Data Integration Market Analysis in the Artificial Intelligence Age

Ahmad Alobaid
Prof. Oscar Corcho

Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

aalobaid@fi.upm.es
[@twitter user]

Date
Place

Papers and Resources

Alobaid, Ahmad, and Oscar Corcho. "Fuzzy semantic labeling of semi-structured numerical datasets." European Knowledge Acquisition Workshop. Springer, Cham, 2018.

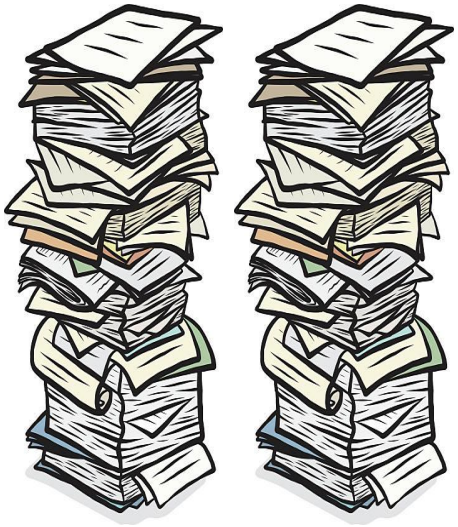
Alobaid, Ahmad, Emilia Kacprzak, and Oscar Corcho. "Typology-based Semantic Labeling of Numeric Tabular Data." Semantic Web (2020).

Table of content

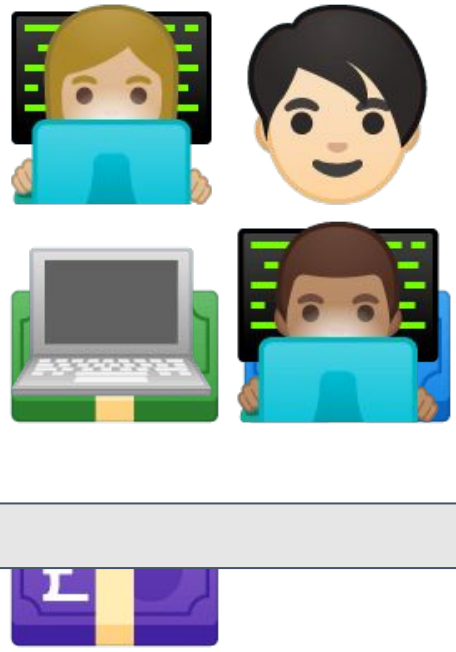
Infobesity?

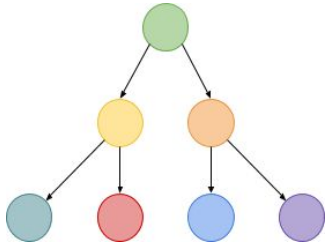
https://youtu.be/o79ngpmfDlc

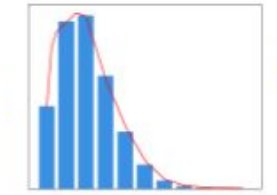
Infobesity



A lot of data







Integration
(and Exploitation)

Generation Speed

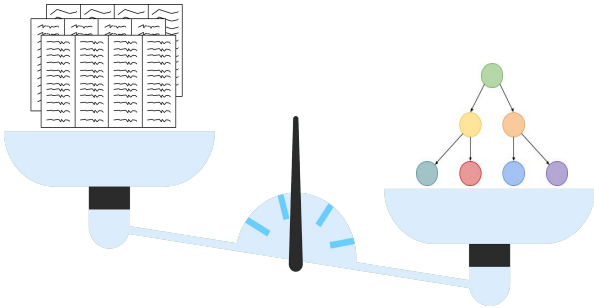
Integration Speed

A lot of data being published today that some referred to it as the age of infobesity. A lot of organisations have **different sources** of data in **different formats**, and sometimes they even differ in their modelling schemes. This process is often carried out by engineers. It can be cumbersome to integrate the data from different sources manually. It is common to have domain experts -who understand the data about a specific topic- are needed to convey the meaning to the engineer. The engineers would then integrate the data by transforming it from one model or schema to another. This can be **time-consuming, error-prone and not easy to maintain**. This also costs organisations **money**, especially the ones with many datasets, which is often the case for medium and large organisations. On the other hand, **Artificial Intelligence** can help **automate** part of this and save money and offer more value than classical data integration. This can saves organisations time and money. However, there is a lack in the literature about the market studies about data integration and the utilisation of Artificial Intelligence. In this work, we perform market analysis studying more than fifty data integration products and services and their relation with Artificial Intelligence. We also show how organisations with a lot of data can **make money** using data integration tools powered by Artificial Intelligence.

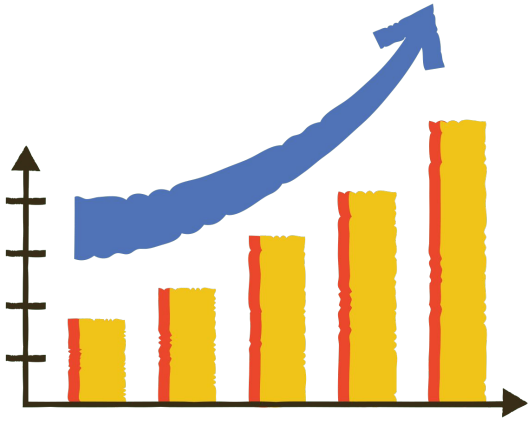
Hypotheses

H1: There is a lack of studies that analyse semantic data integration products and services.

H2: There is a lack of semantics in the majority of data integration products and services.

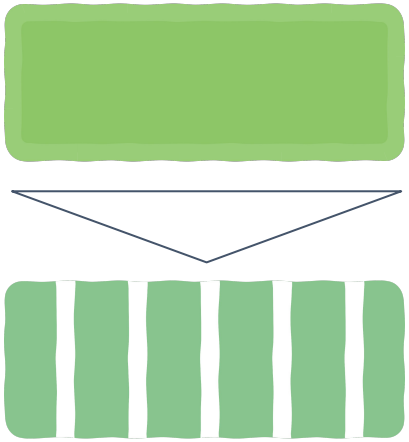


H3: Companies would make a profit by employing semantic semi-automatic data integration as long as the cost of utilising the service is less than the complement of the reciprocal of the speed gain of utilising it.



We do not take into account indirect costs (e.g., office space, energy consumption).

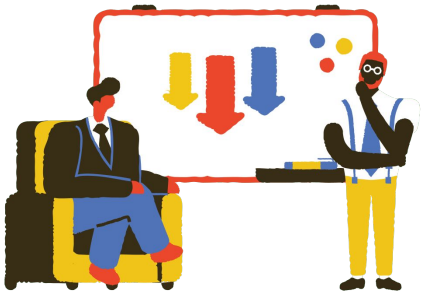
External factors (e.g., natural disasters) are not taken into account.



