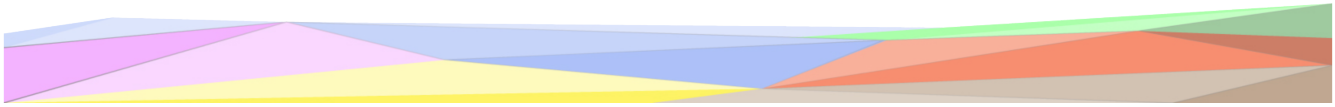
Dipartimento di Ingegneria e Scienza dellInformazione

– KnowDive Group –

Domain Formal Modelling: a Business Ontology

Document Data:	Reference Persons:
	Jacopo Gobbi and Kateryna Konotopska



Index:

Revision History:

Revision	Date	Author	Description of Changes
0.1	xx/10-12/18	J&K	Document Created
0.11	xx/10-12/18	J&K	Added raw scenarios
0.12	xx/10-12/18	J&K	Added ER
0.13	xx/10-12/18	J&K	Added ontology
0.14	xx/10-12/18	J&K	Improved scenarios section
0.14	xx/10-12/18	J&K	Updating ER
0.14	xx/10-12/18	J&K	Updating ontology
0.15	xx/10-12/18	J&K	Added model formalization description
0.16	xx/10-12/18	J&K	Added lexical info description
0.17	xx/10-12/18	J&K	Added list of classes, properties
0.18	xx/10-12/18	J&K	Added top level grounding content and picture
0.18	xx/10-12/18	J&K	Updating ontology
0.18	xx/10-12/18	J&K	Updating ontology grounding content and picture
0.19	xx/10-12/18	J&K	Added Generalized Queries
0.20	xx/10-12/18	J&K	Added final considerations
0.21	xx/10-12/18	J&K	Added visualization pictures
0.21	xx/10-12/18	J&K	Refining pass
0.22	xx/10-12/18	J&K	Added evaluation
0.23	xx/10-12/18	J&K	Refining pass
1.00	16/12/18	J&K	Adding title before submission

1 Scenario Description

Location macro area

Paolo Rossi, 26, just got his master degree from the University of Trento, he now wishes to know more about companies in his area, more specifically, he wants to browse nearby IT companies. He would like to know what is the average salary of those companies, and how many people work there.

Marco Caldi, 18, is preparing a report for its economics class; he is focusing on the Silicon Valley tech boom. He needs to find, for each company with headquarters in the Silicon Valley, its founders, the foundation date, its category and its stock symbol, if it has one.

People macro area

Salvo Montalbano is a fourty years old revenue officer, and wishes to know more about the business life of a certain indivual, Mr. X; more particularly, he wants to know all the companies that have employed him, and the industry they are part of.

Talia Trenit, 31, is leading an investigation on fraud that involves patent admissions, she needs to find all patents that have been granted and that which were examined by Rossi Sara, she would like those patents to be grouped by assignee company.

Alto Garda Assicurazioni wants to keep track of people employed by its competitor, Itas Assicurazioni, it would be nice to have a list of people that are employed in that company, where they live, their degree and their email.

Companies macro area

Billy J. Wood is a 45 years old hedge fund manager, he is looking to make some money by investing in companies that are about to be acquired by big names, such as Google, Nvidia, etc. He wants to browse the acquisitions made by those companies in the last year, and what category would define those companies.

Thomas Brewest, 51, a columnist of the Financial Times, is writing an article about the changing business model and revenue sources of companies in the IT sector, moreover, he wants to see if those business models are similar to the ones of companies that have "subscription" and "trial" in their description.

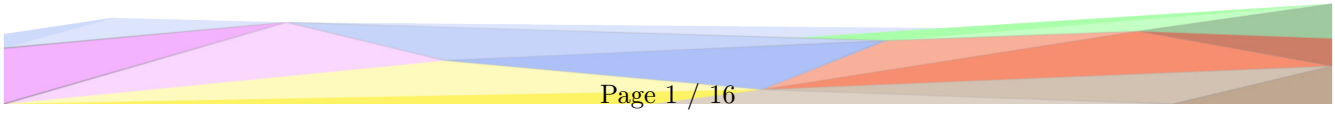
Mixed macro area

Bobby Ryan, 43, runs a pharmaceutical company. He wants the next cycle of investments to be focused on acquiring know-how and patents, to do that, a possible solution would be buying whole companies that either have many patents, many PhDs with the needed knownledge, or both. He needs to find companies with an high number of pharmaceutical patents or people specialized in what he is looking for.

Mario Martini, 36, is currently thinking about opening a company, and he does not like having too many people around; he would like to know, for each industry, whats the category that has the highest profit to employees ratio.

Billy J. Wood, 48, is an hedge fund manager that is now looking to assume someone to help him in financial analysis and automate some work, he's looking for people that got their university degree in Computer Science, Math, Physics or Finance in the past 5 years, the candidate should have problem solving skills; for all returned candidates he wants to know past working positions.

Valeria Marini is a 28 years old secretary in the University of Roma, she has been tasked to find out where graduates in Computer Science are now working, in terms of companies but also where they live now.



2 Model Formalization Description

Please note that not all data was considered, meaning that different columns/fields of different datasets were outright not modeled into our design either because of the documentation accompanying the dataset did not provide any explanation for such fields or it did not exist at all, or because such data was considered too peculiar/specific for the current iteration of the data integration task.
Examples are the founding round characterizing an investment, particular codes related to a patent requiring more than basic knownledge in order to be understood, the year a company has entered the top 500 fortune list, and of course all fields with no explanation and not obvious meaning.
A - rough - list of datasets and their schemas (columns) is provided within the directory containing the report.

