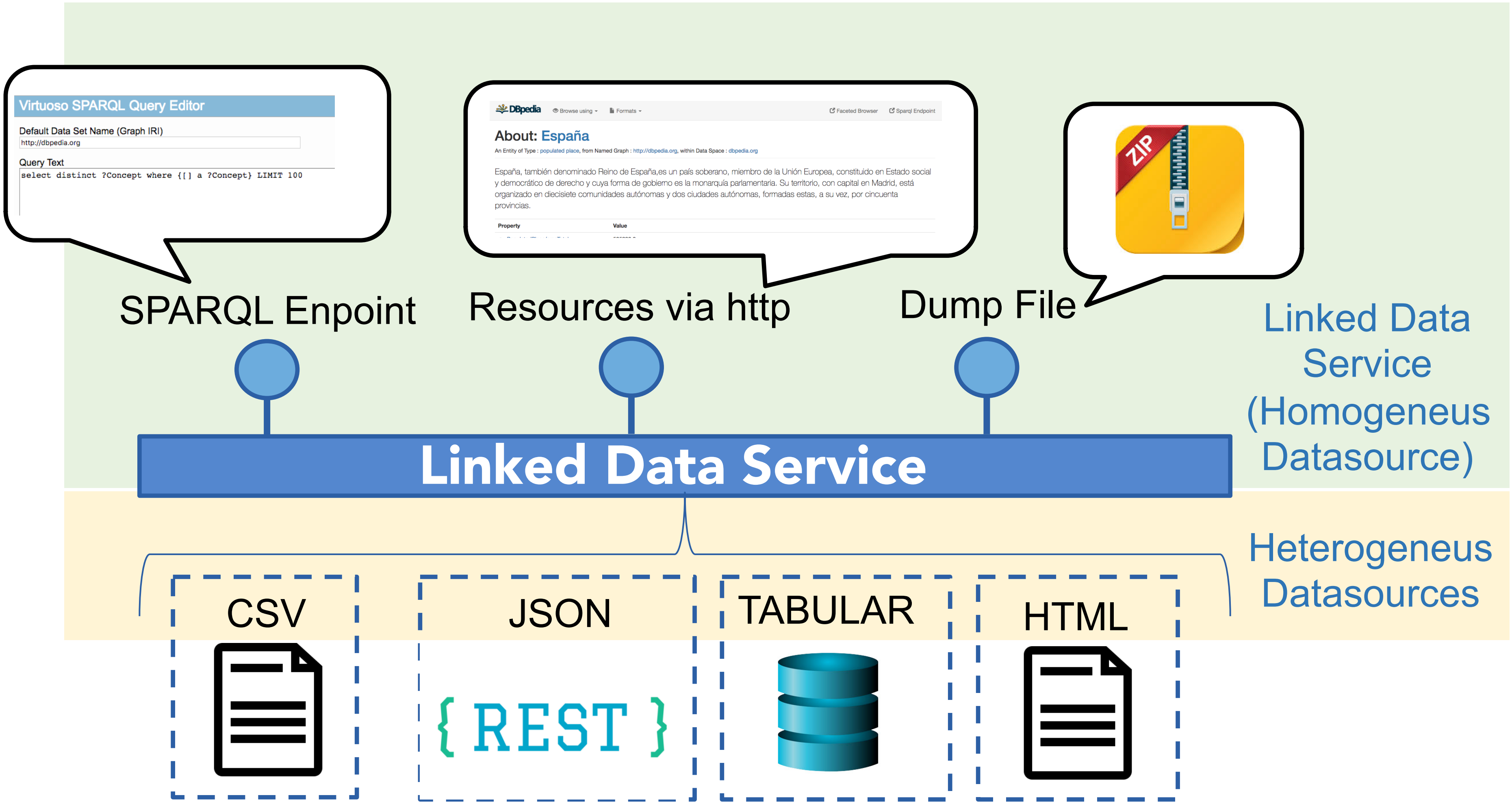
Helio

From Heterogeneus Data Sources to Link Data Services

Andrea Cimmino
Ontology Engineering Group
Universidad Politécnica de Madrid, Spain

✉ cimmino@fi.upm.es

From Heterogeneus Data Sources to Linked Data Services



Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)
http://dbpedia.org

Query Text

```
select distinct ?Concept where {[] a ?Concept} LIMIT 100
```

About: España

An Entity of Type : populated place, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

España, también denominado Reino de España,es un país soberano, miembro de la Unión Europea, constituido en Estado social y democrático de derecho y cuya forma de gobierno es la monarquía parlamentaria. Su territorio, con capital en Madrid, está organizado en diecisiete comunidades autónomas y dos ciudades autónomas, formadas estas, a su vez, por cincuenta provincias.

Property	Value

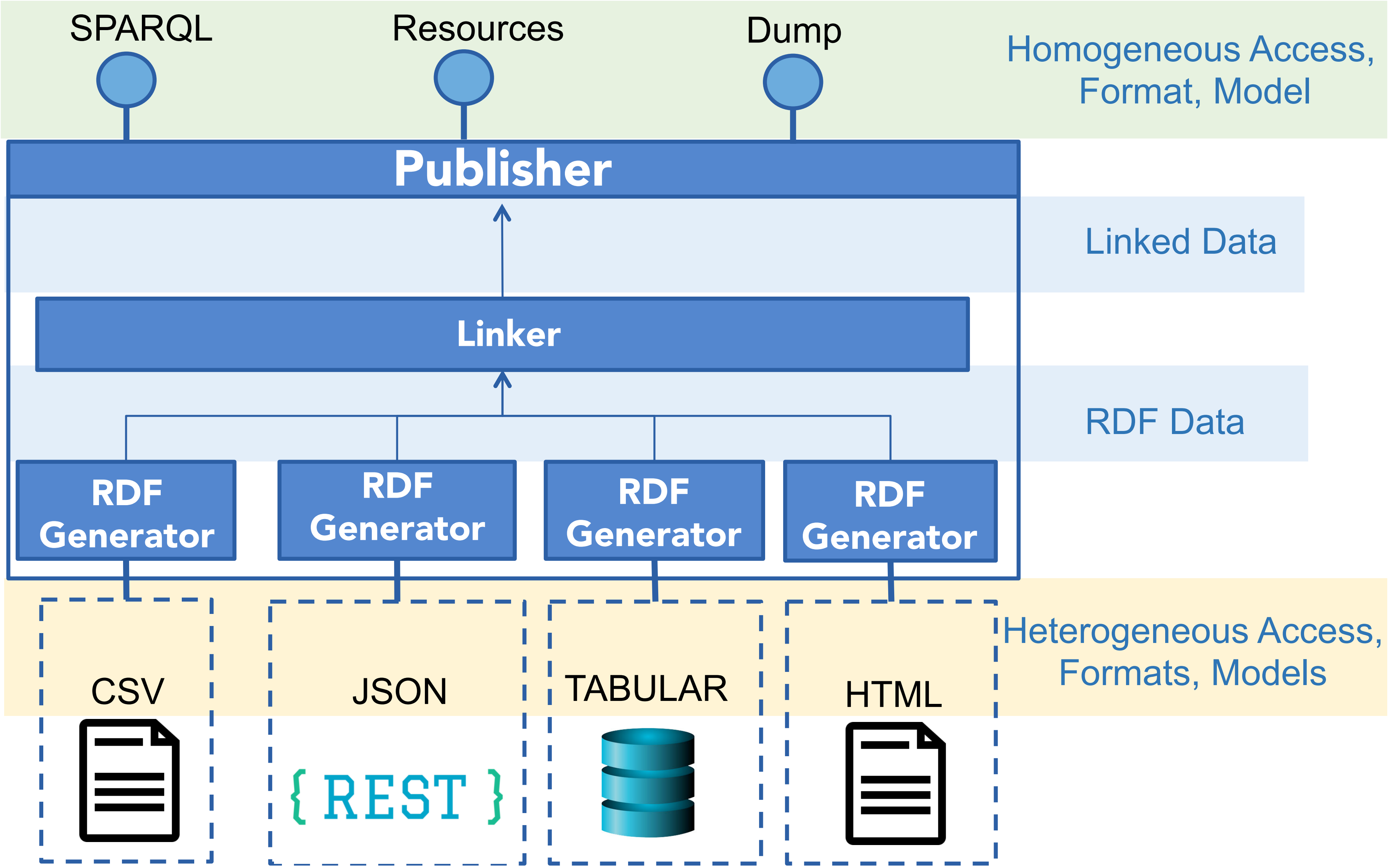
Challenges

- Cope with the different data sources
  - Access methods, i.e., API, file, DB
  - Formats, i.e., JSON, CSV, Tabular, HTML
  - Security, e.g., APIs with Oath, files with passwords
- Clean data
  - Lowercase, missing values
- Relate data
  - Interlink data from differente sources
- Publish as an RDF view the data
  - Enable a SPARQL endpoint
  - Allow resource access
  - Dump generation
- Others
  - Real-time data
  - API restrictions in the number of calls per day
  - Validation of published data

