Prêt-à-LLOD

Transforming Language Resources

Christian Chiarcos
Christian Fäth

FinCan

NUI Galway OÉ Gaillimh

Universidad Zaragoza

POLITÉCNICA

Universität Bielefeld

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

DFKI

SEMALYTIX

OXFORD UNIVERSITY PRESS

SEMANTIC WEB COMPANY Linking data to knowledge

Derilinx

European Commission

HORIZON 2020

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Transforming Language Resources

1. Heterogeneity of Language Resources
2. The Fintan platform
3. Design a conversion pipeline
4. Deploy as a service
5. Contribute

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Heterogeneity of Language Resources

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Varieties of Language Resources

- **Corpora**:
  TSV/CoNLL, Sketch Engine, Toolbox, TIGER-XML, ...
  RDF: NIF, Open Annotation, …

- **Lexical Data**:
  TEI, proprietary XML formats, CSV/TSV, …
  RDF: OntoLex, …

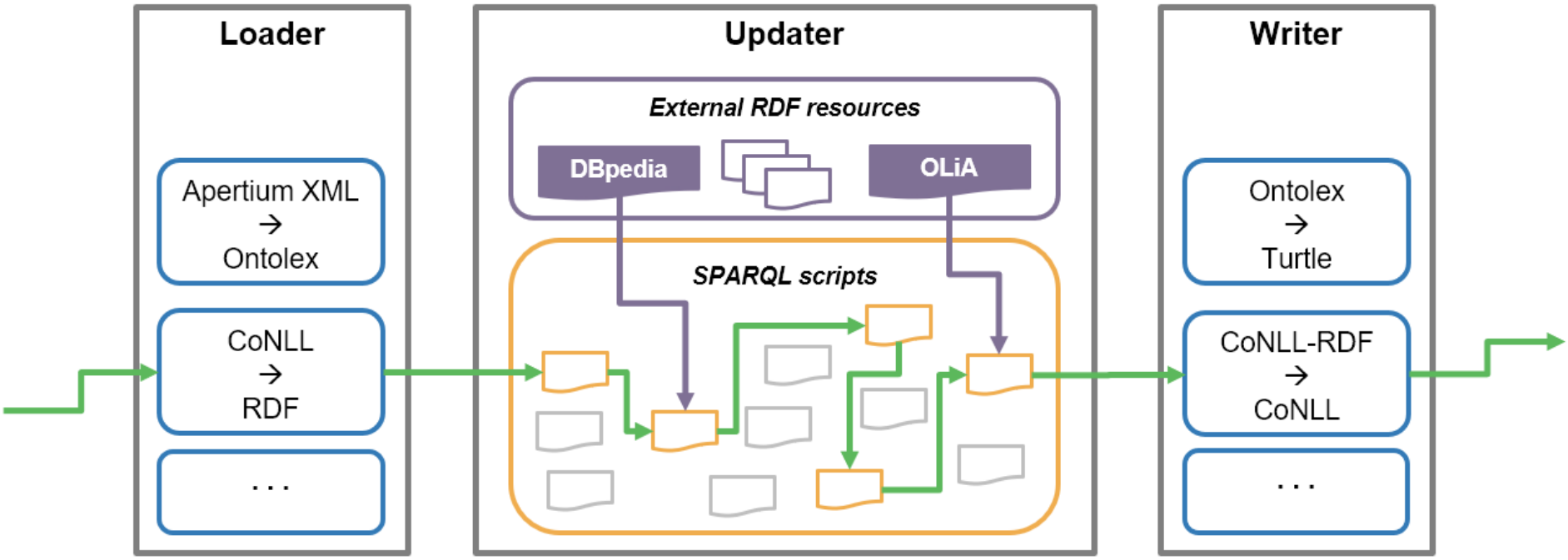
- **Terminological Data**:
  Thesauri and Ontologies: SKOS, OWL, …
  Annotation schemes: GOLD, ISOCat, OLiA, UD, UniMorph, …