* * *

Starting from the scenarios and the data at our disposal, we first iteratively built an ER model [6].
This ER model was designed as if meant for a relational database, without focusing on the future objective of designing an ontology.
While this may seem counter intuitive, it helped by modularizing our task; of course, this also meant that the formalization led to a gap between the ER schema and the ontology bigger than what could have been between the ontology and a simplified version of the ER schema, which we do not provide.
The ER schema is provided as an html file, and can be found in the local directory along with its source file (graphml).
The differences between the informal and formal model are the following:

- We initially modeled industry codes (ISIC, NAICS, SIC) as the class **descriptive code**, using a relation **maps to** to consider the fact this codes could map to other **category** types (also meaning that a code could map to other codes, i.e. representing the same industry).
  This however was not needed in the ontology, where we simply modeled each **industry** as having a **business code** (this actually better "follows" our data), implicitly modeling the mapping between different codes as the codes linked to the same **industry**.

- The **employment** class has been remodeled: we consider a **named entity**, **Role**, that has a **start** and **end date**, and specialize it into an **Employee** entity. An employee is **employed by** an **organization**, is **played by** a **Person** and **receives** a **salary**.
  This allows us to have a more "ontological" representation while simplifying the model by getting rid of the **employment role** that was previously needed to link a **Role** to an **employment**.

- **Had experience in** and its relationships with **Person** and **Category** has been remodeled by using an object property (**hadExperienceIn**) which **Person** to **Sector** or **Industry**.

- **Geospatial coordinates** is not a location anymore, instead we consider anything having a **latitude** and **longitude** as equivalent to an **Address**.

- **Qualification** has been remodeled by considering a **Student** (is a **Role**) which is **played by** a **Person** and receives an **Education**. This makes the model easier to expand and work on, given that **Student** can be specialized into more refined roles and the same can be said for **Education**.
  An example would be the fact that we specialized **Education** into **Institutional Education**, which is **provided by** some **Organization**.

- **Contact** led itself to be naturally transformed into a data property named **Contact** which is later specialized to represent all possible different form of contacts (**email**, **phone number**, etc.). This property maps to a string.
- Initially, we meant the role of founder to be modeled through the **employment** class, which was semantically incorrect. This role is now modeled by a specialization of **Event**, **Foundation**, which is linked to founders (**Person**) via a **founder** object property, and captures the foundation of an **Organization** at a certain **date**.
- All forms of monetary transactions/payments (found in **employment**, **investment**, **acquisition**, etc.) are now modeled by a complex data type, **Money Amount**, characterized by a **currency code** and a **numerical value**.
- The **derived product** property (string) of **acquisition** has been given a full etype to allow for future expandability, for obvious reasons.
- **Financial Information** has been decomposed, meaning that we consider a specialization of **Complex Data Type**, **Financial Data Type**, which has a date, a name and an amount that can either be a **numerical value** or a **Money Amount**. This way companies will be linked to nodes of financial information when it is available, instead of being associated with very sparse rows.
- The **name** class and the **has/had** relation have been made more intuitive by simply having a data property **Name**, among its specializations we now have **Past Name**. The **up to date** property has been dropped given that it is not a real case in our scenario.

* * *

The formalization phase was not simply a translation of the ER model to an ontology, but was also a process in which we took in consideration different priorities: making it easy to use, making it easy to expand, have it model our data and case scenarios.
As an example, we modeled **Category**, **Sector** and **Industry** as etypes, given that it **1)** allows for easy re-use of **part of** object properties, which are common; **2)** it helps future expandability by providing **Industry** as a base class that can be further expanded upon by adding more information, **3)** it does not make our model harder to use, but instead supports looking for equivalence between industry codes, industries, etc.
Another example is the introduction of **Role** that allows a separation between a Person and what it does, allowing future users to easily place more roles (either by specializing **Role**, **Student** or Employee).
Expandability has also been taken in consideration when designing locations (**Address**, **City**, **Region**, **Country**): given our data, we could simply model those as strings, but it would obviously be an obstacle for future improvements of this model, since cities, countries, etc. are common concepts for which data can be found easily.
Instead of designing investors, founders and acquirers as roles, we modeled those are relations in the events **Acquisition**, **Investment** and **Foundation**. This is because, in our opinion, related scenarios are more elegantly modeled this way, especially because those events are characterized by a single **date** instead of a starting and an ending one.
We provide the full list of classes, object properties and data properties in the following table; the data types used in the model are: xsd:dateTime, xsd:dateTimeStamp, xsd:float, xsd:integer, xsd:string.

