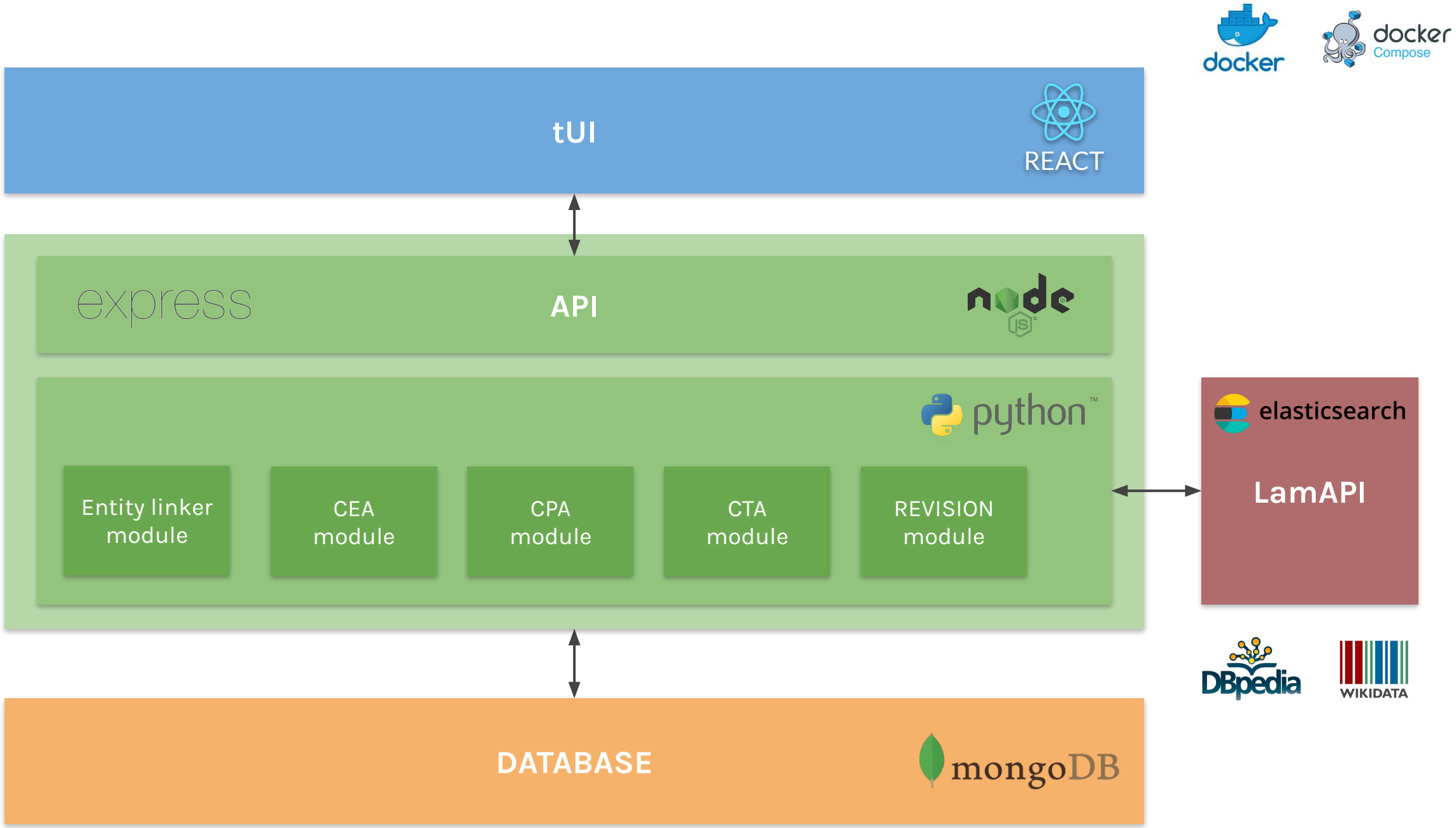
MantisTable V

a **novel** and **efficient** approach to **Semantic Table Interpretation**

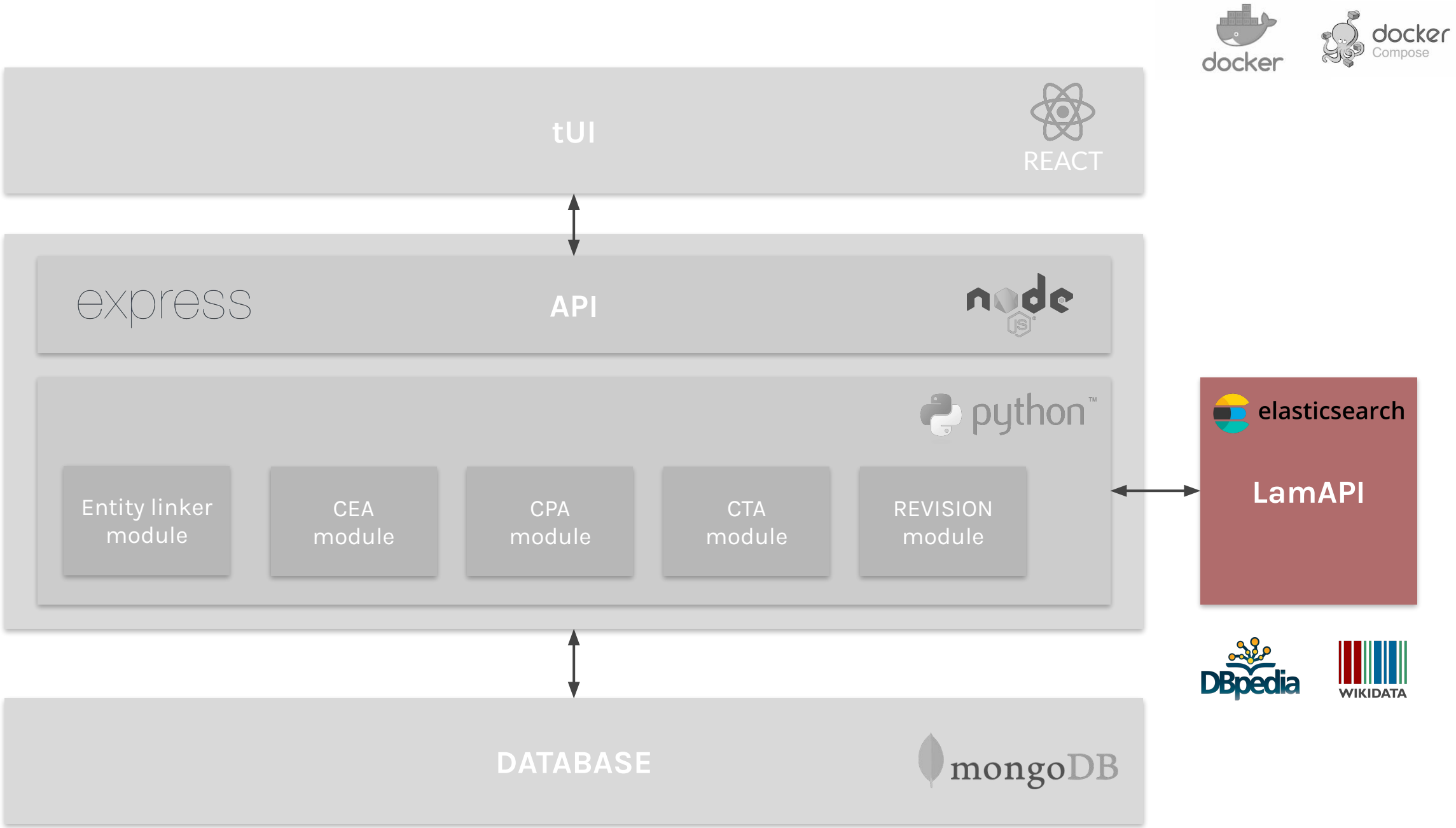
Roberto Avogadro, Marco Cremaschi
Department of Computer Science, Systems and Communication (DISCo)
University of Milano - Bicocca

