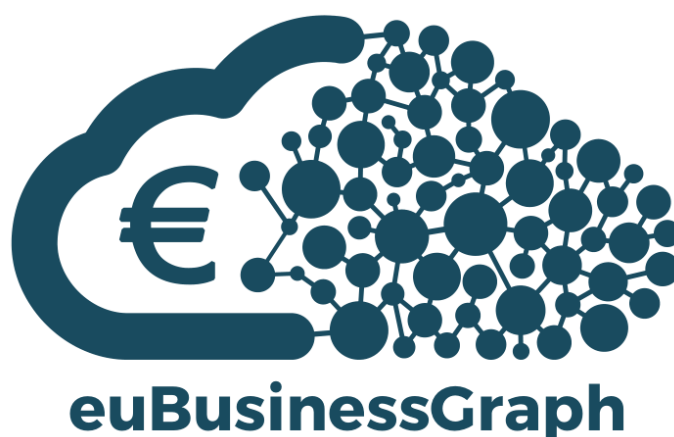


Innovation Action (IA)

ICT-14-2016-2017

H2020-ICT-2016-1

Enabling the European Business Graph for Innovative Data
Products and Services



Deliverable 2.2:

Cross-lingual/Multilingual Data Management Approach for Structured and Unstructured Data

Date	20.12.2018
Author(s)	Inna Novalija (JSI), Miha Jenko (JSI), Luis Rei (JSI), Janez Brank (JSI), Gregor Leban (JSI), M. Beshar Massri (JSI), Javier Paniagua (SDATI), Matteo Palmonari (UNIMIB), Vincenzo Cutrona (UNIMIB), Brian Elvesæter (SINTEF)
Dissemination level	Public (PU)
Work package	WP2
Version	1.0

Document metadata

Quality assurors and contributors

Quality assesor(s)	Divna Djordjevic (Cerved), Kay Macquarrie (Deutsche Welle)
Contributor(s)	euBusinessGraph Consortium

Version history

Version	Date	Description
0.1	01.11.2018	TOC.
0.2	03.12.2018	First round of contributions.
0.3	10.12.2018	Second round of contributions.
0.4	11.12.2018	Editorial updates.
0.5	13.12.2018	Internal review.
0.6	17.12.2018	Updates from contributors.
0.7	19.12.2018	Integrated version.
1.0	20.12.2018	Final formatting and layout.

Table of contents

TABLE OF CONTENTS	3
EXECUTIVE SUMMARY	5
NAMESPACES AND ABBREVIATIONS	6
1 INTRODUCTION	8
1.1 OBJECTIVE	8
1.2 DOCUMENT STRUCTURE	8
2 DATA MANAGEMENT OF CROSS-LINGUAL/MULTILINGUAL DATA	9
2.1 CROSS-LINGUAL/MULTILINGUAL DATA ANNOTATION	9
2.1.1 Objectives	9
2.1.2 Cross-lingual services	9
2.1.2.1 Annotation of text	9
2.1.2.2 Business related vocabularies	12
2.1.2.3 Annotation of structured data	13
3 COMPANY RELATED EVENT AND RELATION EXTRACTION	20
3.1 EVENT TYPES	21
3.1.1 Objectives	21
3.1.2 Data sources	22
3.1.3 Identification of business-related events	23
3.1.3.1 Methodology	23
3.1.3.2 Results	25
3.2 RELATION EXTRACTION	26
3.2.1 Objectives	26
3.2.2 Data sources	26
3.2.3 Tracking business entities through time	27
3.2.3.1 Methodology	27
3.2.3.2 Visualizing results	28
3.3 EUBUSINESSGRAPH EVENT TYPES AND RELATION EXTRACTION TOOLS AND SERVICES	32
4 CROSS-LINGUAL/MULTILINGUAL DATA MANAGEMENT APPLICATION IN EUBUSINESSGRAPH BUSINESS CASES	33
4.1 DJP: DATA JOURNALISM PRODUCT	33
4.1.1 Objectives	33
4.1.2 Data sources	33
4.1.3 Functionalities	33
4.2 ATOKA+	35
4.2.1 Objectives	35
4.2.2 Data sources	35
4.2.3 Functionalities	35
5 UPDATES TO THE EUBUSINESSGRAPH SEMANTIC MODEL (ONTOLOGY)	37
5.1 COMPANY DATA MODEL	37
5.1.1 Datasets	37
5.1.1.1 Dataset hierarchy	38
5.1.1.2 Dataset sources	39
5.1.1.3 Dataset partitions	39
5.1.1.4 Dataset metadata	42
5.1.2 Company officers and memberships	42
5.1.2.1 Officer	44
5.1.2.2 Membership	45
5.1.3 Business graph data onboarding	46
5.2 CROSS-LINGUALITY/MULTILINGUALITY IN EUBUSINESSGRAPH DATA MODEL	47

6	SUMMARY AND OUTLOOK	50
7	REFERENCES	51

Executive summary

The euBusinessGraph project set a goal to tackle semantic annotation and relation extraction tasks for multi-lingual structured and unstructured data.

This deliverable describes the **cross-lingual/multilingual data management approaches for structured and unstructured data annotation** as well as **company-related event and relation extraction**.

First, we provide a description of **annotation services from JSI and UNIMIB**. Data annotation provides functionalities for understanding data sources, such as news, blog posts and tables. As targeted by the euBusinessGraph project, data annotation is performed at content in different languages and annotation results are furthermore used for event and relation extraction.

Second, we present a **methodology for event types categorization** as a procedure that would allow for precise and accurate categorization of textual documents (for instance, news articles in different languages about business, economy and finance) into a taxonomy of business-related event types, such as mergers and acquisitions, bankruptcy and revenues.

Third, we describe the **methodology for tracking business entities through time**, and in particular the "Relation Tracker" tool operating on multi-lingual news data.

Forth, we address the **euBusinessGraph business cases** that are applying tools for cross-lingual and multi-lingual data management. We observe how the cross-lingual and multilingual aims are fulfilled in Data Journalism Product (by DW) and Atoka+ product (by SDATI).

Finally, the **updates to the euBusinessGraph semantic model** are described, with special attention to cross-linguality and multilinguality in the euBusinessGraph data model.

We conclude with a summary and outlook for the development and application of cross-lingual and multilingual data management and relation extraction approaches in the project.

Namespaces and Abbreviations

Table 1: Namespaces used in the document

Schema	Prefix	Namespace
The euBusinessGraph Ontology	ebg	http://data.businessgraph.io/ontology#
DBpedia Ontology	dbo	http://dbpedia.org/ontology/
The Organization Ontology	org	http://www.w3.org/ns/org#
Registered Organization Vocabulary	rov	https://www.w3.org/ns/regorg#
Core Person Vocabulary	person	http://www.w3.org/ns/person#
ISA Programme Location Core Vocabulary	locn	https://www.w3.org/ns/locn#
Schema.org	schema	http://schema.org/
XML Schema	xsd	https://www.w3.org/2001/XMLSchema#
DC Elements	dc	http://purl.org/dc/elements/1.1/
DC Terms	dct	http://purl.org/dc/terms/
IANA language tag list	iana	https://github.com/euBusinessGraph/eubg-data/blob/master/data/iana-lang-tags.csv
EU NUTS classification as Linked Data	nuts	http://nuts.geovocab.org/

Table 1 presents the namespaces used in the document.

Table 2: Abbreviations used in the document

Abbreviation	Description
API	Application Programming Interface
BRC	Brønnøysund Register Centre, a Norwegian data provider in euBusinessGraph
DMOZ	was a multilingual open-content directory of World Wide Web links (from directory.mozilla.org)
DW	Deutsche Welle Innovation, an euBusinessGraph business case partner
euBusinessGraph	euBusinessGraph, a H2020 project to develop a graph of company and related data
EC	European Commission
JSI	Jozef Stefan Institute, an euBusinessGraph partner
JSON	JavaScript Object Notation
IANA	Internet Assigned Numbers Authority. Keeps registers of various entities, including IANA language tags
IPO	Initial public offering
NER	Named-entity recognition
NIF	The NLP Interchange Format
NLP	Natural language processing
OCORP	OpenCorporates, content and use case partner in euBusinessGraph. Has open data on 130M companies sourced from official registers.
ONTO	Ontotext, a semantic technology partner in euBusinessGraph
OWL	The Web Ontology Language
RDF	Resource Description Framework, the graph data model of the semantic web

SDATI	SpazioDati, content and business case partner in euBusinessGraph. Has rich data on Italian companies integrated from various sources (Atoka.io)
SINTEF	A Norwegian research organization, coordinator of euBusinessGraph
TF-IDF	Term frequency–inverse document frequency
UNIMIB	Università degli Studi di Milano, an euBusinessGraph partner
URI	Uniform Resource Identifier
URL	Uniform Resource Locator, a kind of URI that points to a resource using some internet protocol (HTTP, FTP, mailto, etc.)

Table 2 explains abbreviations used in the document.

1 Introduction

1.1 Objective

The euBusinessGraph project aims at simplifying cross-border and cross-lingual collection, reconciliation, aggregation and analysis of company-related information from several authoritative and non-authoritative sources.

In this document, we present the approaches for cross-lingual and multilingual data management of structured and unstructured data, event types categorization and relation extraction.

Cross-lingual and multilingual data management and data annotation of structured and unstructured data provide functionalities for understanding textual sources, such as news and blog posts along with structured data sources, such as tables.

The objectives behind event types categorization is development of the procedure that would allow for precise and accurate categorization of textual documents (for instance, news articles in different languages about business, economy and finance) into a taxonomy of business-related event types, such as mergers and acquisitions, bankruptcy and revenues.

The objective of relation extraction is obtaining relations between main business entities (people and organizations) based on Wikipedia concepts from textual data. In this way, the business entity can be observed in time and generally expressed relations can be monitored and presented to the users.

A methodology for business event categorization along with results of textual categorization of business-related texts is presented.

In addition, we describe the data sources used for relation extraction and show the methodology for tracking business entities through time. Furthermore, we present a tool providing visualizations obtained from the data.

In this document we demonstrate the cross-linguality and multilinguality applications in euBusinessGraph business cases, in order to validate the business value of developed cross-lingual and multilingual tools.

1.2 Document structure

The rest of the document is structured as follows:

- Section 2 describes the data management of cross-lingual and multilingual data, including cross-lingual and multilingual annotation of unstructured data sources (such as news data) and annotation of structured data sources (such as tables).
- Section 3 describes company related event and relation extraction, in particular business event types categorization and tracking business related entities through time.
- Section 4 provides cross-lingual and multilingual data management application in the euBusinessGraph business cases.
- Section 5 presents updates to the euBusinessGraph semantic model, with a subsection dedicated to the cross-linguality and multilinguality in the euBusinessGraph data model.
- Section 6 provides summarising key contributions and describing next steps.

2 Data management of cross-lingual/multilingual data

This section describes the cross-lingual/multilingual company-related data management approaches. Semantic annotation task includes cross-lingual named entity recognition, disambiguation and linking to different knowledge bases. A set of knowledge bases available for semantic annotation includes databases of companies, people and organisations provided by business partners as well as Wikipedia, which is used also for annotating non-entities. Semantic annotation of text in various languages is relevant to several euBusinessGraph business cases (Data Journalism Product from DW, Atoka+ product from SDATI).

2.1 Cross-lingual/multilingual data annotation

2.1.1 Objectives

The objectives of cross-lingual and multilingual data management and data annotation of structured and unstructured data are to provide functionalities for understanding textual sources, such as news and blog posts along with structured data sources, such as tables.

In particular, the data annotation methods developed and utilized within euBusinessGraph Task 2.3 “Cross-lingual Semantic Annotation for Unstructured Data” and Task 2.4 “Event and Relation Extraction from Multi-lingual Unstructured Data” allow for annotating the news articles in different languages with Wikipedia concepts. The annotation results are furthermore used for event extraction, tracking main entities through time and for extraction of relation between entities.

2.1.2 Cross-lingual services

2.1.2.1 Annotation of text

JSI Wikifier¹ is a web service for annotating multilingual textual documents with concepts from Wikipedia. JSI Wikifier targets the issue of semantic annotation (or semantic enrichment) that is a preparatory step in document processing. Given a document and domain ontology the aim of the annotation service is to identify concepts from the ontology that are present in the document, as well as define where in the document these concepts are mentioned.

Wikification is a type of semantic annotation based on Wikipedia [Mihalcea and Csomai] [Zhang and Rettinger]. In the process of wikification each Wikipedia page provides a concept with links to different pages (concepts) expressed as hyperlinks. In such way, Wikipedia categories and cross-language links can be obtained through Wikipedia pages.

The annotation with JSI Wikifier fits to the requirements of euBusinessGraph tasks, since Wikipedia is a freely available source of information and covers a wide number of languages, with cross-language links available to identify pages that refer to the same concept in different languages.

JSI Wikifier currently supports over 130 languages – the languages with at least 1000 pages available. 60 languages have a Wikipedia presence with at least 100.000 pages.

¹ <http://wikifier.org>

Text	Annotations																								
Theresa May announced on Wednesday evening that she had secured the backing of her cabinet for the agreement, after a five hour meeting. But several ministers were understood to have spoken against it. And there are suggestions of moves among Conservative backbenchers to force a no-confidence vote in her. The prime minister is due to make a Commons statement on the Brexit agreement at 1030 GMT. Mr Raab - a Leave supporter who was promoted to the cabinet to replace David Davis when he quit in protest at Mrs May's Brexit plans - is among a group of senior ministers thought to be unhappy with the agreement.	<table> <tr> <th>PR</th><th>Annotation</th></tr> <tr> <td>0.0308</td><td>Theresa May W D</td></tr> <tr> <td>0.0287</td><td>Brexit D</td></tr> <tr> <td>0.0196</td><td>David Davis (British politician) W D</td></tr> <tr> <td>0.0195</td><td>Greenwich Mean Time W D</td></tr> <tr> <td>0.0155</td><td>Conservative Party (UK) W D</td></tr> <tr> <td>0.0140</td><td>Backbencher D</td></tr> <tr> <td>0.0134</td><td>Motion of no confidence D</td></tr> <tr> <td>0.0120</td><td>House of Commons of the United Kingdom W D</td></tr> <tr> <td>0.0106</td><td>Suffrage W D</td></tr> <tr> <td>0.0106</td><td>Cabinet (government) D</td></tr> <tr> <td>0.0094</td><td>Minister (government) W D</td></tr> </table>	PR	Annotation	0.0308	Theresa May W D	0.0287	Brexit D	0.0196	David Davis (British politician) W D	0.0195	Greenwich Mean Time W D	0.0155	Conservative Party (UK) W D	0.0140	Backbencher D	0.0134	Motion of no confidence D	0.0120	House of Commons of the United Kingdom W D	0.0106	Suffrage W D	0.0106	Cabinet (government) D	0.0094	Minister (government) W D
PR	Annotation																								
0.0308	Theresa May W D																								
0.0287	Brexit D																								
0.0196	David Davis (British politician) W D																								
0.0195	Greenwich Mean Time W D																								
0.0155	Conservative Party (UK) W D																								
0.0140	Backbencher D																								
0.0134	Motion of no confidence D																								
0.0120	House of Commons of the United Kingdom W D																								
0.0106	Suffrage W D																								
0.0106	Cabinet (government) D																								
0.0094	Minister (government) W D																								

Figure 1: Annotation of news (in English) with JSI Wikifier

Figure 1 presents an example of semantic news annotation with JSI Wikifier. JSI Wikifier provides the annotations in the original language, annotations in English along with page rank score for each annotation. The original language for text in Figure 1 was auto detected as English. JSI Wikifier identifies the positions of annotations in text.

Parts of Speech
Color Key: verbs , nouns , adjectives , adverbs .
Click on a word to see a list of corresponding synsets in Wordnet.
Theresa May announced on Wednesday evening that she had secured the backing of her cabinet for the agreement, after a five hour meeting. But several ministers were understood to have spoken against it. And there are suggestions of moves among Conservative backbenchers to force a no-confidence vote in her. The prime minister is due to make a Commons statement on the Brexit agreement at 1030 GMT. Mr Raab - a Leave supporter who was promoted to the cabinet to replace David Davis when he quit in protest at Mrs May's Brexit plans - is among a group of senior ministers thought to be unhappy with the agreement.

Figure 2: Parts of Speech from news (in English) with JSI Wikifier

In addition to semantic annotations, JSI Wikifier can provide part of speech tags (for English) based on WordNet [Miller]. Figure 2 shows the part-of-speech tags for news in English.