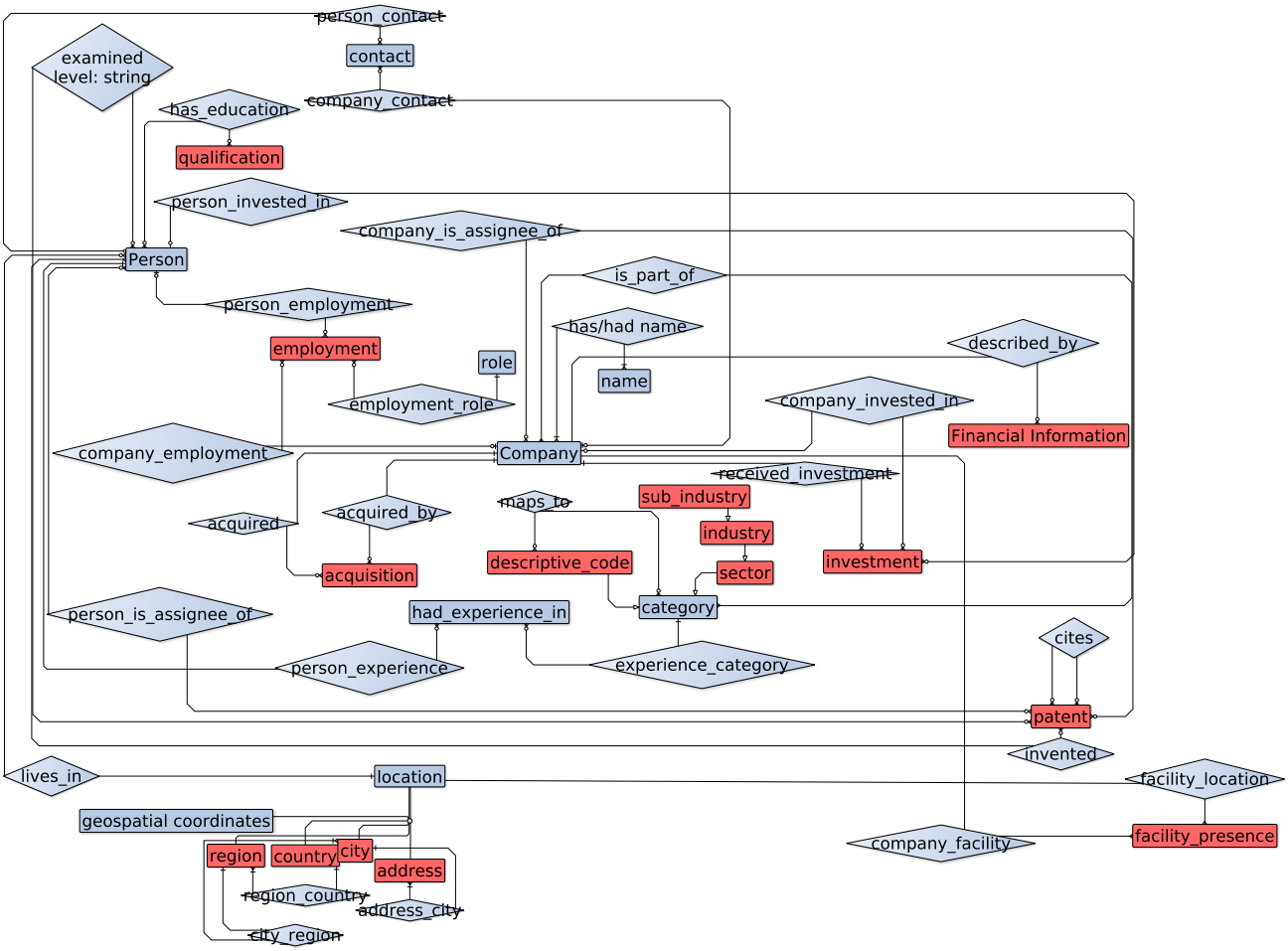


Figure 1: A visualization of the ER model where properties have been removed to allow for easier visualization in the document, as previously stated the full model can be found in the local directory.

Acquisition	Address	Art_Unit	Business_Organization	Category	City
Country	Currency	Derived_Product	Education	Employee	Event
Financial_Descriptor	Foundation	Industry	Institutional_Education	Intellectual_Property	Investment
Location	Money_Amount	Monthly_Recurrence	Named_Entity	Organization	Patent
Person	Product	Recurrence	Region	Role	Salary
Sector	Service	Student	Sub_Industry	Yearly_Recurrence	
acquired	acquirer	amount	assignee	cites	derived
employedBy	examinedBy	founded	founder	hadExperienceIn	hasArtUnit
hasContact	hasFacilityIn	hasFinancialDescriptor	hasHeadquartersIn	hasMember	inventedBy
investor	livesIn	medianWorkerCompensation	memberOf	partOf	playedBy
providedBy	receives	recurs	unitOfMeasure		
abstract	applicationNumber	artUnitCode	birthDate	buildingDescription	buildingNumber
businessCode	businessModel	code	Contact	countryCode	currencyCode
date	description	dissolutionDate	educationName	email	endDate
filingDate	financialConcept	foundationDate	gender	grantDate	ISIC
lastMilestoneDate	latitude	longitude	NAICS	name	numberOfEmployees
numberOfMembers	numericalValue	pastName	personName	phoneNumber	priorityDate
publicationCode	publicationDate	revenueSource	SIC	skills	socialProfileUrl
specializationName	startDate	stockSymbol	surname	title	workingWithDataType
zipCode					

Table 1: Classes (yellow background), object properties (light blue background) and data properties (light red background) of the designed ontology.

3 Lexical Information upload description

We report, for each name used in our model (entities, object properties, data properties), a grounding lexical information, by providing the WordNet [4] ID of the synset in which such word belongs, the BabelNet [5] ID, the word itself and the WordNet synset members, which represent our motivation for the grounding choices.
BabelNet is a multilingual resource covering many (hundreds) languages, and represents a semantic network that can be used to easily find information and meta data across different languages.