ISWC 2021


docker
docker Compose
tUI
REACT
express
API
node
python
Entity linker module
CEA module
CPA module
CTA module
REVISION module
elasticsearch
LamAPI
DBpedia
WIKIDATA
DATABASE
mongoDB

LamAPI

Label Matching API


docker
docker Compose
tUI
REACT
express
API
node
python
Entity linker module
CEA module
CPA module
CTA module
REVISION module
elasticsearch
LamAPI
DBpedia
WIKIDATA
DATABASE
mongoDB

LamAPI

LamAPI (Label Matching API) is a **tool** that aims to provide an easy way to **store and access** Wikidata and DBpedia dumps with the purpose to aid Semantic Table Interpretation tools

The tool allows **Full-text Search** and **Fuzzy matching** over labels and provides **Predicate**, **Entity** and **Concepts Matching** services (using Elasticsearch)

Swagger UI of LamAPI is available at

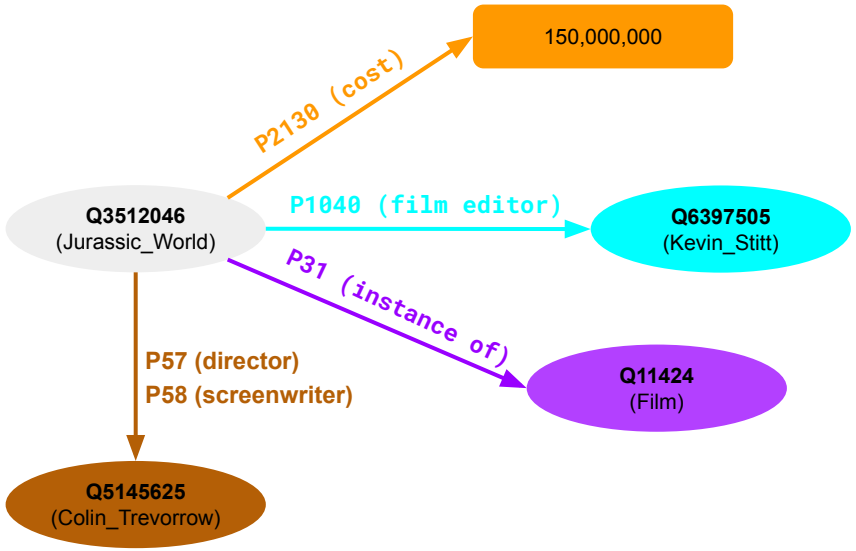
http://149.132.178.112:8097/

elasticsearch

Wikidata

Q3512046

LamAPI




```
type
{
  "Q3512046": {
      "P31": [
            "Q11424",
            "Q229390"
      ]
  }
}
object
{
  "Q3512046": {
      "Q5145625": [
            "P57",
            "P58"
      ],
      "Q155124": [
            "P86"
      ],
      "Q6397505": [
            "P1040"
      ],
      "Q168383": [
            "P272",
            "P750"
      ]
      ...
  }
}
```

```
literal
{
  "Q3512046": {
      "number": {
            "P2130": [
                  "150,000,000"
            ],
            "P2047": [
                  "124"
            ]
            ...
      },
      "string": {
            "P1476": [
                  "Jurassic World"
            ]
            ...
      }
      "date": {
            "P577": [
                  "12 June 2015",
                  ...
            ]
            ...
      }
      ...
  }
}
```

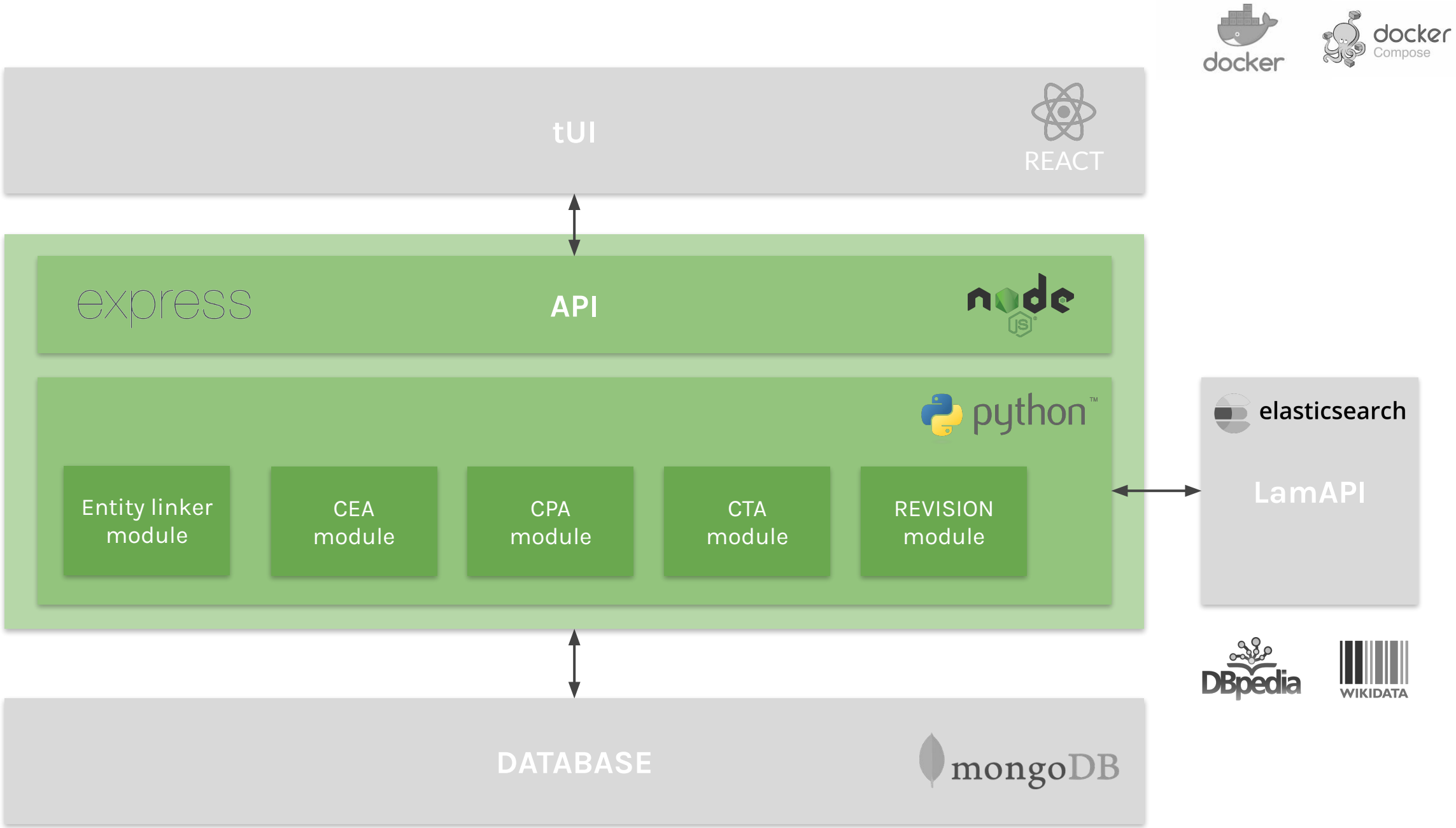
Q3512046 Q5145625

LamAPI

```
predicate
{
  "Q3512046 Q5145625": [
      "P57",
      "P58"
  ]
}
```

