Data extraction and annotation from pdf files

1st AgroHackathon, Hack 14

Montpellier, 2016

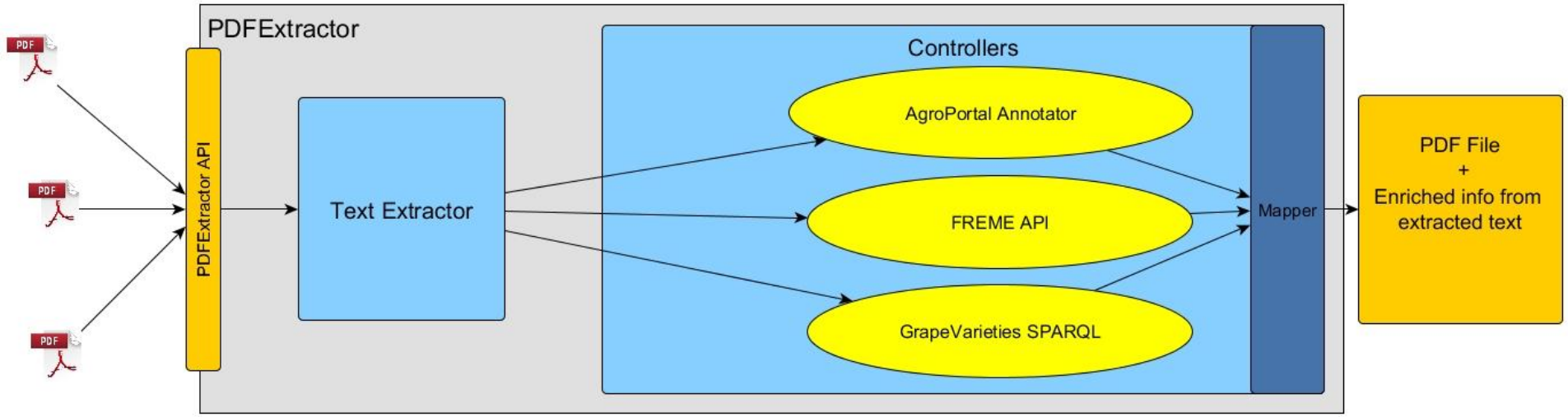
The goal

- Extract textual information contained in pdf files of (agricultural) bibliographic resources.
- Use extracted information against known endpoints for the semantic annotation.
- Design a modular system easily configurable and extendable.
- Have an intuitive way for end users to semantically annotate their pdf files.

