Search-API

Carlos Badenes-Olmedo, Francisco Yedro, Oscar Corcho

Ontology Engineering Group (OEG)

Universidad Politécnica de Madrid (UPM)

Spain

Motivation / Proposal

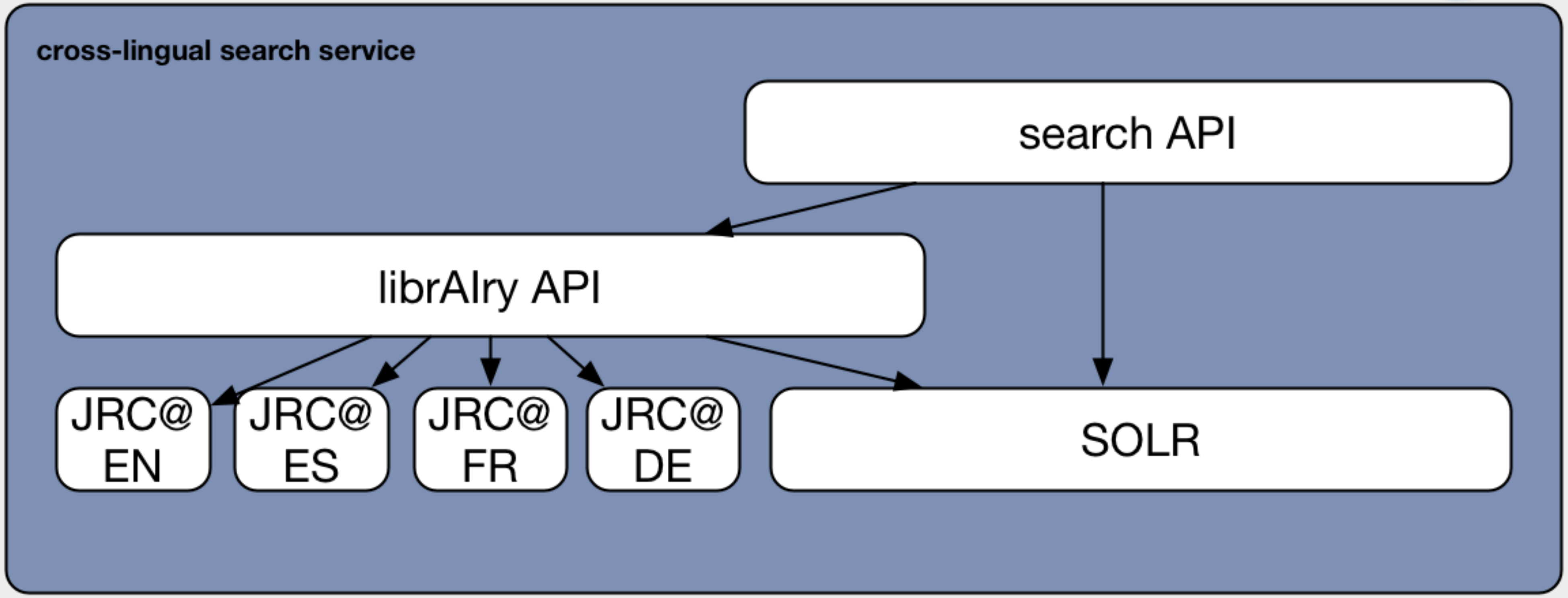
- We got a ***huge*** collection of (un-labelled) documents and we would like to ***explore*** the knowledge inside.

- Imagine that we could run an ***unsupervised***, ***automated*** pipeline to generate ***connections*** between them

- state-of-the art techniques to programmatically generating ***annotations*** for each of the texts inside big collections of documents