MantisTable V

seMantics Table


docker
docker Compose
tUI
REACT
express
API
node
python
Entity linker module
CEA module
CPA module
CTA module
REVISION module
elasticsearch
LamAPI
DBpedia
WIKIDATA
DATABASE
mongoDB

Semantic Table Interpretation: MantisTable V

MantisTable V is a automatic and unsupervised **Semantic Table Interpretation approach**

The approach is performed against **DBpedia** and **Wikidata**, and it can be easily adapted to any other Knowledge Graph

MantisTable V: Key Features

- **Automatic and Unsupervised** approach
- **Conservative** approach
  - All results are sorted and scored but never removed
- **Iterative** approach
  - Tasks are not independent
  - Revision/rescoring of results
- **Row-based** approach
- High **performance** and good **accuracy**
- **Multiple KG support using LamAPI**

MantisTable V: tasks

1. Data preparation and normalisation
2. Column analysis and subject detection
3. Candidate generation
4. Cell Entity Annotation (CEA)
5. Column Predicate Annotation (CPA)
6. Column Type Annotation (CTA)
7. Revision
8. Export

Without loss of generality, let's consider the following example

title	director	release year	domestic distributor	length in min	worldwide gross
Jurassic World	Colin Trevorrow	2015	Universal Pictures	124	1,670,400,637
Superman Returns	Bryan Singer	2006	Warner Bros.	154	391,081,192
Batman Begins	Christopher Nolan	2005	Warner Bros.	140	371,853,783
Avatar	James Cameron	2009	20 Century Fox	162	2,744,336,793

Bring the text to lowercase and apply to each cell the correct transformation based on datatype

The datatypes are identified using regex (e.g., boolean, date, email, geocoords, integer, float, ISBN, URL, XPath, CSS)

title	director	release year	domestic distributor	length in min	worldwide gross
jurassic world	colin trevorrow	2015	universal pictures	124	1670400637
superman returns	bryan singer	2006	warner bros	154	391081192
batman begins	christopher nolan	2005	warner bros	140	371853783
avatar	james cameron	2009	20 century fox	162	2744336793

title	director	release year	domestic distributor	length in min	worldwide gross
jurassic world	colin trevorrow	2015	universal pictures	124	1670400637
superman returns	bryan singer	2006	warner bros	154	391081192
batman begins	christopher nolan	2005	warner bros	140	371853783
avatar	james cameron	2009	20 century fox	162	2744336793

Subject column

Named-Entity column

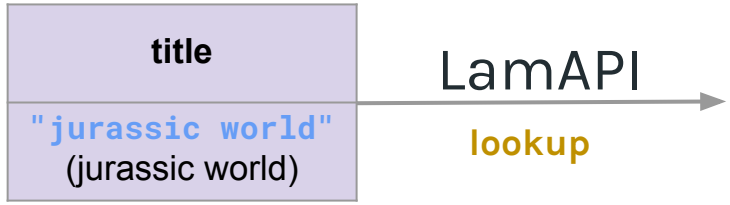
Literal column

1. Detection of LIT-columns uses regular expressions to identify the datatype
2. Others columns are classified as NE-columns
3. Detection of SUBJ-column, among the NE-columns, considering statistic features

For the **NE-columns**, query the KG (LamAPI) with the cells content to retrieve candidate entities and their types

A lookup query with “jurassic world” on **LamAPI Wikidata** returns the list of candidates:

- Q3512046 (“Jurassic World” - Film)
- Q21877685 (“Jurassic World” - Film)
- Q55615459 (“Jurassic World” - Amusement ride)
- …

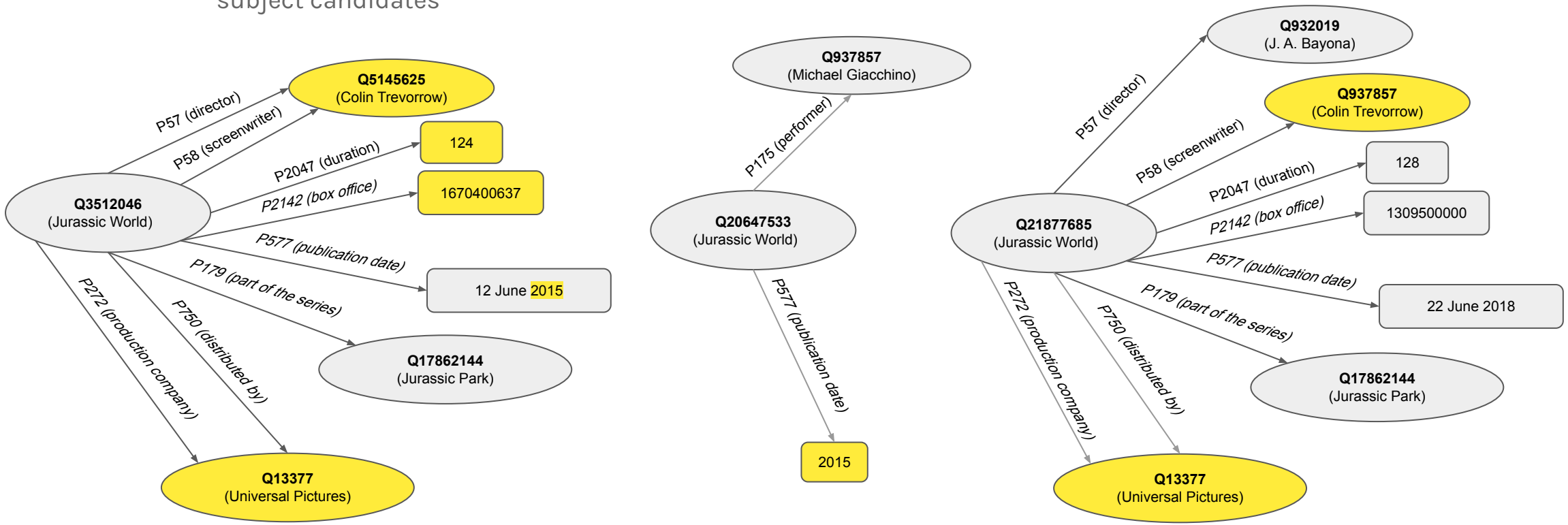



```
{
  "results": [
        {
          "id": "Q3512046",
          "name": "Jurassic World"
          "types" : [
            {
                  "id": "Q229390",
                  "name": "3D film"
             },
             {
                  "id": "Q11424",
                  "name": "film"
            },
          ],
        },
        {
          "id": "Q21877685",
          "name": "Jurassic World"
          "types" : [
            {
                  "id": "Q11424",
                  "name": "film"
            }
          ],
        },
        {
          "id": "Q55615459",
          "name": "Jurassic World"
          "types" : [
            {
                  "id": "Q1144661",
                  "name": "amusement ride"
            }
          ],
        },
      ...
  ]
}
```