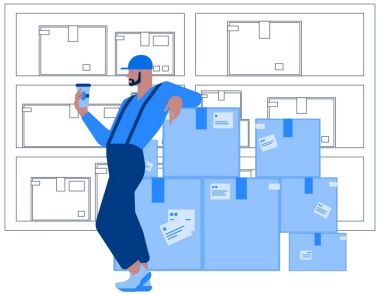
We assume that the data we collected about the data integration products and services are a representative sample.

Assumptions and Limitations

We do not take into account effects related to decisions taken based on insights mined from the data.



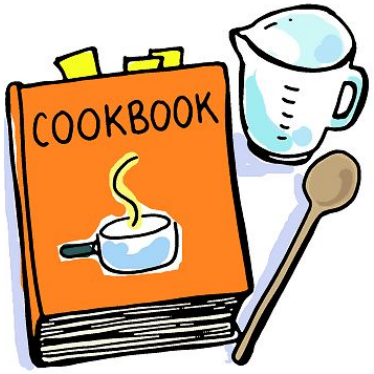
We assume the data to be of a reasonable quality such that the time and effort needed to clean it does not exceed the speed gained from employing semantic semi-automatic data integration.



We assume that there is no reduction in the value proposition in total as the amount of data integrated increase or that the data integration process is performed more quickly.

Goal: evaluate H1

Systematic
Literature Review



Protocol

Keywords:
"Data"
"Integration"
"Market"

Kraus, S., M. Breier, and S. Das´ı-Rodr´ıguez 2020. The art of crafting a systematic literature review in entrepreneurship research. International Entrepreneurship and Management Journal, Pp. 1–20.

Publication Year:

2015-2020

Databases:

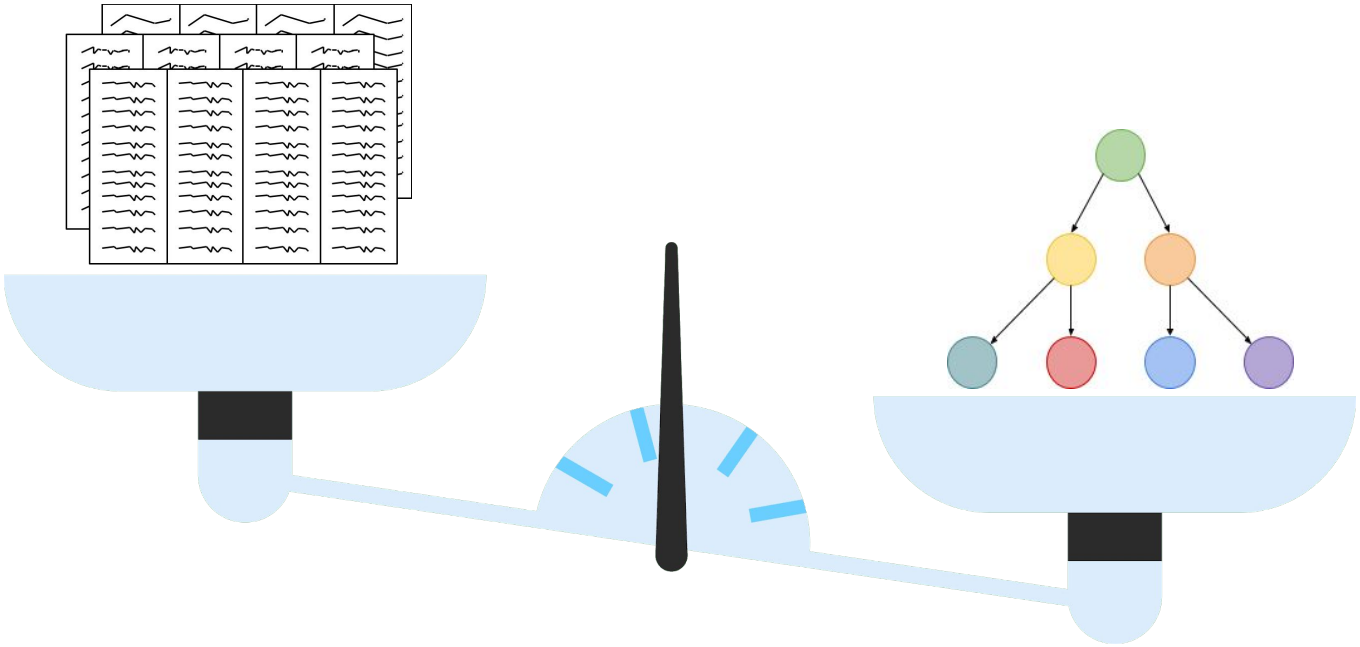
- ABI Inform/ProQuest
- EBSCO/ Business Source Premier
- JSTOR
- MENDELEY
- ScienceDirect
- Scopus
- SpringerLink
- Web of Science

Database	Query	Results	Title	Abstract	Relevant
JSTOR	Data + Integration + Market	3	0	0	0
	Data + Integration + Product	0	0	0	0
	Data + Integration + Service	0	0	0	0
MENDELEY	Data + Integration + Market	7383	-	-	-
	Data + Integration + Market + Product	888	-	-	-
	Data + Integration + Market + Product + Service	165	-	-	-
	Data + Integration + Market + Product + Service + > 2015	75	0	0	0
ScienceDirect	Data + Integration + Market	7	0	0	0
	Data + Integration + Product	21	3	0	0
	Data + Integration + Service	29	5	0	0
Scopus	Data + Integration + Market	20	4	0	0
	Data + Integration + Product	46	3	0	0
	Data + Integration + Service	92	4	0	0
SpringerLink	Data + Integration + Market	0	0	0	0
	Data + Integration + Product	0	0	0	0
	Data + Integration + Service	1	0	0	0
Web of Science	Data + Integration + Market	5	0	0	0
	Data + Integration + Product	109	1	0	0
	Data + Integration + Service	196	8	0	0