Text	Annotations			
<p>Después de una larga espera, los <u>británicos</u> por fin tienen una idea de lo que significará dejar de pertenecer a la <u>Unión Europea (UE)</u> mediante el llamado <u>Brexit</u>. En la noche de este miércoles, luego de cinco horas de discusión con su <u>gabinete de ministros</u>, la <u>primera ministra Theresa May</u> abrió la puerta de su oficina en el <u>número 10 de Downing Street</u> en <u>Londres</u> y habló ante los medios. Lo clave de su breve declaración: el <u>gabinete</u> decidió apoyar un principio de acuerdo para el retiro de <u>Reino Unido</u> de la <u>UE</u>. Aunque el ministro <u>británico</u> para el <u>Brexit</u>, Dominic Raab, anunció este jueves en <u>Twitter</u> que dimitía porque no podía de "buena consciencia" dar apoyo al documento.</p>	PR	Annotation	Annotation (en)	
	0.0271	<u>Unión Europea</u> W D	<u>European Union</u>	>>
	0.0262	<u>10 de Downing Street</u> W D	<u>10 Downing Street</u>	>>
	0.0256	<u>Reino Unido</u> W D	<u>United Kingdom</u>	>>
	0.0256	<u>Theresa May</u> W D	<u>Theresa May</u>	>>
	0.0211	<u>Londres</u> W D	<u>London</u>	>>
	0.0209	<u>Referéndum sobre la permanencia del Reino Unido en la Unión Europea</u> W D	<u>United Kingdom European Union membership referendum 2016</u>	>>
	0.0174	<u>Primer ministro del Reino Unido</u> W D	<u>Prime Minister of the United Kingdom</u>	>>
	0.0139	<u>Twitter</u> W D	<u>Twitter</u>	>>
	0.0138	<u>Europa</u> W D	<u>Europe</u>	>>
	0.0138	<u>Downing Street</u> W D	<u>Downing Street</u>	>>
	0.0125	<u>Consejo de Ministros</u> D	<u>Cabinet (government)</u>	>>

Figure 3: Annotation of news (in Spanish) with JSI Wikifier

Figure 3 presents an example of semantic annotation of news in Spanish. The cross-linguality and multilinguality can be assessed via annotations in Spanish along with annotations in English. "W" presents links to Wikidata classification. "D" provides links to DBPedia entity IRIs and DBPedia properties for specific annotation.

JSI Wikifier's operation is based on PageRank:

- JSI Wikifier identifies phrases (or words) in the input document that refer to a Wikipedia concept.
- JSI Wikifier determines which concept exactly a phrase refers to.
- JSI Wikifier determines which concepts are relevant enough to the document as a whole that they should be included in the output of the system (i.e. presented to the user).

JSI Wikifier uses an approach for processing hyperlinks between Wikipedia pages suggested by Zhang and Rettinger [Zhang and Rettinger]. The challenging task of the wikification process is the disambiguation of the relevant mentioned concepts. For example, in the English-language Wikipedia

article, there are links about »Tesla« that can point to the pages about different entities – such as the inventor, the car manufacturer, units in physics, a band or a film.

Therefore, when »Tesla« phrase occurs in the textual document annotated by JSI Wikifier, a number of candidate annotations are formed and the most relevant annotations are identified.

Local and global approaches can be used for disambiguation – while in the local approach each phrase is disambiguated independently of others, in the global approach the document is taken as a whole. The JSI Wikifier implements the global disambiguation approach based on mention-concept graph, which is a bipartite graph with left set of vertices that correspond to mentions and the right set of vertices that correspond to concepts.

Pagerank values of concepts are used to disambiguate the mentions. If there are edges from a mention 'a' to several concepts 'c', the concept with highest pagerank is chosen as the most relevant for the mention 'a'. This concept is considered supported by the mention 'a'. At the end of the process, concepts that are not supported by any mention are discarded as not being relevant to the input document.

The currently implemented JSI Wikifier processes on average more than 500.000 requests per day. Annotations are provided in JSON format. euBusinessGraph deliverables D3.2 and D3.3 describe the details of JSI Wikifier API.

2.1.2.2 Business related vocabularies

In order to satisfy the requirement of textual annotation in financial and business domains, we have incorporated the functionalities that allow for annotation with extra vocabularies into JSI Wikifier. In particular, the vocabularies based on company registers can be included into the annotation process and as a result, the concepts from these vocabularies will be suggested as annotation results.

Table 3: Extra vocabularies

Vocabulary label	Enabled
angellist	T
argentina	F
australia	T
belgium	F
bolgaria	F
canada	T
jobs_adzuna	T
latvia	F
norway	F
panama_papers	T
slovenia	T
uk	T
usa	T

Table 3 presents a list of possible Extra vocabularies from different sources. Some of the vocabularies are part of JSI Wikifier and are enabled by default ("T"), other can be enabled by request. The extra vocabularies are derived from national company registers (Slovenia, Norway, UK, Australia et al.) and from external collections of documents (Panama Papers, job vacancies data, Angellist²). For instance, job vacancies data collected from Adzuna³ job search engine for the period of 2015-2017 have been analysed and company names have been extracted from job vacancies into extra vocabulary.

In order to produce business related vocabularies, for entity extraction we have introduced **derived entity label generation** as an additional pre-processing step for entity labels. Typically, each business entity comes with one relevant text label. By deriving additional labels from it with a token clustering procedure, we can increase Wikifier's search recall of business entities.

² <https://angel.co>

³ <https://www.adzuna.com>

Derived entity label generation uses tokenization of entity label string and clustering of the tokens into separate groups. TF-IDF approach is applied for token scoring in token clustering.

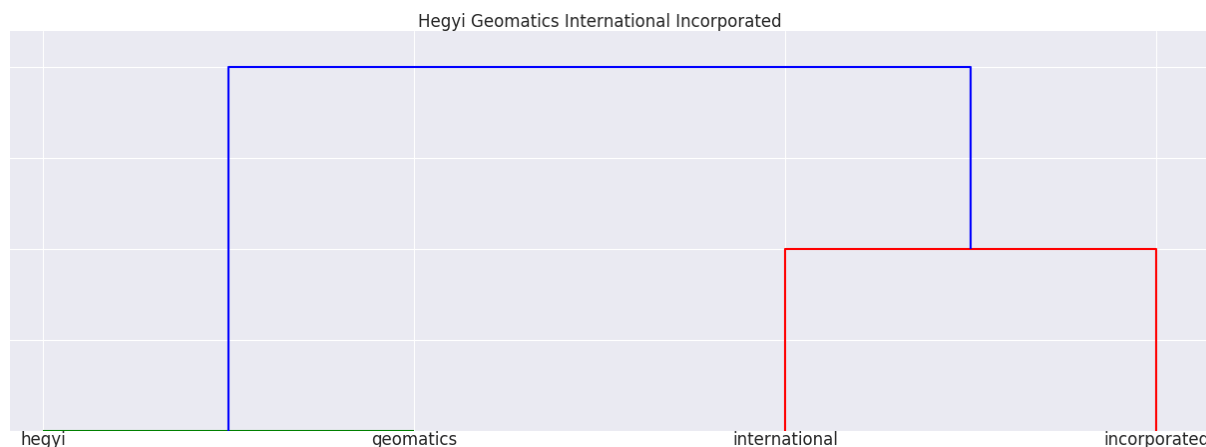


Figure 4: Entity label's tokens clustered into sequential groups (from which derived entity labels can be generated)

It is possible to observe on Figure 4 how hierarchical clustering forms groups of entity label's tokens scored with TF-IDF. Following that, the process involves deriving new labels, whilst keeping sequential order.

The tool provides the users an option to interactively analyse and set a relative token frequency threshold, which discards derived entity label candidates whose tokens have a relative frequency higher than the given threshold. Setting the threshold is useful for ensuring that derived entity labels do not include dataset-specific stop words.

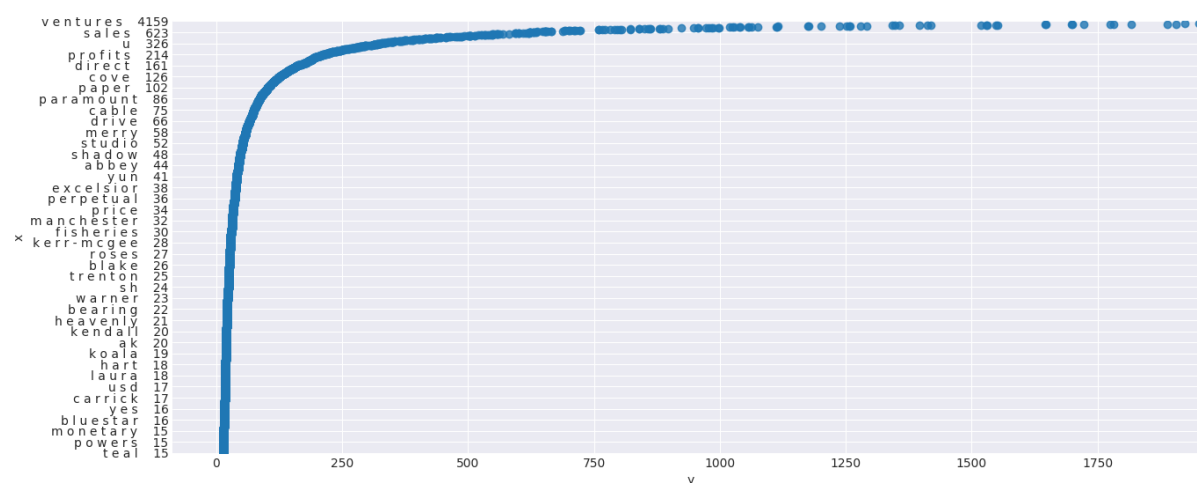


Figure 5: Interactive decision-making utility for setting the token frequency threshold

Figure 5 demonstrates an interactive decision-making utility for setting the token frequency threshold. The horizontal axis shows token frequency in a corpus, the vertical axis displays the tokens and their respective frequencies. The graphical user interface enables the user to explore the token frequency distribution by zooming in and out of selected plot regions, before deciding on a threshold value.

2.1.2.3 Annotation of structured data

Annotation of structured data in euBusinessGraph.

The annotation of structured data provided as tables has the role of supporting data onboarding, under the assumption that tabular formats are frequently used by data providers and it is easy to export legacy data models into tables by using export functionalities available in different database management systems (e.g., relational and NoSql).

Annotations over table elements like columns and cell values can be viewed as implicit mappings from the tabular format to the RDF graph model. In fact, the annotations are translated into specifications of data transformations. When these data transformations are applied to the original table, the facts of the knowledge graph are generated as RDF data. It is useful to distinguish between two kinds of annotations:

- Schema-level annotations, which map the columns and their headers to the terminology of the knowledge graph ontology, in such a way that the generated RDF data will use the desired knowledge graph ontology.
- Instance-level annotations, which map values in the table to shared systems of identifiers, e.g., mapping toponyms to their corresponding identifiers in the Geonames knowledge base. These annotations are used to enrich the content of the original table with shared systems of identifiers.

In euBusinessGraph, ASIA is the annotation tool that extends the functionalities of Grafterizer in such a way that data providers can onboard their data using the reference ontology. ASIA adopts a semi-automatic approach to table annotation, which is explained below.



The screenshot shows the 'Table Annotation panel in Grafterizer 2.0'. It displays a table with 8 rows and 8 columns. The columns are annotated with schema-level information. The annotations are as follows:

	curi	ciduri	name	current_status	jurisdiction_code	incorporation_date	dissolution_date
Type(s)	RegisteredOrganization	Identifier	string	string	string	dateTime	dateTime
Prop	registration	legalName	currentStatus	jurisdiction	foundingDate		
SourceCol	curi	curi	curi	curi	curi	curi	curi

The table data is as follows:

	curi	ciduri	name	current_status	jurisdiction_code	incorporation_date	dissolution_date
1	company/au/620796429	company/au/620796429/id	LOWER RATE FINANCE PTY LTD	FALSE	au	Registered	3
2	company/au/620796492	company/au/620796492/id	THE FOREVER PROJECT PTY LTD	FALSE	au	Registered	3
3	company/au/620796518	company/au/620796518/id	BREAKOUT SCAFFOLDING PTY LTD	FALSE	au	Registered	3
4	company/au/620796545	company/au/620796545/id	GREEN SHOE INVESTMENTS PTY LTD	FALSE	au	Registered	3
5	company/au/620796607	company/au/620796607/id	TURENDO PTY LTD	FALSE	au	Registered	3
6	company/au/620796625	company/au/620796625/id	BENNRHY PTY LTD	FALSE	au	Registered	3
7	company/au/620796634	company/au/620796634/id	KIARRA & SHENAYA PTY LTD	FALSE	au	Registered	3
8	company/au/620796643	company/au/620796643/id	ESHAN & NATASHA PTY LTD	FALSE	au	Registered	3

Figure 6: Schema-level Annotations

Data are annotated with the help of the ASIA user interface. An example of schema-level annotation of a table containing Open Corporate data is depicted in Figure 6; the top grey panel in the table contains the annotations; the name on the top of the panel on the left of the pencil symbol is the header of the column in the table; upon click on the pencil symbol the user can edit the annotations; the annotations are listed right below in the same panel; the green check mark tells that the annotations are valid, which means that valid RDF data can be generated from the annotations. The annotations allow the user to specify the following information.

- Which data type or entity type will be assigned to the values / entities extracted from that column; this annotation is represented as the value of the *Type(s)* attribute; for example, the column with header “curi” is assigned the type RegisteredOrganization from the euBusinessGraph Ontology.
- Whether the values in the column have to be handled as literals or URIs (entities); in the latter case, the pattern used to transform the column values into URIs is also specified; this annotation is implicit in the specification of the type. In case a column is assigned an ontology class as type, the column it will contain entities, otherwise it will contain literals. For example, the column “curi” will contain URIs, i.e., entity identifiers, while the column “name” will contain strings, i.e., literals.
- The role, i.e., subjects and/or objects, which the values in the column will play in the RDF triples generated from the table. *Object columns*: if a column provides triple objects, it means that its values will appear as objects in RDF triples that will be generated from the table. In other words, triples having the form $\langle x_i, p, y_i \rangle$ will be generated for each value y_i that appear in the column. To make sure that the annotation specifies how to generate these triples using the annotated column the annotation specifies also: 1) the source column (*SourceCol* attribute in the annotation panel), i.e., the column from which the triple subjects x_i will be extracted; such a column must be a URI column to comply with RDF model best practices; 2) the RDF property (*prop* attribute in the annotation panel), i.e., the property p that will be used to generate the triples. For example, the column “name” will provide objects in triples whose predicate is legalName from the euBusinessGraph ontology; the triples will have as subjects the values in the column “curi”; in practice the annotation specifies that from each i th row, a triple having the

form $\langle x_i, \text{legalName}, y_i \rangle$ will be generated such that x_i and y_i are taken respectively from the “curi” and “name” column. *Subject* columns: if a column is specified as source column in the annotation of another column, then it is implicitly defined as a subject column. As a result, a column may provide subjects and/or objects in the generated triples. For example, the column “curi” will only play the role of subject column, the column “name” will only play the role of object column, the column “curi” will only play the role of subject column. This example does not report a case when a column is both subject and object, which is possible when one object column is specified as source column in a different column.

In practice, schema-level annotations specify the schema of the RDF data generated from the table, which will use the terminology of the euBusinessGraph ontology.

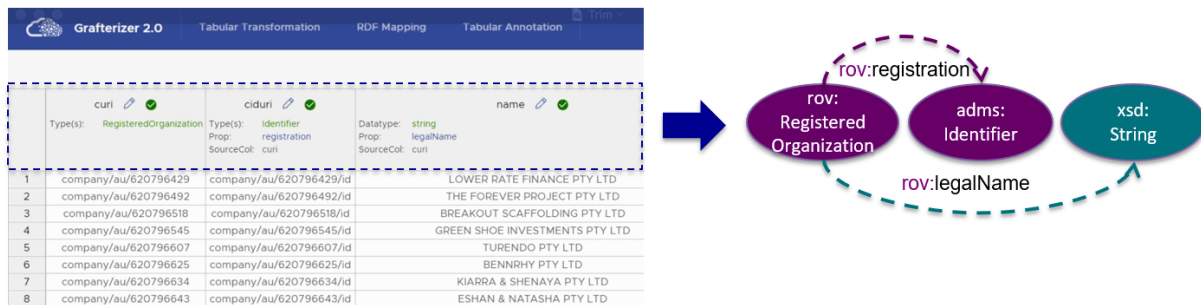


Figure 7: Example of graph schema fragment generated with schema-level annotations

Figure 7 shows an example of schema fragment generated from the table, while Figure 8 shows an example of triples generated from the table according to the specified schema.