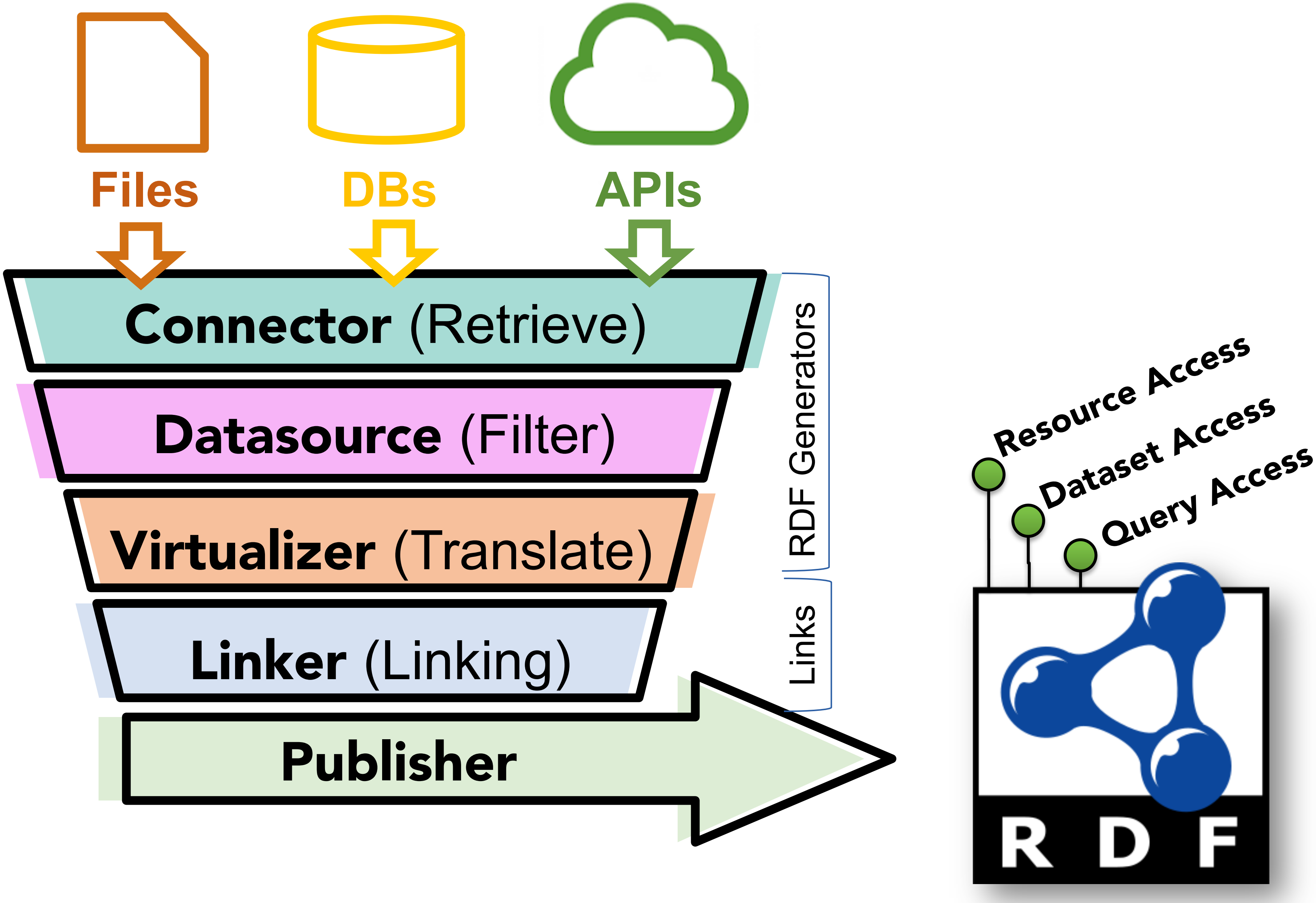
Index

1. **Helio Solution**
2. Use Cases + Challenges
3. Implementation
4. Helio Deployment Scenarios
5. Conclusions

Helio Approach

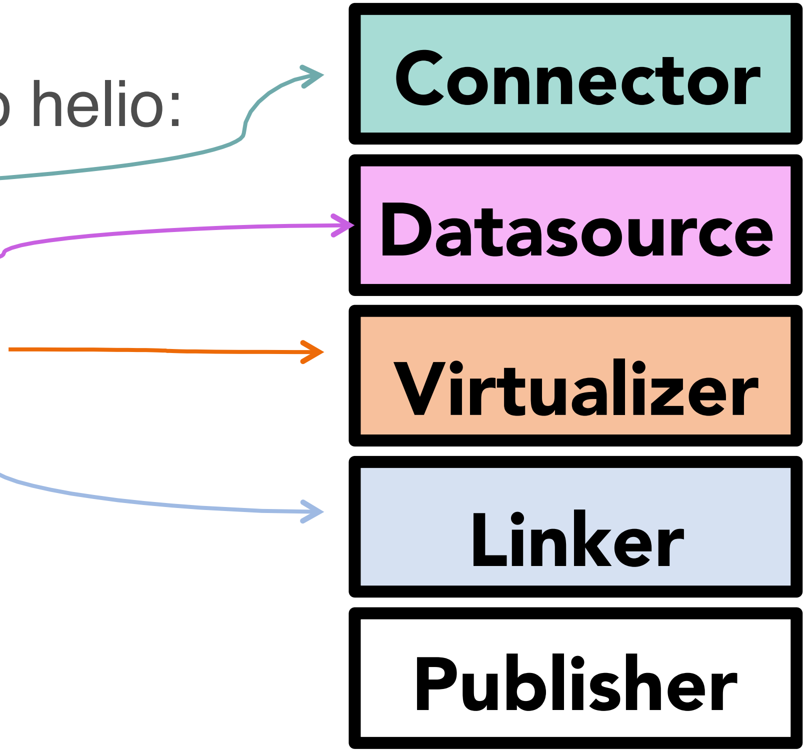


Helio Internal Pipeline



Using Helio

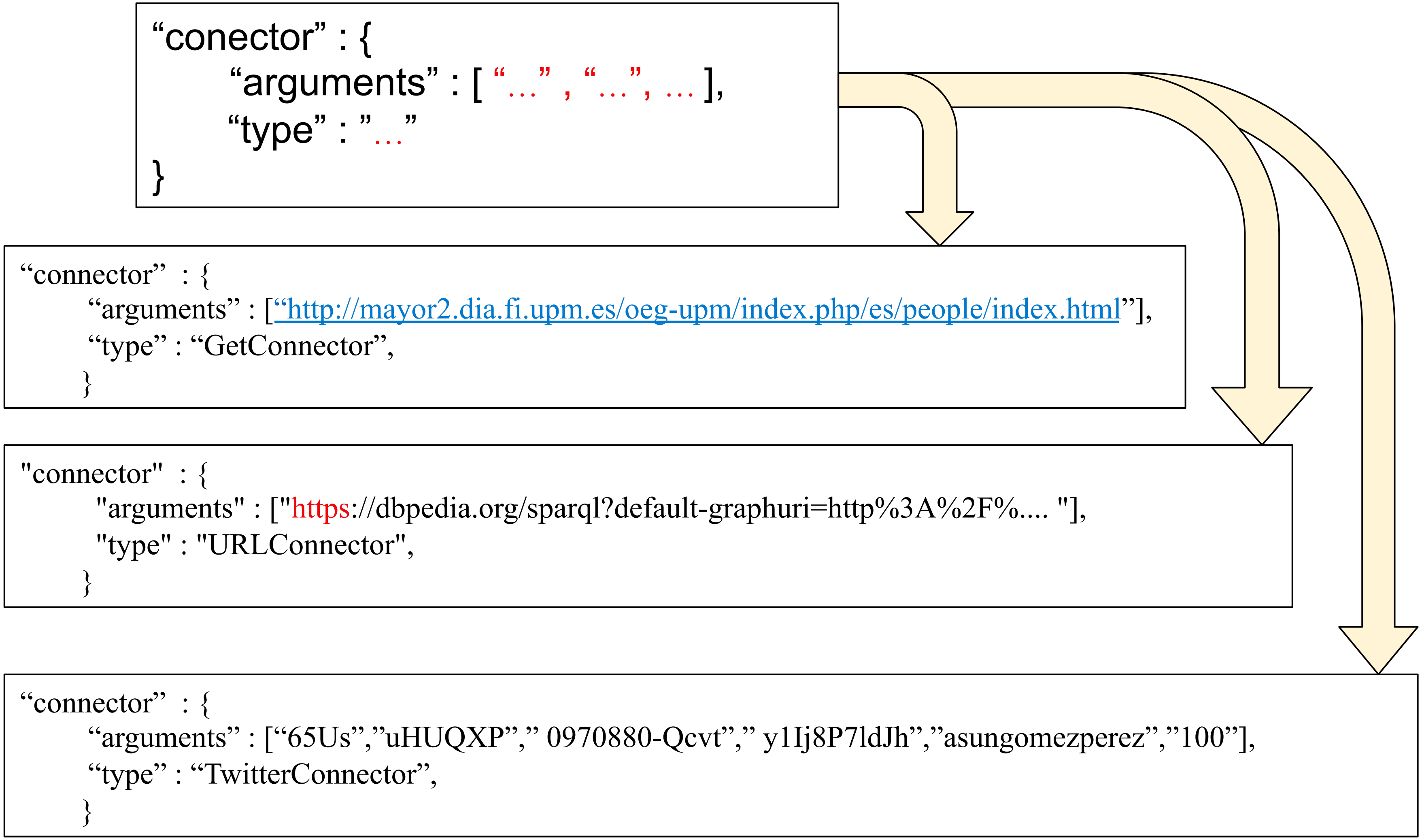
- Download the jar
- Write specifications to setup helio:
  - **Connectors**
  - **Datasources**
  - **Translation** rules (mappings)
  - **Linking rules**



- Run the jar → A service is published automatically

Helio Connector Specification

- Conector specification template



```
“conector” : {
        “arguments” : [ “…” , “…”, … ],
        “type” : ”…”
}
```