Wordnet ID	Babelnet ID	Used in Model	Synset Members
78239	1005n	acquisition	acquisition
8508037	1303n	address	address
8077878	14136n	business_organization	business concern business_concern business_organization business_organisation
5847274	16733n	category	category
8542298	3335997n	city	city metropolis urban_center
6271913	22112n	contact	liaison link contact inter-group_communication
8562388	23235n	country	country state land
13407086	24507n	currency	currency
5993172	26980n	education	education
10073616	14292888n	employee	employee
29677	32021n	event	event
241051	23654n	foundation	initiation founding foundation institution origination creation innovation introduction instauration
8082070	46576n	industry	industry
13266237	47023n	intellectual_property	intellectual_property
1101341	47358n	investment	investing investment
27365	51760n	location	location
256458	116278r	monthly_recurrence	monthly
8024893	59480n	organization	organization organisation
6513132	60984n	patent	patent patent_of_invention
7846	46516n	person	person individual someone somebody mortal soul
3754377	54416n	product	merchandise ware product
7357963	66620n	recurrence	recurrence return
8648560	66884n	region	region
721817	36823n	role	function office part role
13300285	29416n	salary	wage pay earnings remuneration salary
7983333	70206n	sector	sector
578562	70651n	service	service
10685137	29806n	student	student pupil educatee
1976215	97149a	yearly_recurrence	annual yearly
6480622	489n	abstract	outline synopsis abstract precis
8436519	1000n	acquirer	acquirer
5115065	3601n	amount	amount
9834860	6511n	assignee	assignee
15277233	10679n	birthdate	birthday natal_day
1709116	82412v	cites	reference cite
6365341	20353n	code	code
15184543	25336n	date	date day_of_the_month
701707	101215a	derived	derived
6737512	26501n	description	description verbal_description
6289979	29345n	email	electronic_mail e-mail email
10127072	9631n	founder	founder beginner founding_father father
5014082	37634n	gender	sex gender sexuality
13832827	54276n	hasmember	member
10235776	47367n	investor	investor
8613087	50179n	latitude	latitude
1179611	90411v	livesin	live_in sleep_in
8614224	51951n	longitude	longitude
13832827	54276n	memberof	member
6344646	56758n	name	name
13602668	451n	numericalvalue	absolute_value numerical_value
13831419	21395n	partof	part portion component_part component constituent
6437781	58286n	phonenumber	phone_number telephone_number number
5644732	725n	skills	skill science
584498	73205n	specializationname	specialization specialisation specialty speciality specialism
6543318	74360n	stocksymbol	stock_symbol
6348274	20465n	surname	surname family_name cognomen last_name
6357363	77409n	title	title statute_title rubric
13604927	79106n	unitofmeasure	unit_of_measurement unit
6367112	63770n	zipcode	zip_code zip postcode postal_code

Table 2: Note that leading zeros of IDS have been removed for clarity.

For concepts that did not have a direct mapping to pre-existing synsets of WordNet, we tried to either find an already established concept to extend, or to have something close enough to serve as a direction for the way it should be interpreted.
As part of the ontology meta data, for such concepts you can find a description in natural language in the provided owl file.

Wordnet ID	Babelnet ID	Used in Model	Parent Synset Members
6301417	3446421n	art_unit	form word_form signifier descriptor
3754377	54416n	derived_product	merchandise ware product
6365341	20353n	isic	code
5993172	26980n	institutonal_education	education
6365341	20353n	naics	code
1740	31027n	named_entity	entity
6365341	20353n	sic	code
8082070	46576n	sub_industry	industry
6365341	20353n	applicationnumber	code
6365341	20353n	artunitcode	code
6737512	26501n	buildingdescription	description verbal_description
6365341	20353n	buildingnumber	code
6365341	20353n	businesscode	code
5898856	36197n	businessmodel	model theoretical_account framework
6365341	20353n	countrycode	code
15184543	25336n	dissolutiondate	date day_of_the_month
6344646	56758n	educationname	name
15184543	25336n	enddate	date day_of_the_month
15184543	25336n	filingdate	date day_of_the_month
6344646	56758n	financialconcept	name
15184543	25336n	foundationdate	date day_of_the_month
15184543	25336n	grantdate	date day_of_the_month
6301417	26512n	hasartunit	form word_form signifier descriptor
6646883	46698n	hascontact	information info
6301417	46418n	hasfinancialdescriptor	form word_form signifier descriptor
15184543	25336n	lastmilestonedate	date day_of_the_month
13300285	29416n	medianworkercompensation	wage pay earnings remuneration salary
5128718	58285n	numberofemployees	number figure
5128718	58285n	numberofmembers	number figure
6344646	56758n	pastname	name
6344646	56758n	personname	name
15184543	25336n	prioritydate	date day_of_the_month
6365341	20353n	publicationcode	code
15184543	25336n	publicationdate	date day_of_the_month
13276044	41873n	revenuesource	income
15184543	25336n	startdate	date day_of_the_month
2600976	95813v	workingwithdatatype	work

Table 3: Table of parent Synsets for words that did not have any clear match and for which we could find a clear concept to extend. Unsurprisingly, most of these instances are properties.

Wordnet ID	Babelnet ID	Used in Model	Nearest Synset Members
13377127	16459n	financial_descriptor	funds finances monetary_resource cash_in_hand pecuniary_resource
13394134	55644n	money_amount	medium_of_exchange monetary_system
2215637	82276v	acquired	get acquire
13406050	758008n	currencycode	money
2414542	87576v	employedby	hire engage employ
2537291	87731v	examinedby	test prove try try_out examine essay
2431950	85671v	founded	establish set_up found launch
5766056	32306n	hadexperiencein	experience
2133811	32645n	hasfacilityin	located placed set situated
2133811	43332n	hasheadquartersin	located placed set situated
1635953	85743v	inventedby	invent contrive devise excogitate formulate forge
2375741	82287v	playedby	play
2332196	84692v	providedby	supply provide render furnish
2214901	89246v	receives	receive have
343988	92553v	recurs	recur repeat
6370154	13633911n	socialprofileurl	url uniform_resource_locator universal_resource_locator

Table 4: Table of nearest Synsets for words that did not have any clear match and for which we preferred to indicate a close concept instead of a parent. Again, most of these instances are properties.