- in a way that is ***computationally affordable***

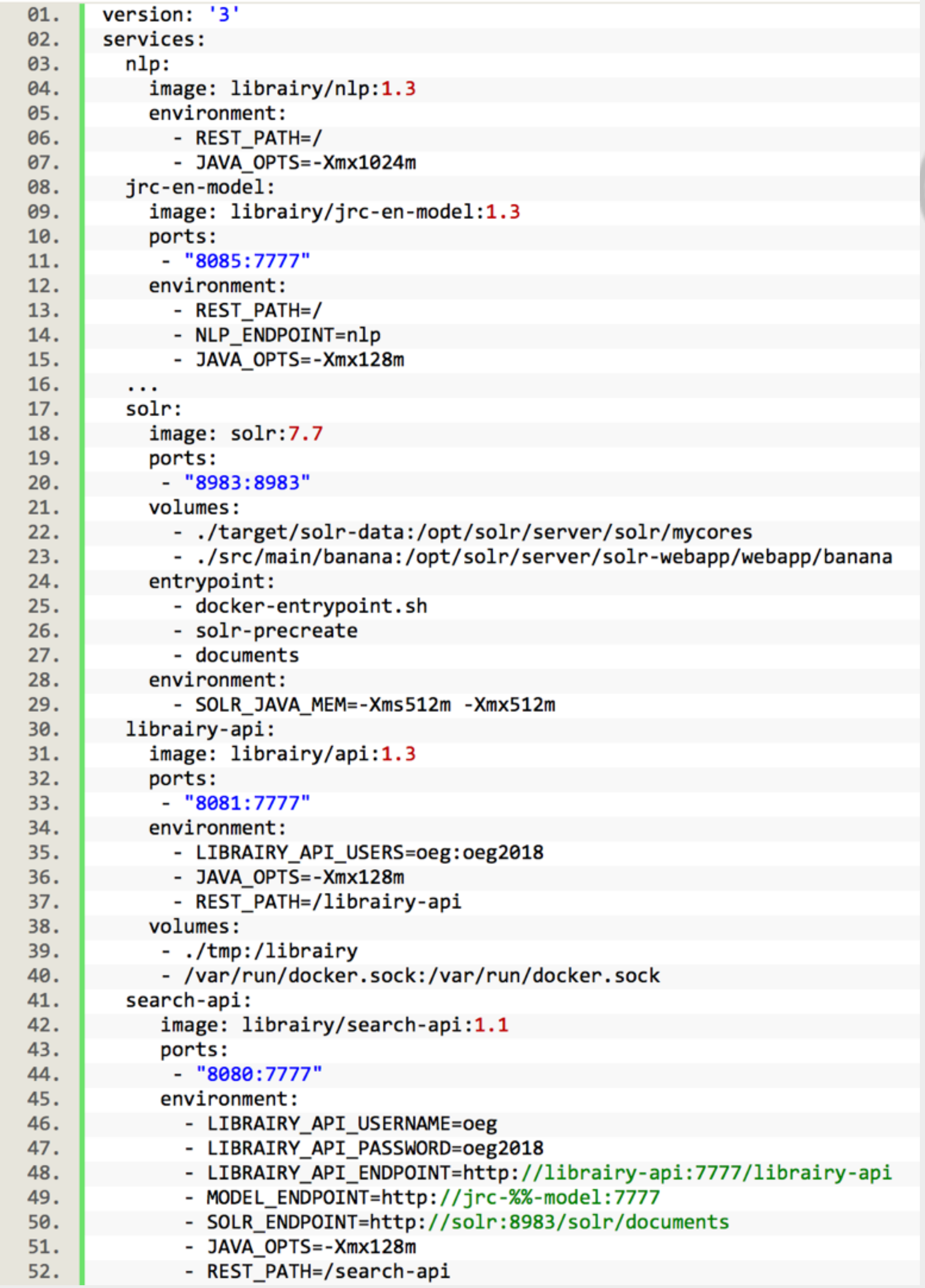
Architecture



Deployment

- https://github.com/TBFY/search-API
- the search-API, librAIry, Solr and all the annotation models are distributed as ***Docker images***
- ***Docker-Engine*** and ***Docker-Compose*** are only required to deploy the whole system
- The service descriptor to run it is:

```
$docker-compose up
```



```
version: '3'
services:
  nlp:
    image: librairy/nlp:1.3
    environment:
      - REST_PATH=/
      - JAVA_OPTS=-Xmx1024m
  jrc-en-model:
    image: librairy/jrc-en-model:1.3
    ports:
     - "8085:7777"
    environment:
      - REST_PATH=/
      - NLP_ENDPOINT=nlp
      - JAVA_OPTS=-Xmx128m
  ...
  solr:
    image: solr:7.7
    ports:
     - "8983:8983"
    volumes:
      - ./target/solr-data:/opt/solr/server/solr/mycores
      - ./src/main/banana:/opt/solr/server/solr-webapp/webapp/banana
    entrypoint:
      - docker-entrypoint.sh
      - solr-precreate
      - documents
    environment:
      - SOLR_JAVA_MEM=-Xms512m -Xmx512m
  librairy-api:
    image: librairy/api:1.3
    ports:
     - "8081:7777"
    environment:
      - LIBRAIRY_API_USERS=oeg:oeg2018
      - JAVA_OPTS=-Xmx128m
      - REST_PATH=/librairy-api
    volumes:
     - ./tmp:/librairy
     - /var/run/docker.sock:/var/run/docker.sock
  search-api:
     image: librairy/search-api:1.1
     ports:
      - "8080:7777"
     environment:
       - LIBRAIRY_API_USERNAME=oeg
       - LIBRAIRY_API_PASSWORD=oeg2018
       - LIBRAIRY_API_ENDPOINT=http://librairy-api:7777/librairy-api
       - MODEL_ENDPOINT=http://jrc-%%-model:7777
       - SOLR_ENDPOINT=http://solr:8983/solr/documents
       - JAVA_OPTS=-Xmx128m
       - REST_PATH=/search-api
```

Deployment

user: tbfy **password:** oeg2019

SERVICE	LOCAL-ENDPOINT	REMOTE-ENDPOINT
Search-API	http://localhost:8080/search-api	http://tbfy.librairy.linkeddata.es/search-api
librAIry-API	http://localhost:8081/librairy-api	http://librairy.linkeddata.es/api
Solr	http://localhost:8983/solr	http://librairy.linkeddata.es/data
Dashboard	http://localhost:8983/solr/banana	http://librairy.linkeddata.es/data/dashboard
JRC-English-Model	http://localhost:8085	http://librairy.linkeddata.es/jrc-en-model/
JRC-Spanish-Model	http://localhost:8086	http://librairy.linkeddata.es/jrc-es-model
JRC-French-Model	http://localhost:8087	http://librairy.linkeddata.es/jrc-fr-model
JRC-German-Model	http://localhost:8088	http://librairy.linkeddata.es/jrc-de-model

Actions

- ***Add new documents***
  - (1) from structured text files
  - (2) from existing Solr indexes
  - (3) from remote APIs
  - (4) from raw files
- ***Annotate existing documents***: *with main topics*

librAIry-API
http://librairy.linkeddata.es/api

- ***Search documents***: *by name, (or) source, (or) language, (or) words*
- ***Read documents***: *by id*
- ***Retrieve similar documents***: *by name, (or) source, (or) language, (or) words*
  - (1) from a document id
  - (2) from free-text

Search-API
http://tbfy.librairy.linkeddata.es/search-api

A1: Add New Documents

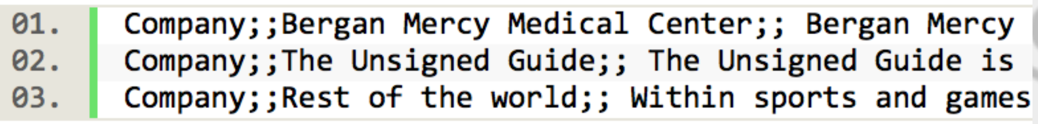
from structured text files

- ***librAIry-api** is the responsible for ingesting documents*

- ***CSV*** or ***JSON-L*** text files (even in ***{tar}.gz*** format) can be used