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Flexible INtegrated Transformation and Annotation eNgineering

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

The Concept



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Core API

Fintan Core API

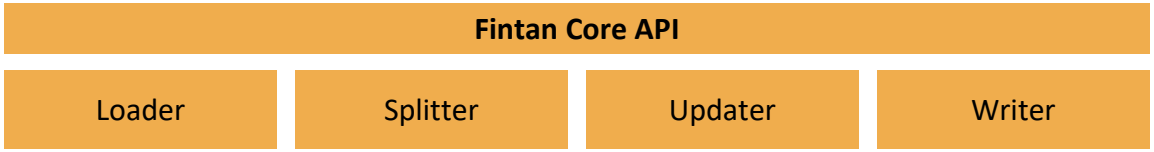
Loader	Splitter	Updater	Writer

• Transform data to RDF
  ◦ XML
  ◦ TSV / CSV
  ◦ CoNLL
  ◦ …
• Stream RDF data
	• Split large RDF datasets into digestible segments
	• Stream RDF segments
		• Load other resources
		• Execute SPARQL updates
		• Parallel processing of RDF segments
			• Serialize RDF
			  ◦ RDF/XML
			  ◦ Turtle
			  ◦ Json-LD
			  ◦ …
			• Other output formats
			  ◦ TSV / CSV
			  ◦ CoNLL
			  ◦ …

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Architecture

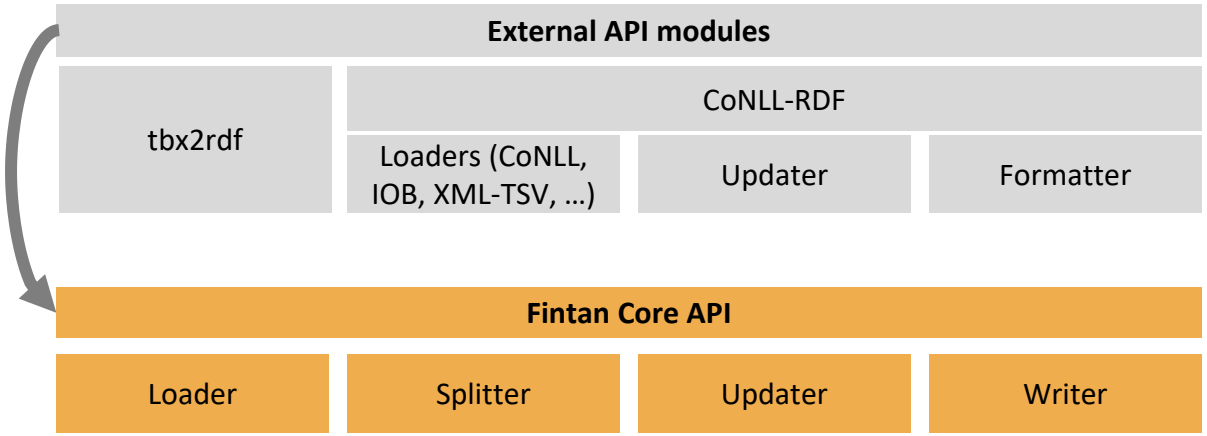
- Core Java API with interoperable interfaces