4 Top-level Grounding

This section provides a description on the top-level mapping of our ontology, the methodology explained during classes [2] has been followed; we tried to map domain roots to top-level leaves while taking into consideration the re-usability of the final outcome and, at the same time, how it fits our scenario.
The information provided by the meta data in CSK, the top level ontology, along with semantic meanings provided by WordNet, have been used as guidelines to connect our design with the upper one.
Given the complexity of this task, and this being the first phase(s) of what is an iterative approach, some issues were inevitable. We tried to deal with them in the best way we could, we highlight some of them:

- The most important mismatch was given by the fact that we modeled strings as data properties, while the upper ontology made use of classes to represent strings such as natural language strings etc., thus having a set of object properties that we had no use of.
- Given the existence of **Action Kind**, **Context Kind** and **Roles**, we found ourselves second guessing our own design. Without being too wordy, an example would be the case of the **Foundation** event: founding is an action, a foundation is indeed an event, and a founder could be considered a role. For such cases, we assumed that "following" the scenarios and the data was the right thing to do.
- We found the distinction between **Concept** and **SString** confusing, probably because of the lack of meta data/documentation.
- It was not completely clear if **QQString** could be used as a generalization of sets and values (i.e. colors represented as the set red, yellow, green, blue, etc.), given that it has a **unitOfMeasure** and that WordNet defines it as "any division of quantity accepted as a standard of measurement or exchange".
- We had doubts about connecting our **Category** entity, given that we wanted to have it as an entity for easier expandability (and not as as a simple string). This is because we believe that if we were to further expand onto **Category** and its current specializations (**Sector**, **Industry**) by adding characteristic specific to particular instances of categories, we would not consider such instances as **Complex data types.**
- Each **Action Kind** had an **Action Property** (complex data type), each **Function Kind** had a **Function Property**, and so on. The reason for this was not clear. While we may guess the reason for such design, more meta data/documentation could have helped.
- Moreover, we had to actually map **endDate** and **startDate** as a specialization of **date**, a data property which we have defined; this was counter intuitive given the direction of the mapping (CSK to ours).
- We expected the object properties **hasPart**, **partOf** to be transitive, but this choice could have been made as general as possible by the authors of CSK.
- in the **Service** class having comment "Anything that is generated by performing a providing action", we expected an object property "provided by" or a provide action.
- the comment of **Destroyer** states "Who perform the act of making", which is obviously wrong. We assume this as not intended and we write about it for the sake of reporting.
- **endDate** and **startDate** were defined as **floats** in the constraints of **Role**, we believe it was a simple mistake and not intended. Again, we write about it for the sake of reporting.

As stated before, we expected some issues during the top level grounding phase, apart from the **QQString** and **Category** problems, we must say that the top level ontology provided a framework which allowed us to connect our design more or less easily.
The mapping is depicted in the following diagram. We did not include properties or relations (apart from is-a) to be concise and clear about the meeting point of our model and the provided one. For the same reason we only considered top-level leaves.

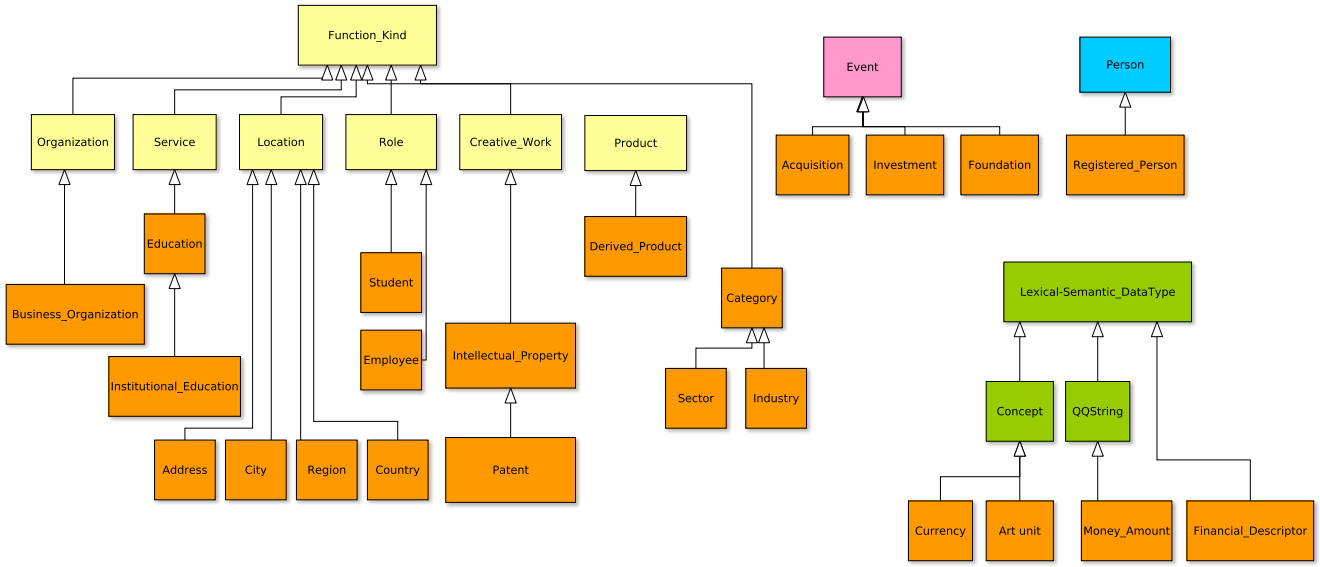


Figure 2: Diagram representing how the top level leaves are mapped to domain roots. Domain nodes are in orange, while yellow, pink, green and blue are used for distinguishing top level concepts belonging to different sub-trees.

* * *

The full and grounded model is provided within the GitHub repository [1], along with the rest of the material.

5 Model Visualization

This section is dedicated to the visualization of our model, we provide 2 different points of view: the OWLViz view on entities, and the WebVOWL view that also allows to appreciate relations and properties.