No relevant papers found

H1 ✓

Hypothesis H2

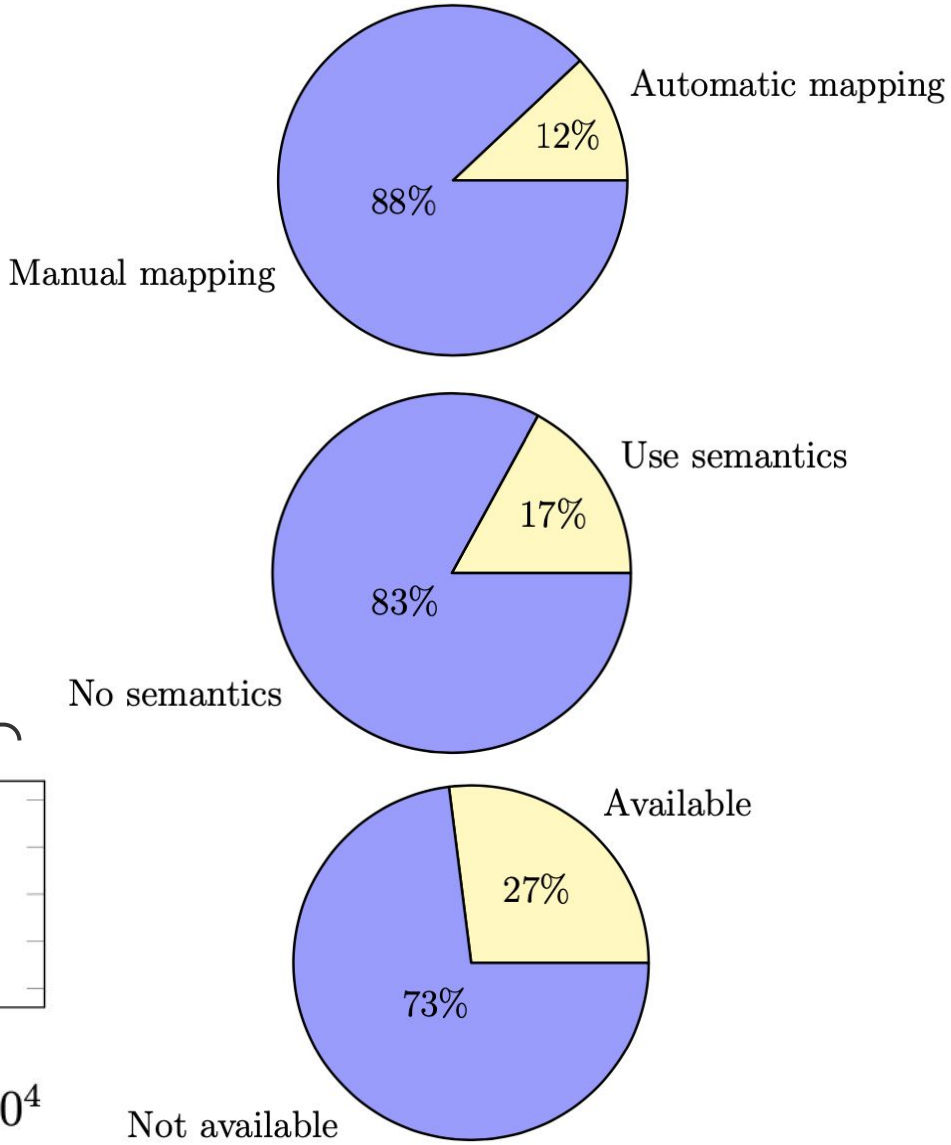


NLP-based

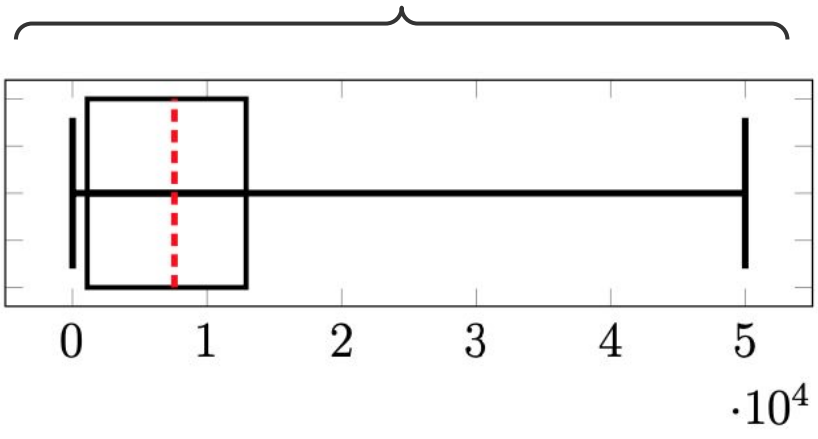
Base	Search Engine	Details
	BOL	Only availa the Brazilian language
	Groovle	No long able
Google	Space Search	Searc y within myspace website
	Seeker	N r available
	Net	rects to AOL
	Ripple	onger hosted
	Startpage	o API is found
	A9	Redirects to amazon
	AOL	API
	Alexa t	rch API
	Ci	on shopping
Bing		No API
		Focus Kurdish language
	tbot	Pivoted ide VPN services
	Msdewey	No longer e
	WebCrawler	No search A
	Ecocho	No longer avai
oo	Forestle	No longer availa
	rectifi	No longer a search engine

Manual data gathering results

H2 is true ✓



Annual price in USD



$$N = V - C$$

Net value gain

Value gain

Cost

Data

$$N(D) = \sum_{i=1}^{||D||-1} V(d_i, d_{i+1}) - \sum_{i=1}^{||D||-1} C(d_i, d_{i+1})$$

$$N(D,a) = \sum_{i=1}^{||D||-1} V(a, d_i, d_{i+1}) - \sum_{i=1}^{||D||-1} C(a, d_i, d_{i+1})$$

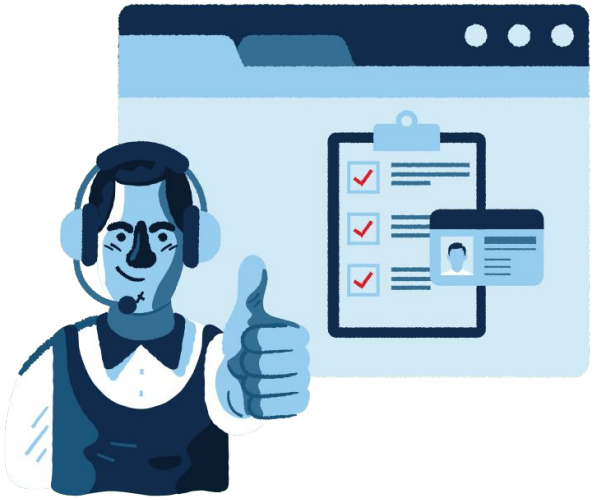
Net value gain

Value gain

Cost

$$N(D,a) = \sum_{i=1}^{||D||} V(a, d_i, M) - \sum_{i=1}^{||D||} C(a, d_i, M)$$

Virtual Assistant



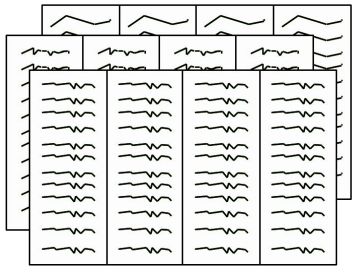
specialized in products

Q:“What is the fastest delivery option for me get one litre of milk”

Example

2800 USD

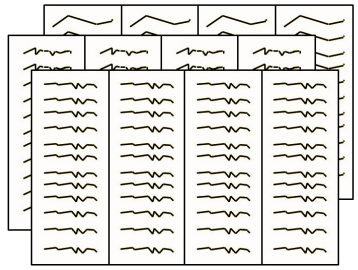
35% revenue



Once

5200 USD

65% revenue



8-12 datasets / month

Integrated data

8K USD

80K Questions

Virtual Assistant

100 USD / 1000 Questions

Revenue

8K per month

5.2K per month

Manual data integration

Duration: 1 month
||D|| = 10 (12+8 /2)