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Architecture

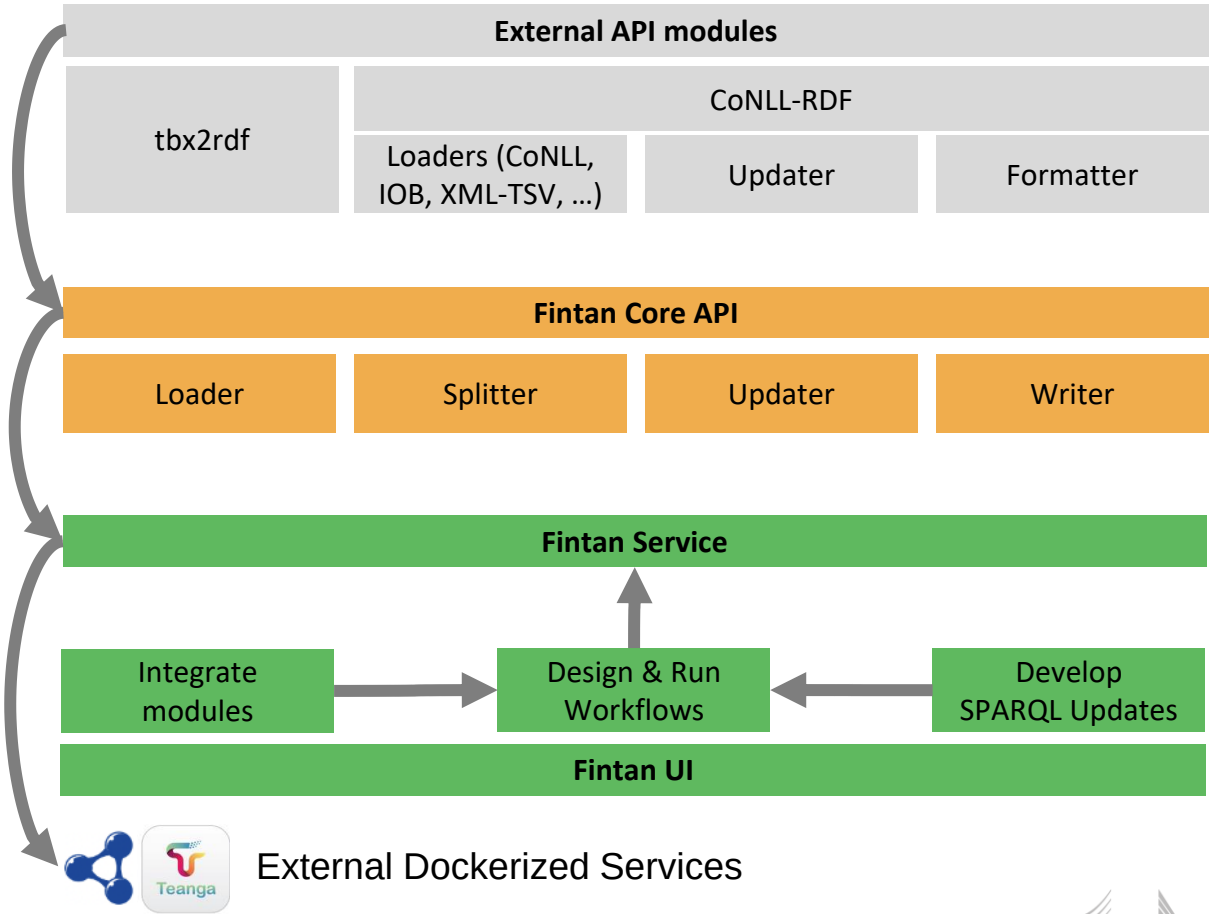
- External modules implement the Core API (Maven dependency)
- Core Java API with interoperable interfaces



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Architecturere

- External modules implement the Core API (Maven dependency)
- Core Java API with interoperable interfaces
- User interface
  - create and run complex workflows
  - export dockerized services
  - compatible to Teanga



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

How to use Fintan

- **Java API**:
  Directly use Fintan as a Java application or as an API in your software (CoNLL-RDF, TBX2RDF)

- **Integrated Workflow Manager**:
  Use Fintan incl. UI as a dockerized service for building and running pipelines