The goals of this tasks are:

1. Annotate each cell belong to a NE-column with a entity from KG
2. Collect **types** (e.g., Q11424:Film) obtained from annotations of cells
3. Collect **predicates** obtained from relations between annotations of two cells

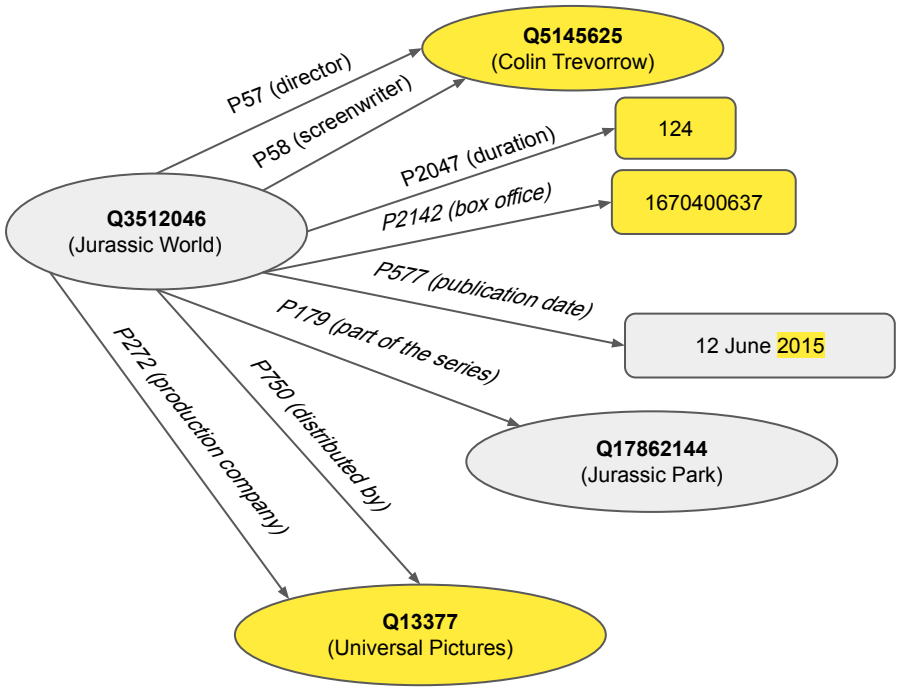
MantisTable calls *object* and *literals services* of LamAPI to obtain a set of literal and object related to subject candidates

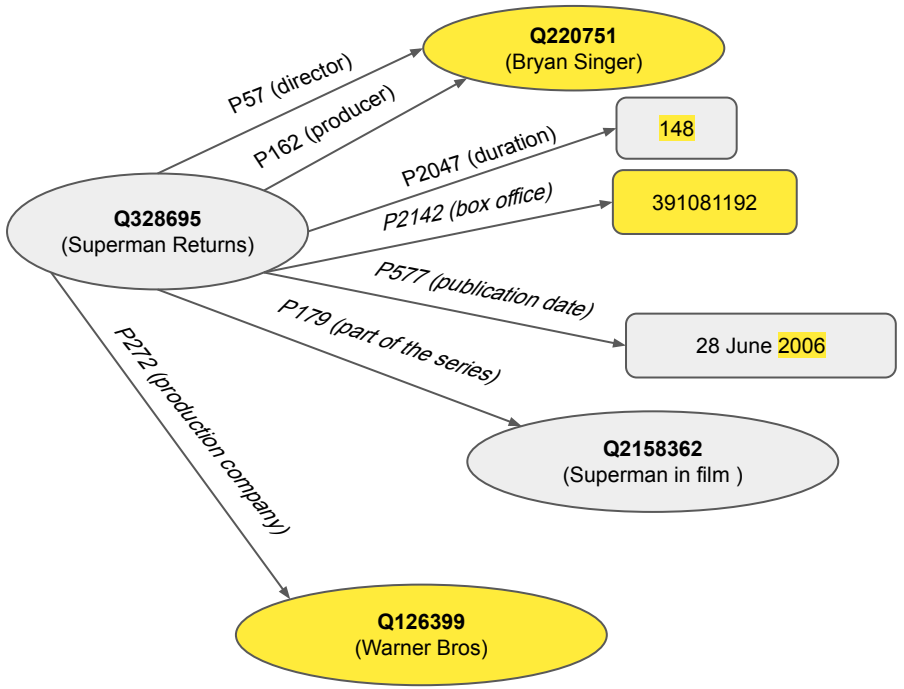


...

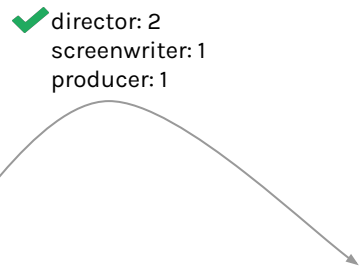
title	director	release year	domestic distributor	length in min	worldwide gross
jurassic world	colin trevorrow	2015	universal pictures	124	1670400637

Considering the winning entities from CEA phase..





title	director	release year	domestic distributor	length in min	worldwide gross
jurassic world	colin trevorrow	2015	universal pictures	124	1670400637
superman returns	bryan singer	2006	warner bros	154	391081192



title	director	release year	domestic distributor	length in min	worldwide gross
jurassic world	colin trevorrow	2015	universal pictures	124	1670400637

Q3512046 (Jurassic World)

Q5145625 (Colin Trevorrow)

P57 (director)
P58 (screenwriter)
?

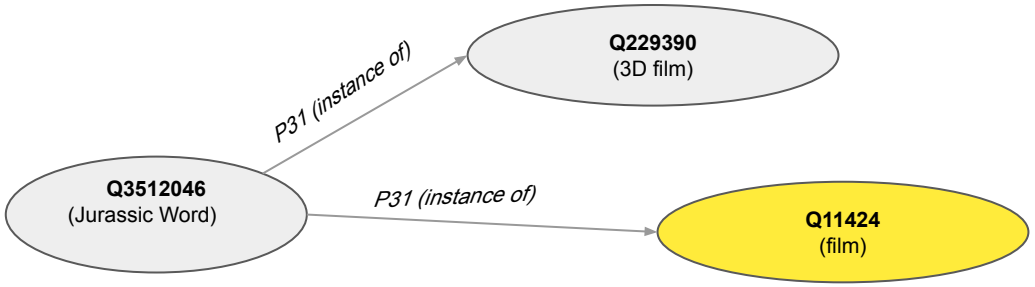
superman returns	bryan singer	2006	warner bros	154	391081192

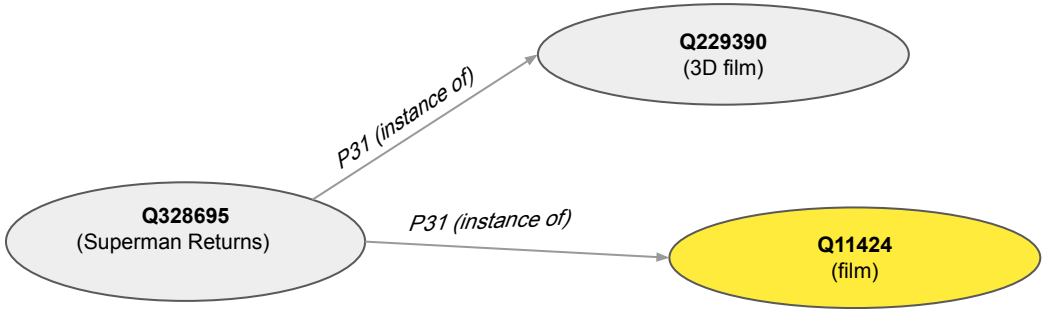
Q328695 (Superman Returns)

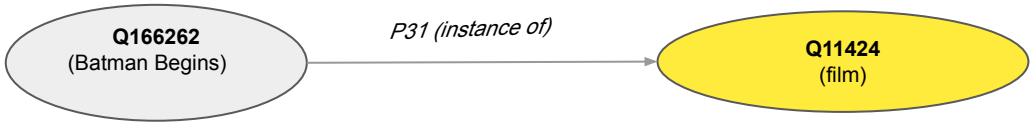
Q220751 (Bryan Singer)

P57 (director)
P162 (producer)
?