Knowledge Engineer salary*

130K USD per year ~ **10.8K** per month

$$N(D,a) = \sum_{i=1}^{||D||} V(a,d_i,M) - \sum_{i=1}^{||D||} C(a,d_i,M)$$

-5.6K = 5.2K - 10.8K => Losing money

* according to indeed.com

Company	Annual Price (in USD)
Altova(MapForce Server)	800*
Blendo	1800
data.world	50000
DIYOTTA(Data Studio)	12000
eltrobot	864
ETLworks	3600
FIVETRAN	12
Google(Cloud Data Fusion)	13200
KEBOOLA	30000
matillion	15000
Microsoft(Azure Data Factory)	5152.8
tableau	840
TALEND(TALEND Data Fabric)	12000
WORKATO	10000

* This is a one time purchase, not annual

Monthly mean:
630 USD

Revenue

8K per month

5.2K per month

Semi-automatic data integration

Duration: **3 days**
||D|| = 10

Knowledge Engineer salary*

110.8K monthly rate/22 working days * 3 days = **1.5K** USD

tool usage cost = **630** USD

$$N(D,a) = \sum_{i=1}^{||D||} V(a, d_i, M) - \sum_{i=1}^{||D||} C(a, d_i, M)$$

3.1K = 5.2K - 2.13K => Making money

* according to indeed.com

Semantic semi-automatic data integration

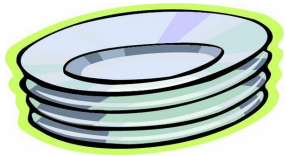
D

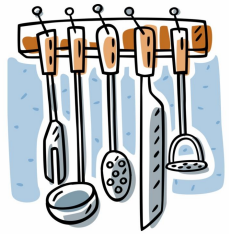
Data from known sources (e.g., databases)

X

Automatic extracted data (e.g., web pages)

Clean up and organise (match to a model)





$$||D|| * 0.2 = ||X||$$

Revenue

8K per month

5.2K * 1.2 = **6.2K** USD per month

Semantic semi-automatic data integration

Duration: ~**3.6 days**
||D|| + ||X|| = 12

Knowledge Engineer salary*

110.8K monthly rate/22 working days * 3 days = 1.5K * 1.2 = **1.8K USD**

tool usage cost = 630 * 1.2 = **756** USD

$$N(D,a) = \sum_{i=1}^{||D+X||} V(a,d_i,M) - \sum_{i=1}^{||D+X||} C(a,d_i,M)$$

3.7K = 6240 - 2556 => Making money

* according to indeed.com

Extra considerations:

1. Extra cost for hourly hire (10% +)
2. Software cost reduction (10% -)
3. Speedups due to utilisation of semantics (standards)

Data integration tool

Data

Semantic (model)

$$C(a_s, D, M) = C_e(a_s, D, M) + C_t(a_s, D, M)$$

Total cost

Engineer cost

Tool cost

$$C(a_m, D, \emptyset) = C_e(a_m, D, \emptyset) + C_t(a_m, D, \emptyset)$$

Manual data integration

Data

No semantics

H: The total cost of employing semantic semi-automatic data integration tool *as* is less than the complement of the reciprocal of the speed gain of utilising it over manual data integration *am*

$$C_t(a_s, D, M) < C_e(a_m, D, \emptyset) * \underbrace{(1 - \frac{1}{r})}$$

The **total cost** of using ***as***

Cost of **manual** data integration am

complement of the reciprocal of the **speed gain (r)**

Semantics cost < Manual cost

Assumption:

Semi-automatic data integration is (*r* times) faster than manual data integration.

We assume that using a_s the engineers finish the data integration faster than doing it manually a_m.

We assume that using a_s is r times faster than a_m, so the cost of the engineers using the semantic semi-automatic data integration tool $C_e(a_s, D, M)$ is $1/r$ of the manual data integration $C_e(a_m, D, \emptyset)$

$C_e(a_s, D, M) = 1/r * C_e(a_m, D, \emptyset)$

We also assume that the value gain from both is the same

$\sum_{i=1}^{||D||} V(a_s, d_i, M) = \sum_{i=1}^{||D||} V(a_m, d_i, \emptyset) = V$

$N(D, a_s) = V - C(a_s, D, M)$

$N(D, a_m) = V - C(a_m, D, \emptyset)$

To see when it make sense to invest in data integration tool, we assume that using the tool a_s results in higher net value gain than manually integrating the data $N(D, a_s) > N(D, a_m)$

$V - C(a_s, D, M) > V - C(a_m, D, \emptyset)$

$-C(a_s, D, M) > -C(a_m, D, \emptyset)$

$C(a_s, D, M) < C(a_m, D, \emptyset)$

$C_e(a_s, D, M) + C_t(a_s, D, M) < C_e(a_m, D, \emptyset) + C_t(a_m, D, \emptyset)$

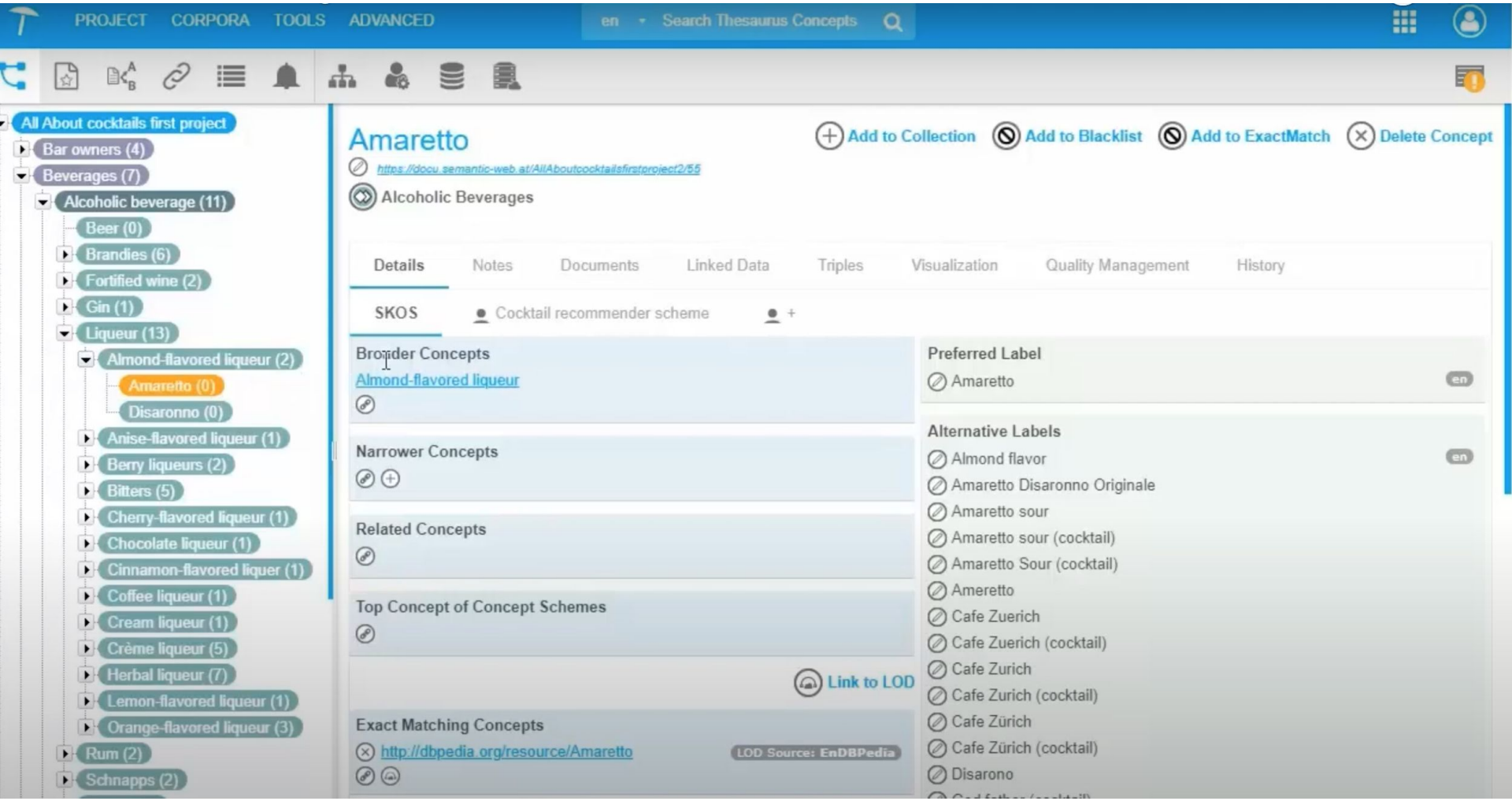
$1/r * C_e(a_m, D, \emptyset) + C_t(a_s, D, M) < C_e(a_m, D, \emptyset)$