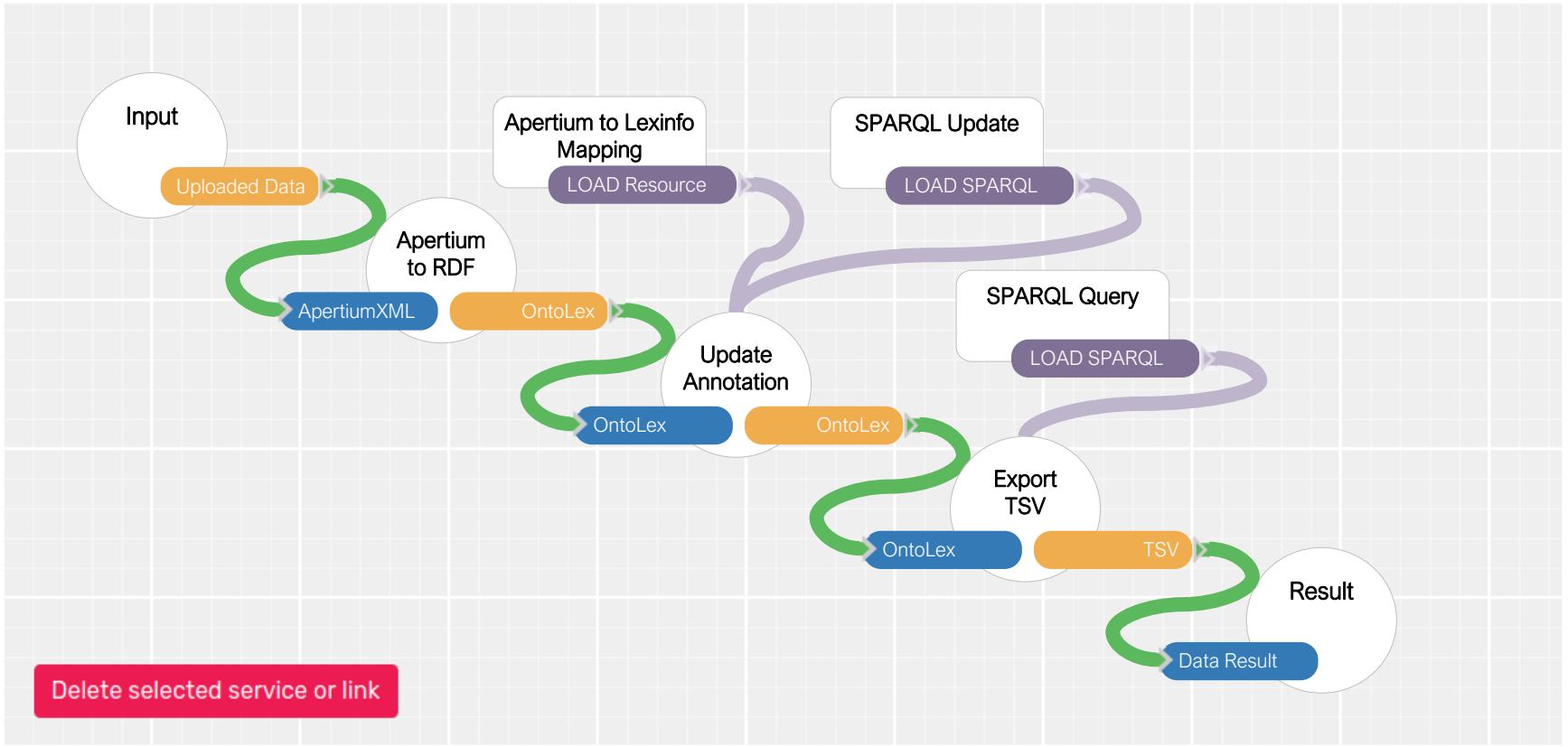
- **Fintan-SaaS**: Deploy Integrated Services to other applications (export Services to Teanga)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Design a pipeline

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Fintan UI



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Fintan Configuration

- All this translates to a JSON configuration file, which could be run from command-line or exported inside a container.

