea').
is_a('CITY', 'gn : populatedPlace').
is_a('COUNTY', 'gn : administrativeDivision').
is_a('STNAME', 'gn : administrativeDivision').
is_a('STALP', 'gn : countryCode').
is_a('ZIP', 'locn : postCode').
is_a('BKCLASS', 'fibo - be - le - lp : BusinessEntity').
is_a('NAME', 'fibo - fnd - org - fm : FormalOrganization').
is_a('CERT', 'fibo - fnd - arr - id : Identifier').
is_a('FI_UNINUM', 'fibo - fnd - arr - id : Identifier').
is_a('UNINUM', 'fibo - fnd - arr - id : Identifier').
is_a('OFFNAME', 'fibo - be - le - fbo : Branch').
is_a('OFFNUM', 'fibo - fnd - arr - id : Identifier').
is_a('MAINOFF', 'fibo - be - le - fbo : HeadOffice').
is_a('SERVTYPE', 'fibo - fbc - fct - fse : FinancialService').
is_a('SERVTYPE_DESC', 'fibo - fbc - fct - fse : FinancialServiceDescription').
is_a('ACQDATE', 'time : Instant').
is_a('ESTYMD', 'time : Instant').
is_a('RUNDATE', 'time : Instant').
is_a('CBSA', 'sdmx - concept : statisticalClassification').
is_a('CBSA_DIV', 'sdmx - concept : statisticalClassification').
is_a('CSA', 'sdmx - concept : statisticalClassification').
is_a('STCNTY', 'sdmx - concept : administrativeClassification').
is_a('ID', 'dct : identifier').
is_a('CBSA_DIV_FLG', 'dcat : Dataset').
is_a('CBSA_METRO_FLG', 'dcat : Dataset').
is_a('CBSA_MICRO_FLG', 'dcat : Dataset').
is_a('CSA_FLG', 'dcat : Dataset').
is_a('MDI_STATUS_CODE', 'skos : Concept').
is_a('MDI_STATUS_DESC', 'skos : prefLabel').
is_a('MDI_STATUS_CODE', 'schema : demographicGroup').
is_a('MDI_STATUS_DESC', 'schema : description').
```

Figura 5: The is a relation (partially) in Prolog

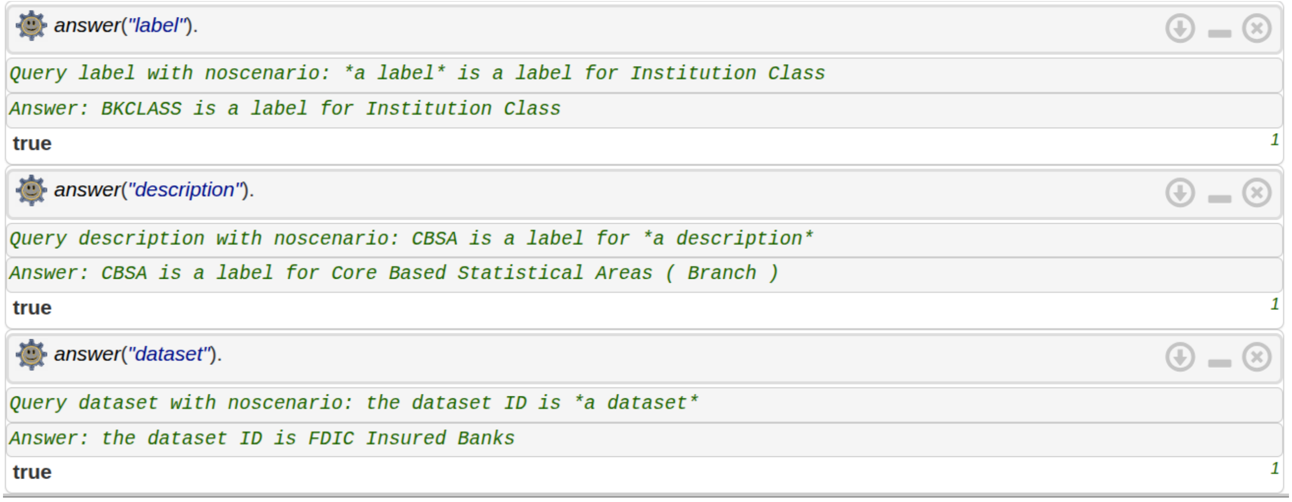
But, in our opinion, a more important contribution of LE to the metadata management is the possibility of quering the document (in English) and obtaining information (also in English) about the related dataset:



```
answer("thing").
Query thing with noscenario: unknown is an LLM
Answer: -122.117956002298 is a geo : long ( longitude coordinate )
true                                                                  1
                                                   0.498 seconds cpu time
Answer: -122.117956002298 is a LONGITUDE
Answer: -122.117956002298 is a Longitude
Answer: 0 is a CBSA_DIV_FLG
Answer: 0 is a Metropolitan Division Number ( Branch )
Answer: 0 is a dcat : Dataset
Answer: 0 is a CBSA_NO
Answer: 0 is a MAINOFF
Answer: 0 is a CSA_NO
Answer: 0 is a CSA_FLG
Answer: 0 is a CBSA_DIV_NO
```

Figura 6: Answering the question "which thing is of which type"

Figure 6 shows one of such type of interaction in which a user ask the document to if it knows about the type of things in it. But, of course, more specific questions are also possible, like in Figure 7:



```
answer("label").
Query label with noscenario: *a label* is a label for Institution Class
Answer: BKCLASS is a label for Institution Class
true                                                                    1
answer("description").
Query description with noscenario: CBSA is a label for *a description*
Answer: CBSA is a label for Core Based Statistical Areas ( Branch )
true                                                                    1
answer("dataset").
Query dataset with noscenario: the dataset ID is *a dataset*
Answer: the dataset ID is FDIC Insured Banks
true                                                                    1
```

Figura 7: Extracting information from the metadata (in English)

Describe any features or limitations with respect to the mapping file.

LE is work in progress and there a few details of operation that must be addressed before it can be offer as companion for any dataset. The software is, however, already open source and can be obtained from: https://github.com/LogicalContracts/LogicalEnglish

We have not tested the translation into RDF yet.

The processing environment(s) in which it was run, including versions of software

Logical English has been developed on SWI-Prolog (threaded, 64 bits, version 9.3.3-200-g7fae34c05). It requires tabling to operate, particularly with the is_a/2 relation. The tests for this submission where done on the swish environment of SWI-Prolog adapted for LE (https://le.logicalcontracts.com/). A serverless version is being tested as well.

Comments on how the format scales with respect to larger datasets.

We could load the whole dataset requested (one CSV file) into Prolog in about 8 seconds:

?- time(load_file('FDIC_Insured_Banks.csv')).
% 65,506,655 inferences, 8.164 CPU in 8.216 seconds (99% CPU, 8023544 Lips)
true.

And then answer (Prolog) queries from the actual data without any overhead (On a regular Dell Intel® Core™ i7-8650U × 8 PC: with 32,0 GiB RAM, running Ubuntu 24.04.2 LTS)

```
?- is_a('St. Louis-St. Charles-Farmington, MO-IL', Type).
Type = 'CSA' .

?- is_a('18001 Saint Rose Rd', Type).
Type = 'ADDRESS' .
```

References

Kowalski, R., Dávila, J., Sartor, G., Calejo, M. (2023). Logical English for Law and Education. In: Warren, D.S., Dahl, V., Eiter, T., Hermenegildo, M.V., Kowalski, R., Rossi, F. (eds) Prolog: The Next 50 Years. Lecture Notes in Computer Science(), vol 13900. Springer, Cham. https://doi.org/10.1007/978-3-031-35254-6_24