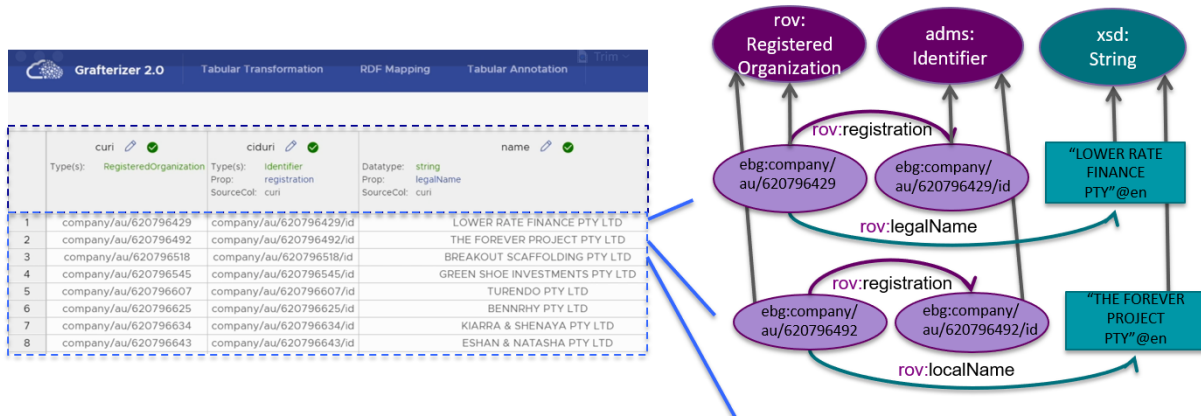


Figure 8: Example of RDF triples generated from the table according to the specified graph schema fragment.

Instance-level annotations are identifiers of entities in a knowledge base attached to the values in the table cells. For convenience, when a user wants to add instance-level annotations to values of a given column, ASIA creates a new column on its right-hand side, which will contain the annotations. An example of column reporting instance-level annotations is given in Figure 9.

Graftizer 2.0 Tabular Transformation RDF Mapping Tabular Annotation						
UPDATE RDF MAPPINGS SETTINGS						
current_status	current_alternative_legal_name	registered_addressstreet_address splitted_0	registered_addressstreet_address splitted_1	locality_GN	registered_ad	
Active		24 Oswald Road, Chorlton Cum	Hardy, Manchester	11599281	Gree	
Active		20 Enfield Road	Swinton	2636371	Gree	
Active		286 Chorley Old Road	Bolton	2655236	Gree	
Active		532 Stockport Road	Longsight		Gree	
Active		86 Brightmet Drive	Bolton	2655236	Gree	
Active		4 Nutt Lane	Prestwich	2639897	Gree	
Active		1st Floor 5 Tottington Road	Harwood	2163842	Gree	
Active		24 Oswald Road, Chorlton Cum	Hardy, Manchester	11599281	Gree	
Active		4 Goulden Road	Manchester	3383611	Gree	

Figure 9: The column `registered_addressstreet_address splitted_1` is reconciled against GeoNames. Annotations are in column `locality_GN`.

Annotations and reconciliation services.

ASIA provides functionalities to support schema and instance-level annotations. These functionalities are based on reconciliation services (also referred to as interlinking services in D3.2 and D3.3) that are integrated into ASIA and support the user in adding an annotation.

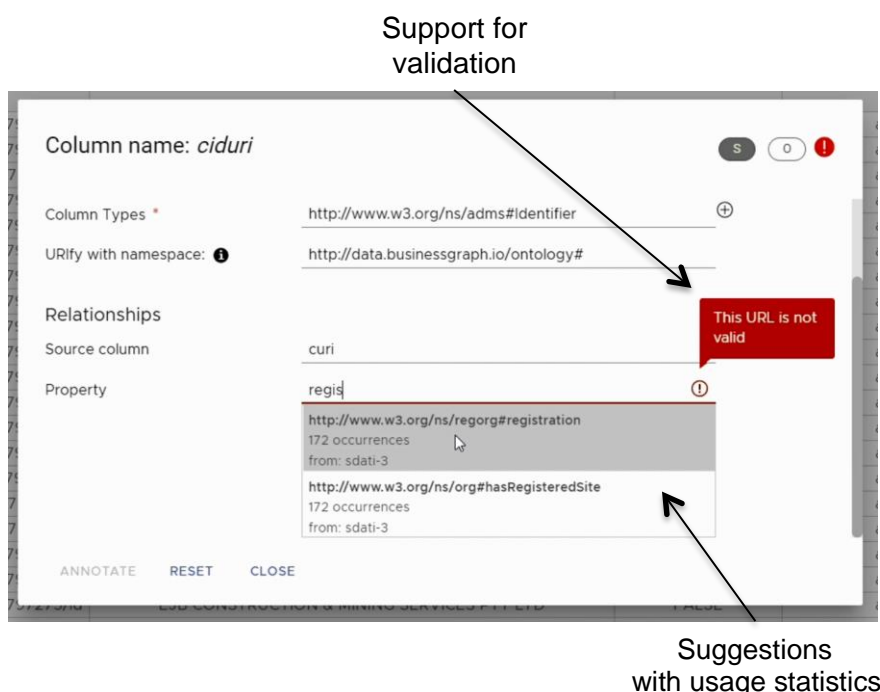


Figure 10: Widget for schema-level column annotation

For *schema-level annotations*, ASIA provides an annotation widget, which is depicted in Figure 10. The widget provides suggestions of types and properties using the vocabulary suggestion service offered by ABSTAT (see D3.2 and D3.3). This suggestion is one of the interlinking services offered by the euBusinessGraph platform. When the widget is open, a suggestion for the type and the property is prompted by matching the column header to ABSTAT profiles, which describe the usage of properties and types (ontology classes and data types) in the knowledge graph. Most importantly, because of the limited information provided in the column header, an autocomplete service helps the user in refining the suggested ontology terms by autocompleting the string typed by the user.

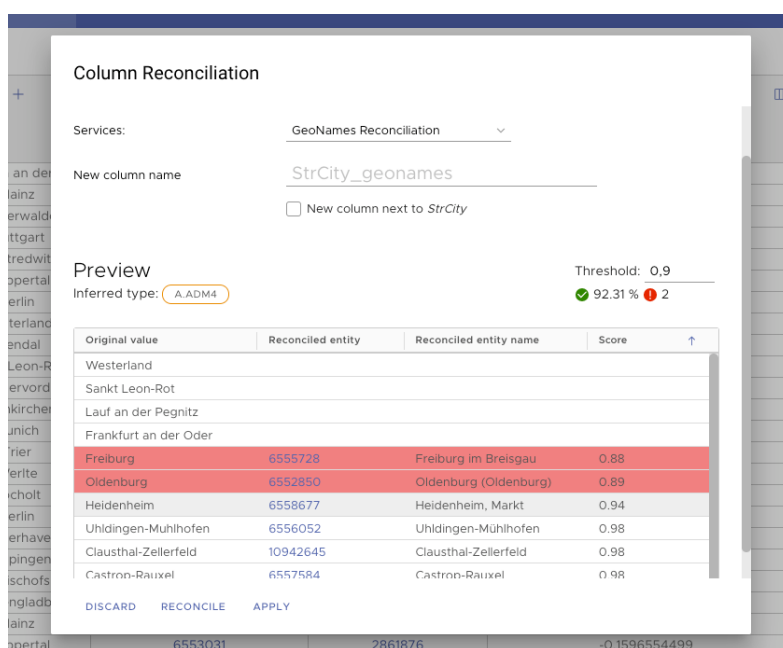
For *instance-level annotation*, ASIA incorporates different reconciliation services, among which Wikifier, GeoNames, Wikidata and Open Corporates. More reconciliation services can be added modularly to the application to serve reconciliation against a desired system of identifier or knowledge base. Thus, instance-based annotations are provided by the reconciliation service that is configured by the user (e.g., by selecting a threshold over the similarity computed by the reconciliation algorithm). Figure 11 shows the instance-level annotation widget applied to a column including toponyms that use the GeoNames reconciliation service.

Portability of annotations to new tables.

While annotations are defined and controlled using a user interface, the specifications defined using the interface are translated into executable data transformations. Thus, once a user has annotated a table, the annotations can be ported and attached to a new table that has the same schema. For data onboarding, this feature means that once a data provider has annotated one table and the data transformations are generated, these transformations can be applied to other data that are provided with a table that has the same schema.

Cross-lingual Data Annotation Methodology.

Now that we have clarified annotations are supported using vocabulary suggestion services and reconciliation services (respectively, schema-level and instance-level interlinking services), we explain the methodology applied to deal with cross-linguality in structured data annotation. First, in euBusinessGraph, the problem of cross-linguality in data onboarding refers to the case where data about companies that refer to a given jurisdiction, or which are provided by a party, appear in a language different than English. In other words, the case we consider is the one when a table contains data lexicalized using one language different from English. Otherwise, there is no evidence that the problem of annotating a table lexicalized in more than one language different than English is relevant for euBusinessGraph and data onboarding. We now explain strategies for cross-lingual schema-level and instance-level annotation and discuss the impact of the problem.



Original value	Reconciled entity	Reconciled entity name	Score
Westerland			
Sankt Leon-Rot			
Lauf an der Pegnitz			
Frankfurt an der Oder			
Freiburg	6555728	Freiburg im Breisgau	0.88
Oldenburg	6552850	Oldenburg (Oldenburg)	0.89
Heidenheim	6558677	Heidenheim, Markt	0.94
Uhlinden-Mühlhofen	6556052	Uhlinden-Mühlhofen	0.98
Clausthal-Zellerfeld	10942645	Clausthal-Zellerfeld	0.98
Castro-Aluxel	6557584	Castro-Aluxel	0.98

Figure 11: Widget to control a reconciliation service. The example refers to the GeoNames reconciliation service.

Cross-lingual Schema-level Data Annotation.

Vocabulary suggestions come under two functionalities of ASIA for schema-level table annotation, namely, type / property suggestion based on column headers, and type / property autocomplete. Cross-lingual autocomplete does not make sense and we can assume that the provider must use strings that represent words or word fragments in English. Otherwise, automatic type / property suggestion based on column headers require translation of column headers in English.

To achieve this, we extend ASIA vocabulary suggestion with a **cross-lingual suggestion feature** that takes as input three parameters:

- Input language (required), which contains the language of the input text.

- Column header (required), which contains the column header as a string.
- Context (optional), which contains a bag of word that can help the translation. The context can be fed in different ways. At the moment, we automatically add to the context a set of words in English that are expected to represent the domain of companies (Company, Organization, etc.).

Based on this input ASIA collects a set of possible translations of the header into English using an automatic translation service, in particular at the moment we use Microsoft translator⁴ (which is free for 2M characters per month). The output of the vocabulary suggestion service is a set of words in English. We then select the top-k words that are most related to words characterizing the company domain using a pre-trained Word2vec model to compute word similarity. We invoke the ABSTAT service with these top-k words, where each invocation returns a ranking. We fuse the types / properties in the ranking and pick up the most frequent as a suggestion. The ranking is kept so that the user can see alternatives.

We observe that while cross-lingual schema-level annotation can be viewed as a nice functionality of the tool, the impact on data onboarding in euBusinessGraph is limited for two reasons:

- Several table column headers, e.g., “ciduri”, do not use intuitive naming conventions, which, in the mono-lingual case make the matching difficult and in the cross-lingual case do not yield translations.
- Table columns are limited in number (so do the columns with intuitive naming conventions that may yield better automatic matching results). They can be annotated quite efficiently using the English-fed autocomplete functionality. In addition, it is reasonable to assume that a provider to a pan-European company graph would be knowledgeable of the English language enough to feed the autocomplete service with English word fragments.
- Because of portability of annotations, a user onboarding different data into the knowledge graph using a common schema (e.g., when updates of information need to be committed), can manually annotate the table once and then reuse these annotations.

Cross-lingual Instance-level Data Annotation.

While we can fairly assume that schema-level annotation can be performed manually with limited impact on a data onboarding task, instance cannot be manually annotated for tables of even limited size. However, ASIA can support cross-lingual instance-level annotations by plugging in a cross-lingual reconciliation service. Cross-lingual reconciliation services are based on multilingual indexes for the reference data (the data used for reconciliation). Different services are already available as of today, and, in particular:

- Wikifier, which covers Wikipedia entities in 130 languages (described above in the deliverable).
- GeoNames, which covers labels of spatial entities in a large variety of languages. The covered languages change from entity to entity, but usually include the local language of a toponym. This means that a data provider that provides data of companies for a given jurisdiction where toponyms are named using the local language would be able to reconcile the toponym against GeoNames.
- Wikidata, which provides different reconciliation services (one per language) to import as needed.

⁴ <https://azure.microsoft.com/id-id/pricing/details/cognitive-services/translator-text-api/>

Preview
Inferred type: S.FRM

Original value	Reconciled entity	Reconciled entity name
SAUDA	9827139	Sauda
KLEPP	9827600	Klepp
LIER	9416637	Lier
BERGEN	9463139	Bergen
SANDNES	3413718	Sandnes
OS	9825690	Os
BÆRUM	9477586	Bærum
GJØVIK	9451697	Gjøvik
TØNSBERG	9476937	Tønsberg
ÅSNES	9822237	Åsnes

Figure 12: A list of Norway's cities reconciled against GeoNames

Figure 12 shows an example of Norway's toponyms lexicalized in Norwegian and reconciled using the GeoNames service.

In summary, cross-lingual instance-level annotation is covered by two general purpose reconciliation services and one spatial reconciliation services. At the moment we are not aware of requirements for data onboarding that require additional cross-lingual instance-level services. However, whenever this need would emerge, new cross-lingual reconciliation services can be modularly added to ASIA.

3 Company related event and relation extraction

This section in detail describes the cross-lingual/multilingual company-related event and relation extraction.

Event Registry⁵ is a system for cross-lingual and multilingual news tracking and analysis, event identification and observation. Event Registry provides a set of functionalities related to media monitoring and media intelligence. In particular, the user can:

- Observe current articles and events.
- Follow current topics.
- View currently trending entities from news – people, organizations, locations.
- Obtain information with respect to specific concept, such as:
 - List of events/articles; Contains the list of relevant events or articles containing the query concept, which can be sorted by date, relevance, source importance and virality.
 - Top concepts from the events/articles; Provides a chart of top concepts (entities and non-entities) from the relevant events/articles that best describe what the events or articles are about. Figure 13 presents an example of top concepts chart for search query “Bitcoin”.
 - Tag cloud; Presents a tag cloud of keywords from the events or articles.
 - Timeline; Provides a view on when the articles/events that match the search criteria were published/occurred.
 - Event locations; Provides a map that shows the geographic locations where the resulting events occurred. The chart shows the time distribution of the events based on the dates.
 - Article authors; Provides a chart of article authors sorted by the number of publications.
 - Concept graph; Provides a graph of frequent concepts in the results. The edges between the nodes of concepts are displayed if the pair of concepts frequently co-occur in the results.
 - Concept trends; Provides a view on how frequently the selected concepts are mentioned in the resulting events/articles on a daily basis.
 - Date mentions; Presents a chart displaying how frequently individual dates are mentioned in the news articles about the resulting events/articles.
 - Event clusters; Provides a tree displaying how individual resulting events can be organized into groups and subgroups based on their relatedness and similarity.
 - Categories; Visualizes the various categories and sub-categories for the resulting events/articles.
 - Language statistics for articles. Provides a chart displaying the languages in which the resulting articles are written. Figure 14 presents an example of languages chart for search query “Bitcoin”.