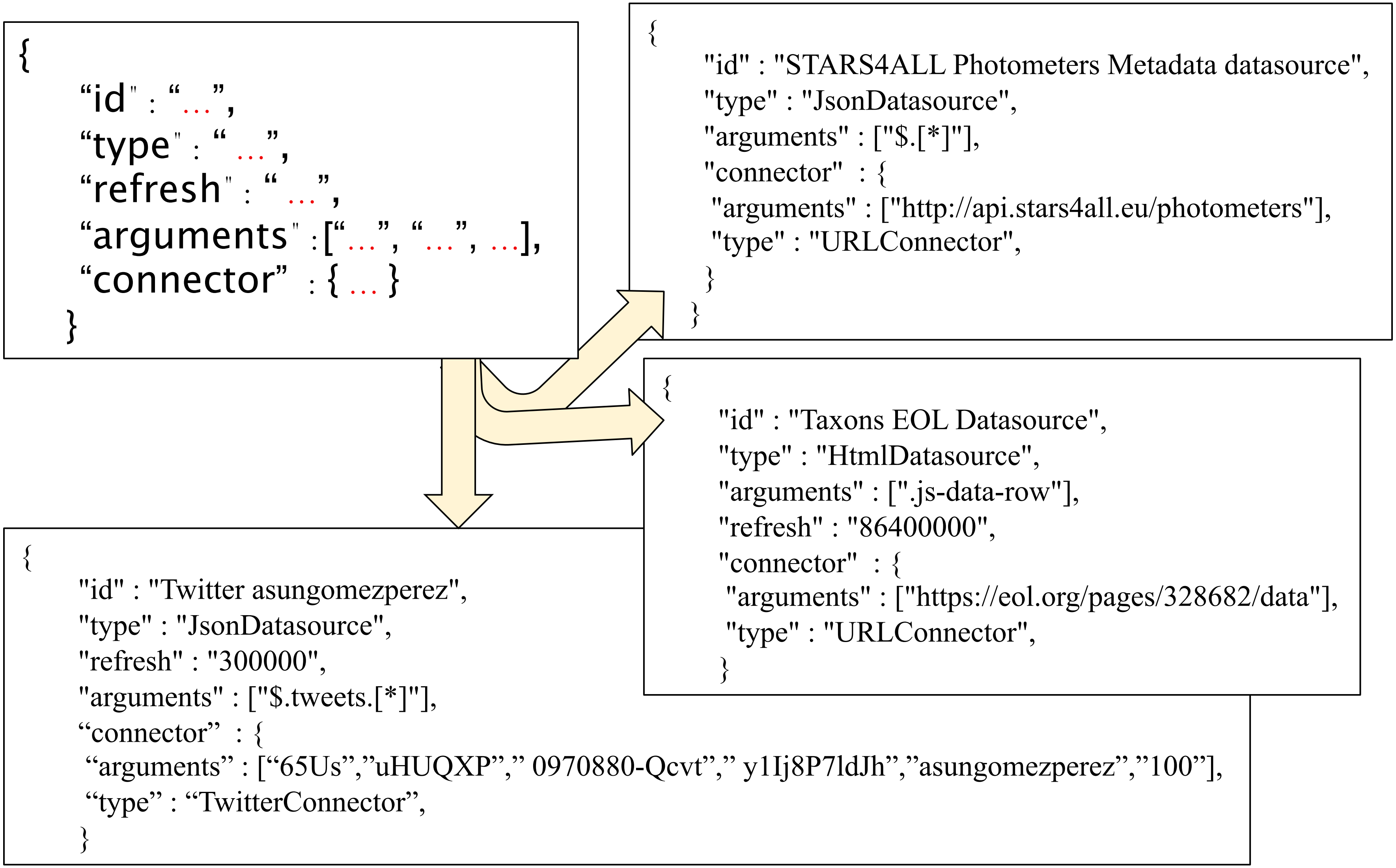
```
“connector”  : {
        “arguments” : [“http://mayor2.dia.fi.upm.es/oeg-upm/index.php/es/people/index.html”],
        “type” : “GetConnector”,
        }
```

```
"connector"  : {
         "arguments" : ["https://dbpedia.org/sparql?default-graphuri=http%3A%2F%.... "],
         "type" : "URLConnector",
        }
```

```
“connector”  : {
        “arguments” : [“65Us”,”uHUQXP”,” 0970880-Qcvt”,” y1Ij8P7ldJh”,”asungomezperez”,”100”],
        “type” : “TwitterConnector”,
        }
```

Helio Datasources Specification

- Datasource specification template



```
{
    “id" : “…”,
    “type" : “ …”,
    “refresh" : “ …”,
    “arguments" :[“…”, “…”, …],
    “connector” : { … }
  }
```

```
{
    "id" : "STARS4ALL Photometers Metadata datasource",
    "type" : "JsonDatasource",
    "arguments" : ["$.[*]"],
    "connector"  : {
     "arguments" : ["http://api.stars4all.eu/photometers"],
     "type" : "URLConnector",
    }
  }
```

```
{
    "id" : "Taxons EOL Datasource",
    "type" : "HtmlDatasource",
    "arguments" : [".js-data-row"],
    "refresh" : "86400000",
    "connector"  : {
     "arguments" : ["https://eol.org/pages/328682/data"],
     "type" : "URLConnector",
    }
```

```
{
    "id" : "Twitter asungomezperez",
    "type" : "JsonDatasource",
    "refresh" : "300000",
    "arguments" : ["$.tweets.[*]"],
    “connector”  : {
     “arguments” : [“65Us”,”uHUQXP”,” 0970880-Qcvt”,” y1Ij8P7ldJh”,”asungomezperez”,”100”],
     “type” : “TwitterConnector”,
    }
```

Helio Translation Rules Specification

- Translation rules template

```
{
    “id” : “…”,
    “datasource_ids : [“…”, “…”, …]
    “subject” : “…”,
    “properties” : [
        {
            “predicate” : “…”
            “object” : “…”,
            “is_literal” : ”True/False”,
             “datatype” : “…”,
             “lang” : “…”
        }
      ]
}
```

Helio Translation Rules Specification

```
{
    "id" : "STARS4ALL Photometers metadata",
    "datasource_ids" : ["STARS4ALL Photometers Metadata datasource"],
    "subject" : "http://helio.linkeddata.es/stars4all/photometers/{$.name}",
    "properties"  : [
        {
          "predicate" : "http://www.w3.org/2003/01/geo/wgs84_pos#location",
          "object" : "http://helio.linkeddata.es/stars4all/photometers/{$.name}/location",
          "is_literal" : "False"
        },{
          "predicate" : "http://schema.org/location",
       "object" : "http://helio.linkeddata.es/stars4all/locations/countries/[lower(regexp_replace(escapeHtml4(stripAccents({$.country})), '\\s+',
'_'))]",
          "is_literal" : "False"
        },{
          "predicate" : "http://schema.org/location",
          "object" : "http://helio.linkeddata.es/stars4all/locations/cities/[lower(regexp_replace(escapeHtml4(stripAccents({$.city})), '\\s+', '_'))]",
          "is_literal" : "False"
        },{
          "predicate" : "http://schema.org/location",
          "object" : "http://helio.linkeddata.es/stars4all/locations/places/[lower(regexp_replace(escapeHtml4(stripAccents({$.place})), '\\s+', '_'))]",
          "is_literal" : "False"
        }
```

Helio Linking Rules Specification

- Linking Rules template(s)

```
{
   “condition” : “…”,
   “predicates” : [“…”],
  “inverse_predicates” : [“…”],
   “source_resource_rule_id” : “…”,
   “target_resource_rule_id” : “…”
  }
```

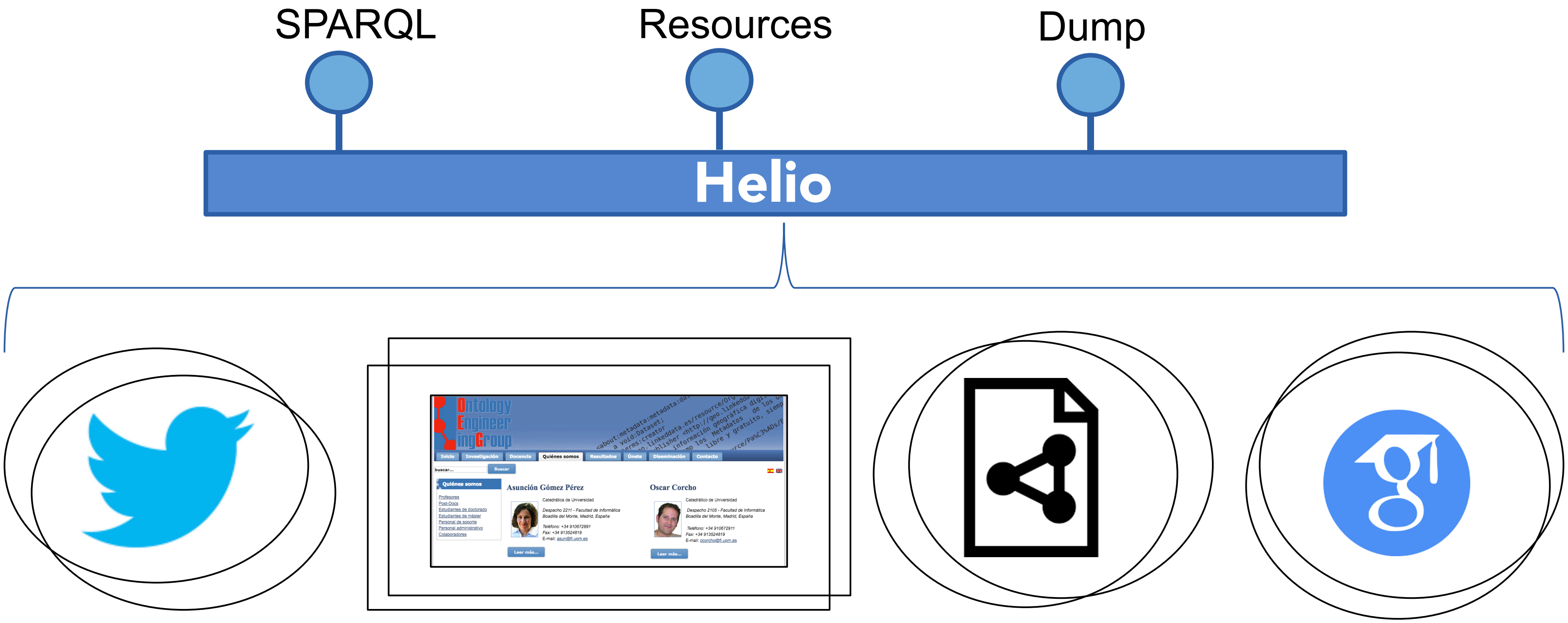
```
{
    "condition" : "cosine(stripAccents(T({$.name})),
                   stripAccents(trim(regexp_replace(replace(S({a.PostHeader}),'&nbsp;','’),'<br>.*',"))))) > 0.70",
    "predicates" : ["http://www.schema.org/tweets"],
    "inverse_predicates" : ["http://www.schema.org/writtenBy"],
    "source_resource_rule_id" : "OEG People",
    "target_resource_rule_id" : "UPM Tweets"
  }
```

Index

1. Helio Solution
2. **Use Cases + Challenges**
3. Implementation
4. Helio deployment scenarios
5. Conclusions

OEG People Use Case

	Connector	Datasource	Translation	Linking
Challenges	✓	✓	✓	✓



SPARQL

Resources

Dump

Helio

Tweets of OEG Members

OEG Members

Static data of OEG Group

Scholar indexes

Connector Challenge

- Twitter API requires credentials
  - Our connector passes them as argument in the specification

```
{
        "id" : "Twitter asungomezperez",
        "type" : "JsonDatasource",
        "refresh" : "300000",
        "arguments" : ["$.tweets.[*]"],
        "connector"  : {
         "arguments" :["65UsI12RvUVH”,"uHUQcp9YXP”, "100880-QcvtT3”,
                                     ”o4SZmiRfTh6","asungomezperez","100"],
         "type" : "TwitterConnector",
        }
      }
```