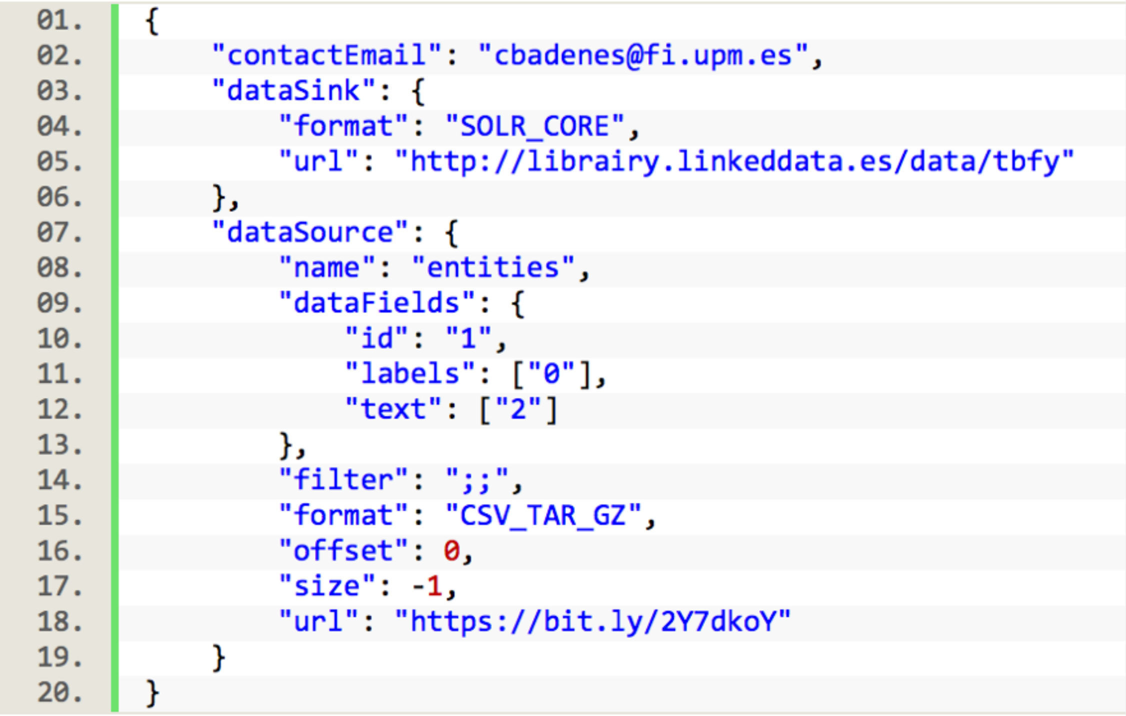
- A HTTP_POST request to:
  http://librairy.linkeddata.es/api/documents

csv file



```
Company;;Bergan Mercy Medical Center;; Bergan Mercy
Company;;The Unsigned Guide;; The Unsigned Guide is
Company;;Rest of the world;; Within sports and games
```

json request



```
{
    "contactEmail": "cbadenes@fi.upm.es",
    "dataSink": {
        "format": "SOLR_CORE",
        "url": "http://librairy.linkeddata.es/data/tbfy"
    },
    "dataSource": {
        "name": "entities",
        "dataFields": {
            "id": "1",
            "labels": ["0"],
            "text": ["2"]
        },
        "filter": ";;",
        "format": "CSV_TAR_GZ",
        "offset": 0,
        "size": -1,
        "url": "https://bit.ly/2Y7dkoY"
    }
}
```