⁵ Eventregistry.org

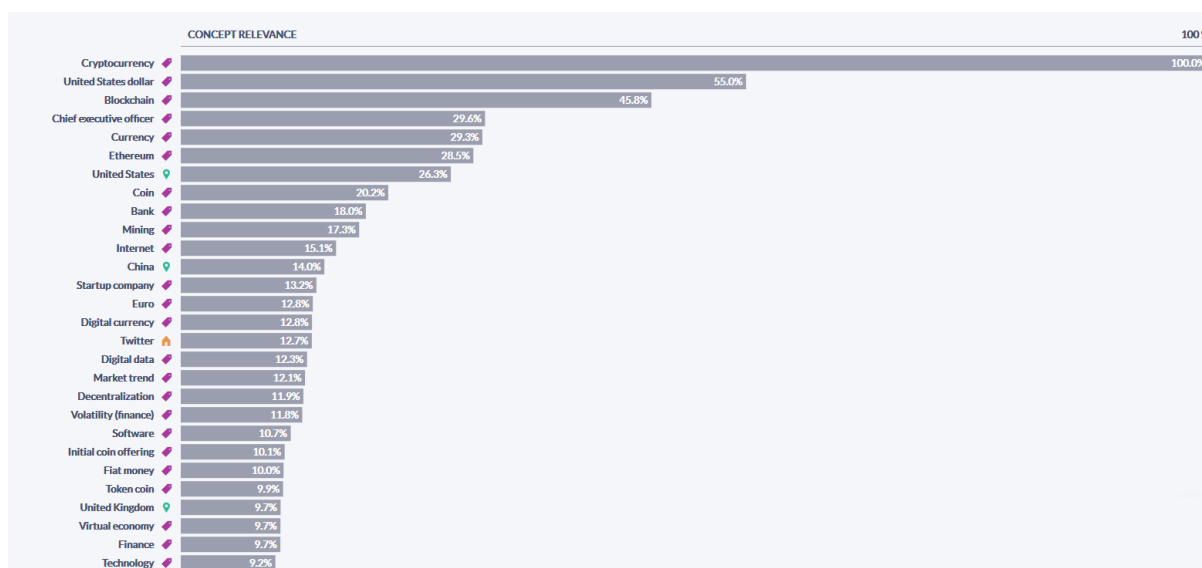


Figure 13: Top concepts (for search query "Bitcoin")

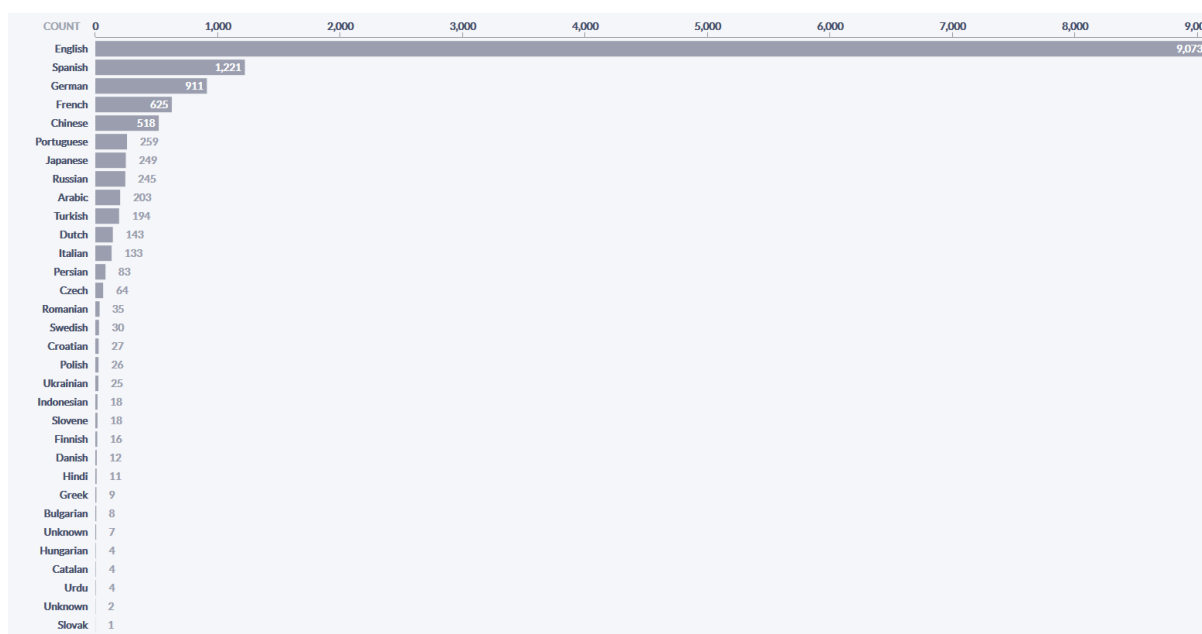


Figure 14: Languages (for search query "Bitcoin")

3.1 Event types

3.1.1 Objectives

Event Registry categorization is currently based on DMOZ taxonomy that was a multilingual open-content directory of World Wide Web links. However, DMOZ is no longer maintained and euBusinessGraph tasks and scenarios require an efficient categorization mechanism for detecting business events of particular types.

The objectives behind event types categorization is development of the procedure that would allow for precise and accurate categorization of textual documents (for instance, news articles in different languages about business, economy and finance) into a taxonomy of business related event types, such as mergers and acquisitions, bankruptcy and revenues.

3.1.2 Data sources

For building event types models we have used a dataset of business relations produced under weak supervision with data from 2016 to 2017.

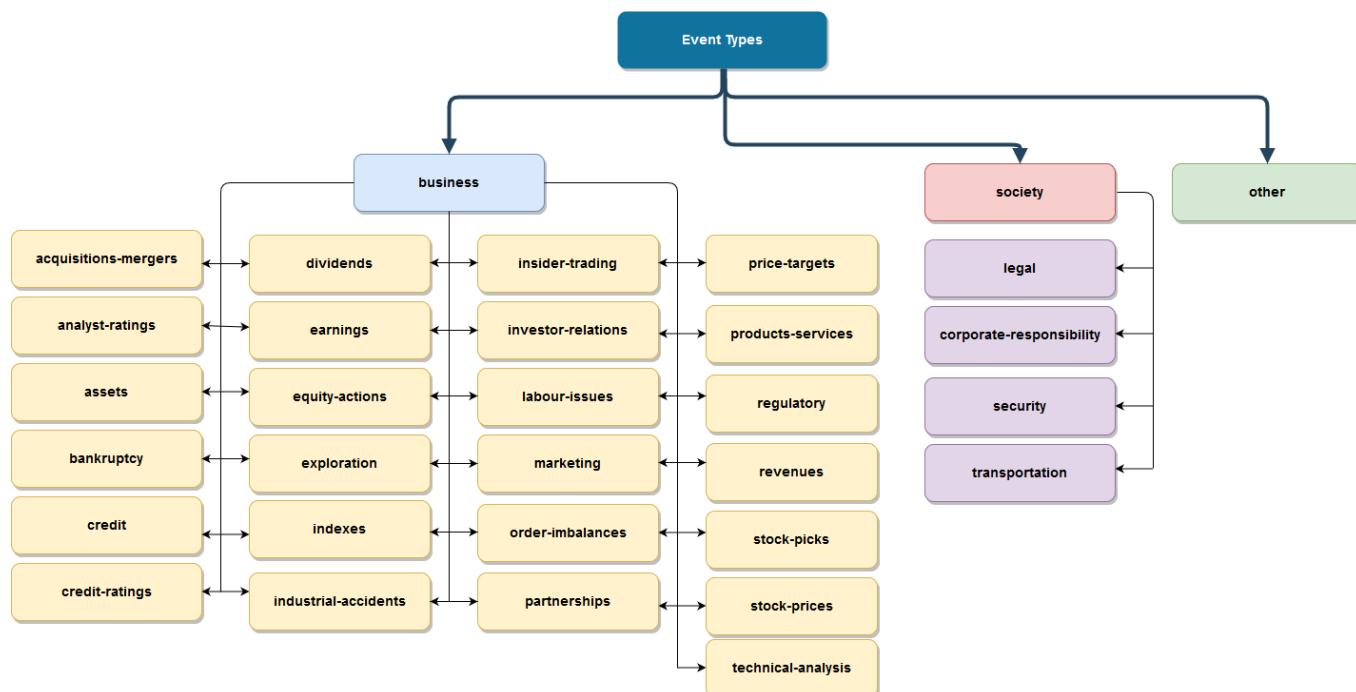


Figure 15: Event types by groups

Figure 15 presents event types by groups – the high level groups includes events related to business, society and other group. Second level events types that belong to business and society groups are as well displayed on the picture. ‘Other’ event types represent a class of events not related to business or society groups (we include society events into the categorization, since events from society categories have influence on business related events).

Table 4: Event types data description

Label	Train	Development	Test	Total
acquisitions-mergers	480000	60000	60000	600000
analyst-ratings	480000	60000	60000	600000
assets	480000	60000	60000	600000
bankruptcy	154338	19292	19293	192923
credit	480000	60000	60000	600000
credit-ratings	480000	60000	60000	600000
dividends	480000	60000	60000	600000
earnings	480000	60000	60000	600000
equity-actions	480000	60000	60000	600000
exploration	211843	26480	26481	264804
indexes	40286	5036	5036	50358
industrial-accidents	66324	8291	8291	82906
insider-trading	480000	60000	60000	600000
investor-relations	480000	60000	60000	600000
labour-issues	480000	60000	60000	600000
marketing	480000	60000	60000	600000

order-imbalances	380139	47517	47518	475174
partnerships	480000	60000	60000	600000
price-targets	480000	60000	60000	600000
products-services	480000	60000	60000	600000
regulatory	138876	17359	17360	173595
revenues	480000	60000	60000	600000
stock-picks	69059	8632	8633	86324
stock-prices	480000	60000	60000	600000
technical-analysis	480000	60000	60000	600000
other	1920000	240000	240000	2400000
corporate-responsibility	65638	8205	8205	82048
legal	480000	60000	60000	600000
security	40056	5007	5007	50070
transportation	48351	6044	6044	60439
Total	12254910	1531863	1531868	15318641

Table 4 shows a number of event types with relevant data used for training, testing and further development. The dataset has been split into train, development, and test subsets. The split has been a stratified split where each subset contains the same proportion of each label as in the original dataset. The original dataset was shuffled (randomized order) and each subset has been created by extracting examples randomly from the original set. The train set, comprising 80% of all examples, has been used to train the supervised text classification algorithm; The test set, comprising 10% of all examples, has been used to evaluate the algorithm; Finally, the development set, comprising the remaining 10%, has been set aside for future parameter tuning.

Table 5: Labelled examples

Text	Label
Lumenpulse scheduled a conference call to	investor-relations
Bank Rakyat Indonesia Reports Q3 Pre-Tax Profit IDR 4828.10B Vs Consensus IDR 3904.00B	earnings
Mondelez International Inc.: The company standardized its powdered beverage	products-services

Table 5 shows labelled examples from the dataset.

3.1.3 Identification of business-related events

The process of relevant event types identification involves several steps, such as **pre-processing**, **unsupervised representation of words**, and **supervised text classification**, described below.

3.1.3.1 Methodology

Pre-processing

Text pre-processing consists of an implementation of Matt Mahoney's perl script⁶ commonly used for pre-processing language modelling data (e.g. fastText⁷). It includes removal of some special characters, such as numbers and case-folding. The exact code used is shown in Figure 16.

⁶ <http://matmahoney.net/dc/textdata.html>

⁷ <http://fasttext.cc>

```
import html

# replace html encoded chars
s = html.unescape(s)

s = s.replace(""", "").replace("'", "").replace("'''", " ") \
    .replace("'''", " ' ").replace('""', "").replace('""', "") \
    .replace('""', ' ').replace(".", ". ").replace(", ", ", ") \
    .replace("(", "( ").replace(")", ") ").replace("!", "! ") \
    .replace("?", "? ").replace("-", "- ").replace(";", "; ") \
    .replace(":", ": ").replace("=", "= ").replace("*", "* ") \
    .replace("|", "| ").replace("</", "< ").replace('[', '[ ') \
    .replace(']', '] ')

# replace all whitespaces characters with a space
s = ' '.join(s.split())

# remove numbers (digits)
s = ''.join(filter(lambda c: not c.isdigit(), s))

# strip and lower
s = s.strip().lower()
```

Figure 16: Pre-processing code example

Table 6: Pre-processing text examples

Original Text	Pre-processed Text
Lumenpulse scheduled a conference call to	lumenpulse scheduled a conference call to
Bank Rakyat Indonesia Reports Q3 Pre-Tax Profit IDR 4828.10B Vs Consensus IDR 3904.00B	bank rakyat indonesia reports q pre - tax profit idr . b vs consensus idr . b

Table 6 demonstrates the examples of the original text and pre-processed text.

Unsupervised representation of words

Using both the text in our dataset and an additional corpus of news text from Even Registry, we created an unsupervised corpus of text (total: 7.8GB). We used the Continuous Bag-of-Words model [Mikolov et al.] to create 300-dimensional unsupervised representations (embeddings) of over 2.7M tokens using the fastText tool. The parameters used were the following: a vector size of 300, a word context of 7, a minimum count of 15, no subword information was used, 8 negative samples were used per example and we did 6 passes (epochs) over the training data.

Supervised text classification

Supervised text classification was performed using a Bag-of-Vectors (also known as Embedding Bag or Bag of Embeddings) neural network. First a context embedding is created by looking up the embeddings for the words (initialized to the previously described embeddings) in the given text and averaging them together. The context embedding is then fed into the classifier (a softmax layer) which produces a prediction for the label. During training we do backpropagation to all layers (including the embeddings).

We trained our model on the training dataset for 20 epochs using Adam [Kingma and Ba] with a learning rate of 0.001. The algorithm was implemented using Keras⁸ with the Tensorflow⁹ backend.

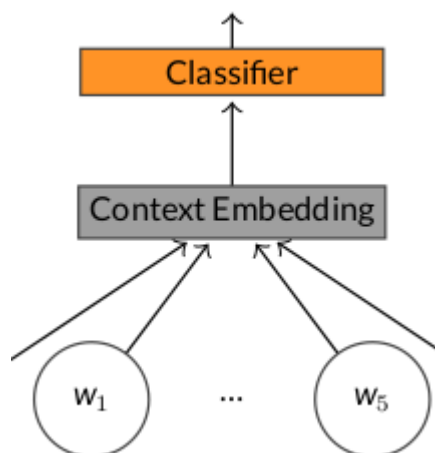


Figure 17: Supervised text classification

Figure 17 presents a view on supervised text classification using context embeddings.

3.1.3.2 Results

The results of event types identification are presented in Table 7.

Table 7: Results of event types identification

Label	Precision	Recall	f1-score	support
acquisitions-mergers	0.95	0.94	0.94	60000
analyst-ratings	0.99	0.99	0.99	60000
assets	0.96	0.96	0.96	60000
bankruptcy	1.00	1.00	1.00	19293
credit	0.99	0.99	0.99	60000
credit-ratings	0.99	0.99	0.99	60000
dividends	1.00	1.00	1.00	60000
earnings	0.99	0.99	0.99	60000
equity-actions	0.95	0.96	0.95	60000
exploration	0.99	1.00	1.00	26481
indexes	0.99	1.00	0.99	5036
industrial-accidents	0.98	0.99	0.99	8291
insider-trading	0.99	0.99	0.99	60000
investor-relations	1.00	1.00	1.00	60000
labor-issues	0.98	0.99	0.98	60000
marketing	1.00	1.00	1.00	60000
order-imbalances	1.00	1.00	1.00	47518

⁸ <http://keras.io>

⁹ <https://www.tensorflow.org>

partnerships	0.98	0.99	0.99	60000
price-targets	1.00	1.00	1.00	60000
products-services	0.95	0.94	0.94	60000
regulatory	0.97	0.97	0.97	17360
revenues	0.99	0.99	0.99	60000
stock-picks	0.98	0.98	0.98	8633
stock-prices	0.99	0.99	0.99	60000
technical-analysis	1.00	1.00	1.00	60000
other	0.99	0.99	0.99	240000
corporate-responsibility	1.00	0.99	0.99	8205
legal	0.99	0.99	0.99	60000
security	0.98	0.98	0.98	5007
transportation	0.99	1.00	0.99	6044
Average				
micro avg	0.99	0.99	0.99	1531868
macro avg	0.99	0.99	0.99	1531868
Accuracy	0.98			

All results are obtained by evaluating the classifier on the test subset of the data.

3.2 Relation extraction

3.2.1 Objectives

The objective of relation extraction is obtaining relations between main business entities (people and organizations) based on Wikipedia concepts from textual data.

In this way, the business entity can be observed in time and generally expressed relations can be monitored and presented to the users.

Following that, in this section we describe the data sources used for relation extraction, the methodology for tracking business entities through time and present a tool providing visualizations obtained from the data.