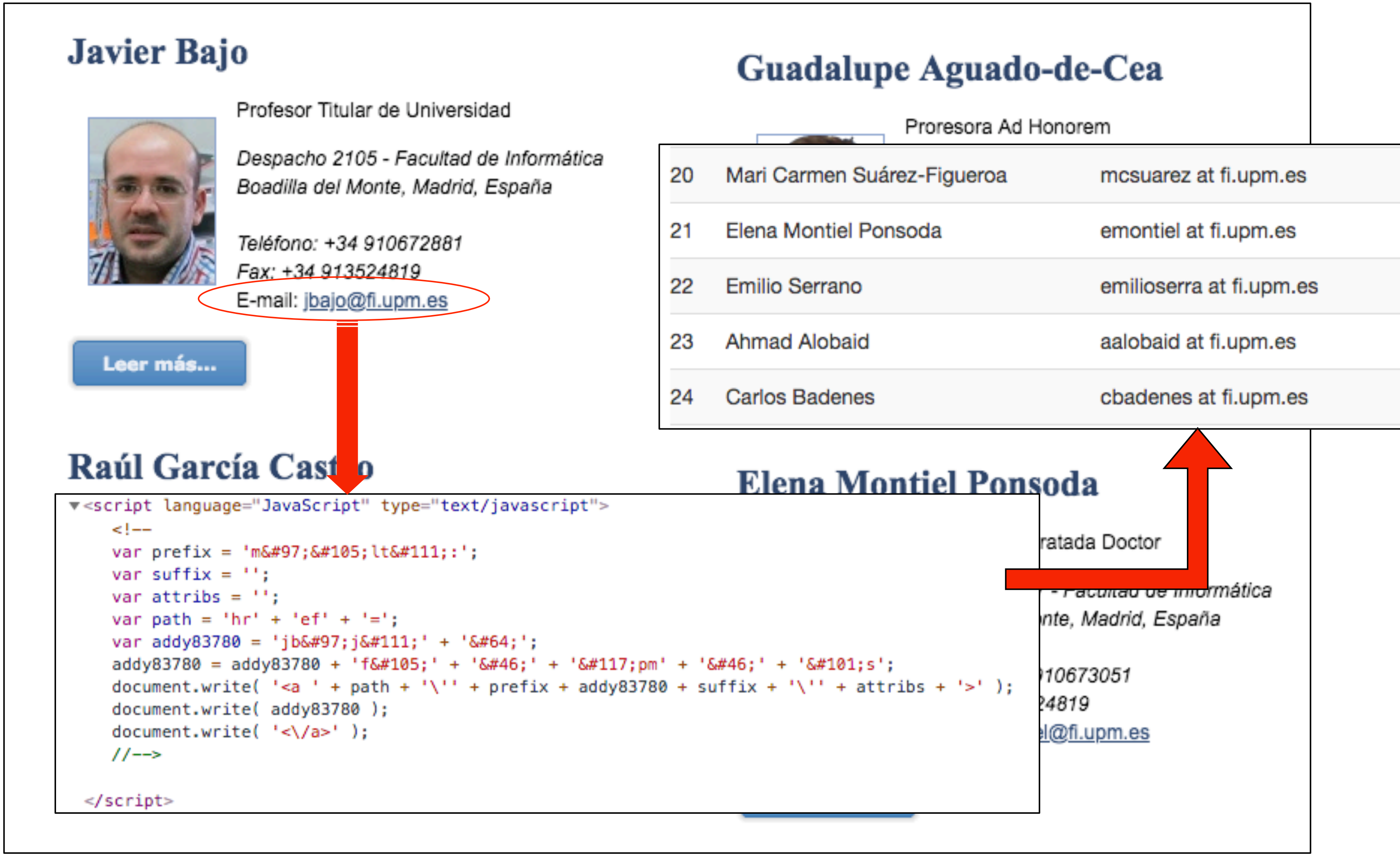
Datasource Challenge

- Twitter API has a limitation of the number of calls
  - Our specification updates the data asyncronously from user requests

```
{
       "id" : "Twitter asungomezperez",
       "type" : "JsonDatasource",
       "refresh" : "300000",
       "arguments" : ["$.tweets.[*]"],
       "connector"  : {
        "arguments" :["65UsI12RvUVH”,"uHUQcp9YXP”, "100880-QcvtT3”,
                              ”o4SZmiRfTh6","asungomezperez","100"],
        "type" : "TwitterConnector",
       }
     }
```

Translation Challenge

- Data cleaning and transformation [link]



Javier Bajo

Profesor Titular de Universidad

Despacho 2105 - Facultad de Informática
Boadilla del Monte, Madrid, España

Teléfono: +34 910672881
Fax: +34 913524819
E-mail: jbajo@fi.upm.es

Leer más...

Guadalupe Aguado-de-Cea

Proresora Ad Honorem

20	Mari Carmen Suárez-Figueroa	mcsuarez at fi.upm.es
21	Elena Montiel Ponsoda	emontiel at fi.upm.es
22	Emilio Serrano	emilioserra at fi.upm.es
23	Ahmad Alobaid	aalobaid at fi.upm.es
24	Carlos Badenes	cbadenes at fi.upm.es

Raúl García Cast o

Elena Montiel Ponsoda

ratada Doctor

- Facultad de Informática
nte, Madrid, España

10673051
24819
l@fi.upm.es

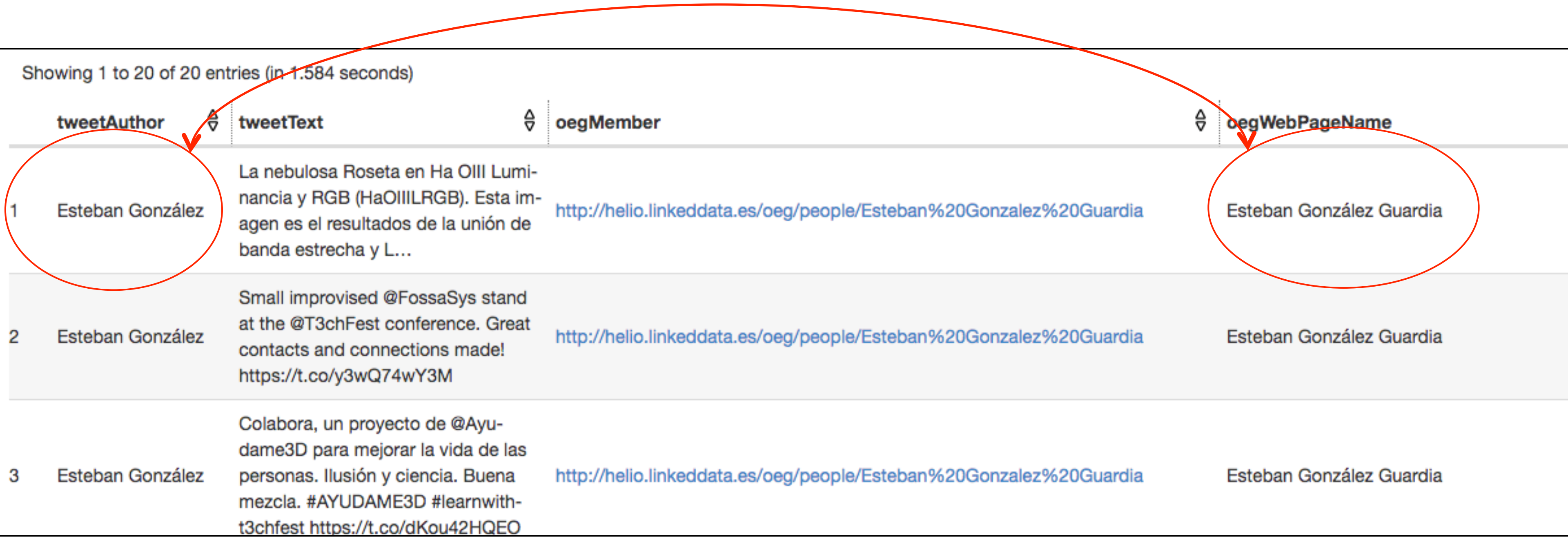
```
▼<script language="JavaScript" type="text/javascript">
    <!--
    var prefix = 'm&#97;&#105;lt&#111;:';
    var suffix = '';
    var attribs = '';
    var path = 'hr' + 'ef' + '=';
    var addy83780 = 'jb&#97;j&#111;' + '&#64;';
    addy83780 = addy83780 + 'f&#105;' + '&#46;' + '&#117;pm' + '&#46;' + '&#101;s';
    document.write( '<a ' + path + '\'' + prefix + addy83780 + suffix + '\'' + attribs + '>' );
    document.write( addy83780 );
    document.write( '<\/a>' );
    //-->

  </script>
```

Linking Challenge

- Relate the author name in a tweet with his/her name in the OEG web [link]
- **In addition we had to cope with Twitter API request limitations**