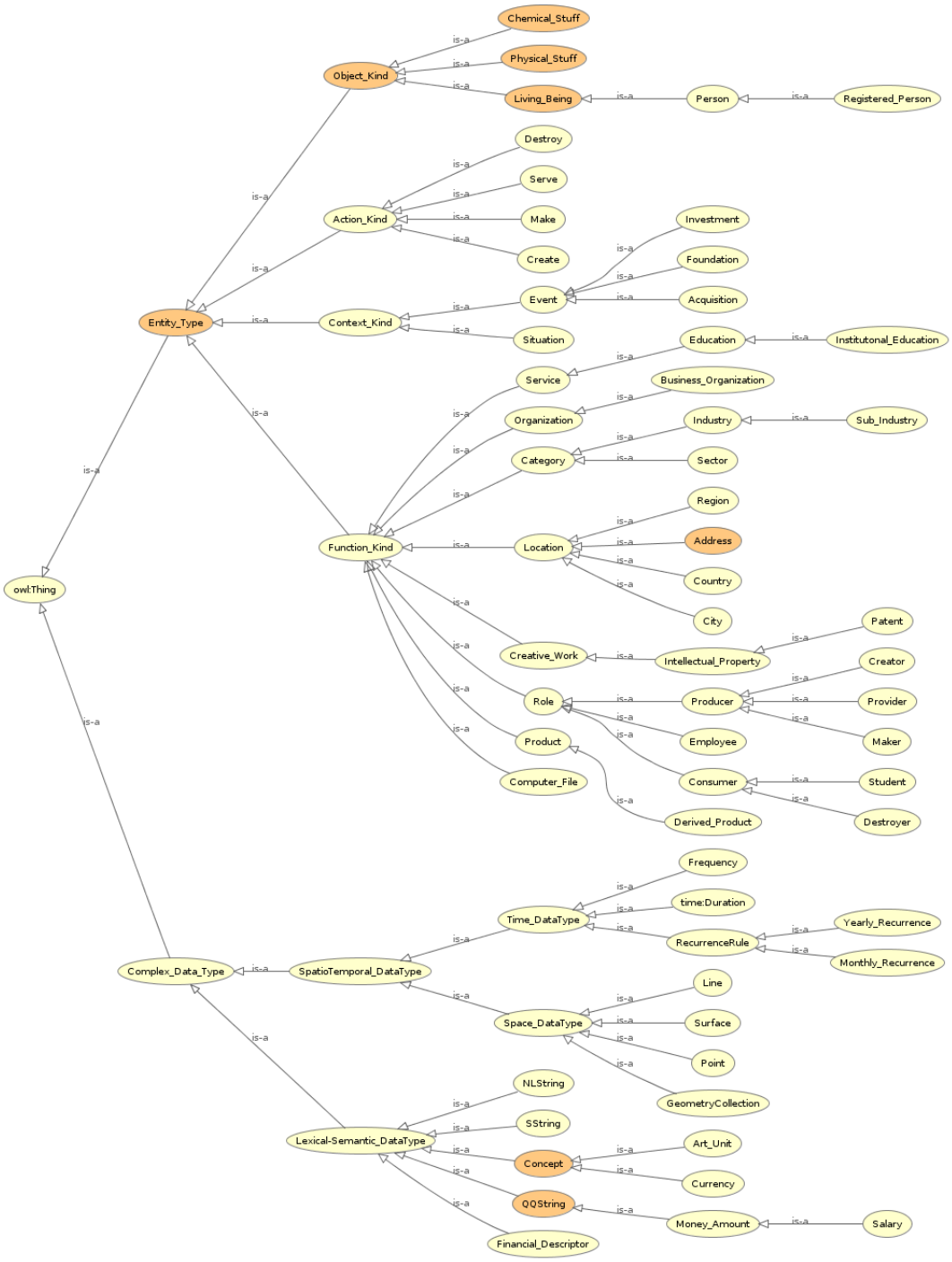


Figure 3: OWLViz diagram showing the class hierarchy of the final output ontology, some classes were kept even if they were not actually used by us because they were part of CSK.