3.2.2 Data sources

The relation extraction is based on Event Registry, and uses its events data as its main data source. The developed tool is seeded with approximately 5.7m events covering the range between January 2014 and July 2017. Each event consists of general information like title, event date and total article count and a list of concepts that characterize the event, which is split into entity concepts and non-entity concepts. Entity concepts are people, organizations, and locations related to the event. Whereas non-entity concepts represent abstract terms that define the topic of the event, like technology, education, and investment. Those concepts were extracted using JSI Wikifier, which is a service that enables semantic annotation of the textual data in different languages. In addition, each concept has a score that represents the relevancy of that concept to the event.

3.2.3 Tracking business entities through time

For the euBusinessGraph project we have developed a tool called “Relation Tracker” that tracks main business entities (people and organizations) within each topic through time. The main types of relations between the entities are detected and observed in time. The tool provides multiple ways of visualizing this information with different scales and durations. It uses events data from Event Registry as a source of information, with the aim of getting holistic insights about the searched topic.

3.2.3.1 Methodology

Clustering and Formatting Data

To process the data, events are first grouped into topics using a clustering algorithm. Each event is represented as a sparse vector of the non-entity concepts it has, with the weights equal to their scores in that event. The constant number of topics is set experimentally to be 80 clusters, in a balance between mixed clusters and repeated clusters.

Each cluster describes a set of events that fall under the same topic, whereas the centroid vector of each cluster represents the main characteristics of it. The names of the clusters were determined using a category classifier service from Event Registry, which uses DMOZ Taxonomy, that is used to classify texts and webpages into different categories. The text for each cluster was formed from the components of its centroid vector, taking into account their weights within the vector. The resulted cluster names range from technology and business to refugees and society, and clusters were exported as a JSON file for processing them in the visualization part.

Choosing the Main Entities

Under any topic, the top entities at each duration of time have to be chosen. At first, the concepts were filtered from outliers like publishers and news agencies. Then, an initial importance value has been set for each concept based on two parameters: the TF-IDF score of concept with respect to each event, and the number of articles each event contains. If we denote the set of events that occur in the interval of time D by E_D , the number of articles that event e contains is A_e , the TF-IDF score of concept c at event e by $S_{c,e}$, then the importance value of each item with respect to the interval D is calculated by the formula:

$$Imp_{init}(c)_D = \sum_{\substack{e \in E_D \\ e \text{ has concept } c}} S_{c,e} * \sum_{\substack{e \in E_D \\ e \text{ has concept } c}} A_e \quad (1)$$

The TF-IDF function is used to give importance to the concept based on its relevance to the events, and the number of articles is used to give more importance to the events that have more articles talking about it, and hence, more importance to the concepts that it has. The product of summation of the two terms has been used rather than the summation of their product because of its computation efficiency while still producing good results. However, to prevent the case where all the chosen entities get nominated because of one or two big events (which results in a bias towards those few events), a modification to the importance value formula has been made by introducing another parameter, which is the links between concepts (whenever two concepts occur in the same event, there is a link between them). Each concept now affects negatively the other concepts it is linked to by an amount equal to the initial importance value divided by the number of neighbours. If we denote the set of neighbors of concept c during the interval of time D by $N_{c,D}$, then the negative importance value is defined by:

$$Imp_{neg}(c)_D = \sum_{c' \in N_{c,D}} \frac{Imp_{init}(c')_D}{|N_{c',D}|} \quad (2)$$

The final score is just the initial importance value minus the negative importance value, which is then used to sort and nominate the top entities.

$$Imp_{final}(c)_D = Imp_{init}(c)_D - Imp_{neg}(c)_D \quad (3)$$

Detecting the Characteristics of Relationship

The main goal is to model the relationship between any two entities through a vector of words where two entities are collocated. Since the relationship between two entities at any given time is based on the shared events between them, and each event is characterized by a set of concepts, those concepts - specifically the abstract or the non-entity concepts - were used to characterize such relationships. For each pair, all the **non-entity concepts** from the shared events between them were aggregated, and each one of them was assigned a value based on the number of events it is mentioned in and its score in those events. Those concepts were sorted and ranked depending on their values, and the top ones were chosen as the main features of the relationship.

In addition, the values of the non-entity concepts were used to rank the shared events and extract the most descriptive ones. In the ranking procedure, each event was given a value equal to the aggregated values of all non-entity concepts it has.

To summarize the set of characteristics, DMOZ category classifier has been used again in a similar way to what has been done in determining the names of the clusters. These categories were used to label the relationship between the entities, indicating the main topic of the shared events between them.

3.2.3.2 Visualizing results

To access a topic, a search bar is provided to select among the list of extracted topics from clustering step. Once the user selects a topic, a default date is chosen and a network graph is shown explaining the topic.

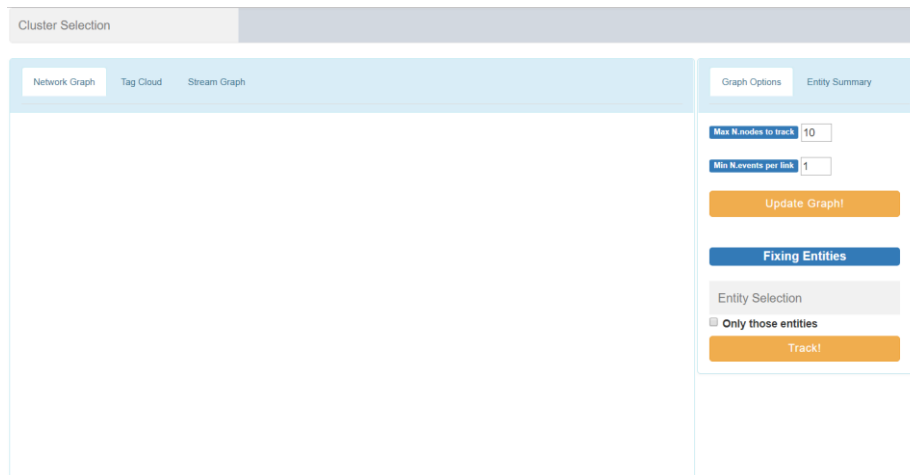


Figure 18: Main interface of Relation Tracker

Figure 18 presents the main interface of the tool, with the cluster selection tab at the top.

Characteristics of the Main Graph

Since the tool's main goal is to show the top entities and their relations, the network graph is the best choice for this matter. Following that, an interactive network graph has been built, which has the following features:

- The main entities within that topic at the selected interval of time are represented by the vertices of the graph.
- The size of the vertices reflects the importance value of each entity, scaled to a suitable ratio to fit in the canvas.
- The colors represent the type of the entity, whether it is a person (red) or an organization (blue).
- The links between the entities represent the existence of shared events in that interval of time between them under that topic, and hence indicating some form of relations. The thickness of the links is proportional to the number of shared events, whereas the labels are the ones calculated in previous section.

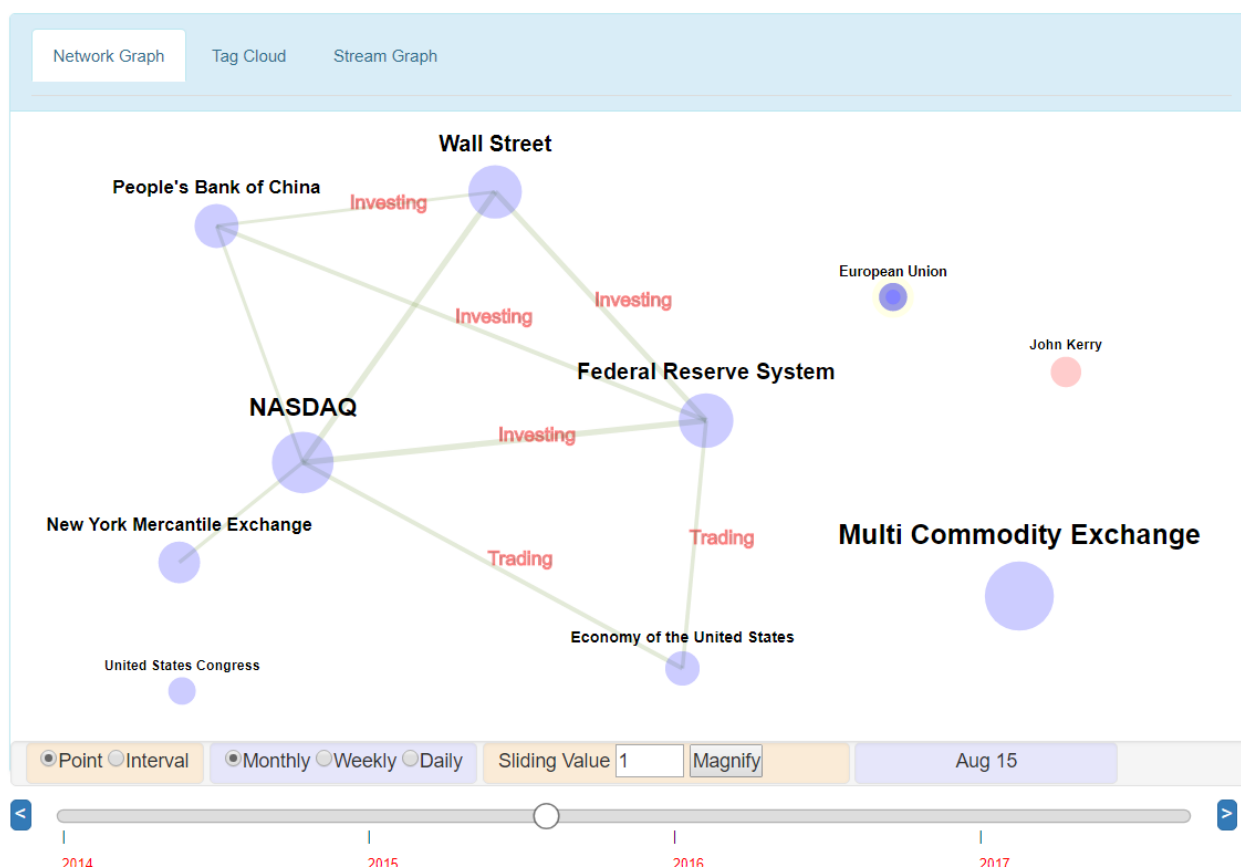


Figure 19: Top people and organization in August 2015 and their relations under the business topic.

Figure 19 presents top people and organizations with relevant relations in Aug 2015 found among business news.

Main Functionality

As the tool is concerned about tracking the changes with time, the graph is supported with a slide bar that allows the user to choose from the dates where there is at least one event occurred with respect to the selected topic. Different scales for moving dates are also provided; the user can choose to move

day by day, week by week, or month by month and see the changes accordingly. In addition, the user can choose a specific interval of time, and track how the entities and their relations are changing when the interval moves slightly with respect to its length. An interval magnifier is also given if the user wants to get a closer look at the changes that happen in a small interval.

Apart from the top entities, a user can select a specific entity and track how it's interacting with the top entities of a certain topic, or select a set of entities and track their interactions among those set only.

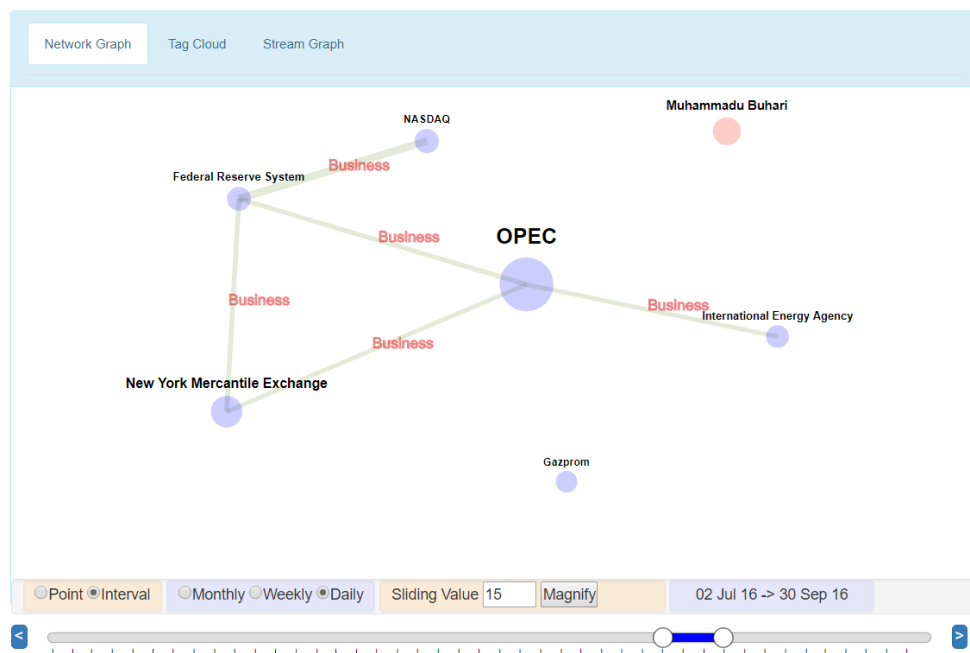


Figure 20: Top entities for the third quarter of 2016 under the energy topic.

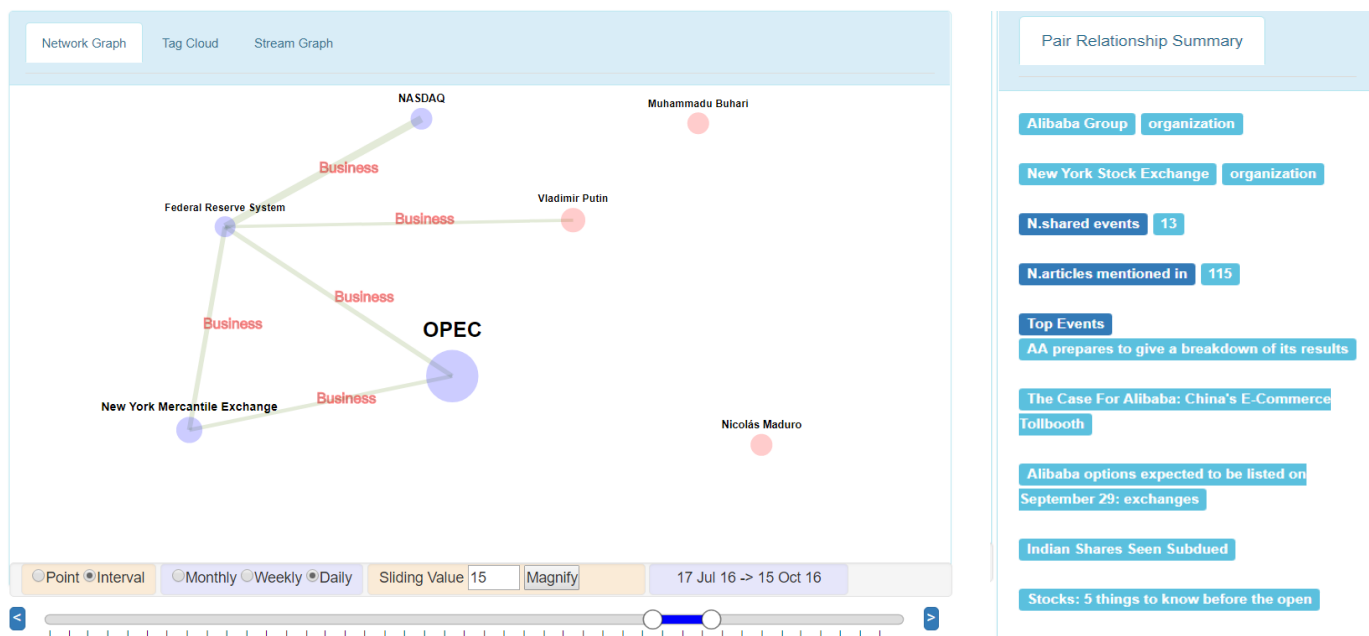


Figure 21 (on the left): The changes in top entities under the same topic after moving the interval for 15 days

Figure 22 (on the right): Relationship summary about Alibaba Group and New York stock exchange in September 2015 under the business topic.

An example illustrating that can be seen in Figure 20 and 21. In Figure 20, we see the top 7 entities under the energy topic in the third quarter of 2016. When the interval is moved by 15 days, we notice that some of the entities disappear, like “Gazprom” and “Internal Energy Agency”, indicating that they are no longer among the top 7 entities, whereas “Vladimir Putin” and “Nicolás Maduro” entities emerges and “Vladimir Putin” is connected with “Federal Reserve System” indicating a shared events between them. The change in size indicates the change in the importance value of each one, while Business is the general theme among all labels.