Equality of names does not solve this problem, Fuzzy rules required

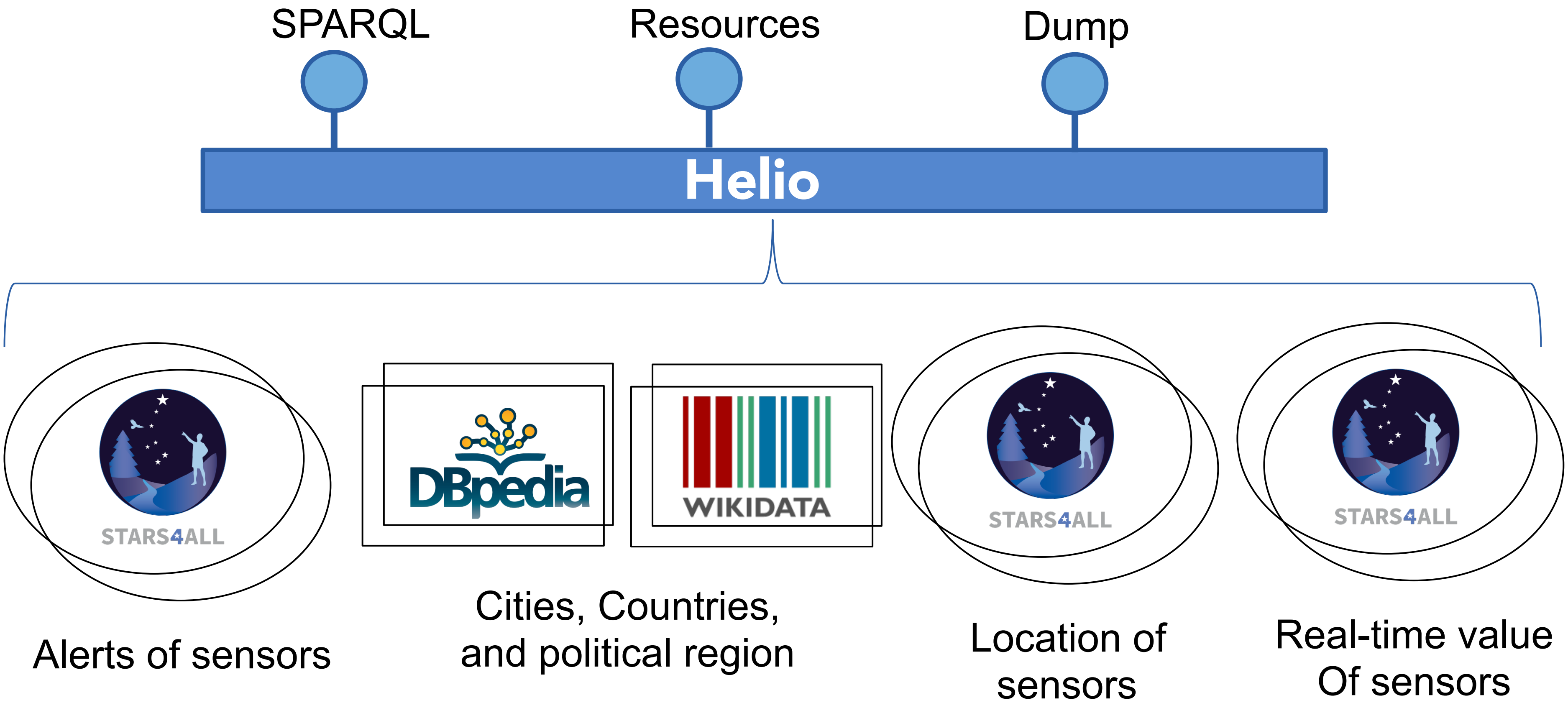


Showing 1 to 20 of 20 entries (in 1.584 seconds)

	tweetAuthor	tweetText	oegMember	oegWebPageName
1	Esteban González	La nebulosa Roseta en Ha OIII Luminancia y RGB (HaOIIILRGB). Esta imagen es el resultados de la unión de banda estrecha y L...	http://helio.linkeddata.es/oeg/people/Esteban%20Gonzalez%20Guardia	Esteban González Guardia
2	Esteban González	Small improvised @FossaSys stand at the @T3chFest conference. Great contacts and connections made! https://t.co/y3wQ74wY3M	http://helio.linkeddata.es/oeg/people/Esteban%20Gonzalez%20Guardia	Esteban González Guardia
3	Esteban González	Colabora, un proyecto de @Ayudame3D para mejorar la vida de las personas. Ilusión y ciencia. Buena mezcla. #AYUDAME3D #learnwitht3chfest https://t.co/dKou42HQEO	http://helio.linkeddata.es/oeg/people/Esteban%20Gonzalez%20Guardia	Esteban González Guardia

Stars4ALL Use Case

	Connector	Datasource	Translation	Linking
Challenges	✗	✗	✓	✓



Alerts of sensors

Cities, Countries, and political region

Location of sensors

Real-time value Of sensors

Translation challenge

- One field in JSON contained more than one RDF property

```
"name” : "stars001 - Coslada, Spain”
```

```
{
        "predicate" : "http://stars4all.es/ontology#referTo",
         "object" : "http://helio.linkeddata.es/stars4all/photometers/[trim(regexp_replace({$.name}, '-.*', "))]",
         "is_literal" : "False"
       }
```

```
{
        "predicate" : "http://schema.org/name",
        "object" : "[trim(regexp_replace(regexp_replace({$.name}, '.*\\s+-', "), '.*,', "))]",
        "is_literal" : "True"
       }
```

```
{
        "predicate" : "http://schema.org/name",
        "object" : "[trim(regexp_replace(regexp_replace({$.name}, '.*\\s+-', "), ',.*', "))]",
        "is_literal" : "True"
       }
```

Linking Challengee

- Cities & Countries were tricky to [link]

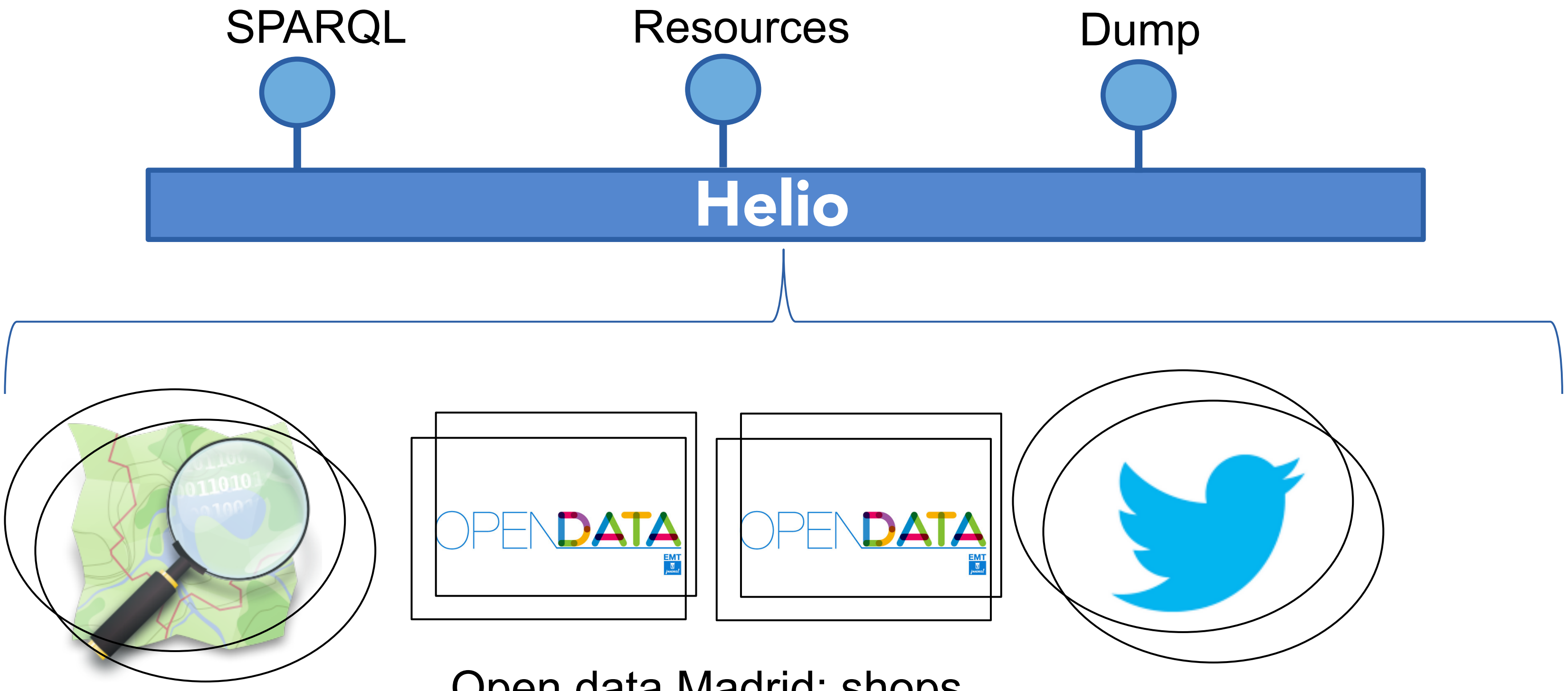
```
"name" : "stars001 - Coslada, España"
```

```
"country":"España","city":"Coslada","place":"Coslada",
```

```
"relationships": [
   {
    "condition" : "levenshtein(lower(regexp_replace(trim(escapeHtml4(stripAccents(regexp_replace(
                  regexp_replace(S({$.name}), '.*\\s+-', "), ',.*', ")))), '\\s+', '_')),
                                  lower(regexp_replace(escapeHtml4(stripAccents(T({$.city}))), '\\s+', '_'))) < 3",
    "predicates" : ["http://www.w3.org/2002/07/owl#sameAs"],
    "inverse_predicates" : ["http://www.w3.org/2002/07/owl#sameAs"],
    "source_resource_rule_id" : "STARS4ALL Alerts Cities metadata",
    "target_resource_rule_id" : "STARS4ALL Cities metadata"
   }
 ]
```

Climathron Use Case

	Connector	Datasource	Translation	Linking
Challenges	✓	✓	✗	✓



SPARQL

Resources

Dump

Helio

Street Lamps

Open data Madrid: shops, bus stops (without downloading)

Twitter

Connector Challenge

- Twitter API required credentials
  - Our connector passes them as argument in the specification

```
{
      "id" : "Twitter climapatron",
      "type" : "JsonDatasource",
      "arguments" : ["$.tweets.[*]"],
      "connector"  : {
       "arguments" :
["65UsIe34FvDoT","uHmwcp9YXP","100434VjLypDt",”oZmiRfTh6","climapatron","
100"],
       "type" : "TwitterConnector",
      }
    }
```

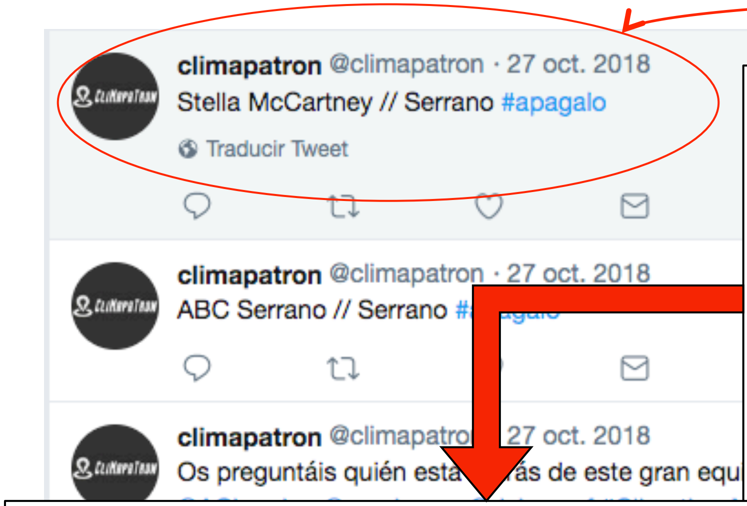
Datasource Challenge

- OpenStreet Map have a limitation for the number of calls
  - Our specification updates the data asyncronously from user requests

```
{
        "id" : "OpenStreetMaps Lamps Datasource",
        "type" : "XmlDatasource",
        "refresh" : "36000000",
        "arguments" : ["//node"],
        "connector"  : {
         "arguments" : ["https://www.overpass-api.de/api/interpreter?
data=[out:xml];node[highway=street_lamp]
(40.1497785,-4.1736937,40.6159541,-3.2877552);out%20meta;"],
         "type" : "GetConnector",
        }
      }
```