A1: Add New Documents

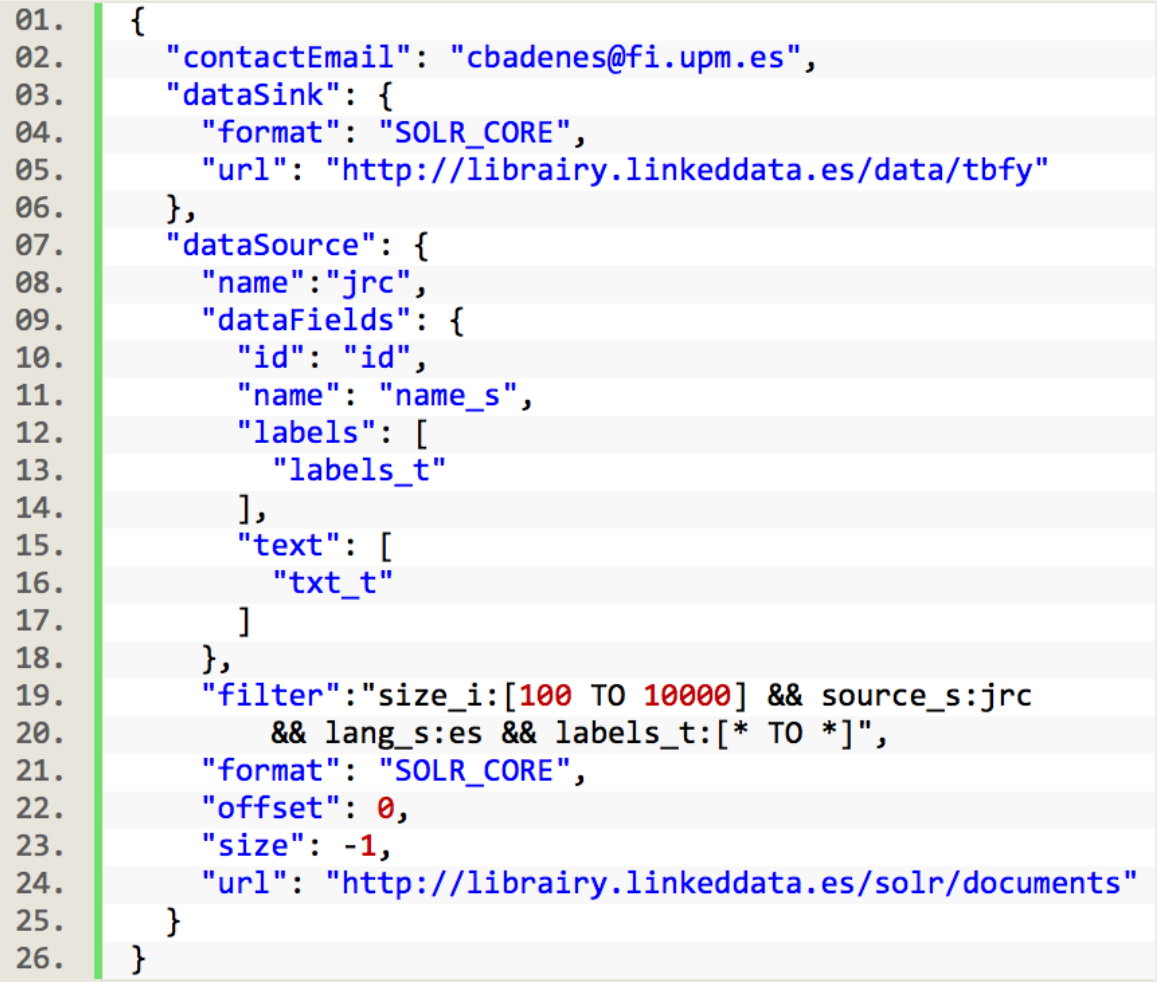
from Solr indexes

- ***librAIry-api** is the responsible for ingesting documents*

- The documents can already be hosted in a ***Solr collection***

- A HTTP_POST request to:
  http://librairy.linkeddata.es/api/documents

json request



```
{
  "contactEmail": "cbadenes@fi.upm.es",
  "dataSink": {
    "format": "SOLR_CORE",
    "url": "http://librairy.linkeddata.es/data/tbfy"
  },
  "dataSource": {
    "name":"jrc",
    "dataFields": {
      "id": "id",
      "name": "name_s",
      "labels": [
        "labels_t"
      ],
      "text": [
        "txt_t"
      ]
    },
    "filter":"size_i:[100 TO 10000] && source_s:jrc
        && lang_s:es && labels_t:[* TO *]",
    "format": "SOLR_CORE",
    "offset": 0,
    "size": -1,
    "url": "http://librairy.linkeddata.es/solr/documents"
  }
}
```

A1: Add New Documents

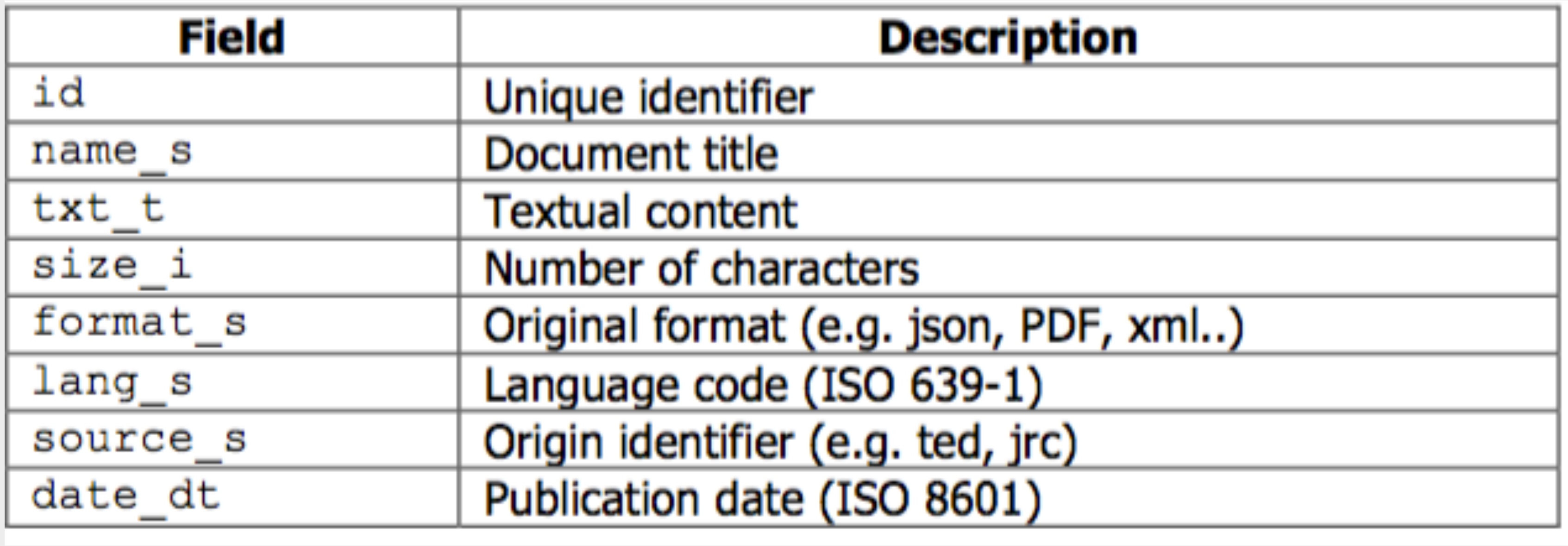
from remote APIs

- documents can be available from a remote ***HTTP_Restful API*** (e.g OpenOpps-API)
- or stored on remote ***storage systems*** (e.g. Amazon S3)
- a **Harvester** client has been created to handle these sources
- It ***download***, ***parse*** and ***store*** the resources from external sources into Solr collection with a predefined format
- GitHub project: https://github.com/TBFY/harvester

A1: Add New Documents

from raw files

- There may be exceptional circumstances that do not conform to the above scenarios
- Since the search service uses Solr internally to store its documents, the ***Solr API*** can be used directly to add new documents: http://librairy.linkeddata.es/data/#/tbfy
- The only requirement in this scenario is to use the following ***fields*** to describe the documents:



Field	Description
id	Unique identifier
name_s	Document title
txt_t	Textual content
size_i	Number of characters
format_s	Original format (e.g. json, PDF, xml..)
lang_s	Language code (ISO 639-1)
source_s	Origin identifier (e.g. ted, jrc)
date_dt	Publication date (ISO 8601)