Displaying Relation Information

Whenever the user selects a pair of entities, detailed information about their relationship in the selected interval of time is given, such as the number of shared events and articles, along with the top events both concepts were mentioned in. Also, the top shared characteristics that shape the relationship between them at this period is shown and sorted by percentage of importance.

As seen in Figure 22; when selecting “Alibaba Group” and “New York stock exchange” under the business topic in September 2015, we see a list of the top events that involve both of them during this period. We see also that the relationship between them is mainly about IPO of Alibaba group that happened in September that year and the company’s prices in the stock market, as it can be understood from the top events and top shared characteristics.

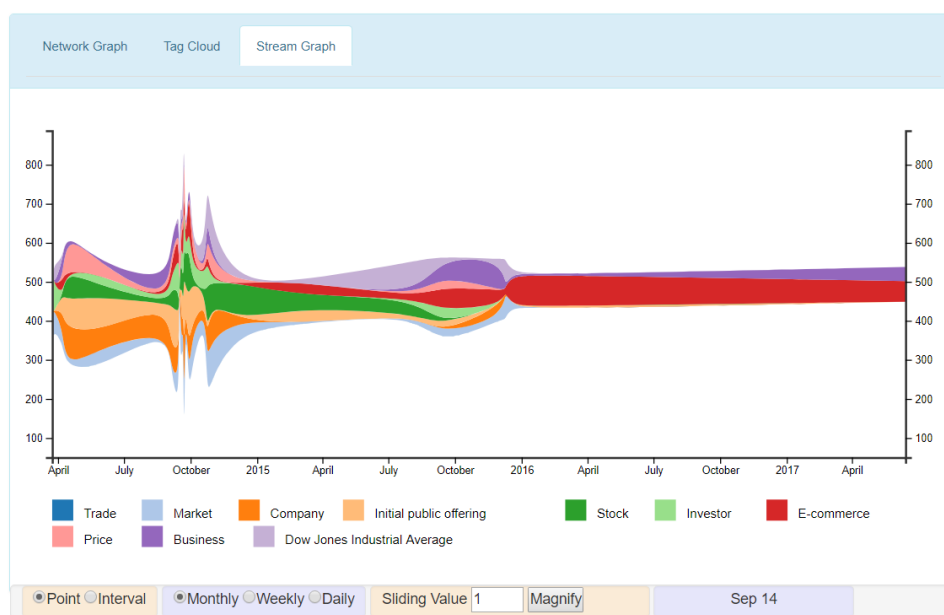


Figure 23: Stream graph showing how the effect of the main features on the relationship between Alibaba Group and New York stock exchange is changing through time.

To illustrate how the importance of those top features with respect to the relationship is changing through time, a stream graph is used as shown in Figure 23. A spike in the graph can be detected at September 2014 when the IPO happened.

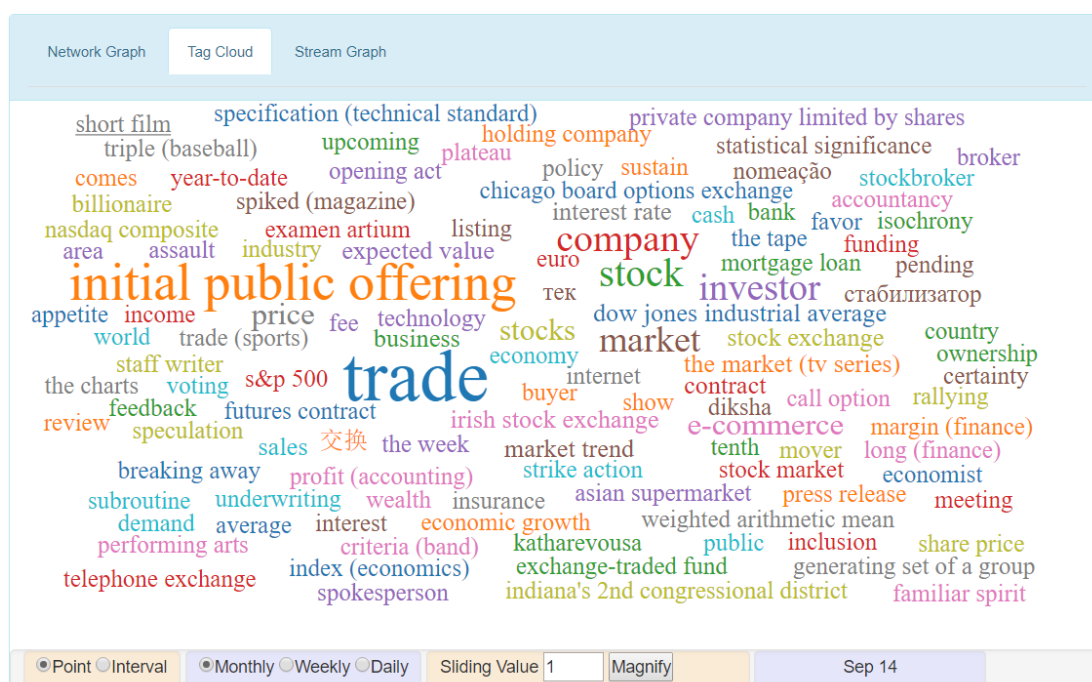


Figure 24: Tag cloud illustrating a general view about all the characteristics that affects the relationship between Alibaba Group and New York stock exchange under the business topic.

Also, the set of all characteristics that affect the relationship is visualized in a tag cloud to give a big picture about it. Figure 24 shows the tag cloud of the same relationship mentioned above.

3.3 euBusinessGraph event types and relation extraction tools and services

Data in Event Registry can be accessed through the web interface or directly through the available API. In order to access data in Event Registry an API key is required. Accessing data through the API can be done by issuing HTTP GET requests with specific parameters.

Event Registry API and services for event categorization are described in euBusinessGraph D3.4. Accessing data through the API can be done by issuing HTTP GET requests with specific parameters. The resulted methodology of business type event identification is planned to be incorporated in Event Registry – in this way, it would be possible to obtain business event types within Event Registry API.

The “Relation Tracker” tool is currently accessible on the following website:

<http://connection.ijs.si>

4 Cross-lingual/multilingual data management application in euBusinessGraph business cases

This section in detail describes data management application in euBusinessGraph business cases. In particular, the cross-linguality and multilinguality aspects from Data journalist product from DW and Atoka+ product from SDATI are discussed below.

4.1 DJP: Data Journalism Product

4.1.1 Objectives

The objectives of the data journalism Screener product¹⁰ from DW business case are related to the main tasks addressed by modern journalists. In particular, Screener tool allows journalists to find valuable information, to monitor and publish business related data.

An important feature of data journalism product is the possibility of obtaining and monitoring information in different languages. For instance, the journalist is able to find, fact-check and view the company data along with business related news from Event Registry tool described above.

Consequently, the Screener tool covers the following cross-lingual and multilingual aims:

- Find information related to a company in news feeds from international publishers, filter by source and by language.
- Detect cross-lingual news events around a specific company name.
- Set alerts to automatically get multilingual news articles for one or more companies.

4.1.2 Data sources

The Screener Tool news search is based on Event Registry system that enables cross-lingual and multilingual news tracking and analysis, event identification and observation. The number of collected articles ranges between 100.000 and 200.000 articles per day. The collected articles are in various languages. Most represented languages are English (50% of all articles), German (10%), Spanish (8%) and Chinese (5%). More details about Event Registry can be found in Section 3 (Company related event and relation extraction) of this deliverable.

4.1.3 Functionalities

Overall, the Screener Tool supports the following features:

- Company related search across multiple company registration and news data.
- Monitoring of company news with alerts.
- Template-based creation of associated digital content items.
- Company name identification support.
- Direct access to wider data sources and tools.

¹⁰ <http://screenertool.com>

Select articles for your news collection

175 results

Internet of Things Market 2021 | Google Inc. (United States), Cisco Systems Inc. (United States), Apple Inc. (United States), Microsoft Corp. (United States), Intel Corp. (United States)

Published: 2018-11-19
Event Registry

Source: openpr.com

Microservices in Healthcare Market - Emerging Trends and Growth Opportunities for End Users Like Health care Provider, Payer and Life Science

Published: 2018-11-19
Event Registry

Source: openpr.com

Клавиатуры и мыши Corsair подружатся с Xbox One

Published: 2018-11-19
Event Registry

Source: iXBT.com

Stora problem med Office 365 i Europa - användare kan inte logga in

Published: 2018-11-19
Event Registry

Source: Computer Sweden

Figure 25: Screener tool: find news by keyword

Filter by Language

Please select one

- English
- Russian
- Swedish
- Chinese
- German
- Croatian
- Arabic
- Spanish; Castilian
- Serbian
- Dutch; Flemish
- Italian
- Hungarian
- Turkish
- Finnish

Figure 26: Screener tool: filter by language

Figure 25 demonstrates the functionalities related to the multilingual news search, while the list sample of available languages for filtering is provided at Figure 26.

4.2 Atoka+

4.2.1 Objectives

Atoka¹¹ aggregates business information from official and unofficial sources and presents all this data through a uniform interface. Atoka+ is the business case of SDATI in the euBusinessGraph project and is about extending the Atoka product to work in new jurisdictions, namely the United Kingdom and Norway.

Extending Atoka to a new jurisdiction implies on the one hand work to integrate new data sources, normalising and curating data, and on the other hand extending automatic analysis tools that operate on unstructured texts to annotate these texts with business-related entities (i.e., companies and company officers).

In particular, Atoka builds a company-centric news stream that is presented to the user as one of the dimensions that they can explore. Through this functionality, users can get up-to-date information about what a particular company is doing to complement the official information from the business registers that is, by its own nature, more static.

The company-centric news stream requires the following cross-lingual/multilingual features to be implemented:

- Annotation of long texts (e.g., news articles) from several sources in several languages.
- Detection of news articles that depict the same underlying newsworthy event.

4.2.2 Data sources

Similarly to DW's Screener Tool (see Section 4.1), Atoka builds company-centric news streams linking company instances in the business graph with annotated news items from Event Registry on the UK and Norwegian jurisdictions to leverage its multilingual annotation capabilities.

4.2.3 Functionalities

Company-centric news streams are implemented in Atoka to closely follow an existing implementation that already is offered for companies in the Italian jurisdiction. The user selects a company that is of interest and Atoka presents several topics that organise all the information available for the company selected.

¹¹ Atoka — <https://atoka.io/>

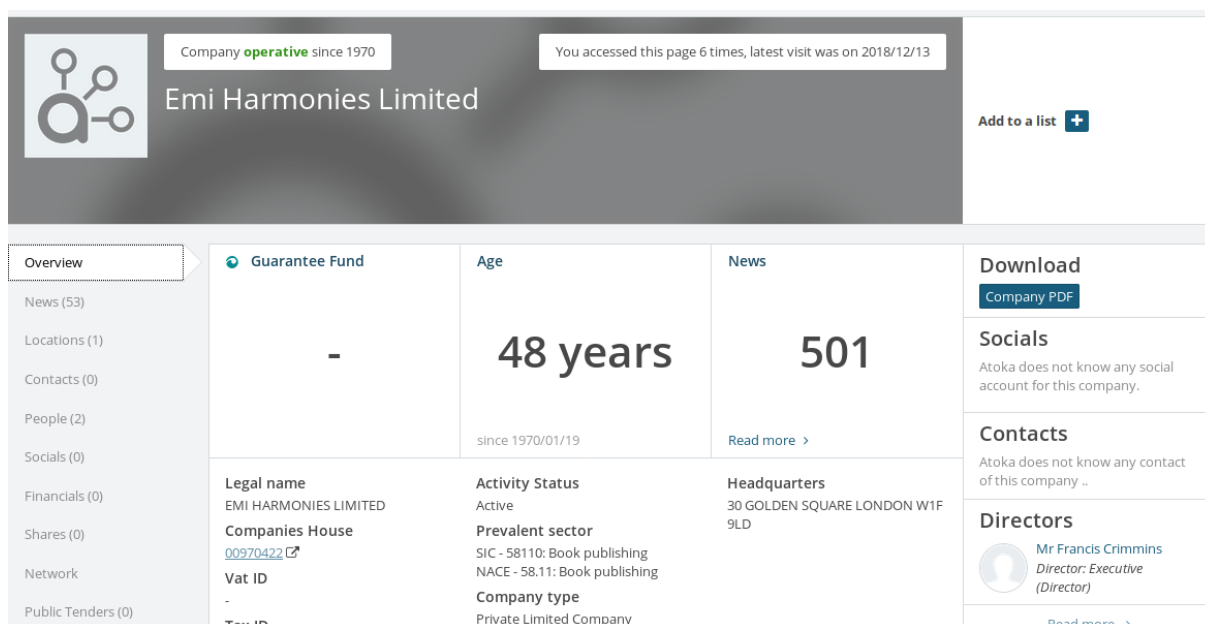


Figure 27: Example Company Overview Page in Atoka

Figure 27 shows an example in which all the available information can be seen.

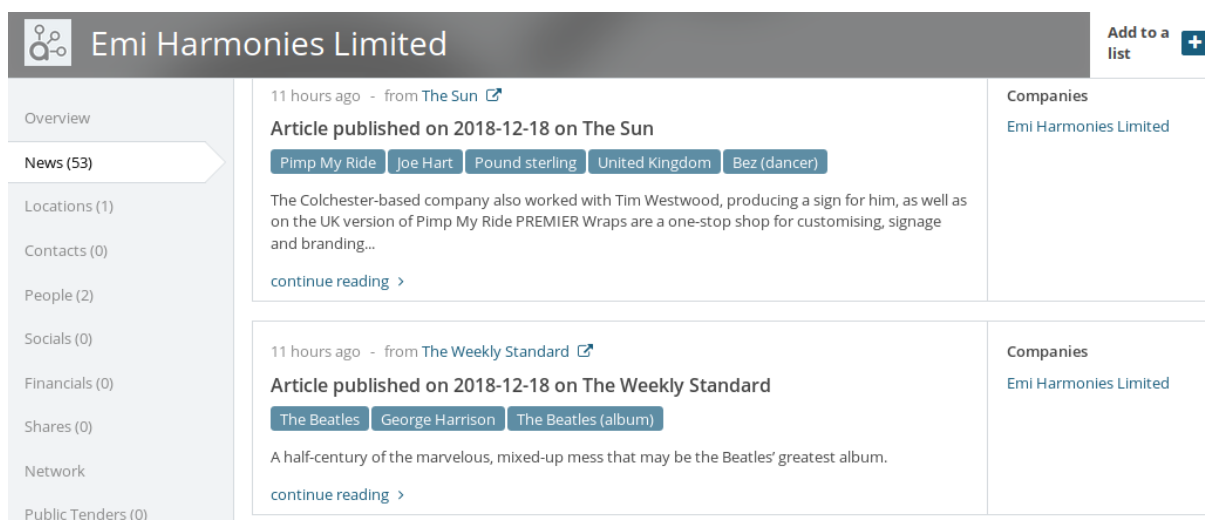


Figure 28: Company-centric news stream functionality

One of the topics, labelled "News", presents the user with the latest news articles in which the company is mentioned. Figure 28 shows an example news stream.

5 Updates to the euBusinessGraph Semantic Model (Ontology)

This section describes in detail the updates to the euBusinessGraph Common Semantic Model¹² since the first release, including important remarks regarding the use of the semantic model to onboard the data coming from the different data providers. Additionally, the section presents work done to address cross-lingual and multi-lingual challenges that arise from the task of integrating data from these different data providers.

5.1 Company data model

The work described in this section is a continuation of the work related to the first release of the **common semantic model**, presented in Deliverable D2.1 "System of identifiers, Ontologies and Vocabularies".

The first release of the common semantic model focused on representing company attributes sourced from different national business registers as well as information that is constantly gathered from the corporate web and maintained by the companies themselves. As such, the goal was to represent companies and their attributes in a consistent way, addressing the challenge of combining several representations of company entities as sourced from the different data providers.

Special attention was given to the representation of company identifiers that are fundamental to model company links in the graph and also to connect the graph to external sources. Work was done to represent not only the identifiers but also their properties in aspects such as uniqueness, immutability and official character in the issuing jurisdiction. Finally, the issuing authorities and rules were also modelled including rules for validation of identifier instances.

In this deliverable, we build on the previous work and describe the second release of the common semantic model that can be summarised as follows:

- Datasets and dataset partitions to describe what entities are present in a dataset, which attributes are included and jurisdictions that are covered, and
- Company officers and the nature of the relationships with the companies they work for.