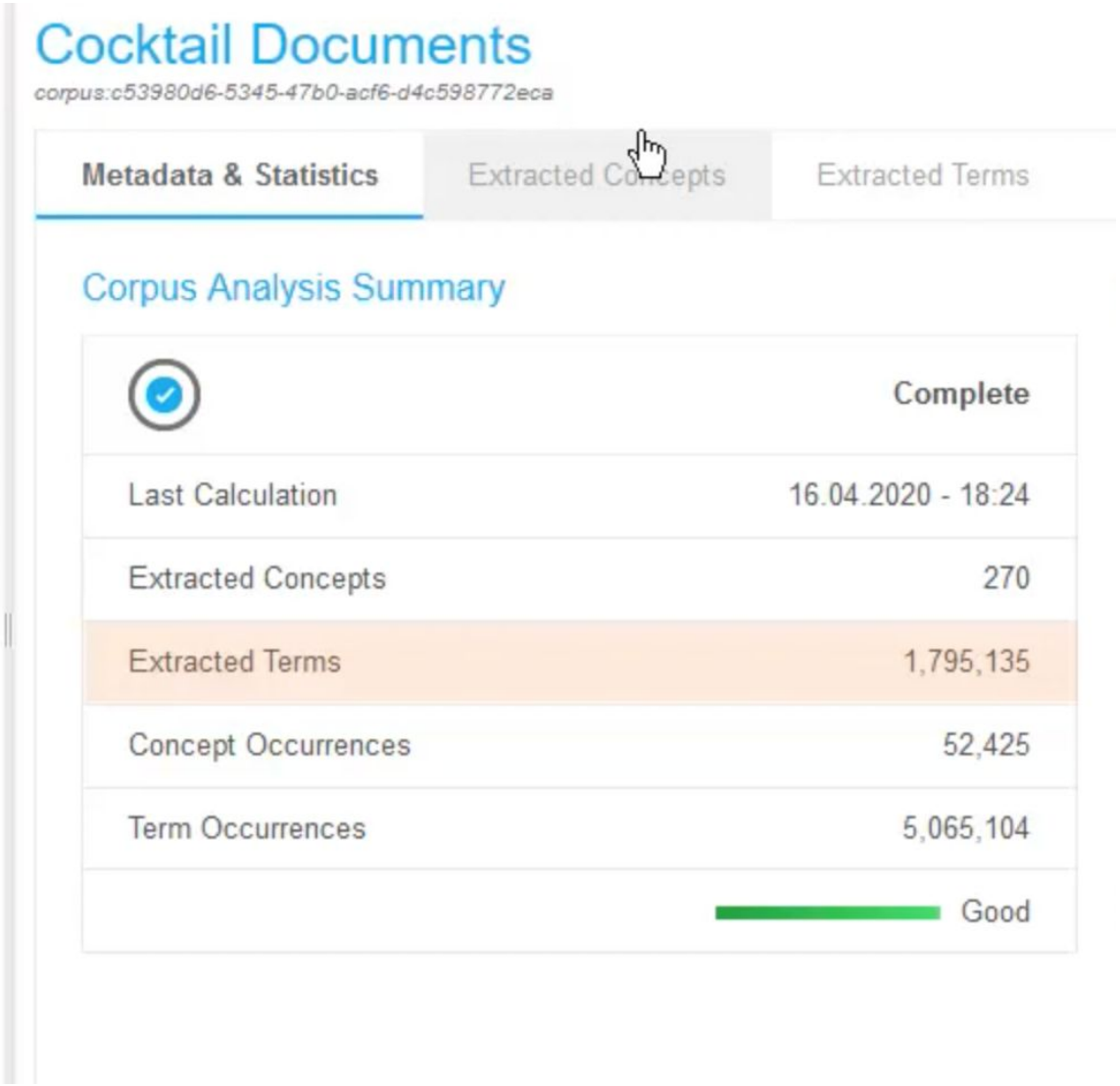
$C_t(a_s, D, M) < C_e(a_m, D, \emptyset) - 1/r * C_e(a_m, D, \emptyset)$