A2: Annotate Documents

- ***librAIry-api** is the responsible for annotate documents*

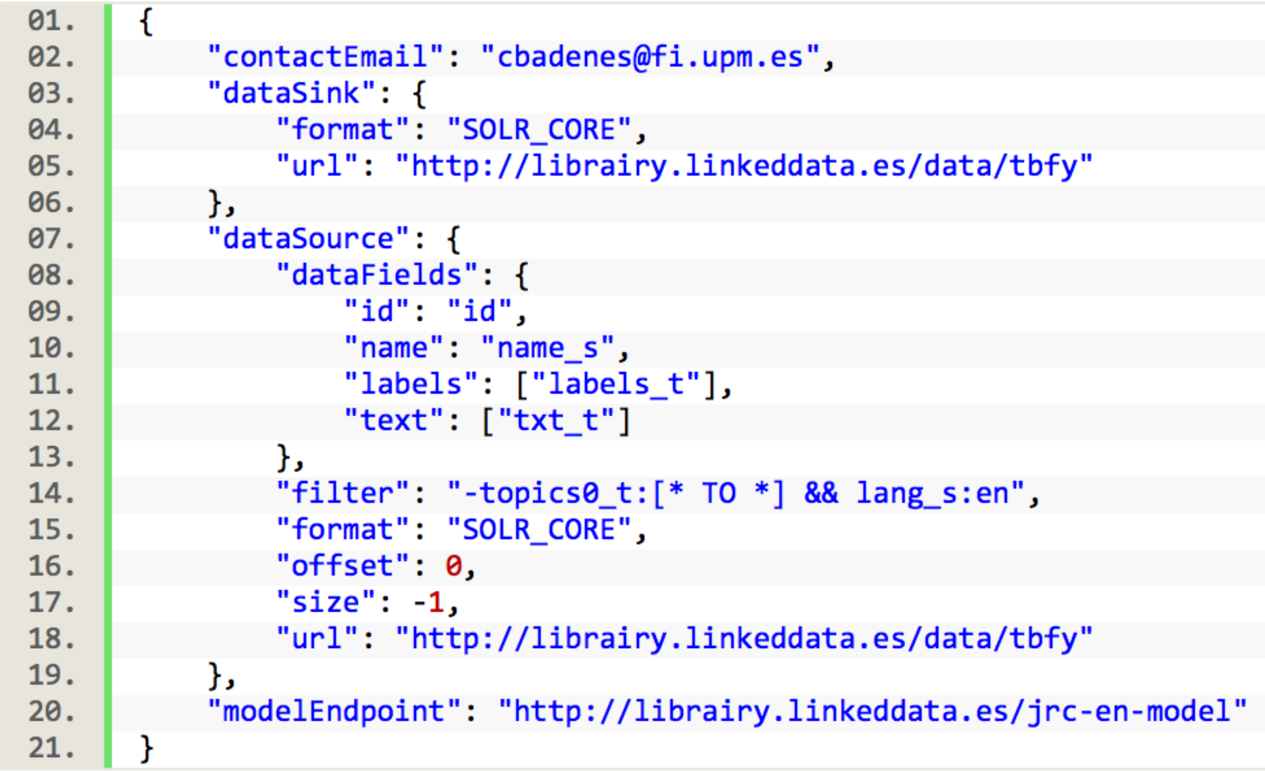
- The documents should be hosted in a ***Solr collection***

- The similarity calculation is based on these annotations.

- A HTTP_POST request to:

  http://librairy.linkeddata.es/api/annotations

json request



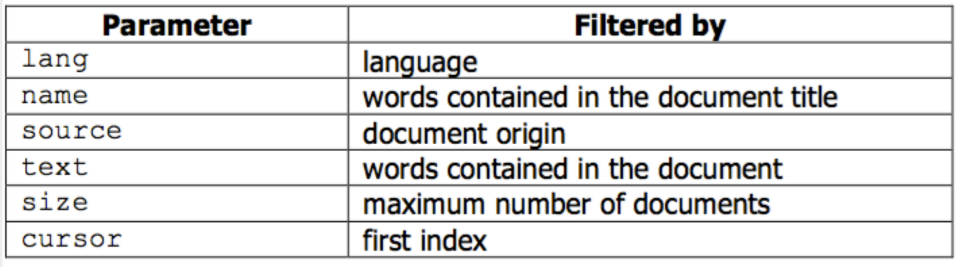
```
{
     "contactEmail": "cbadenes@fi.upm.es",
     "dataSink": {
         "format": "SOLR_CORE",
         "url": "http://librairy.linkeddata.es/data/tbfy"
     },
     "dataSource": {
         "dataFields": {
             "id": "id",
             "name": "name_s",
             "labels": ["labels_t"],
             "text": ["txt_t"]
         },
         "filter": "-topics0_t:[* TO *] && lang_s:en",
         "format": "SOLR_CORE",
         "offset": 0,
         "size": -1,
         "url": "http://librairy.linkeddata.es/data/tbfy"
     },
     "modelEndpoint": "http://librairy.linkeddata.es/jrc-en-model"
}
```

A3: Search Documents

- ***search-api** is the responsible for searching documents*
- Swagger documentation is available at:

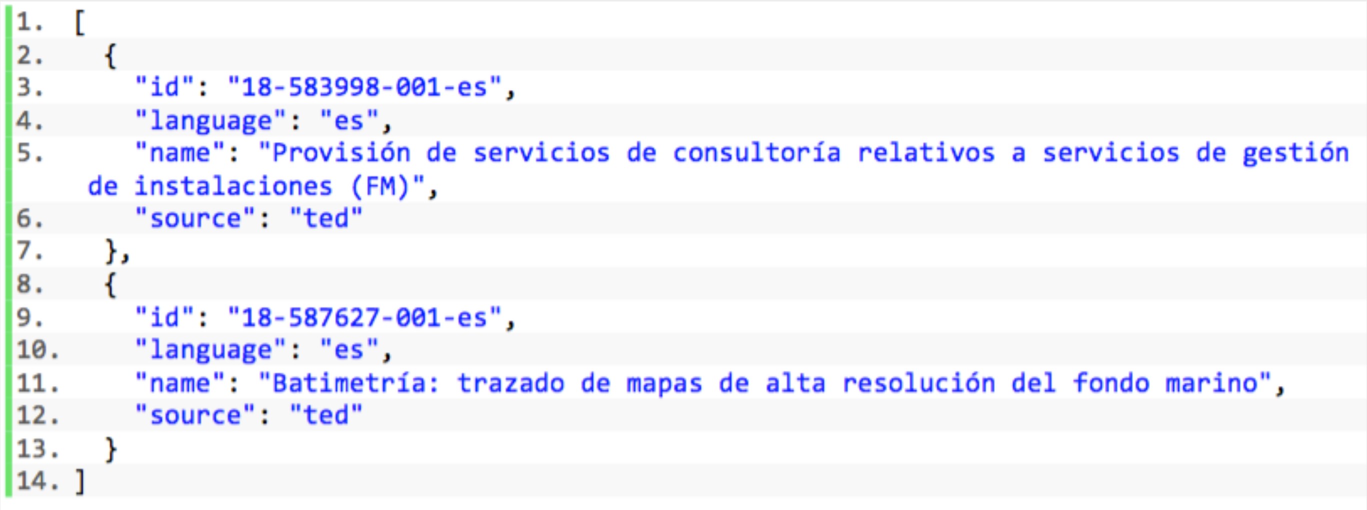
  http://tbfy.librairy.linkeddata.es/search-api

- Document retrieval can be done via HTTP_GET requests filtered by query parameters:



Parameter	Filtered by
lang	language
name	words contained in the document title
source	document origin
text	words contained in the document
size	maximum number of documents
cursor	first index

http://tbfy.librairy.linkeddata.es/search-api/documents?lang=es&size=2&source=ted



```
[
  {
    "id": "18-583998-001-es",
    "language": "es",
    "name": "Provisión de servicios de consultoría relativos a servicios de gestión de instalaciones (FM)",
    "source": "ted"
  },
  {
    "id": "18-587627-001-es",
    "language": "es",
    "name": "Batimetría: trazado de mapas de alta resolución del fondo marino",
    "source": "ted"
  }
]
```

http://tbfy.librairy.linkeddata.es/search-api/documents?lang=en&source=oo-api&name=system

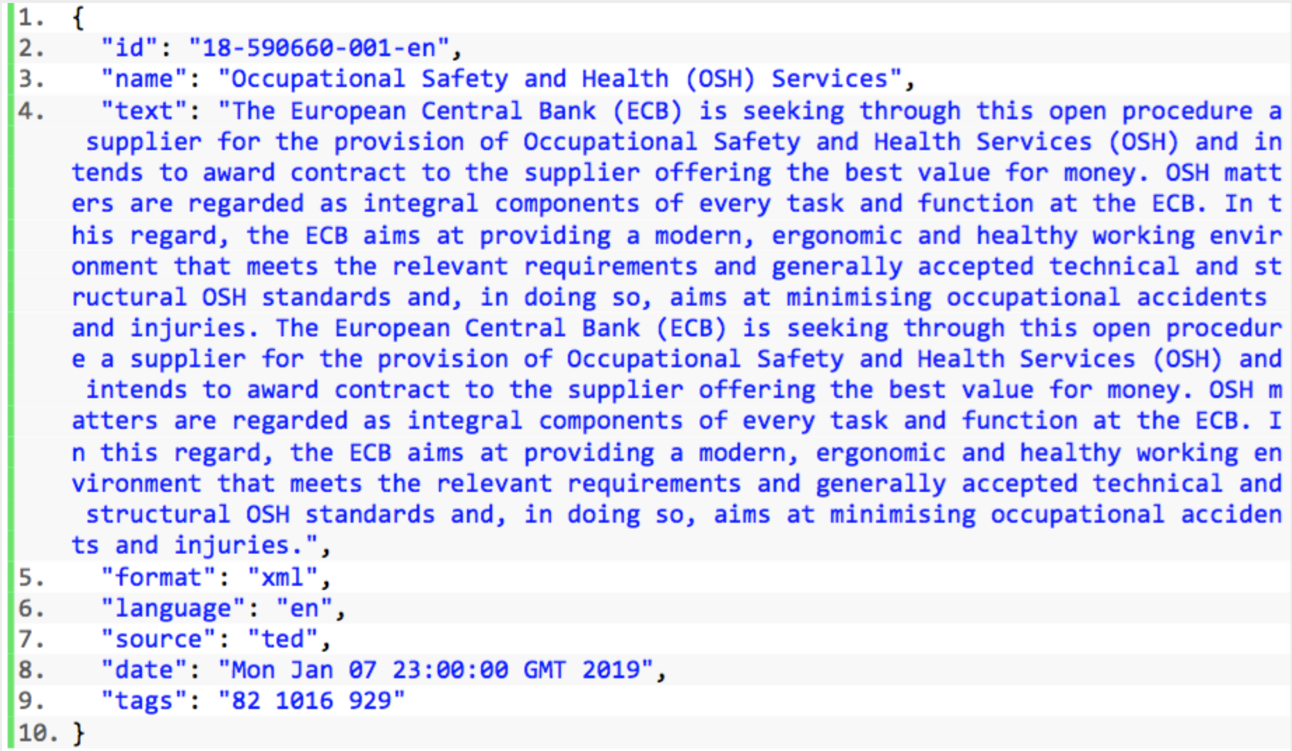
A4: Read Documents

- ***search-api** is the responsible for reading documents*

- Swagger documentation is available at: http://tbfy.librairy.linkeddata.es/search-api