5.1.1 Datasets

Dataset modelling is a key concern of the euBusinessGraph project. Data providers contribute subsets of their datasets, offering different properties and covering different jurisdictions. An analysis of the dataset offerings from the data providers established the need to cover the following requirements:

- Aggregating company datasets from various providers.
- Advertising dataset partitions: by jurisdiction, covered fields and others.
- Describing the different subsets that come from different data providers.
- Describing internal dataset structure and content (e.g. number of companies per jurisdiction).
- Describing rich dataset metadata such as source, publisher, dates and license.

euBusinessGraph data consumers need to know how many companies are included in a dataset, from which jurisdictions, and what depth of data is included (e.g., which properties, addresses with what geographic resolution). Consequently, there is a need to express both metadata about the dataset itself, and fine-grained statistics about the content of a dataset. Examples include:

- Publisher, source, last modified, license, home page, download distribution.
- Subset description per provider.
- Subset description by kind of entity (e.g., Companies versus Addresses), field coverage (e.g., which fields are included in which subsets), entity characteristics (e.g. Italian companies, Startups, Startups in Italy).

¹² EBG Common Semantic Model: https://docs.google.com/document/d/1dhMOTIIOC6dOK_jksJRX0CB-GIRoiYY6fWtCnZArUhU/

- Count of entities in a dataset or subset.

After researching numerous dataset description ontologies, we settled on using VOID with some extensions. VOID describes RDF datasets in terms of entities, property/class partitions and number of triples, among others. Although the original datasets by most providers are not in RDF, parts of datasets that are aggregated in euBusinessGraph are, making the case for the applicability of VOID.

We find it useful to describe the kinds of entities and available characteristics using URLs rather than strings. In this way we harmonise properties across providers, can include extra linked information about them, reuse definitions from other ontologies, and are prepared to capture them in RDF when the opportunity arises.

5.1.1.1 Dataset hierarchy

euBusinessGraph aggregates data from several data providers. These data providers select parts of their datasets to share with the project. To represent this dataset polyhierarchy, we use the void:subset relation as follows:

- Each provider describes their full dataset (e.g., <dataset/OCORP>, the full OCORP dataset), and the respective subset provided to EBG (e.g., <dataset/OCORP/EBG>). The full dataset may include parts that are not provided to EBG and are only "advertised" in the EBG marketplace application.
- The EBG dataset (<dataset/EBG>) points to all subsets it aggregates from different providers (e.g., <dataset/OCORP/EBG>, <dataset/SDATI/EBG>, <dataset/BRC>, <dataset/ONTO>). Since the last two in the example list are fully provided to EBG, there is no need to describe subsets (e.g., see <dataset/ONTO/EBG>).

For each dataset we specify:

- dct:publisher
- dct:type to indicate the kind of dataset:
 - <dataset/type/provider>: a provider dataset.
 - <dataset/type/provided-to-EBG>: subset provided to EBG.
 - <dataset/type/EBG>: the root EBG dataset (aggregation).
 - <dataset/type/partition>: describes a jurisdiction and property partition.

Each dataset must specify a dct:license.

A dataset covers one or more jurisdictions. A void:subset is created for each jurisdiction, proceeding in the same way even if there is only one subset in the jurisdiction for uniformity reasons.

Examples:

- <dataset/OCORP> includes <dataset/OCORP/UK>, <dataset/OCORP/HR> and so on.
- <dataset/ONTO> includes only <dataset/ONTO/BG>.



5.1.1.2 Dataset sources

- A *brief* style using `dct:source`, without distinguishing the nature of the source.
- Further description of source details using `ebq:WebResource`.



- <http://brra.bg>: Registry Agency.
- <http://www.registryagency.bg/bg/registri/targovski-registar>: Trade Register home page.
- <http://opendata.government.bg/dataset/tbprobckn-pernctbp>: Trade Register open data.

5.1.1.3 Dataset partitions

Page 39 / 51

For the following examples we resort to *partitions*, that is, subsets of a dataset constructed in specific ways, using blank nodes to represent each partition and `dct:type <dataset/partition>` to distinguish them from other kinds of datasets (e.g., a downloadable).

We use the following mechanisms:

- To specify the **jurisdictions** that a dataset covers, we use `dct:spatial` with a NUTS node for European countries (e.g., `nuts:IT` for Italy) or a literal in other cases (e.g., "US-CA" California). Literals are also useful to describe subsets covering only a province (e.g., `nuts:ITH31` Verona).
- To specify the **number** of core entities (e.g., companies) included in a dataset, we use `void:entities`.
- To describe several **kinds of entities** we use `void:classPartition` with `void:class` giving the entity class, and `void:entities` to provide the number.
- To specify **properties** (features) included in a dataset, we use `void:propertyPartition` with `void:property` listing all relevant properties. This specifies the dataset has at least some triples using each of those properties. Although `void:` suggests to use only one `void:property` per partition, we allow multiple uses to simplify the representation. No claim is implied that every entity in the dataset has all these properties. A superset should list the union of properties included in its subsets (and may list more, if the subsets don't partition it in full).
- To specify the **number of statements with given property**, we use `void:propertyPartition` with `void:property` specifying the prop, and `void:triples` giving the number of statements. If we need to count several specific properties, we need to use several `propertyPartitions`, with one `void:property` per partition.

Figure 31 shows:

- Total number of companies (main entities) in the overall EBG dataset (`<dataset>`), its provided subsets (e.g., `<dataset/ONTO>`, `<dataset/OCORP/EBG>`), as well as provider datasets (e.g., `<dataset/OCORP>`, `<dataset/OCORP/GB>`).
- Number of entities per class: `<dataset/companies>` (`rov:RegisteredOrganization`) as main entity, and optionally `<dataset/persons>` (`schema:Person`), `<dataset/addresses>` (`locn:Address`) as subsidiary entities.
- Spatial coverage (jurisdictions) of the overall EBG dataset (e.g., `nuts:BG`, `nuts:GB`, `nuts:IT`, `nuts:NO`), its subsets and provider datasets. We show only two for `<dataset/OCORP>` (`nuts:GB`, `nuts:HR`) but in fact there are more.

Features included in each dataset and/or subset can also be seen. Some examples follow:

- `<dataset/OCORP/EBG>` includes `ebg:orgTypeText` (legal form free text), `locn:adminUnitL1` (address country), `locn:fullAddress` (address free text).
- `<dataset/EBG>` includes also `rov:orgType` (legal form nomenclature), `locn:adminUnitL2`, `ebg:adminUnitL3`, `ebg:adminUnitL4` (administrative place hierarchy), `locn:postName`, `locn:thoroughfare`, `locn:postCode` (address fields). This wider set of features comes from various subsets: there is no implication that all entities have these features.
- `<dataset/SDATI/IT>` includes specific Boolean properties `ebg:isStartup`, `ebg:isStateOwned`, `ebg:isPubliclyTraded`, which are not provided to the EBG dataset.
- We use the fact that `void:propertyPartition` is a sub-property of `void:subset` thus compatible with it, and use the same node to express both jurisdiction and property partition, connecting that node by both `void:propertyPartition` and `void:subset` to the main dataset.

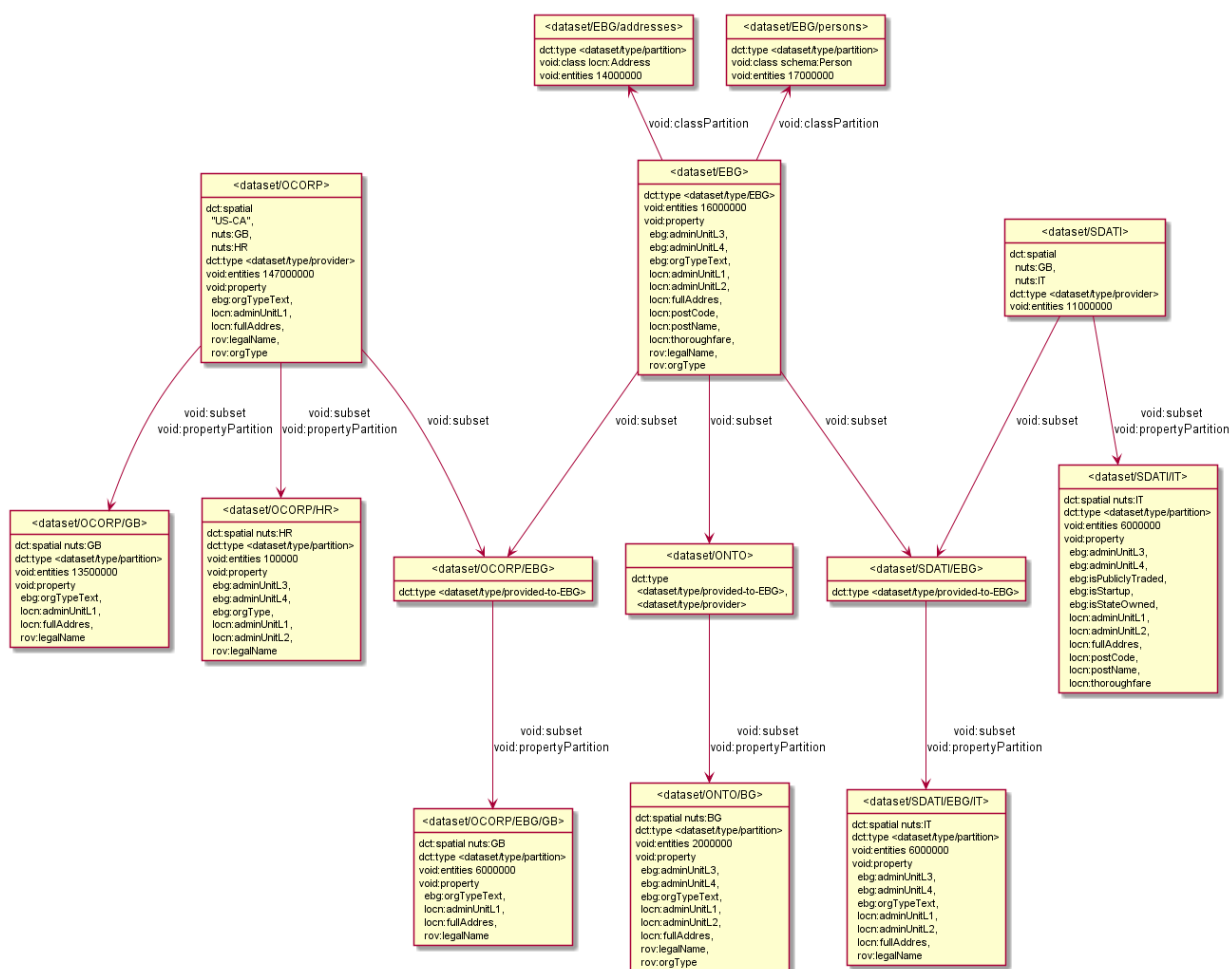


Figure 31: EBG Dataset Partitions

Finally, we introduce one extension of void: from the VOID-ext (vext:) ontology that lets us specify the number of statements with given property and object. In such cases we use void:propertyPartition and then vext:objectPartition.

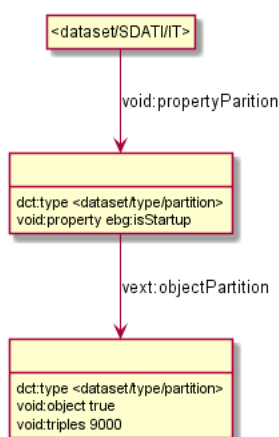


Figure 32: Using [void:] and [vext:] to specify number of statements with given property and object

Figure 32 shows <dataset/SDATI/IT> including 9k startup companies. We could use similar mechanisms to express even more complex combinations (e.g., "number of startup companies in the province of Verona" or what exact properties are provided for these companies).

5.1.1.4 Dataset metadata

Well-described datasets are easier to discover and use by consumers. To this end, the semantic model makes use of void: descriptive properties that include but are not restricted to:

- dct:type, dct:license, dct:publisher as described in the previous “Dataset hierarchy” subsection.
- dct:creator: possibly being the same as dct:publisher or another party who helped with creating the data.
- Source info as described in “Dataset sources”.
- Statistical/partition info as described in “Dataset partitions”.
- dct:title: name (mandatory).
- dct:description: description (optional).
- dct:subject: topic of the dataset (e.g., "companies"). We express it using several LOD resources:
 - <<http://dbpedia.org/resource/Company>>
 - <<https://www.wikidata.org/entity/Q783794>>
 - <<http://vocab.getty.edu/aat/300160084>>
 - <<http://eurovoc.europa.eu/4189>>
- dct:date: dataset date. If more specific dates are desired: dct:created (creation), dct:published (publication), dct:modified (last update of the underlying data).
- dct:accrualPeriodicity: how often the dataset is updated.
- foaf:homepage, foaf:page: homepage or another web page describing the dataset.
- void:sparqlEndpoint: SPARQL endpoint that can be used to query the dataset. Subsets are assumed to be accessible from the same one.
- void:dataDump: to associate and describe dumps in various formats.
- void:exampleResource: to describe examples of core entities in the dataset.
- void:uriSpace: to define a common prefix of the core resources in the dataset. For the EBG dataset, these are <http://data.businessgraph.io/company/> and <http://data.businessgraph.io/person/>, inherited by subsets.
- void:features: RDF formats the dataset is available in (format:N-Triples, format:RDF_XML, format:Turtle).
- void:vocabulary: ontologies used in the dataset. For the EBG dataset, these are [ebg:], [foaf:], [locn:], [ngeo:], [org:], [ramon:], [rov:], [schema:], [sioc:], [time:]. As a convention, we don't list ontologies used only for dataset metadata (e.g., [void:], [format:], [freq:]).

For more details, see the EBG Dataset Description¹³.

5.1.2 Company officers and memberships

To represent company officers and their relationships with the companies they work for, we use the W3C Org Membership model in a straightforward way.

Consider Figure 33 that shows membership examples with the following Officer records:

- One director from OCORP UK: Kiryakov at LDBC (see Figure 33a).
- Two directors from BG: Kiryakov and Momtchev at ONTO, using a shared org:Membership node (see Figure 33b). This works only if the whole board of directors is appointed and discharged at

¹³ EBG Dataset Description: <https://github.com/euBusinessGraph/eubg-data/blob/master/data/dataset/dataset-EBG.ttl>

once (as it is at AGMs). To reflect individual directors coming and going, individual Membership records are needed.

- One person, Michele Barbera, with two roles in the same organization, SDATI, with different roles and intervals (see Figure 33c). In this case we use two different Membership nodes.

We describe the classes and properties needed to satisfy these examples in the following subsections.