Linking Challengee

- Some Tweets create relationships in the dataset



climapatron @climapatron · 27 oct. 2018
Stella McCartney // Serrano #apagalo
Traducir Tweet

climapatron @climapatron · 27 oct. 2018
ABC Serrano // Serrano #apagalo

climapatron @climapatron · 27 oct. 2018
Os preguntáis quién esta ás de este gran equ

```
{
    "condition" :
"(levenshtein(trim(regexp_replace(regexp_replace(S({$.text}),'^[^#]
+',''),'[^/]+[/]+','')), '#apagalo') < 1) AND
(cosine(regexp_replace(regexp_replace(regexp_replace(S({$.text}),'
@[^\\s]+\\s+',''),'#.+',''),'[/]+.+',''), T({//basicData//name})) > 0.4)
AND
(cosine(regexp_replace(regexp_replace(regexp_replace(S({$.text}),'
@[^\\s]+\\s+',''),'#.+',''),'[^/]+[/]+',''),T({//geoData//address})) > 0.4)",
redicates" : ["http://www.schema.org/light-overkill", "http://
schema.org/isrelatedto"],
ource_resource_rule_id" : "Climapathron Tweets",
rget_resource_rule_id" : "Tiendas madrid"
```

```
SELECT DISTINCT ?text ?shopName ?address {

 ?tweet sch:light-overkill ?shop .
 ?tweet sch:text ?text .
 ?shop sch:legalName ?shopName .
 ?shop sch:address ?address .
}
```

```
  "text": { "type": "literal" , "value": "Stella McCartney // Serrano #apagalo" } ,
  "shopName": { "type": "literal" , "value": "Stella McCartney" } ,
  "address": { "type": "literal" , "value": "Serrano, 62" }
 } ,
```

Other Use Cases

- Taxons:
  - Sources: Wikidata, custom csv, patheria csv
  - Challenges: Linking
- VICINITY:
  - Sources: Google weather, RDF files about sensors, Helio Stars4all
  - Challenges: translation and linking

Index

1. Helio Solution
2. Use Cases + Challenges
3. **Implementation**
4. Helio deployment scenarios
5. Conclusions

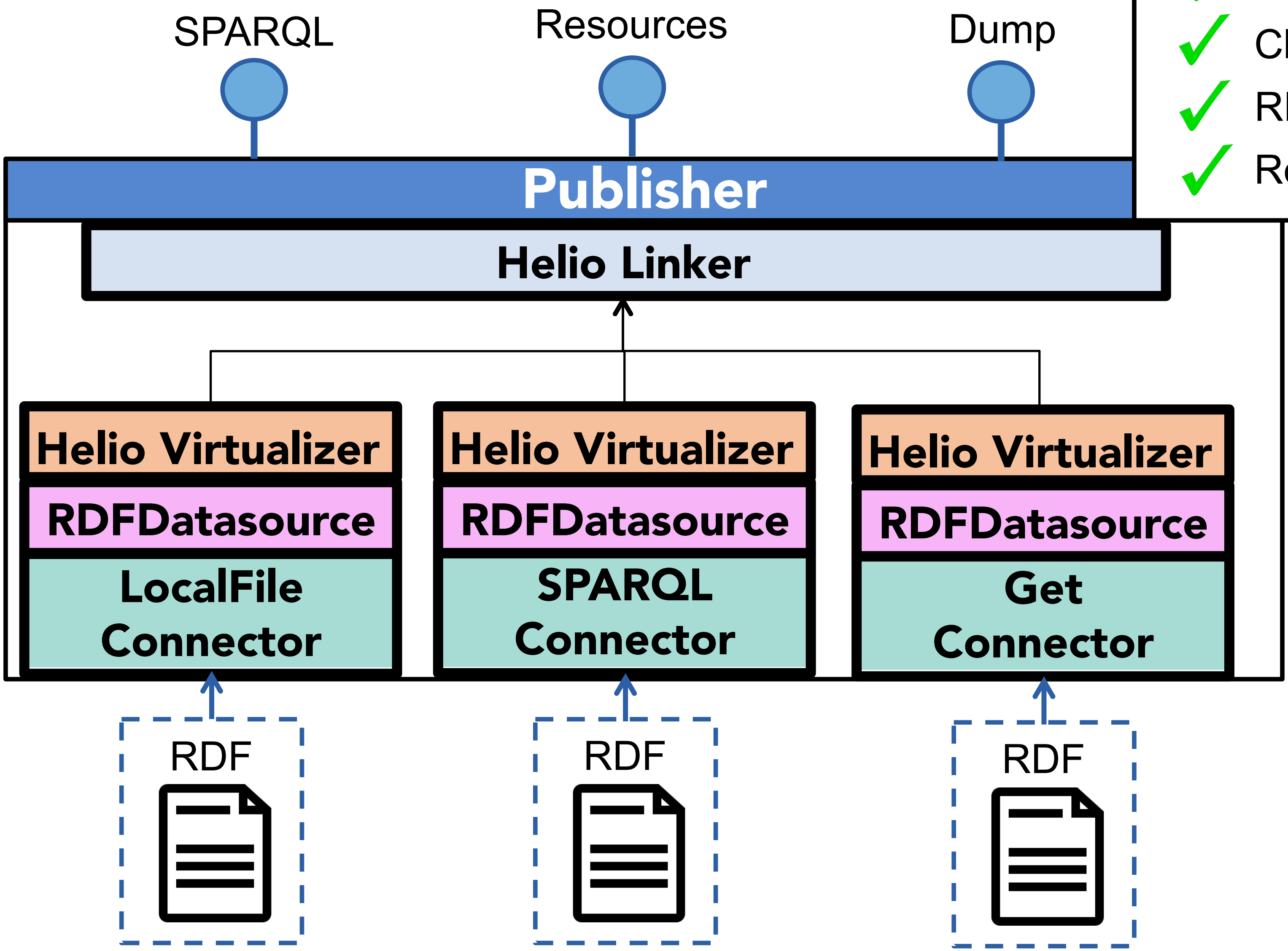
Helio current status

Specification	Elements	Implemented
Helio Specification	Connector	FederatedSparqlConnector, GetConnector, LocalFileConnector, TwitterConnector, URLConnector
	Datasource	CsvDatasource, HtmlDatasource, JsonDatasource, RDFDatasource, TextDatasource, XmlDatasource
	Translator	Helio Virtualizer
	Linking	Helio Linker
RML* (proof of concept)	Connector	LocalFileConnector
	Datasource	JsonDatasource
	Translator	Helio Virtualizer
	Linking	Helio Linker

Index

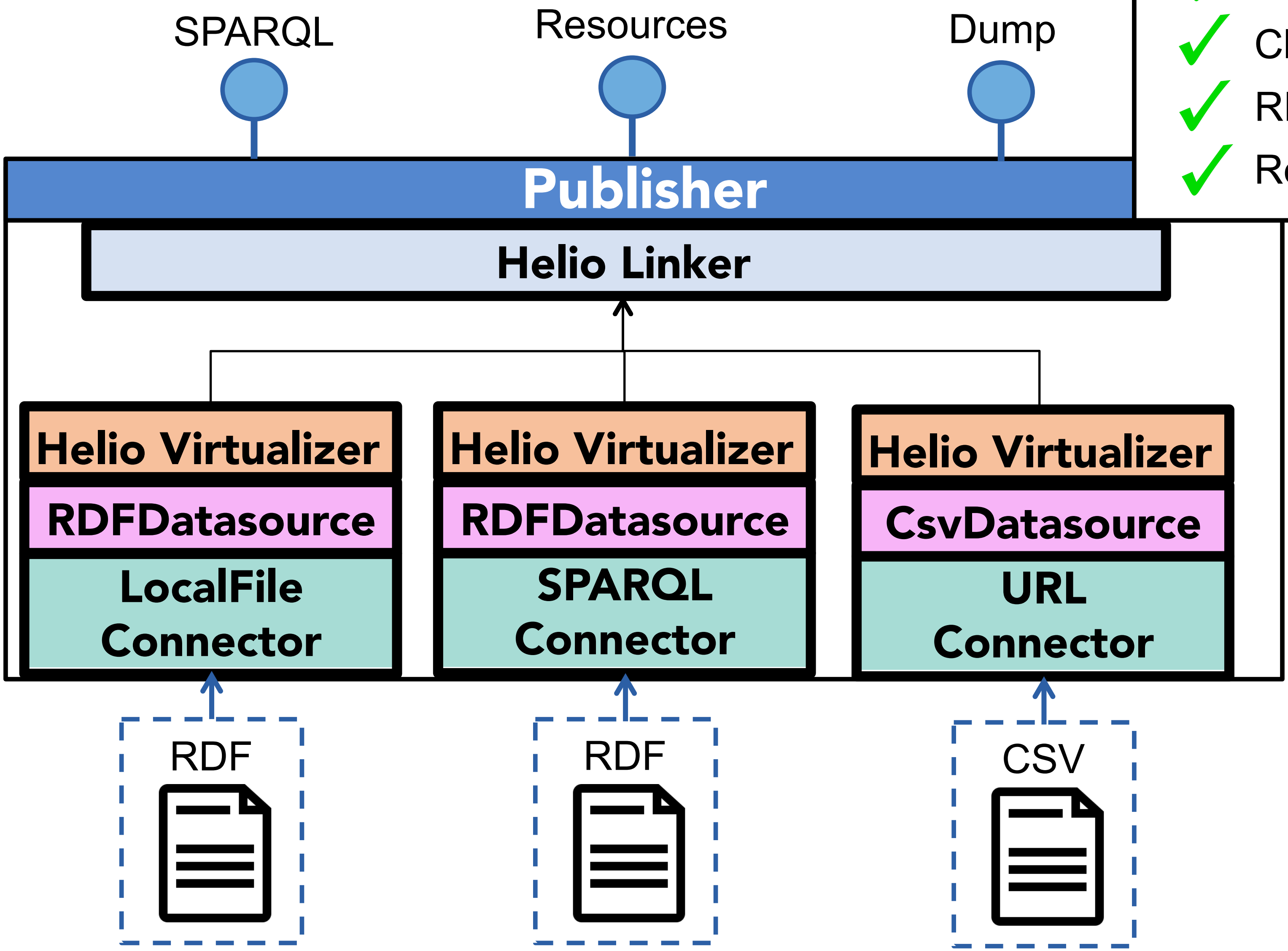
1. Helio Solution
2. Use Cases + Challenges
3. Implementation
4. **Helio deployment scenarios**
5. Conclusions