$C_t(a_s, D, M) < (1 - 1/r) * C_e(a_m, D, \emptyset)$

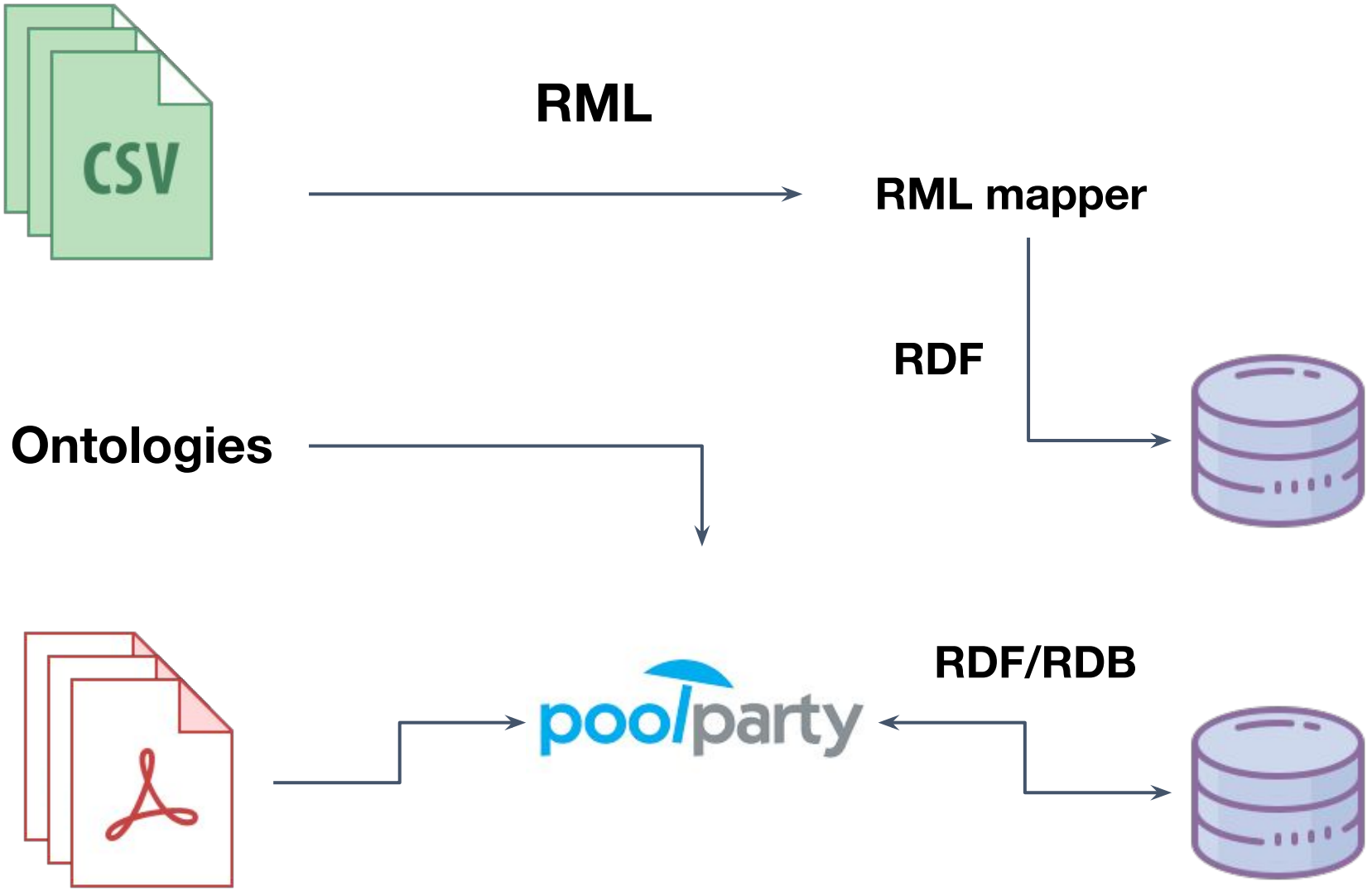
So, it makes sense to invest in data integration tool if Inequality 6 holds □