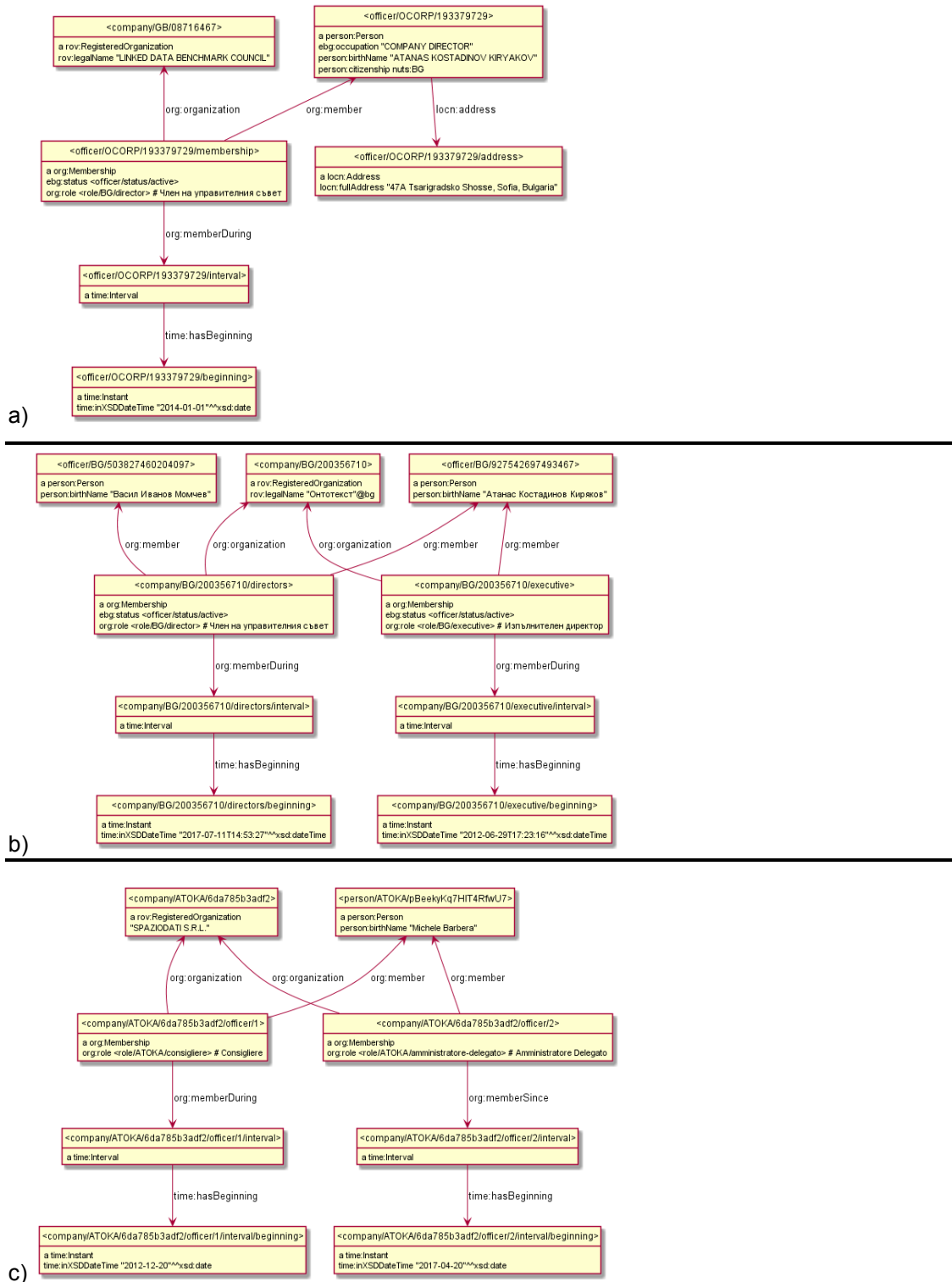


Figure 33: Membership examples

5.1.2.1 Officer

To represent officer data we follow the Person Core Vocabulary [person:]. Only the officer's name and identifier are mandatory. The identifier may come from official registries or be derived from these. Additionally, other properties may be present such as address, birth date and citizenship.

Class: Officer	
Description	An officer is a natural person (as opposed to a legal person) [person:] that has a high-level management role in a Company (e.g., the CEO, treasurer and chief financial officer). Despite their high status, they typically serve at the will of the corporate directors, who can fire or replace them. Officers can also be shareholders and directors but do not necessarily have to be. They have the authority to act on behalf of the corporation, including contract authority.
Examples	Whether a person is an officer or not is determined by the role they fulfill which must entail management attributes. Among the roles that determine that a person is an officer, we may find: <ul style="list-style-type: none"> • Chief Executive Officer: the head of management on an organisation. • Chief Operating Officer: responsible for the day-to-day operations of an organisation. • Chief Financial Officer: in charge of a corporation's financial risk.
RDF	person:Person
Data Property: officer name	
Description	The legal name of the person which is usually determined at or around birth, usually persistent and for this reason usually recorded by some official registers. Although birth names tend to be persistent we do not assume this has to be the case.
Rules	Leading, trailing and consecutive spaces are not allowed.
Examples	<ul style="list-style-type: none"> • "Michele Barbera" • "Atanas Kostadinov Kiryakov"
RDF	person:birthName
Data Type	xsd:string
Cardinality	1
Data Property: officer address	
Description	Full address as free text.
Examples	"47A Tsarigradsko Shosse, Sofia, Bulgaria"
RDF	locn:fullAddress
Data Type	rdf:langString or xsd:string
Cardinality	0 or 1

URI Construction:

- If we have the official registration ID of the officer, we use that **officer/<jurisdiction>/<id>**
Example: the officer officer/BG/927542697493467 "Atanas Kostadinov".
- Otherwise we use an ID from a data provider (alternative business register) **officer/<provider>/<id>**
Example: company/ATOKA/pBeekyKq7HIT4RfwU7 "Michele Barbera".

5.1.2.2 Membership

A membership describes the relation between a company officer and the company in which the officer fulfils a role.

Class: Membership	
Description	A membership describes the relation between an organisation and a member of that organisation.
Scope Note	It is acceptable to use the same Membership node for several officers, if they had exactly the same role and membership interval (e.g. in many countries the Board is appointed and discharged at the same interval, namely between AGMs of the company)
Example	Michele Barbera became a director at SpazioDati on April 20th, 2012.
RDF	org:Membership
Object Property: Membership Company	
Description	Points to the company side of the membership relation.
Examples	
RDF	org:organization
Range	Company
Cardinality	1
Object Property: Member Person	
Description	The officer that the membership is about.
Examples	<ul style="list-style-type: none"> <officer/ATOKA/pBeekyKq7HIT4RfwU7> for "Michele Barbera". <officer/BG/927542697493467> for "Атанас Костадинов Киряков".
RDF	person:person
Range	Officer
Cardinality	1
Object Property: Membership Interval	
Description	The interval (beginning/end) of the membership relation. May be an open interval.
Scope Note	For the common case of missing time, we prefer to use xsd:date rather than completing it with a fake null time (e.g., "<date>T00:00:00Z"^^xsd:dateTime).
Examples	
RDF	org:memberDuring
Range	time:Interval
Cardinality	1
Object Property: Membership Role	
Description	The role that the officer fulfils according to the membership.
Scope Note	EBG does not introduce a global set of officer roles. These may vary per jurisdiction and/or provider
Examples	<role/BG/director>
RDF	org:role

Range	org:Role
Cardinality	0 or more

URL Construction

Membership

Memberships are expressed from the company side of the relation like this:

<company url>/officer/<id>,

where <id> is a sequential number or type of officer (e.g. company/BG/927542697493467/executive and company/ATOKA/6da785b3adf2/officer/1).

This works only if the most recent officer records are RDFized (i.e., we have a single Executive and Board of Directors per company).

In the general case each Membership depends on person (org:member), organization (org:organization), and interval (org:memberDuring), so a GUID should be used.

Interval

Interval URLs are built according to **<membership_URL>/interval** (e.g., company/BG/200356710/executive/interval).

This node includes relations time:hasBeginning and/or time:hasEnd that point to further nodes **<membership_URL>/interval/beginning** and **<membership_URL>/end** having attributes time:inXSDDateTime.

Role

Roles are attached to the jurisdiction following **role/<jurisdiction>/<role_label>** or provider as **role/<provider>/<role_label>** to which they apply to (e.g., <role/BG/director> for the role of director in the Bulgarian jurisdiction).

5.1.3 Business graph data onboarding

The Company Data Model is used to onboard and integrate company data from the different data providers (i.e., OCORP, SDATI and BRC) into a semantic knowledge graph referred to as the **business graph**. Figure 34 illustrates how company data from different data providers are onboarded to the business graph. Grafterizer¹⁴ is used together with ASIA to onboard the company data in the project, i.e. mapping the company data to the euBusinessGraph company data model (euBiz ontology¹⁵) in order to publish the data as a knowledge graph in GraphDB. The onboarding process involves the following three steps:

1. *Tabular transformation:* **Grafterizer** simplifies the data cleaning and transformation processes for onboarding business graph data. It provides suggestion-based data cleaning and transformation, and visual data profiling.
2. *Tabular annotation:* **ASIA** is semantic table enrichment tool which provides Assisted Semantic Interpretation and Annotation of tables (ASIA), simplifying the tabular to RDF mapping. **ABSTAT** computes summaries of existing data onboarded and gives mapping suggestions when onboarding new datasets.
3. *RDF mapping:* **Grafterizer** also provides a graphical tree-based interface that supports mapping of tabular CSV data to RDF knowledge graphs.

¹⁴ <https://datagraft.io/>

¹⁵ <https://github.com/euBusinessGraph/eubg-data/blob/master/model/ebg-ontology.ttl>

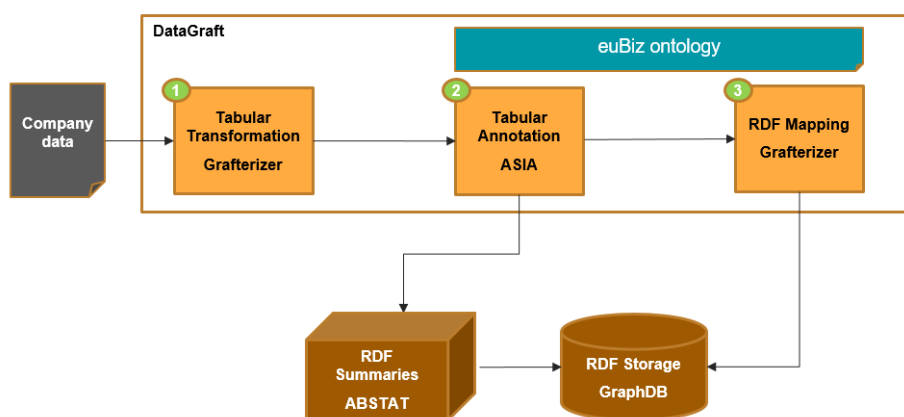


Figure 34: Business graph data onboarding with Grafterizer and ASIA

5.2 Cross-linguality/multilinguality in euBusinessGraph data model

Deliverable D2.1 describes the euBusinessGraph company data model, while Section 5 of the current document provides the updates to the euBusinessGraph semantic model.

EuBusinessGraph partners defined several properties that can be represented in the multilingual mode. For instance, euBusinessGraph company data model captures company names in different languages:

- Each company name should include an optional language tag (xml:lang).

The data properties *type text*, *status text* as well includes valid language tag (from [iana:]). An important feature of the semantic model is the physical representation of companies with data properties (locality/city/settlement) providing language tags where possible.

While Task 2.1 and Task 2.2 of WP2 of the project deal with system of Identifiers for corporate-related data and entities and with development of agreed and shared data models, Task 2.3 presents cross-lingual semantic annotation for unstructured data. Below in the document we look into the outputs of the cross-lingual semantic annotations.

The NLP Interchange Format (NIF) is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. NIF ontology can be used to present the cross-lingual semantic annotations in RDF.

JSI Wikifier returns JSON response of the following form:

Table 8: JSI Wikifier output

Field	Type
annotations	array of objects
spaces	array of strings
words	array of strings
normWords	array of strings
ranges	array of objects
verbs	array of objects
nouns	array of objects
adjectives	array of objects

adverbs	array of objects
language	string
languageAutodetectDetails	object

The `spaces` and `words` arrays show how the input document has been split into words. `annotations` is an array of objects of the following form:

Table 9: JSI Wikifier annotations

Field	Example
<code>title</code>	"New York City"
<code>url</code>	"http://en.wikipedia.org/wiki/New_York_City"
<code>lang</code>	"en"
<code>pageRank</code>	0.102831
<code>cosine</code>	0.662925
<code>secLang</code>	"en"
<code>secTitle</code>	"New York City"
<code>secUrl</code>	"http://en.wikipedia.org/wiki/New_York_City"
<code>wikiDataClasses</code>	[{"itemId":"Q515", "enLabel":"city"}, {"itemId":"Q1549591", "enLabel":"big city"}, ...]
<code>wikiDataClassIds</code>	["Q515", "Q1549591", ...]
<code>dbPediaTypes</code>	["City", "Settlement", "PopulatedPlace", ...]
<code>dbPediaIri</code>	"http://dbpedia.org/resource/New_York_City"
<code>supportLen</code>	2.000000
<code>support</code>	[{"wFrom":0, "wTo":1, "chFrom": 0, "chTo": 7, "pMentionGivenSurface":0.122591, "pageRank":0.018634}, {"wFrom":0, "wTo":2, "chFrom": 0, "chTo": 12, "pMentionGivenSurface":0.483354, "pageRank":0.073469}]

`url` is the URL of the Wikipedia page corresponding to this annotation, and `title` is its title.

`lang` is the language code of the Wikipedia from which this annotation is taken.

`secUrl` and `secTitle` refer to the equivalent page of the Wikipedia in the language `secLang`.

`wikiDataClasses` and `wikiDataClassIds` are lists of the classes to which this concept belongs according to WikiData (using the `instanceOf` property, and then all their ancestors that can be reached with the `subclassOf` property).

`dbPediaIri` is (one of) the DBPedia IRIs corresponding to this annotation, and `dbPediaTypes` are types to which this DBPedia IRI is connected via the `http://www.w3.org/1999/02/22-rdf-syntax-ns#type` property.

`support` is an array of all the subranges in the document that support this particular annotation. `wFrom` and `wTo` are the indices (into words).

JSI Wikifier provides annotations linked to **Wikipedia** concepts and **DBpedia** types, and in this way covers NER individuals.

NIF ontology¹⁶ contains classes such as **Annotation**, **Word**, **String**, **Paragraph** and **Context** that can be used for JSI Wikifier output representations.

NIF properties, such as `nif:beginIndex` and `nif:endIndex` can be associated with information presented in support field.

The Wikifier can additionally provide four arrays called `verbs`, `nouns`, `adjectives`, and `adverbs`. Each of these arrays is of the following form:

Table 10: JSI Wikifier part-of-speech output (verbs)

Field	Example
<code>iFrom</code>	27
<code>iTo</code>	32
<code>normForm</code>	"offer"
<code>synsetIds</code>	["200706557", "200871623", ..]

For each entry, `iFrom` and `iTo` are the indices of the first and last character of that verb. The indices refer to the input text as a sequence of Unicode codepoints (i.e. not as a sequence of bytes that is the result of UTF-8 encoding). You can use these indices to recover the surface form of this verb as it appears in the input text. By contrast, `normForm` is the lemmatized form (e.g. *have* instead of *has*). `synsetIds` is a list of all the Wordnet synsets that contain this verb. Part-of-speech tags from WordNet can be associated with lemmas from **Wordnet RDF**¹⁷.

¹⁶ <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html>

¹⁷ <http://wordnet-rdf.princeton.edu>

6 Summary and Outlook

This document reports on the development and application of cross-lingual and multilingual tools for data management, event categorization and relation extraction.

We presented semantic annotation tools and methods for unstructured data sources (such as news and blog posts) and structured data sources (such as tables).

JSI Wikifier is a web service for annotating multilingual textual documents with concepts from Wikipedia. JSI Wikifier currently supports over 130 languages and includes additional functionalities that allow for annotation with extra vocabularies in financial and business domains.

The annotation of structured data provided as tables has the role of supporting data onboarding. **ASIA** is the annotation tool that extends the functionalities of **Grafterizer** in such a way that data providers can onboard their data using the reference ontology. ASIA adopts a semi-automatic approach to table annotation and can support cross-lingual instance-level annotations by plugging in a cross-lingual reconciliation service.

Event Registry is a system for cross-lingual and multilingual news tracking and analysis, event identification and observation. Event Registry provides a set of functionalities related to media monitoring and media intelligence.

The developed methodology for **business event types categorization** covers over 30 business event types, such as acquisitions and mergers, bankruptcies, earnings, partnerships, products and services etc.

For the euBusinessGraph project we have developed a **“Relation Tracker”** tool that operates based on Event Registry data and tracks main business entities (people and organizations) within each topic through time. The main types of relations between the entities are detected and observed in time. The tool provides multiple ways of visualizing this information with different scales and durations.

Deliverable D2.2 contains the examples of **cross-linguality and multilinguality applications in euBusinessGraph business cases**, in particular in Data Journalism product from DW and Atoka+ product from SDATI.

In addition, this document provides the **updates for euBusinessGraph semantic data model**, initially described in euBusinessGraph D2.1. In particular, the second release of the common semantic model provides the datasets and dataset partitions to describe what entities are present in a dataset, which attributes are included and jurisdictions that are covered, and company officers, including the nature of the relationships with the companies they work for.

The project foresees future application of the developed cross-lingual and multilingual tools and services inside of euBusinessGraph marketplace and individual business cases.

The future work lies in the extension and deep analysis of event types, along as in development of additional business-related features based on news data.

7 References

[Mihalcea and Csomai] Rada Mihalcea, Andras Csomai. Wikify! Linking Documents to Encyclopedic Knowledge, CIKM'07.

[Zhang and Rettinger] Zhang, Lei & Rettinger, Achim. (2014). Semantic Annotation, Analysis and Comparison: A Multilingual and Cross-lingual Text Analytics Toolkit. 13-16. 10.3115/v1/E14-2004.

[Miller] George A. Miller. (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[Mikolov et al.] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

[Kingma and Ba] Kingma, Diederik P., and Jimmy Ba. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980