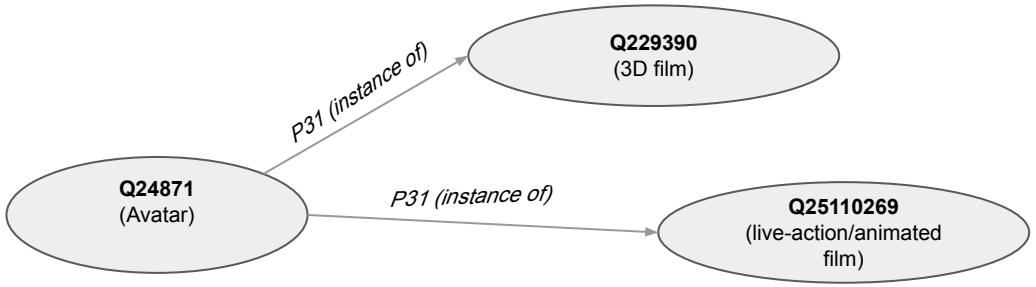
Considering the types of winning entities from the CEA phase..

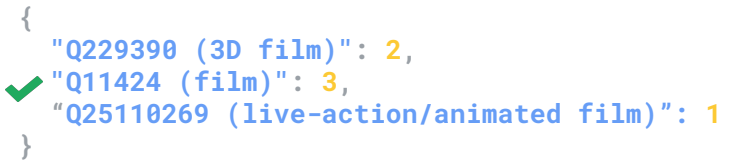






title
jurassic world
superman returns
batman begins
avatar





Consistency check the cell annotations against final results

=> recompute annotations on inconsistent annotations

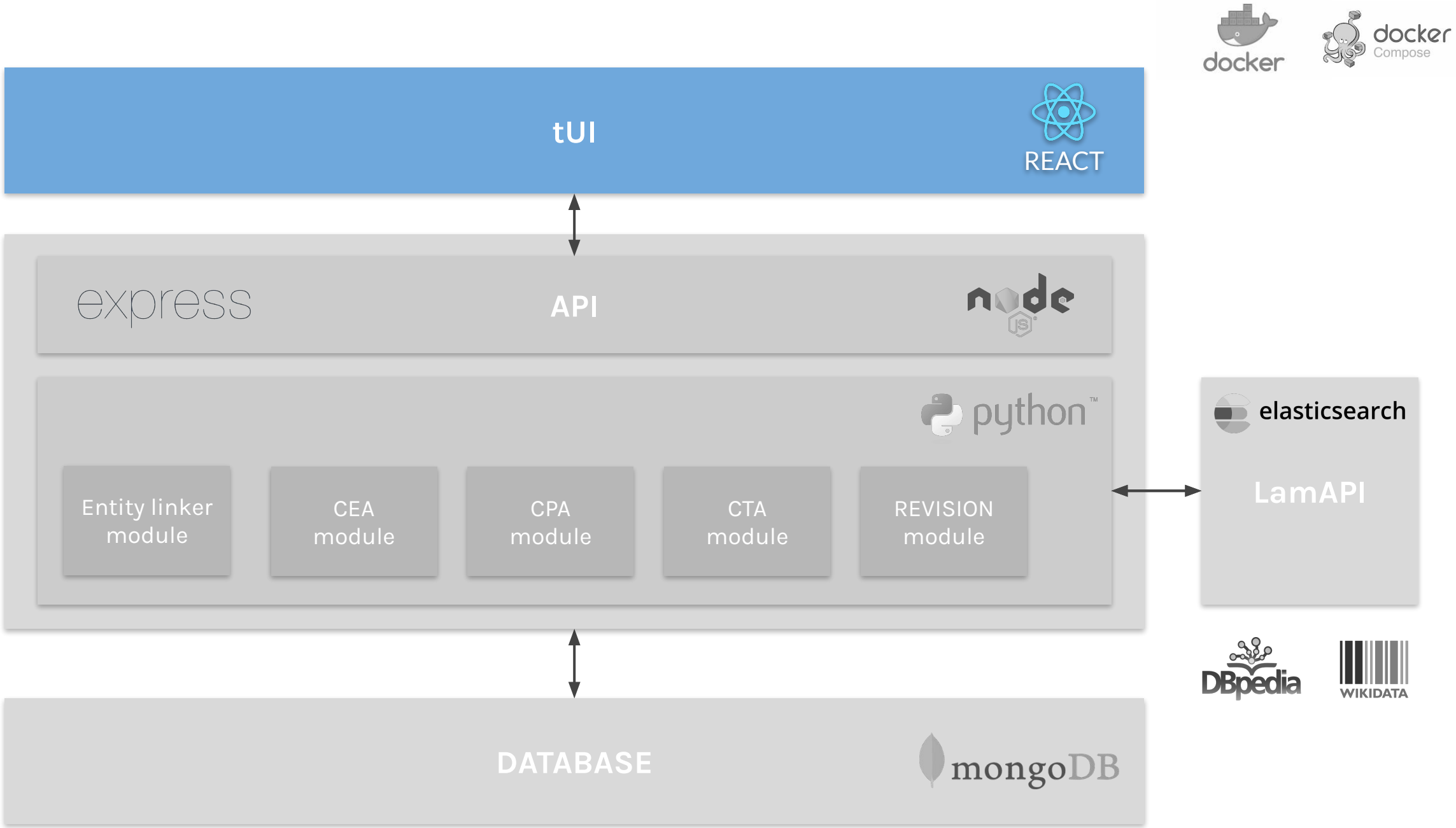
e.g., a predicate of a pair of cells annotation that does not match with the selected column predicate (CPA) may lead to a different selection of matching entities among candidates

Export of table with annotation in different formats:

- JSON-LD
- RDF
- Turtle

tUI

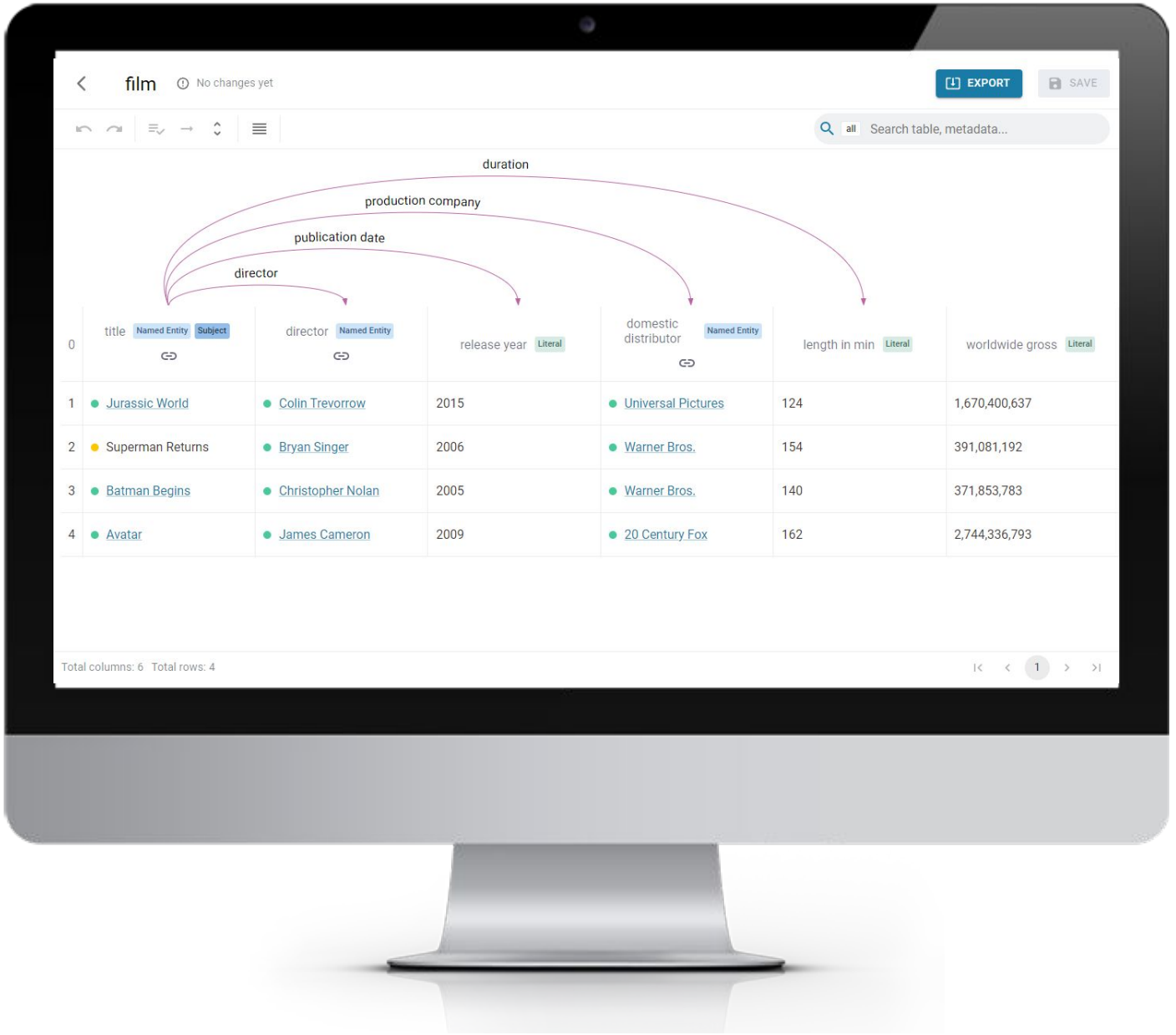
table User Interface


docker
docker Compose
tUI
REACT
express
API
node
python
Entity linker module
CEA module
CPA module
CTA module
REVISION module
elasticsearch
LamAPI
DBpedia
WIKIDATA
DATABASE
mongoDB

tUI

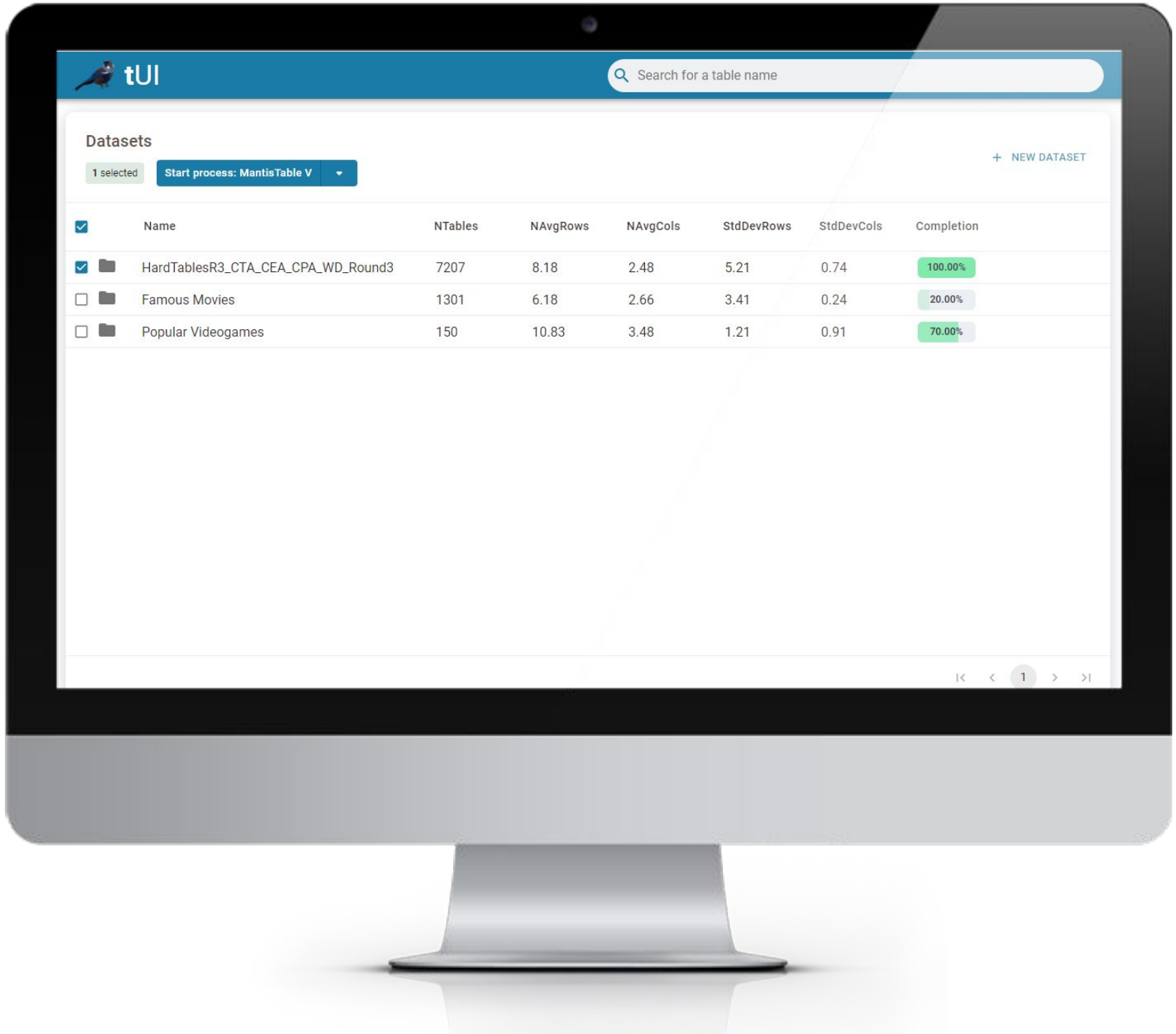
tUI is a **Web application** that aims to provide a visualisation tool for **STI approaches**