Helio Scenario 1: Refining RDF



- ✓ Linking
- ✓ Cleaning
- ✓ RDF generation
- ✓ Real Time data

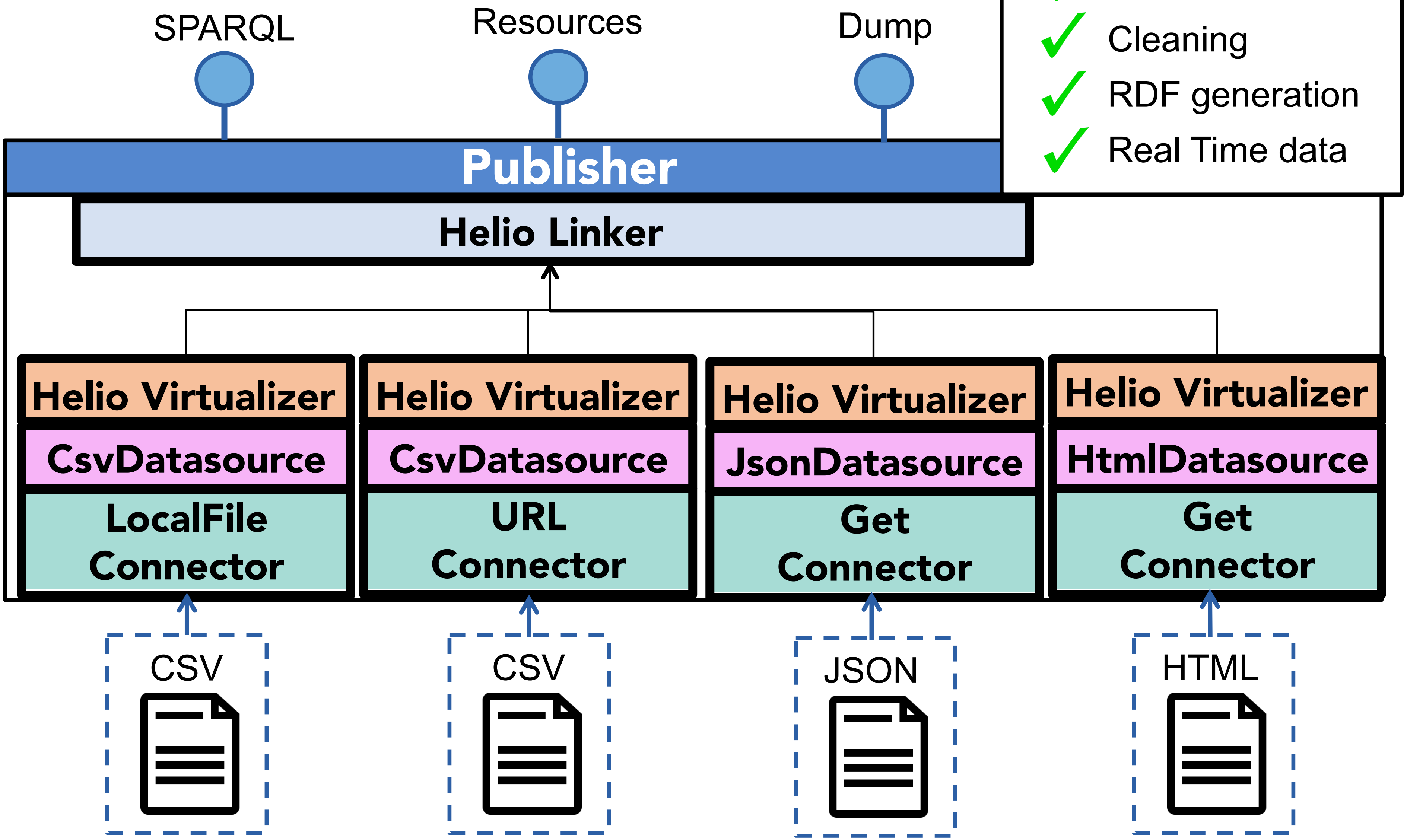
Helio Scenario 2: Enhancing RDF



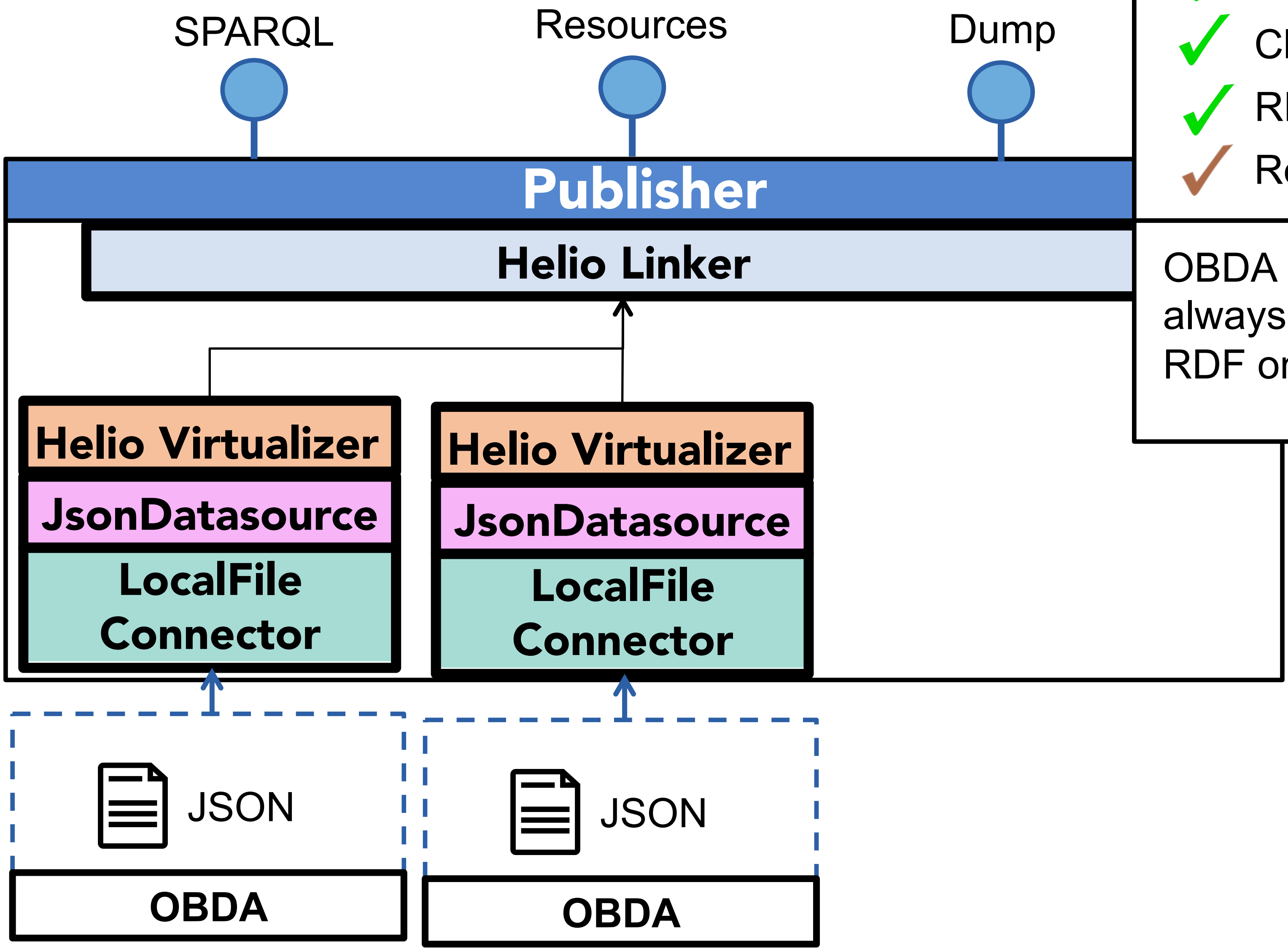
- ✓ Linking
- ✓ Cleaning
- ✓ RDF generation
- ✓ Real Time data