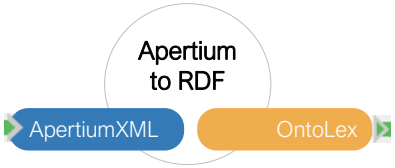
```
{
"input" : "samples/xslt/apertium/data/apertium-en-es.en-es.dix"
, "output" : "System.out"
, "pipeline" : [

    {
      "componentInstance" : "Apertium to RDF",
      "class" : "XSLTStreamTransformer",
      "delimiterIn" : null,
      "delimiterOut" : null,
      "xsl" : "samples/xslt/apertium/dix2src-ttl.xsl <$param0> <$param1>"
    } ,
    {
      "class" : "SegmentedRDFStreamLoader",
      "lang" : "ttl",
      "delimiter" : ""
    } ,
    {
      "componentInstance" : "Update Annotation",
      "class" : "RDFUpdater",
      "models" : [
        {
          "source":"samples/xslt/apertium/apertium-lexinfo-enrichment-update.ttl",
          "graph":"http://apertium-lexinfo-enrichment-update.ttl"
        }
      ],
      "updates" : [
        { "path":"samples/xslt/apertium/apertium-pos.sparql", "iter":"1" }
      ]
    } ,
    {
      "componentInstance" : "Export TSV",
      "class" : "TSVStreamWriter",
      "delimiter" : "",
      "select" : "samples/xslt/apertium/apertium-tiad.sparql"
    }
  ]
}
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Fintan Configuration

- All this translates to a JSON configuration file, which could be run from command-line or exported inside a container.

- First: create Apertium RDF
  - Apply XSLT transformation
  - Load the resulting OntoLex entries as segmented RDF-Stream

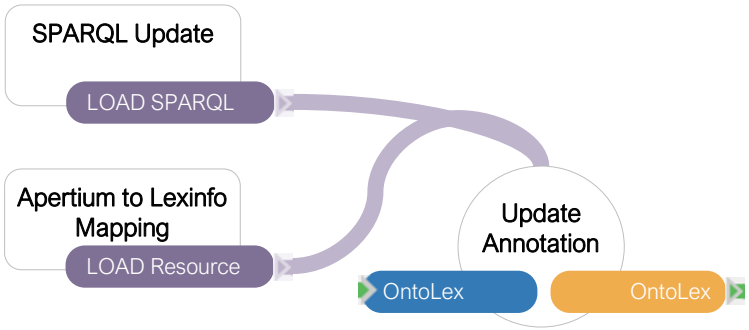



```
{
"input" : "samples/xslt/apertium/data/apertium-en-es.en-es.dix"
, "output" : "System.out"
, "pipeline" : [
    {
      "componentInstance" : "Apertium to RDF",
      "class" : "XSLTStreamTransformer",
      "delimiterIn" : null,
      "delimiterOut" : null,
      "xsl" : "samples/xslt/apertium/dix2src-ttl.xsl <$param0> <$param1>"
    } ,
    {
      "class" : "SegmentedRDFStreamLoader",
      "lang" : "ttl",
      "delimiter" : ""
    } ,
    {
      "componentInstance" : "Update Annotation",
      "class" : "RDFUpdater",
      "models" : [
        {
          "source":"samples/xslt/apertium/apertium-lexinfo-enrichment-update.ttl",
          "graph":"http://apertium-lexinfo-enrichment-update.ttl"
        }
      ],
      "updates" : [
        { "path":"samples/xslt/apertium/apertium-pos.sparql", "iter":"1" }
      ]
    } ,
    {
      "componentInstance" : "Export TSV",
      "class" : "TSVStreamWriter",
      "delimiter" : "",
      "select" : "samples/xslt/apertium/apertium-tiad.sparql"
    }
  ]
}
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Fintan Configuration

- Second: RDF Updater
  - Load the mapping file
  - Load the SPARQL update
  - Apply LexInfo Mapping
  - Parallel processing per segment