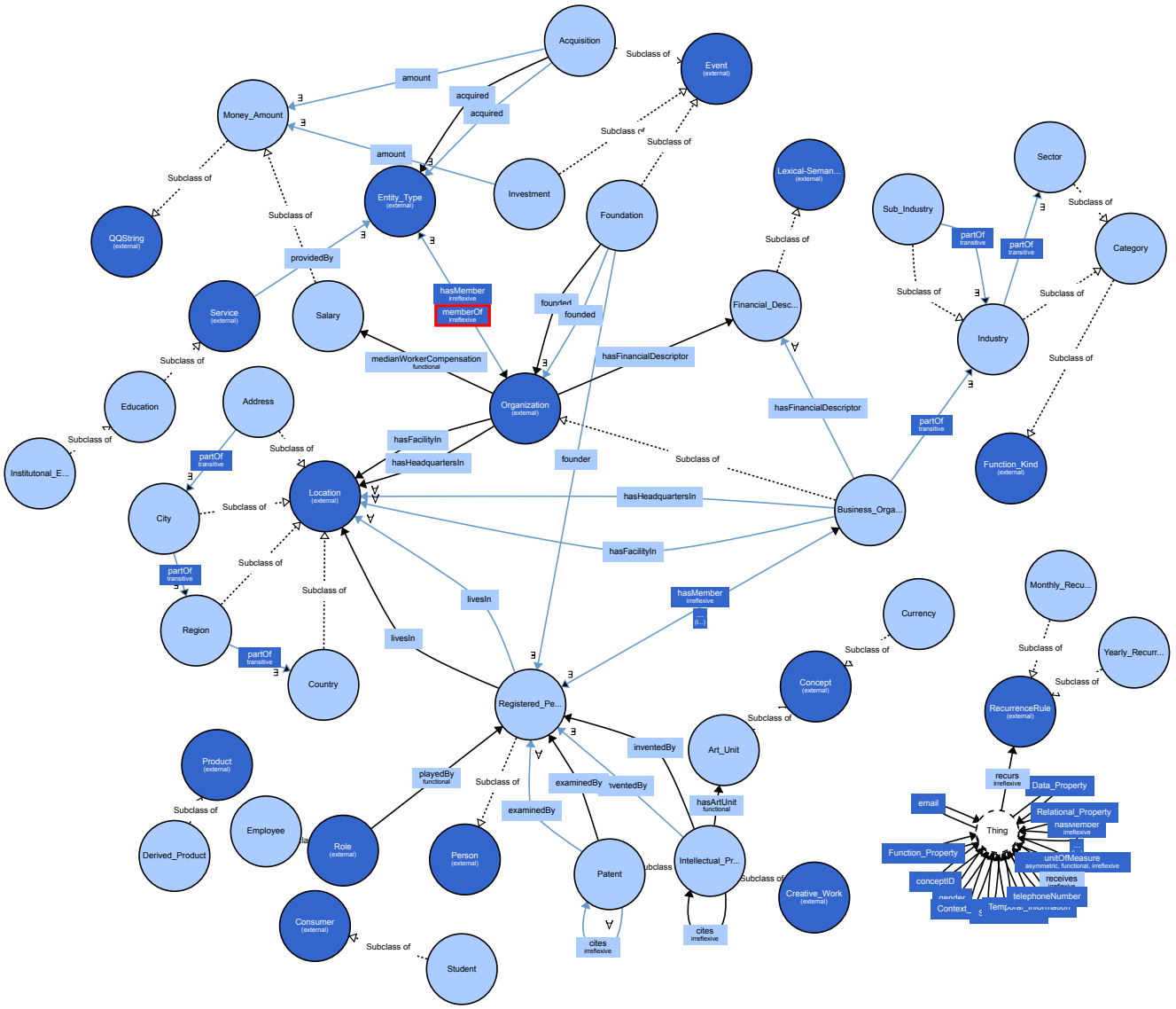


Figure 4: WebVOWL [3] visualization of the model (grounded with the top level), note that we did not include the full details (data properties and their ranges, etc.) to allow an easier visualization. For example, you may notice that isA relations between nodes of the top level (Dark blue) are missing.

6 Generalized Queries

We provide a list of query examples and expected results. We preferred to keep these queries (and their results) as near as possible to natural language to provide a better presentation in the report.

Person	Generalized query	Expected result
Paolo Rossi, 26, looking for a job in IT	Give me all IT companies with offices in Trentino	A list of company names
	Give me the number of workers and medium salary of company X	A row containing two requested numbers
	Give me the locations of all facilities of company X	A list of locations
	Find the city containing the majority of the offices of company X	A name of city, its region and country
Marco Caldi, 18, student	Find companies with headquarters in Silicon Valley	A list of company names
	Find the foundation date, category and stock symbol of company X	Foundation date, category name and stock symbol strings
	Find founders of company X	A list of founders names
	Find geographical coordinates of companies founded by X	A list of geographical coordinates for each company with company name
	Find address of the headquarter of company X	String containing country, region, city and address of the headquarter
Salvo Montalbano, 40, financial policeman	Find all companies where Mr. X was employed in	A list of company names
	Find an industry company X is part of	An industry name
	Find all contacts of Mr. X	A list of contacts
	Find all qualifications of Mr. X	A list of the qualifications of Mr.X
	Find which roles Mr. X had in companies in years 2016 - 2018	A list of roles
Talia Trenit, 31, investigator	Find all granted patents examined by Rossi Sara	A list of patent titles
	Group patents for assignee company	A list of lists of patents and their assignee company names
	Find patents cited by patents owned by company X.	A list of patent titles
	Find patents containing in their description/title/abstract words "pharmaceutical" and "Alzheimer"	A list of patents with their titles, abstracts and descriptions
	Find the inventor of patent having publication number X	A name of inventor
	Find art units of patents granted in last 5 years	A list of art unit codes
Alto Garda Assicurazioni	Find emails of people employed by Itas Assicurazioni	A list of names with contacts
	Find where a person X lives	A complete address or any information related to the person's living place
	Find education of person X	Name of degree and university
Anonymous hacker	Find names, birth dates and all contacts of people employed by Valve Corporation	A list with names, birth dates,telephone numbers, emails and social profiles
	Find past names of Valve corporation	A list with past names
	Find names of all CEOs (current and past of Valve corporation	A list of CEOs' names

Person	Generalized query	Expected result
Billy J. Wood, 45, hedge fund manager	Find all acquisitions of company X in last year	A list of company names
	Find categories of companies acquired by Google	A list of industries
	Find all investments of company X	A list of investment amounts, investors and investment dates
	Find derived products of Google	A list with names of derived products and names of correspondent acquired companies
	Find for how much company X was sold to company Y	A money amount
Thomas Brewest, 51, columnist of the Financial Times	Find business model and revenue source of companies in IT sector	List of business models and revenue sources
	Find business model and revenue source of companies having "subscription" and "trial" in their description	List of business models and revenue sources
	Find cash flow and sales revenue of company X	Cash flow and sales revenue amounts
	Find assets and sales to calculate the book value of the company and compare it to market value of company X	Assets, sales and market value amounts
	Find the profit and profit change of a company X	Profit and profit change amounts
	Find stock symbol of company X	Stock symbol string
Bobby Ryan, 43, runs a pharmaceuti-	List companies sorted by number of patents, containing "pharmaceutical" in their description	A list with company names and numbers of correspondent patents
	List companies sorted by number of employed PhDs and with knowledge in pharmaceutical field	A list of company names and numbers of searched PhDs
Mario Martini, 36, entrepreneur	Find average profit and number of employees for each sub-industry of industry X	A list of sub-industries with average number of employees and average profit
	Sort industries and sub-industries by the highest profit/employee ratio	A list of industries with the average profit/employee ratio and list of sub-industries with the average profit/employee ratio
Billy J. Wood, 48, hedge fund	Find people having a university degree in Computer Science, Math, Physics or Finance in the past 5 years	A list of people names with degree names and graduation dates
	Check if person X has "problem solving" skills	A boolean value
	Find past employment roles of person X	A list with role names
Valeria Marini, 28, secretary	Find the company which employed Mr. X	Company name
	Find all locations where Mr. X is considered to be living in.	A list of countries, regions, cities and addresses