Helio Scenario 3: Data Integration

- ✓ Linking
- ✓ Cleaning
- ✓ RDF generation
- ✓ Real Time data



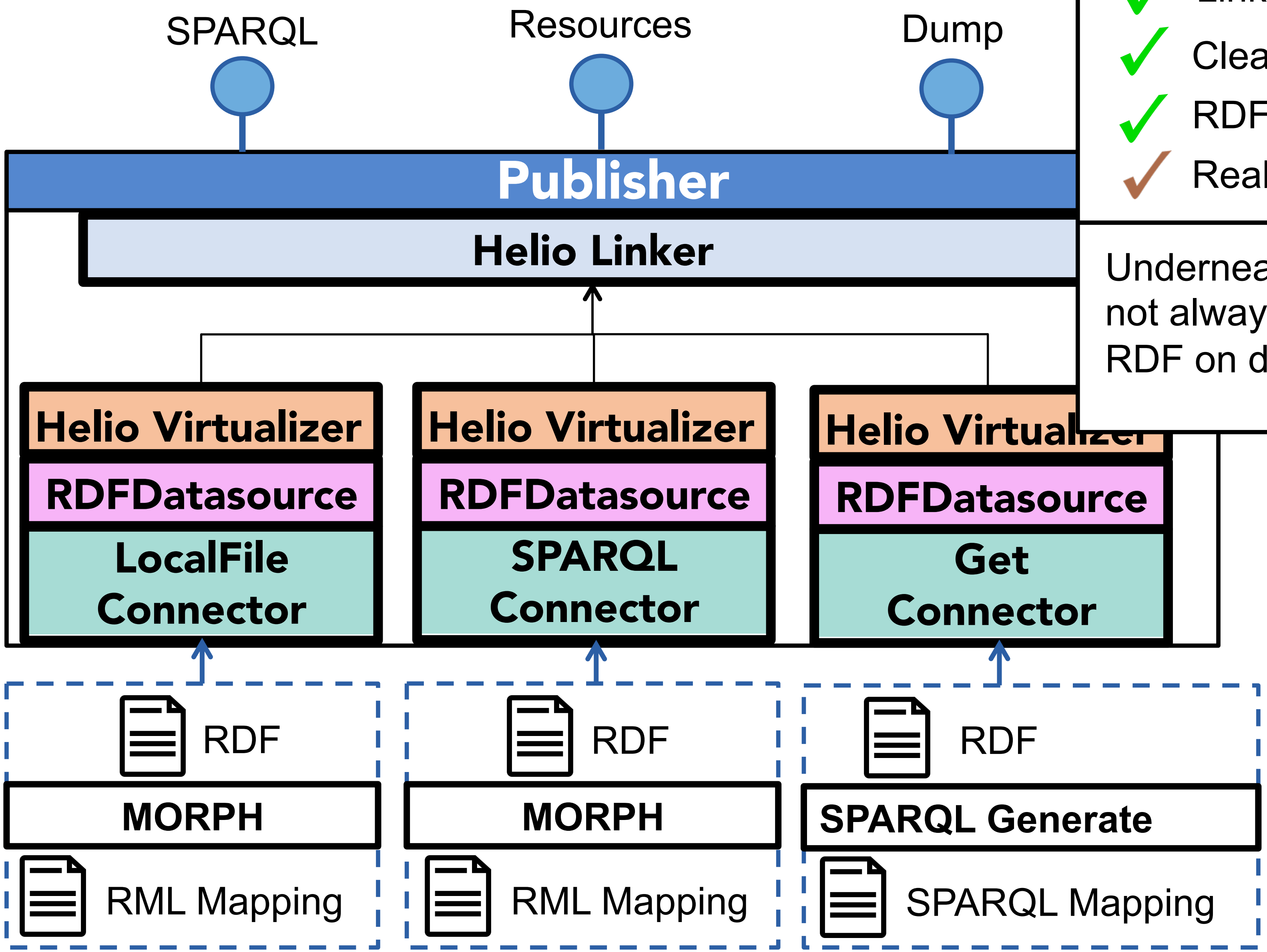
Helio Scenario 4: OBDA



- ✓ Linking
- ✓ Cleaning
- ✓ RDF generation
- ✓ Real Time data

OBDA software not always generates RDF on demand

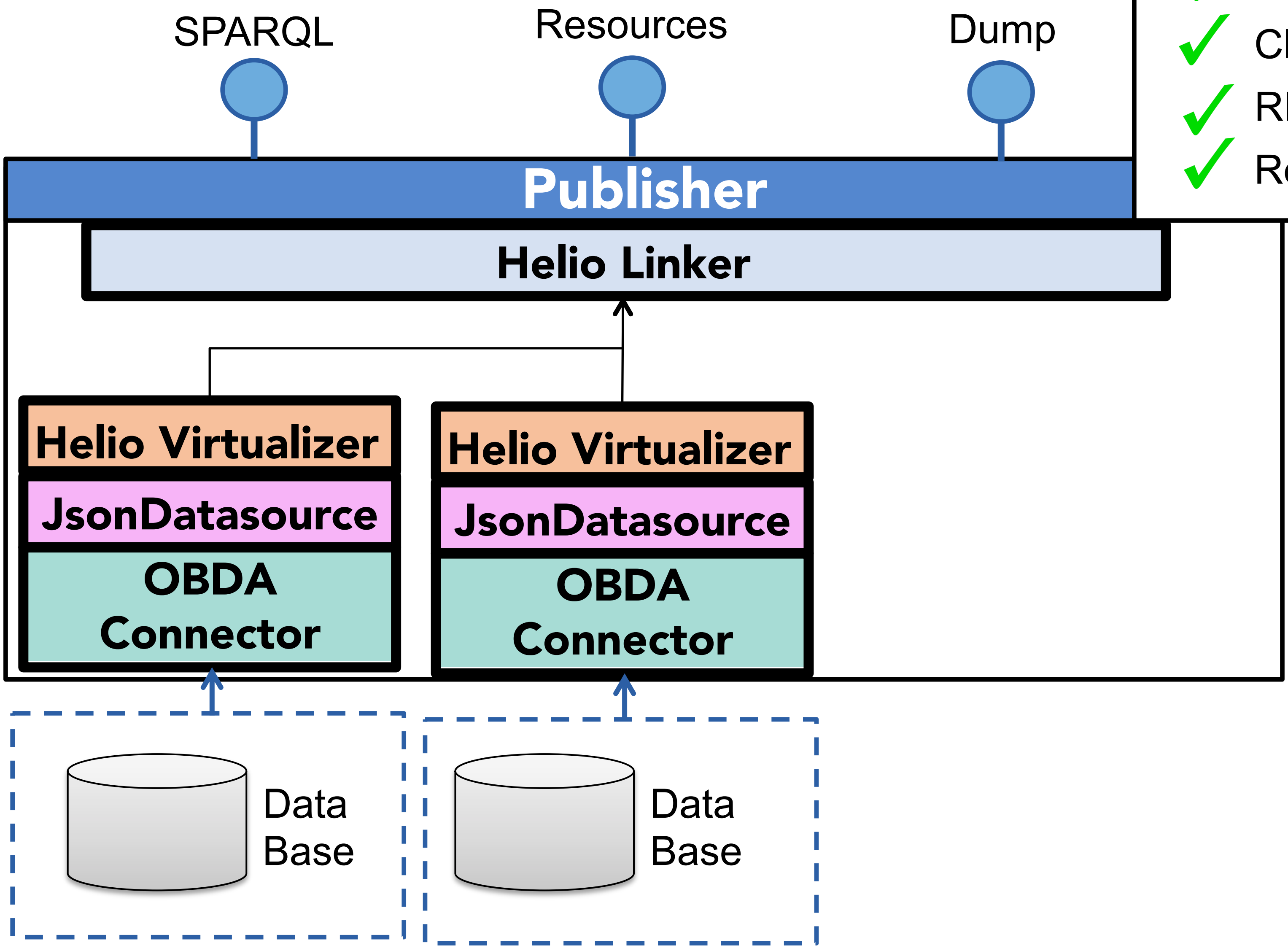
Helio scenario 5: Using third-part RDF generators



- ✓ Linking
- ✓ Cleaning
- ✓ RDF generation
- ✓ Real Time data

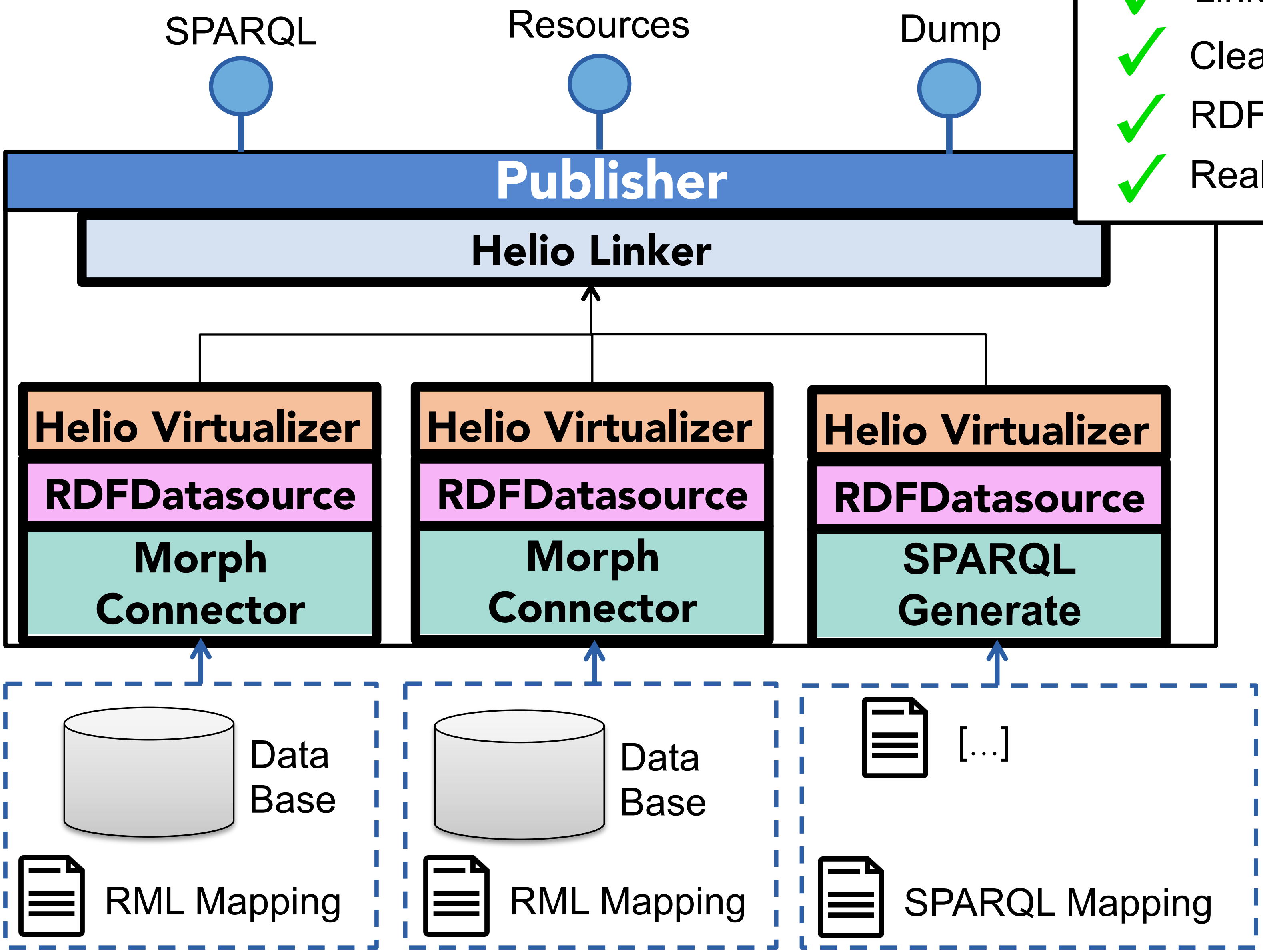
Underneath software not always generates RDF on demand

Helio Scenario 6: Integrating OBDA



- ✓ Linking
- ✓ Cleaning
- ✓ RDF generation
- ✓ Real Time data

Helio scenario 7: Integrating third-part RDF generators



- ✓ Linking
- ✓ Cleaning
- ✓ RDF generation
- ✓ Real Time data

Index

1. Helio Solution
2. Use Cases + Challenges
3. Implementation
4. Helio deployment scenarios
5. **Conclusions**

Conclusions

- Publish data from heterogeneus datasources
  - Clean & transform
  - Data interlinking
- Integrate existing technoligies to generate RDF
- Helio is meant to be pluggable
- Specifications for the pipeline with:
  - RML
  - SPARQL-Generate
  - ….

- Helio can validate published data with Shapes in different levels of pipleline due to its modularity
- Helio aim at integrating current technoliges

Thanks to all