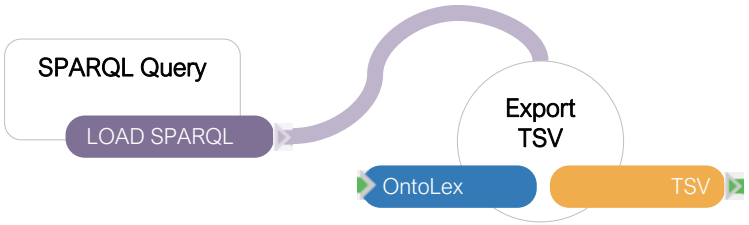



```
{
"input" : "samples/xslt/apertium/data/apertium-en-es.en-es.dix"
, "output" : "System.out"
, "pipeline" : [

    {
      "componentInstance" : "Apertium to RDF",
      "class" : "XSLTStreamTransformer",
      "delimiterIn" : null,
      "delimiterOut" : null,
      "xsl" : "samples/xslt/apertium/dix2src-ttl.xsl <$param0> <$param1>"
    } ,
    {
      "class" : "SegmentedRDFStreamLoader",
      "lang" : "ttl",
      "delimiter" : ""
    } ,
    {
      "componentInstance" : "Update Annotation",
      "class" : "RDFUpdater",
      "models" : [
        {
          "source":"samples/xslt/apertium/apertium-lexinfo-enrichment-update.ttl",
          "graph":"http://apertium-lexinfo-enrichment-update.ttl"
        }
      ],
      "updates" : [
        { "path":"samples/xslt/apertium/apertium-pos.sparql", "iter":"1" }
      ]
    } ,
    {
      "componentInstance" : "Export TSV",
      "class" : "TSVStreamWriter",
      "delimiter" : "",
      "select" : "samples/xslt/apertium/apertium-tiad.sparql"
    }
  ]
}
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Fintan Configuration

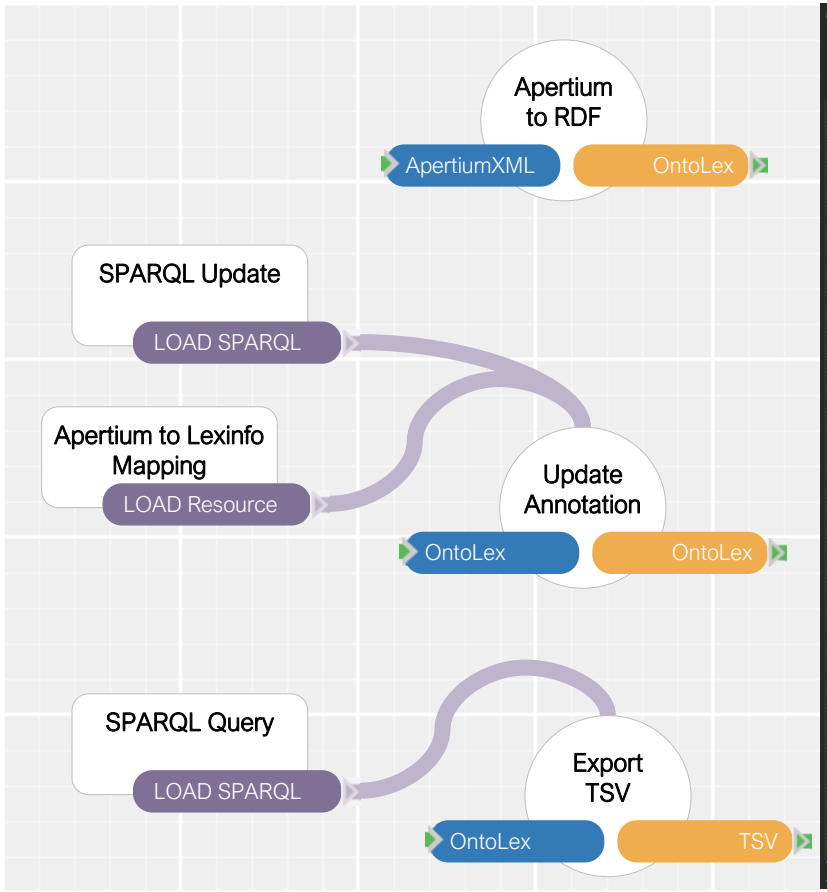
- Third: Export TIAD TSV
  - Load SPARQL query
  - Write TSV data