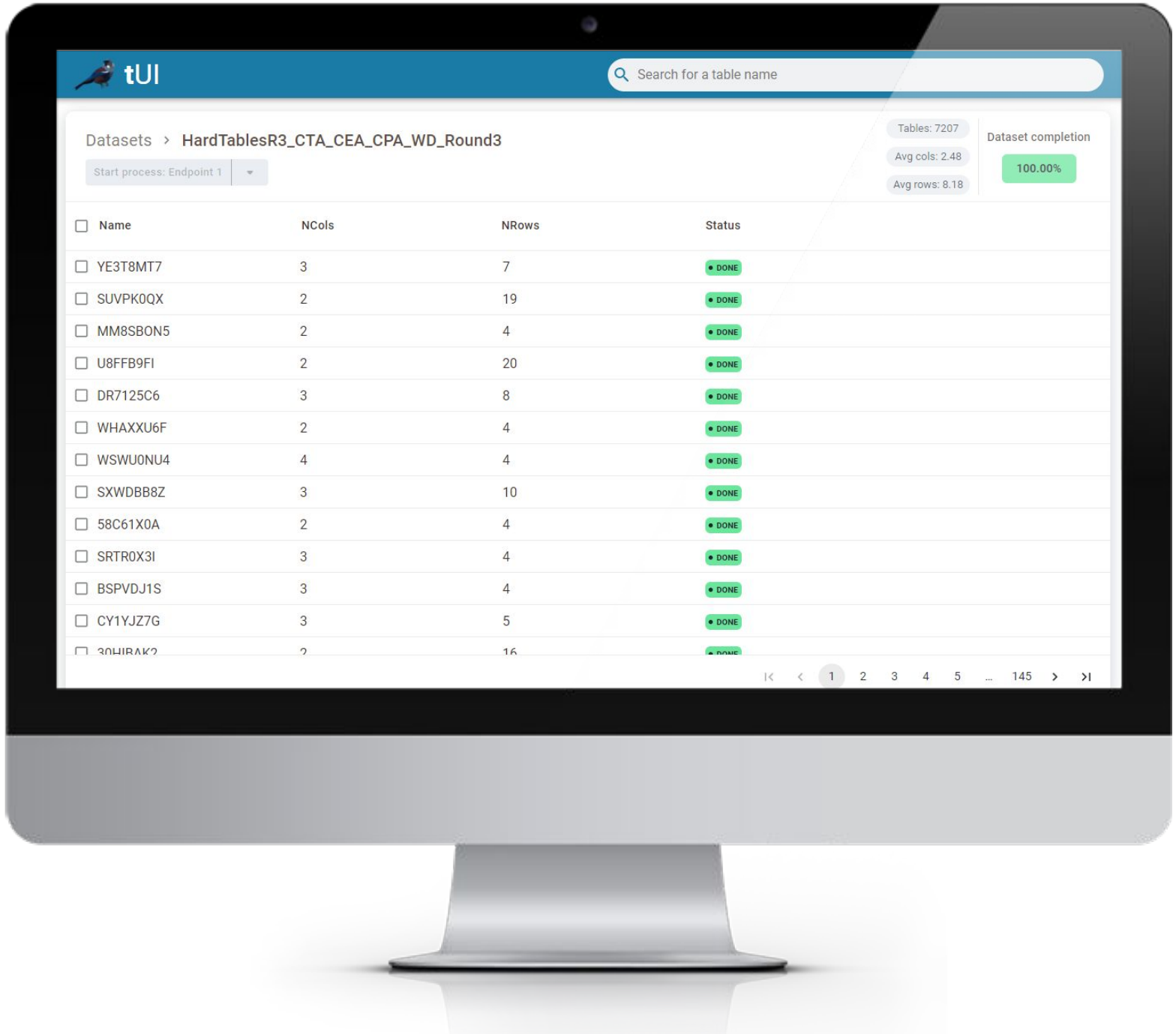
http://149.132.178.112:81

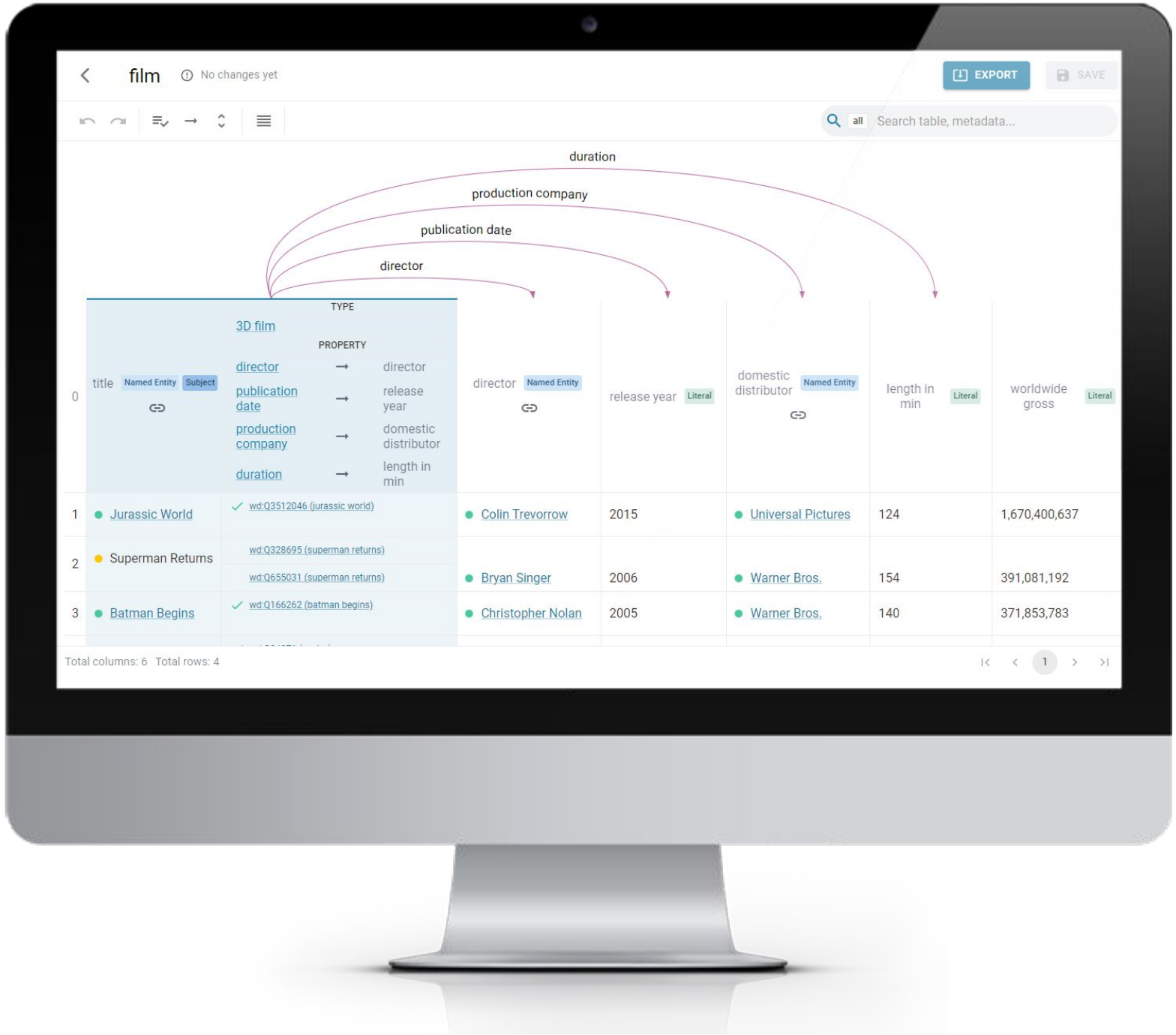


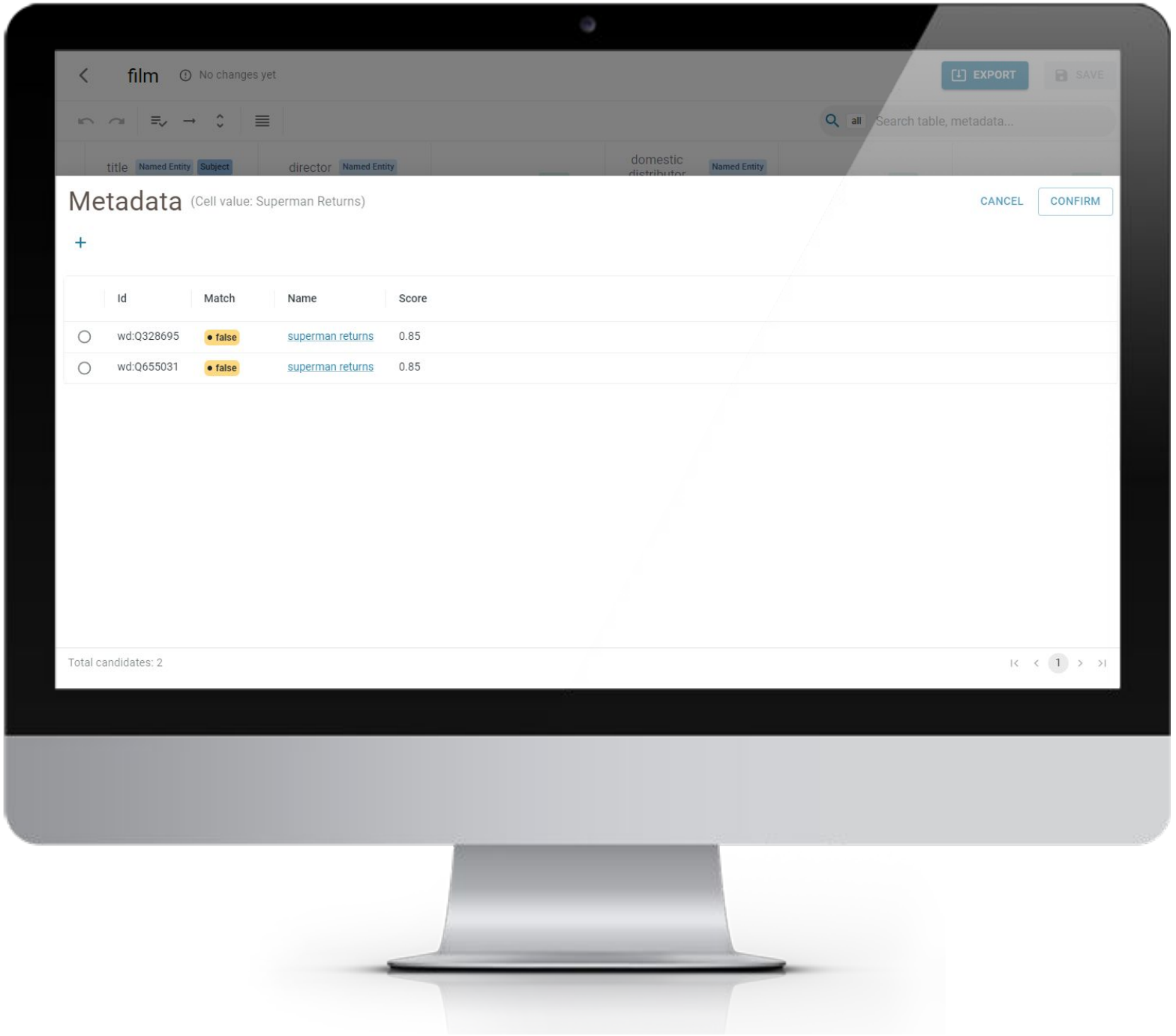
tUI: Key Features

- It works with any backend
- Data are retrieved through API endpoint
- All endpoint can be customize in a configuration file (YAML)
- Data and functionalities are based on available APIs endpoints
- All the three main tasks of the STI (**CTA**, **CPA**, **CEA**) are supported and annotations can be viewed directly inside the table


tUI
Search for a table name
Datasets
1 selected
Start process: MantisTable V
+ NEW DATASET
Name
NTables
NAvgRows
NAvgCols
StdDevRows
StdDevCols
Completion
HardTablesR3_CTA_CEA_CPA_WD_Round3
7207
8.18
2.48
5.21
0.74
100.00%
Famous Movies
1301
6.18
2.66
3.41
0.24
20.00%
Popular Videogames
150
10.83
3.48
1.21
0.91
70.00%
1


tUI
Search for a table name
Datasets › HardTablesR3_CTA_CEA_CPA_WD_Round3
Start process: Endpoint 1
Tables: 7207
Avg cols: 2.48
Avg rows: 8.18
Dataset completion
100.00%
Name
NCols
NRows
Status
YE3T8MT7 3 7 DONE
SUVPK0QX 2 19 DONE
MM8SBON5 2 4 DONE
U8FFB9FI 2 20 DONE
DR7125C6 3 8 DONE
WHAXXU6F 2 4 DONE
WSWU0NU4 4 4 DONE
SXWDBB8Z 3 10 DONE
58C61X0A 2 4 DONE
SRTR0X3I 3 4 DONE
BSPVDJ1S 3 4 DONE
CY1YJZ7G 3 5 DONE
1 2 3 4 5 ... 145


film
No changes yet
EXPORT
SAVE
all
Search table, metadata...
duration
production company
publication date
director
TYPE
3D film
PROPERTY
director → director
publication date → release year