- Following the Restful principles, a document can be read by making a HTTP_GET request to the URI containing the document identifier

http://tbfy.librairy.linkeddata.es/search-api/documents/18-590660-001-en

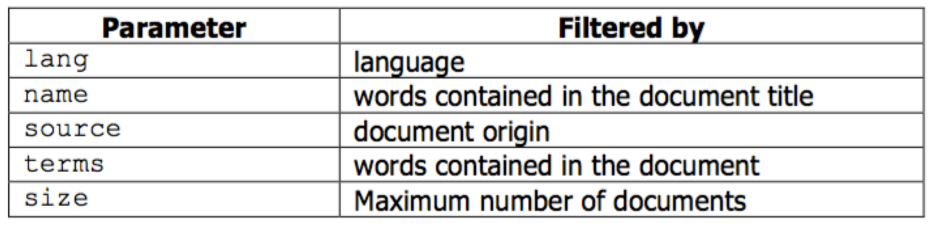


```
{
  "id": "18-590660-001-en",
  "name": "Occupational Safety and Health (OSH) Services",
  "text": "The European Central Bank (ECB) is seeking through this open procedure a supplier for the provision of Occupational Safety and Health Services (OSH) and intends to award contract to the supplier offering the best value for money. OSH matters are regarded as integral components of every task and function at the ECB. In this regard, the ECB aims at providing a modern, ergonomic and healthy working environment that meets the relevant requirements and generally accepted technical and structural OSH standards and, in doing so, aims at minimising occupational accidents and injuries. The European Central Bank (ECB) is seeking through this open procedure a supplier for the provision of Occupational Safety and Health Services (OSH) and intends to award contract to the supplier offering the best value for money. OSH matters are regarded as integral components of every task and function at the ECB. In this regard, the ECB aims at providing a modern, ergonomic and healthy working environment that meets the relevant requirements and generally accepted technical and structural OSH standards and, in doing so, aims at minimising occupational accidents and injuries.",
  "format": "xml",
  "language": "en",
  "source": "ted",
  "date": "Mon Jan 07 23:00:00 GMT 2019",
  "tags": "82 1016 929"
}
```

A5: Retrieve Similar Documents

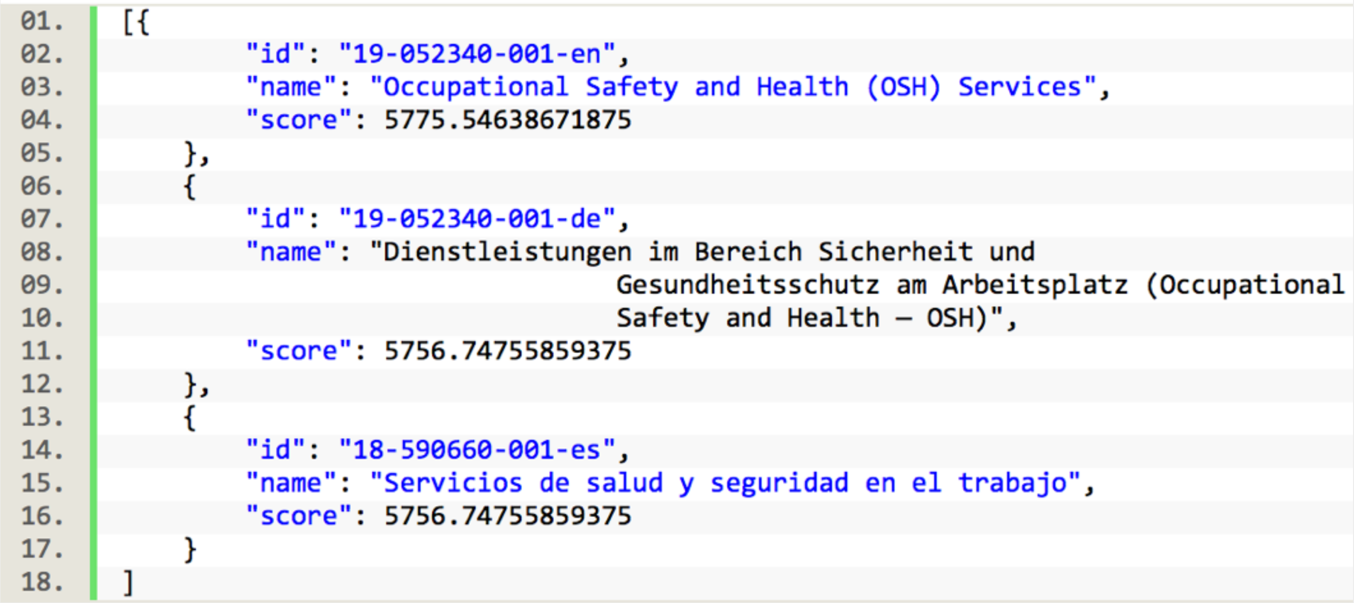
from a document

- ***search-api** is the responsible for retrieve similar documents*
- documents are sorted according to their content similarity to the reference
- the request can be filtered by:



Parameter	Filtered by
`lang`	language
`name`	words contained in the document title
`source`	document origin
`terms`	words contained in the document
`size`	Maximum number of documents

http://tbfy.librairy.linkeddata.es/search-api/documents/18-590660-001-en/items



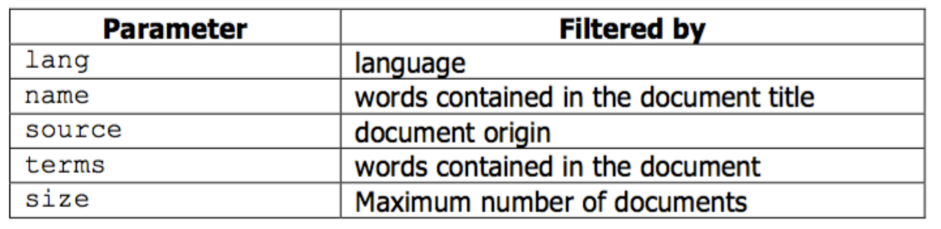
```
[{
        "id": "19-052340-001-en",
        "name": "Occupational Safety and Health (OSH) Services",
        "score": 5775.54638671875
    },
    {
        "id": "19-052340-001-de",
        "name": "Dienstleistungen im Bereich Sicherheit und
                             Gesundheitsschutz am Arbeitsplatz (Occupational
                             Safety and Health – OSH)",
        "score": 5756.74755859375
    },
    {
        "id": "18-590660-001-es",
        "name": "Servicios de salud y seguridad en el trabajo",
        "score": 5756.74755859375
    }
]
```