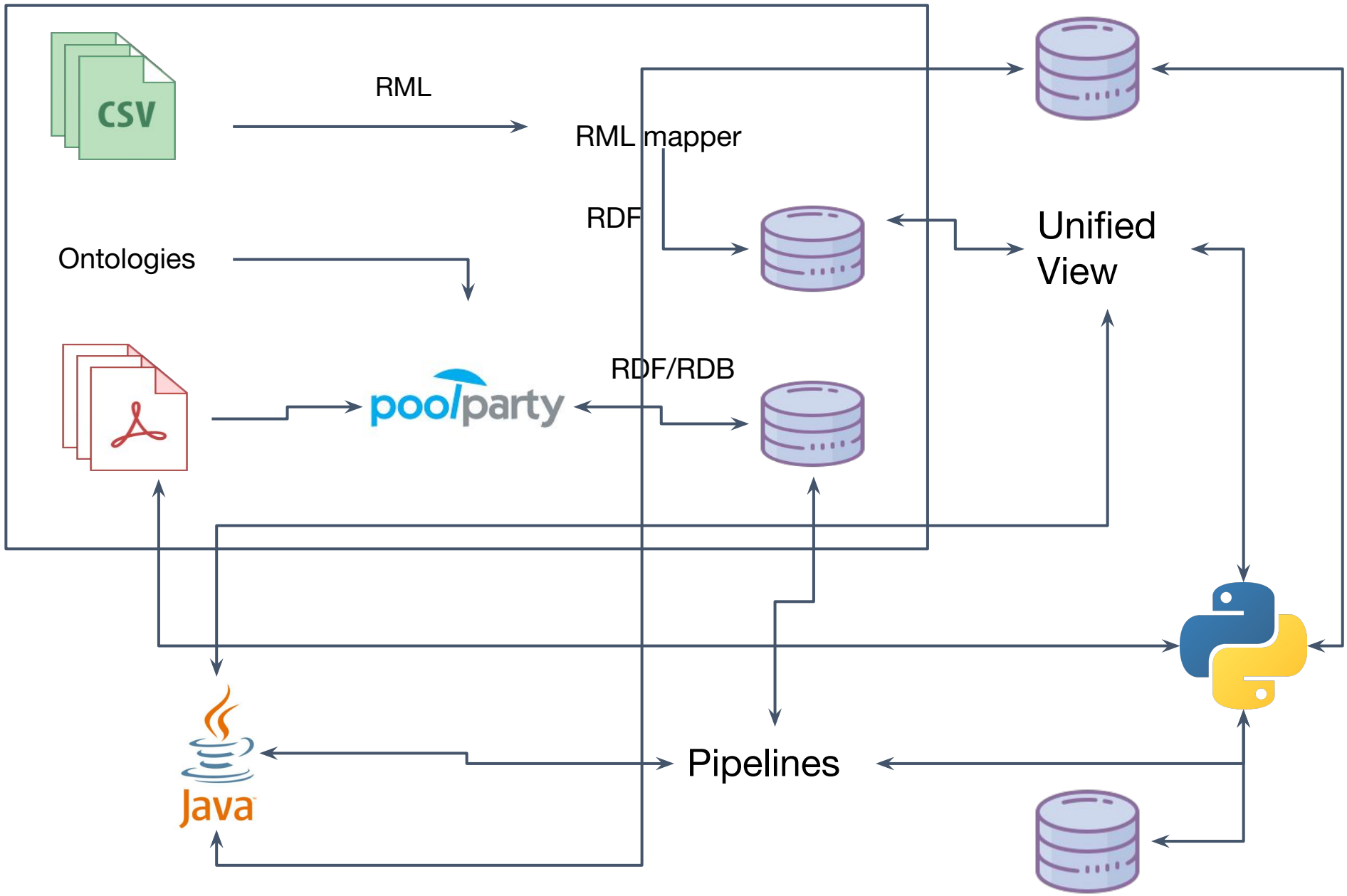
H3 is true ✓


PROJECT
CORPORA
TOOLS
ADVANCED
en
Search Thesaurus Concepts
All About cocktails first project
Bar owners (4)
Beverages (7)
Alcoholic beverage (11)
Beer (0)
Brandies (6)
Fortified wine (2)
Gin (1)
Liqueur (13)
Almond-flavored liqueur (2)
Amaretto (0)
Disaronno (0)
Anise-flavored liqueur (1)
Berry liqueurs (2)
Bitters (5)
Cherry-flavored liqueur (1)
Chocolate liqueur (1)
Cinnamon-flavored liquer (1)
Coffee liqueur (1)
Cream liqueur (1)
Crème liqueur (5)
Herbal liqueur (7)
Lemon-flavored liqueur (1)
Orange-flavored liqueur (3)
Rum (2)
Schnapps (2)
Amaretto
Add to Collection
Add to Blacklist
Add to ExactMatch
Delete Concept
https://docu.semantic-web.at/AllAboutcocktailsfirstproject2/55
Alcoholic Beverages
Details
Notes
Documents
Linked Data
Triples
Visualization
Quality Management
History
SKOS
Cocktail recommender scheme
Broader Concepts
Almond-flavored liqueur
Narrower Concepts
Related Concepts
Top Concept of Concept Schemes
Link to LOD
Exact Matching Concepts
http://dbpedia.org/resource/Amaretto
LOD Source: EnDBPedia
Preferred Label
Amaretto
en
Alternative Labels
Almond flavor
Amaretto Disaronno Originale
Amaretto sour
Amaretto sour (cocktail)
Amaretto Sour (cocktail)
Ameretto
Cafe Zuerich
Cafe Zuerich (cocktail)
Cafe Zurich
Cafe Zurich (cocktail)
Cafe Zürich
Cafe Zürich (cocktail)
Disarono

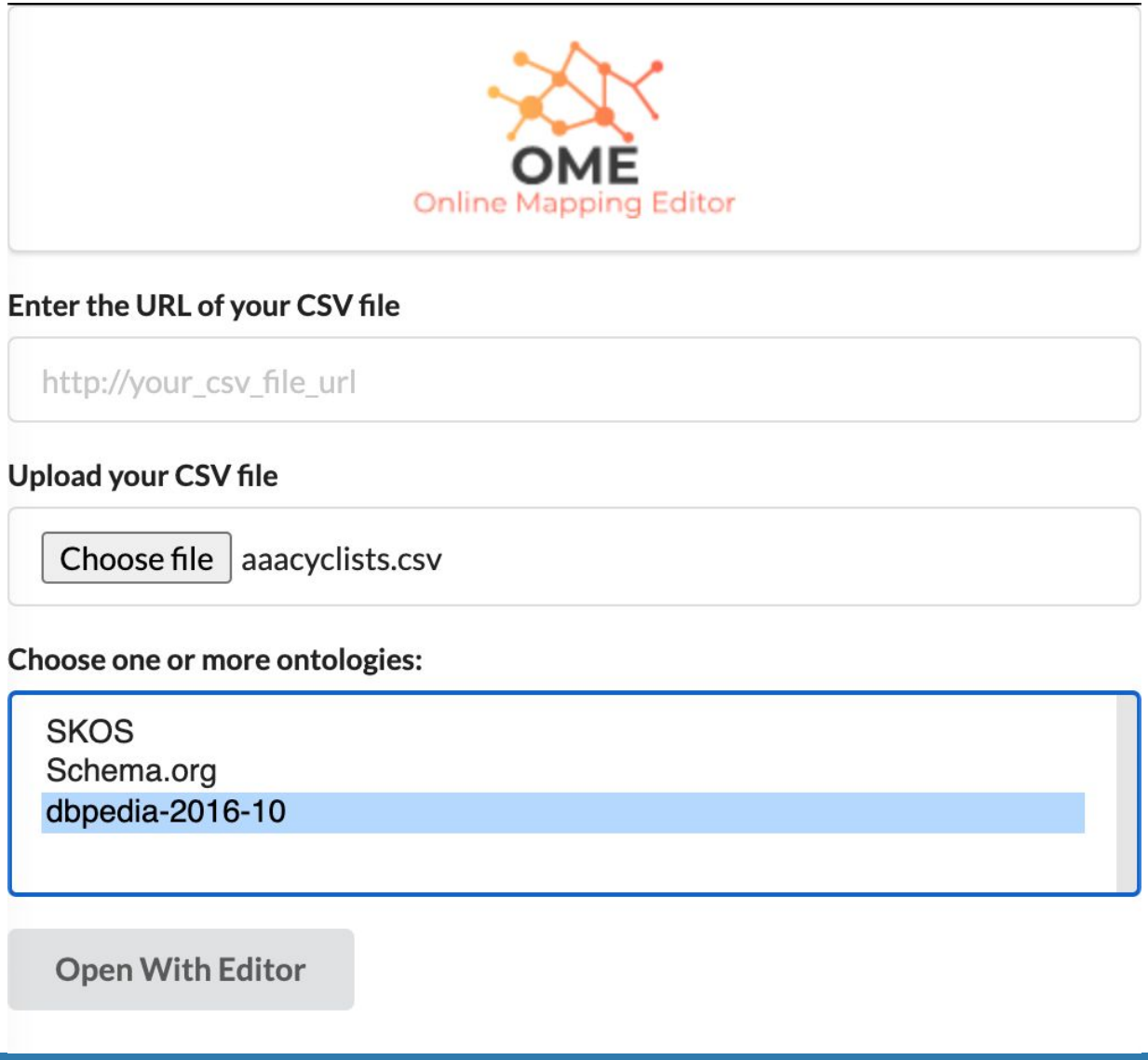

Cocktail Documents
corpus:c53980d6-5345-47b0-acf6-d4c598772eca
Metadata & Statistics
Extracted Concepts
Extracted Terms
Corpus Analysis Summary
Complete
Last Calculation
16.04.2020 - 18:24
Extracted Concepts
270
Extracted Terms
1,795,135
Concept Occurrences
52,425
Term Occurrences
5,065,104
Good