production company → domestic distributor
duration → length in min
0
title Named Entity Subject
director Named Entity
release year Literal
domestic distributor Named Entity
length in min Literal
worldwide gross Literal
1
Jurassic World
wd:Q3512046 (jurassic world)
Colin Trevorrow
2015
Universal Pictures
124
1,670,400,637
2
Superman Returns
wd:Q328695 (superman returns)
wd:Q655031 (superman returns)
Bryan Singer
2006
Warner Bros.
154
391,081,192
3
Batman Begins
wd:Q166262 (batman begins)
Christopher Nolan
2005
Warner Bros.
140
371,853,783
Total columns: 6 Total rows: 4
1


film
No changes yet
EXPORT
SAVE
all
Search table, metadata...
title
Named Entity
Subject
director
Named Entity
domestic
distributor
Named Entity
Metadata
(Cell value: Superman Returns)
CANCEL
CONFIRM
Id
Match
Name
Score
wd:Q328695
false
superman returns
0.85
wd:Q655031
false
superman returns
0.85
Total candidates: 2
1

Final Remarks and Future Plans

Open research problem

- **Novel entities**
  - Identification of missing entities in the KG
  - KG evolution
- **Streaming data**
  - Row-based approach opens to incremental annotation of data streams
- **Human in the loop**
  - Revision task with domain experts to address ambiguities

Open research problem

- **Knowledge Graph**
  - Create an additional layer to unify the different KGs, and treat them as a single dataset
  - Improve the efficiency and effectiveness of candidate generation
  - Define a strategy to unify the different KGs schemas