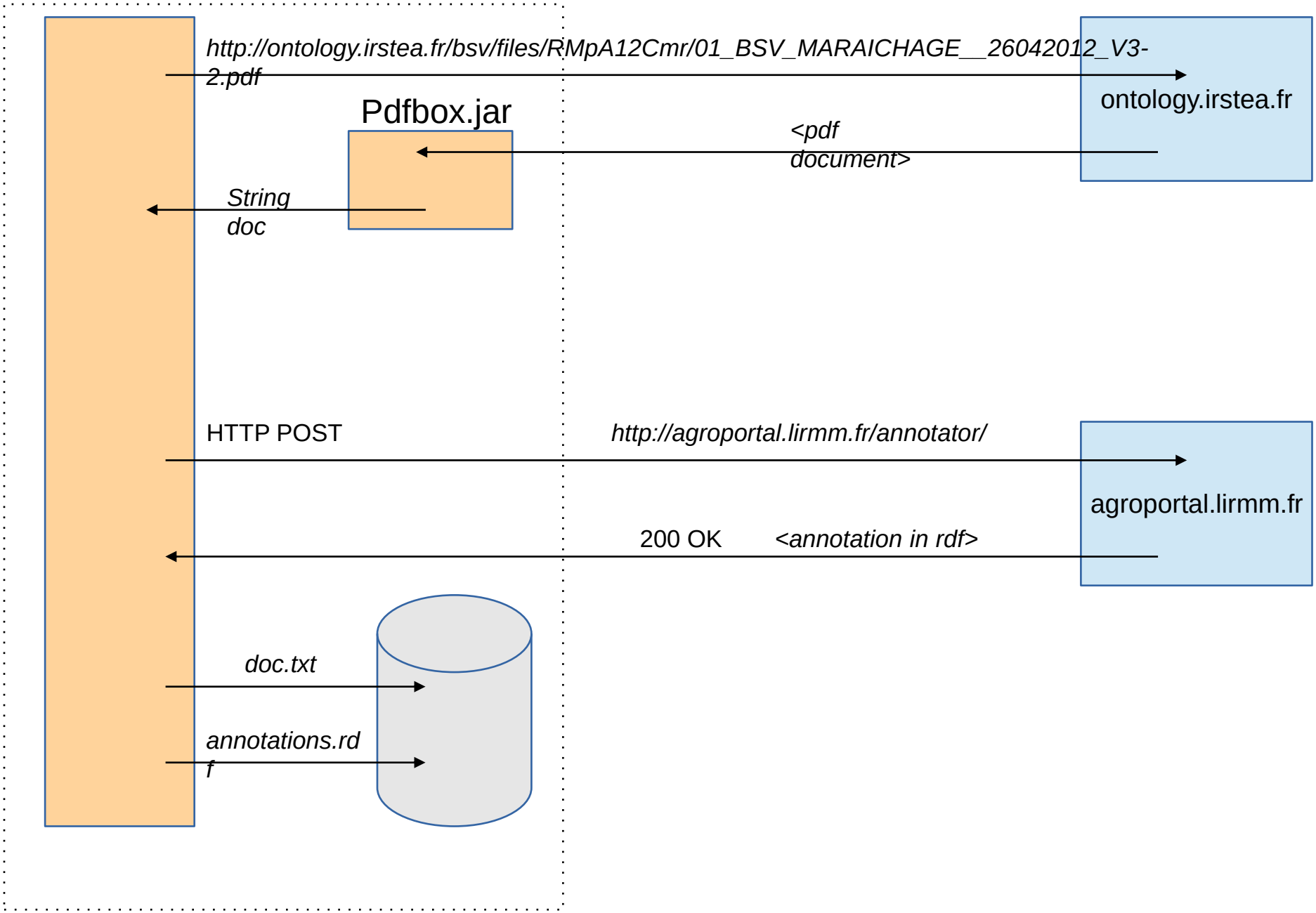
The process – Key Features

- We divided the system into 4 separate components:
  - API Endpoint, where all the calls are made, using as input the url of a pdf file.
  - TextExtractor, where the actual extraction of textual information takes place.
  - Controllers, where the extracted information is passed and then external endpoints are called.
  - Mapper, where the returned annotated result is presented back in a unifying manner.