```
{
"input" : "samples/xslt/apertium/data/apertium-en-es.en-es.dix"
, "output" : "System.out"
, "pipeline" : [

    {
      "componentInstance" : "Apertium to RDF",
      "class" : "XSLTStreamTransformer",
      "delimiterIn" : null,
      "delimiterOut" : null,
      "xsl" : "samples/xslt/apertium/dix2src-ttl.xsl <$param0> <$param1>"
    } ,
    {
      "class" : "SegmentedRDFStreamLoader",
      "lang" : "ttl",
      "delimiter" : ""
    } ,
    {
      "componentInstance" : "Update Annotation",
      "class" : "RDFUpdater",
      "models" : [
        {
          "source":"samples/xslt/apertium/apertium-lexinfo-enrichment-update.ttl",
          "graph":"http://apertium-lexinfo-enrichment-update.ttl"
        }
      ],
      "updates" : [
        { "path":"samples/xslt/apertium/apertium-pos.sparql", "iter":"1" }
      ]
    } ,
    {
      "componentInstance" : "Export TSV",
      "class" : "TSVStreamWriter",
      "delimiter" : "",
      "select" : "samples/xslt/apertium/apertium-tiad.sparql"
    }
  ]
}
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

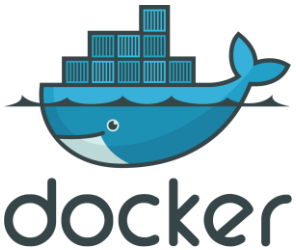
Fintan Configuration




```
{
"input" : "samples/xslt/apertium/data/apertium-en-es.en-es.dix"
, "output" : "System.out"
, "pipeline" : [

    {
      "componentInstance" : "Apertium to RDF",
      "class" : "XSLTStreamTransformer",
      "delimiterIn" : null,
      "delimiterOut" : null,
      "xsl" : "samples/xslt/apertium/dix2src-ttl.xsl <$param0> <$param1>"
    } ,
    {
      "class" : "SegmentedRDFStreamLoader",
      "lang" : "ttl",
      "delimiter" : ""
    } ,
    {
      "componentInstance" : "Update Annotation",
      "class" : "RDFUpdater",
      "models" : [
        {
          "source":"samples/xslt/apertium/apertium-lexinfo-enrichment-update.ttl",
          "graph":"http://apertium-lexinfo-enrichment-update.ttl"
        }
      ],
      "updates" : [
        { "path":"samples/xslt/apertium/apertium-pos.sparql", "iter":"1" }
      ]
    } ,
    {
      "componentInstance" : "Export TSV",
      "class" : "TSVStreamWriter",
      "delimiter" : "",
      "select" : "samples/xslt/apertium/apertium-tiad.sparql"
    }
  ]
}
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

meets



Deploy as a service

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Fintan pipeline as a Docker container

- **Preconfigured pipeline**
  - Contains everything required to run Fintan for a specific task
  - Exposes rudimentary API to provide the input and collect the output
    - Plain text
    - Gzipped text
    - JSON
  - Can be called from Teanga, curl, Swagger UI (included)

```
docker build --tag fintan-api .
docker run -d -p 8080:8080 --name fintan-api
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Contribute

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

How to contribute

- **Fintan is an extensive converter suite**
  - Add Fintan to your projects: Simply define the Maven dependency
  - Directly benefit:
    - Fomalized pipeline development
    - Parallel streamed graph processing
    - Preconfigured converter pipelines and components

- **Fintan is an extensible converter suite**
  - Implement the Core API
    - Use your tools directly in the workflow manager
  - Deploy to central Fintan repository
    - Gain visibility for your own converters

- **What's next?**
  - Fintan will be published as a software deliverable of the Prêt-à-LLOD project on Sep. 30
  - We will actively work on adding more components and preconfigured pipelines.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Get in touch!

https://github.com/pret-a-LLOD/Fintan
https://github.com/acoli-repo/conll

faeth@em.uni-frankfurt.de
chiarcos@cs.uni-frankfurt.de

Thank you!

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182