Table 5: Note that colors refer to macro areas as for the scenario section: red for the location area, yellow for people, light purple for companies area and brown for mixed.

7 Final considerations and open issues

The provided methodology embeds the data integration procedure in a controllable framework. Each task is somewhat insulated from the others (hence easier to manage) but still allows to propagate changes through the whole pipeline in a manageable manner.
CSK was reasonably easy to use, and the encountered issues might be more related to our inexperience than to CSK itself, with this said, we think that the top level grounding has been successful.
Most of our problems were related to developing our own ontology, given the ambiguity of natural language and many different ways in which we could describe something, it is not always obvious how to model a particular context.
Should someone investing in a company be considered as having an investor role (thus concretely having such class), or should we consider an investment event that "connects" people and businesses through object properties? We tried to "steer" based on the data at our disposal (and its sparsity) and scenarios. However, trying to keep an ontology simple, easy to expand, but also correct under all possible considerations is indeed a hard task that is going to require more iterations over this methodology and surely more experience.
Another relevant issue is the sparseness of our data, because we were often unsure on how to model constraints that would make sense (and actually provide some constraint) while taking into consideration sparsity, that could make many of such constraints break.

* * *

The current output is surely raw, our hope is to provide a good and documented input that someone could iterate over during successive phases of this course.
Everything mentioned in this report can be found in the directory accompanying it, more specifically it should contain: data (in the form of links and explanations), all scenarios, the ER schema (both the source and the html provided for easier visualization) and the ontology.
As a final note, providing the evaluation guidelines earlier, like during the period of the development of the ontology, would have contributed to its correctness.

7.1 Submitted Material

This report has been submitted as part of a directory containing all the output material. At the date of submission (16 December 2018) both the ontology before mapping and the final output mapped to CSK run through reasoners with no issues.
Data sets are provided as a - rough - document containing links, given that their size would exceed what should be reasonably uploaded in a git repository. Moreover, a more complete, sometimes rough, list of scenarios is provided.
The ER model is provided both as a source code and as a html page for visualization. The ontologies are provided as owl files.

8 Evaluation

	Class Coverage	Class flexibility	Attribute Coverage	Attribute Flexibility
CQ-Model	$\simeq$ 0.8 (Ideal) 1.00	$\simeq$ 0.2 (Ideal) 0.08	$\simeq$ 0.8 (Ideal) 1.00	$\simeq$ 0.5 (Ideal) 0.23
DS-Model	$\simeq$ 0.8 (Ideal) 0.70	$\simeq$ 0.2 (Ideal) 0.00	$\simeq$ 0.8 (Ideal) 0.52	$\simeq$ 0.5 (Ideal) 0.00
CQ-DE	$\simeq$ 0.8 (Ideal) 1.00	$\simeq$ 0.2 (Ideal) 0.37	$\simeq$ 0.8 (Ideal) 1.00	$\simeq$ 0.5 (Ideal) 0.57

Table 6: Table defining ideal and resulting values for different metrics, as suggested by the methodology.

Schema Level

Does the model including cycles in the class hierarchy? **No**

Does the model uses any polysemous terms for its class or property name? **No**

Is Multiple Domain / Range defined for any property? **Yes**

Does any class have more than one direct parent class? **No**

Does the Model include multiple classes which have same meaning? **No**

Is the class Hierarchy over specified? **No**

Does the model use isA as a object Property or relation? **Relation**

Does the model have any leaf class for which there is no relation with the rest of the model? **Yes**

Did you use miscellaneous or others as one of the class name? **No**

Does the model have any chain of Inheritance in class hierarchy? **Partially**

Do all properties have explicit domain and range declarations? **No**

Does the model have any classes or properties which are not used? **Yes**

Are a collection of elements included as a group in a number of class/attribute ? **No**

Linguistic Level

Does all elements of the model (i.e. class and property) have human readable annotations? **Yes**

Do all elements of the model follow the same naming convention? **Yes**

Metadata Level

Is provenance information (Creator, Version, Date) available for the final protege model? **Yes**

Is provenance information available for any property or class which is taken from some reference standard or ontology? **Yes**

References

[1] *Github repository*, `https://github.com/UniTN-KDILab/Business-2018-19`.

[2] *Kdi course materials*, `http://knowdive.disi.unitn.it/wordpress/kdi/`, Accessed: 2018-12-16.

[3] Steffen Lohmann, Vincent Link, Eduard Marbach, and Stefan Negru, *WebVOWL: Web-based visualization of ontologies*, Proceedings of EKAW 2014 Satellite Events, LNAI, vol. 8982, Springer, 2015, pp. 154–158.

[4] George A. Miller, *Wordnet: A lexical database for english*, Commun. ACM **38** (1995), no. 11, 39–41.

[5] Roberto Navigli and Simone Paolo Ponzetto, *Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network*, Artif. Intell. **193** (2012), 217–250.

[6] Peter Pin shan Chen, *The entity-relationship model: Toward a unified view of data*, ACM Transactions on Database Systems **1** (1976), 9–36.