A6: Retrieve Similar Documents

from free-text

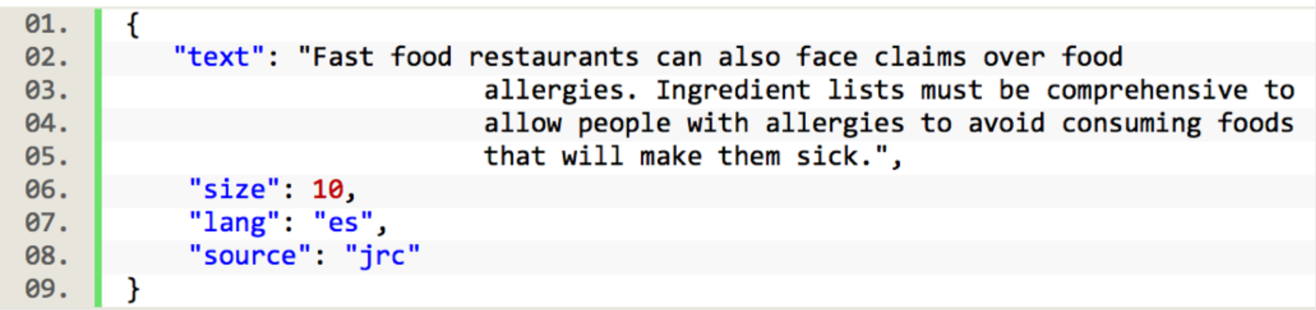
http://tbfy.librairy.linkeddata.es/search-api/items

- ***search-api** is the responsible for retrieve similar documents*
- Now a HTTP_POST request to http://tbfy.librairy.linkeddata.es/search-api/items



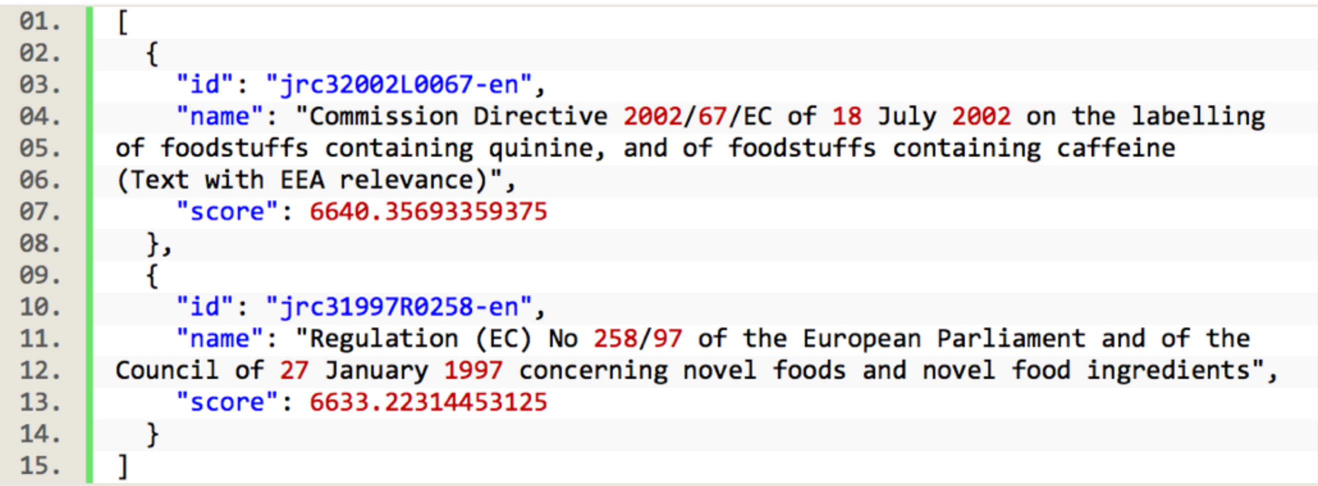
Parameter	Filtered by
`lang`	language
`name`	words contained in the document title
`source`	document origin
`terms`	words contained in the document
`size`	Maximum number of documents

JSON request



```
{
   "text": "Fast food restaurants can also face claims over food
                    allergies. Ingredient lists must be comprehensive to
                    allow people with allergies to avoid consuming foods
                    that will make them sick.",
    "size": 10,
    "lang": "es",
    "source": "jrc"
}
```

JSON response



```
[
  {
    "id": "jrc32002L0067-en",
    "name": "Commission Directive 2002/67/EC of 18 July 2002 on the labelling
of foodstuffs containing quinine, and of foodstuffs containing caffeine
(Text with EEA relevance)",
    "score": 6640.35693359375
  },
  {
    "id": "jrc31997R0258-en",
    "name": "Regulation (EC) No 258/97 of the European Parliament and of the
Council of 27 January 1997 concerning novel foods and novel food ingredients",
    "score": 6633.22314453125
  }
]
```

Next Steps

- develop new data-source ***adapters*** if required (e.g *Elasticsearch)*
- automatic documents ***uploading*** (daily from OpenOpps-API)
- improve ***quality*** of models (mainly french and german) for short texts
- create ***notebooks*** that show how:
  - to search/read documents about public procurement
  - to explore a multilingual collection of legal documents
  - to retrieve similar contracts/news/.. from a given text

Search-API

Carlos Badenes-Olmedo, Francisco Yedro, Oscar Corcho

Ontology Engineering Group (OEG)

Universidad Politécnica de Madrid (UPM)

Spain