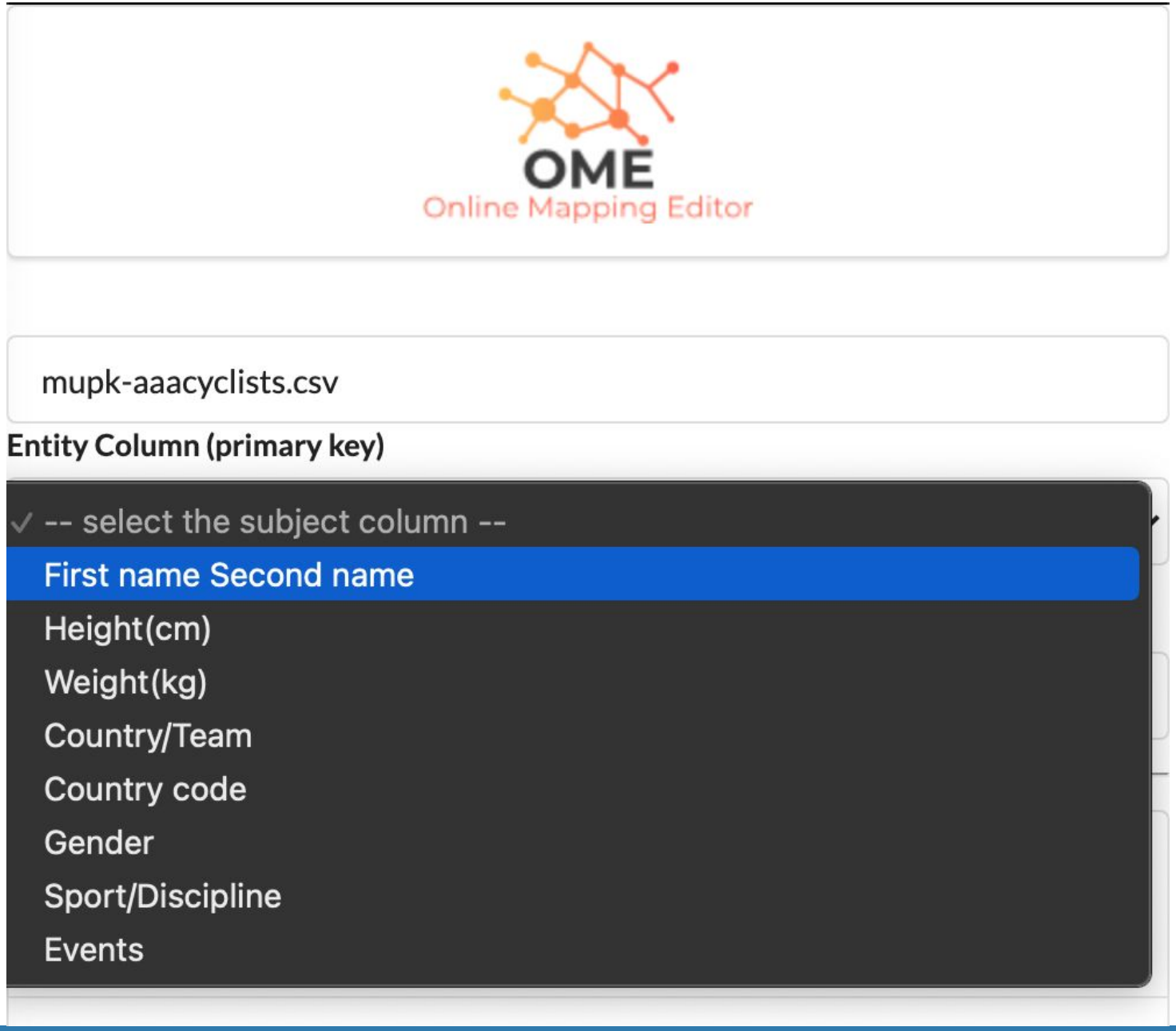
Preferred Label	Frequency	Relevance	Most Frequent Label	Broader Concepts	Concept Scheme
Singapore Sling	20	0.26	singapore sling	Mixed drink, Contemporary Classics	Cocktails
single malt	40	56.02	single malt	Scotch whisky	Beverages
Soft drink	204	16.88	soft drink	Non-alcoholic beverage	Beverages
Sour cocktails	1	0.25	sour cocktail		Cocktails
Prosecco	262	21.59	sparkle	Sparkling wine	Beverages
Sparkling wine	130	22.04	sparkling wine	Wine	Beverages
Spritz Veneziano	7	2.93	spritz	New Era Drinks, Apéritif and digestif, Wine cocktails	Cocktails
Stemware	2	6.24	stemware		Glassware
Stevia	4	2.68	stevia	Sugar	Ingredients & Garnish
Stinger	22	1.5	stinger	The Unforgettables, Duo and trio cocktails	Cocktails


CSV
RML
RML mapper
RDF
Ontologies
poolparty
RDF/RDB


CSV
RML
RML mapper
RDF
Ontologies
Unified View
RDF/RDB
poolparty
Java
Pipelines

OME: https://ome.linkeddata.es/




OME
Online Mapping Editor
mupk-aaacyclists.csv
Entity Column (primary key)
-- select the subject column --
First name Second name
Height(cm)
Weight(kg)
Country/Team
Country code
Gender
Sport/Discipline
Events

Online Mapping Editor

mupk-aaacyclists.csv

Entity Column (primary key)

First name Second name

Choose the concept of this file:

http://dbpedia.org/ontology/Cyclist

Header

Schema

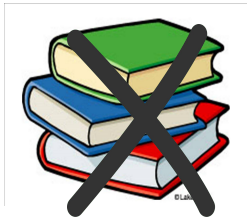
First name Second name

http://www.w3.org/2000/01/rdf-schema#label

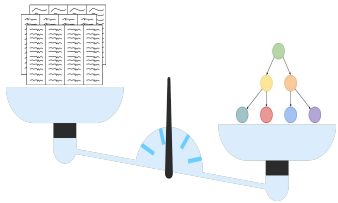

✓ R2RML
RML
YARRRML
R2RML
Generate Mappings

```
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix rml: <http://semweb.mmlab.be/ns/rml#> .
@prefix ql: <http://semweb.mmlab.be/ns/ql#> .
@prefix mail: <http://example.com/mail#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix ex: <http://www.example.com/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix transit: <http://vocab.org/transit/terms/> .
@prefix wgs84_pos: <http://www.w3.org/2003/01/geo/wgs84_pos#>.
@prefix schema: <http://schema.org/>.
@prefix gn: <http://www.geonames.org/ontology#>.
@prefix geosp: <http://www.telegraphis.net/ontology/geography/geography#> .
@base <http://mappingpedia.linkeddata.es/resource/> .
<fynpiubfhd>
rml:logicalSource [
    rml:source  "mupk-aaacyclists.csv";
    rml:referenceFormulation ql:CSV

];
rr:subjectMap [
    rml:reference "First name Second name";
    rr:class <http://dbpedia.org/ontology/Cyclist>
];

        rr:predicateObjectMap [
          rr:predicate <http://www.w3.org/2000/01/rdf-schema#label>;
          rr:objectMap    [ rml:reference "First name Second name" ]
        ];
.
```

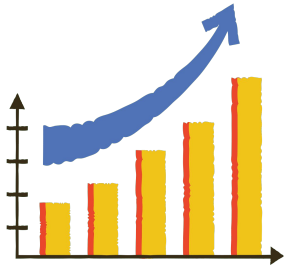
The literature lacks studies that analyses data integration market.



There is a lack of automated and semantic-based data integration products and services in the market.



Companies can make money using AI-based data integration solutions using semantic and automatic mapping.



Why Semantics? => Make money

Vision

Online Chowlk Editor?

Vocab

Ontology Engineering

DATA, NLP and Linguistic

Law corpus

Med corpus

Corpus profiling

Suggest classes, properties

Editor

Knowledge graph enrichment

Annotation Improvement & ML models

Mirror

Morph-CSV Mapeator

Morph-GraphQL Morph-RDB

Knowledge Graph Construction

TTLA

TADA-GAM TADA-Entity

TADA-NUM TADA-HDT

Out of the scope:

- Pipeline
- ETL and Data management
- Streaming and IOT
- Adapters to known systems (e.g., shopify)

References

Images are taken from

- http://clipart-library.com/
- https://icons8.com/