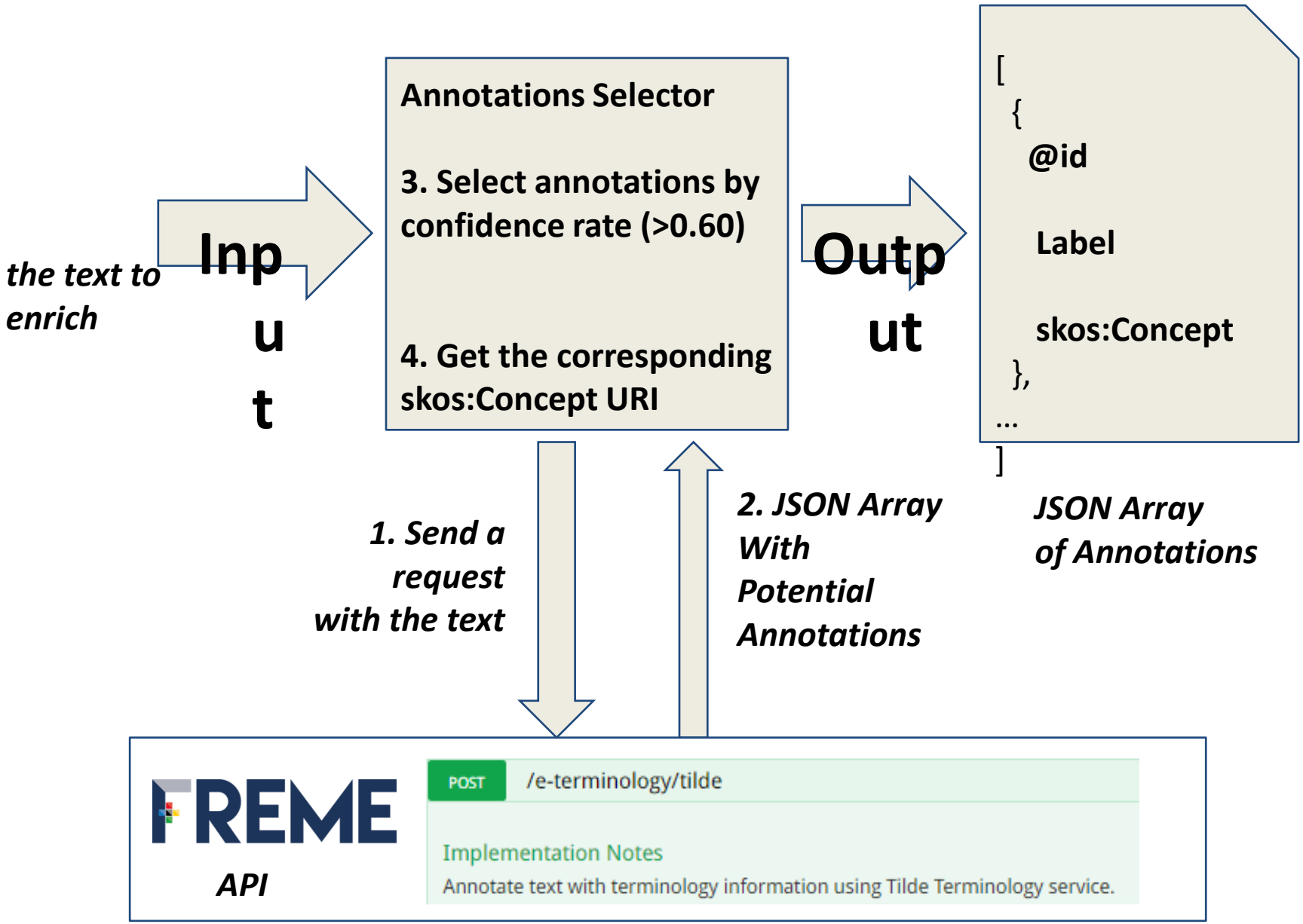
PDFExtractor Workflow



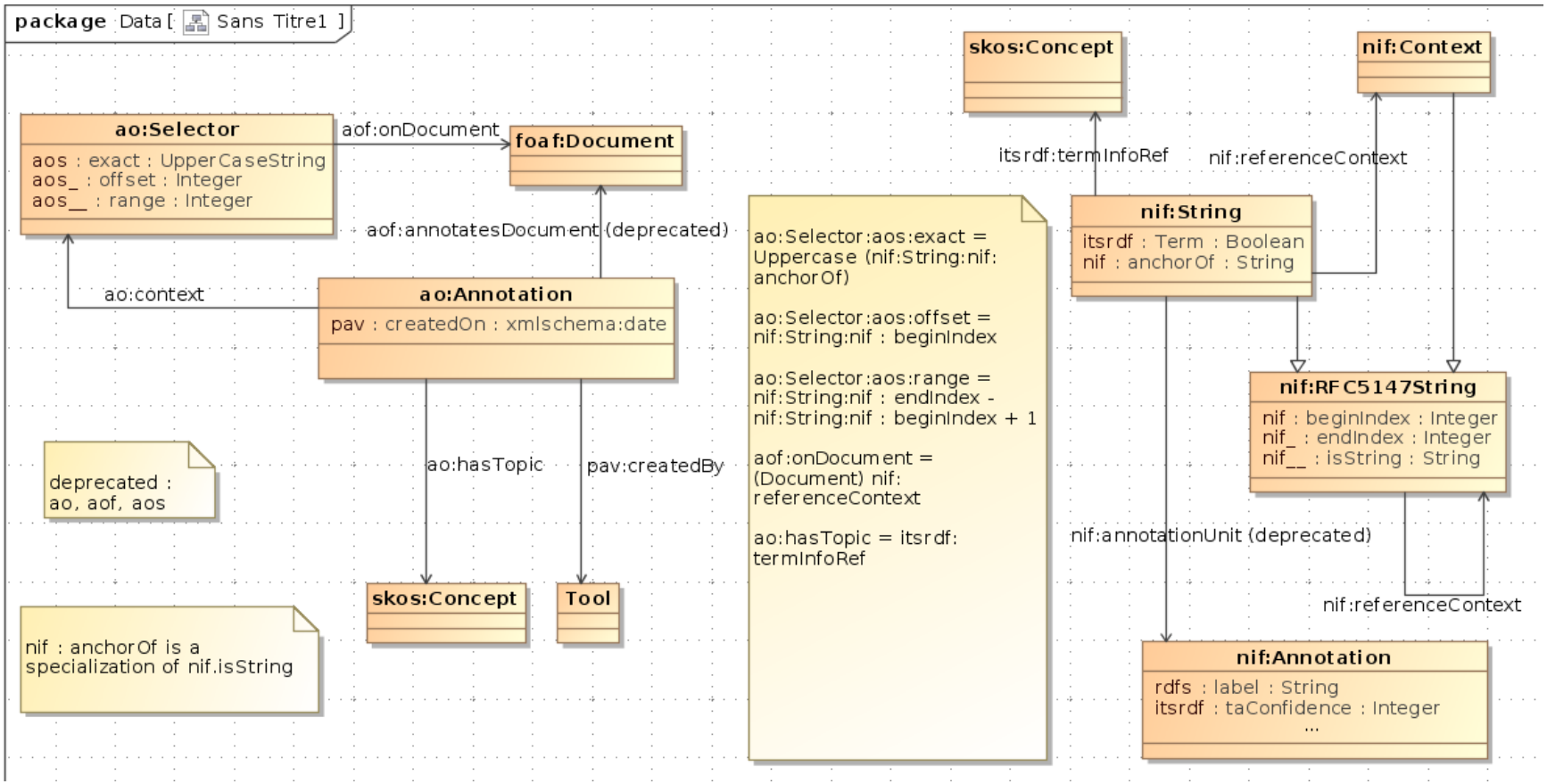
Text Annotation using {agro, bio}Portal



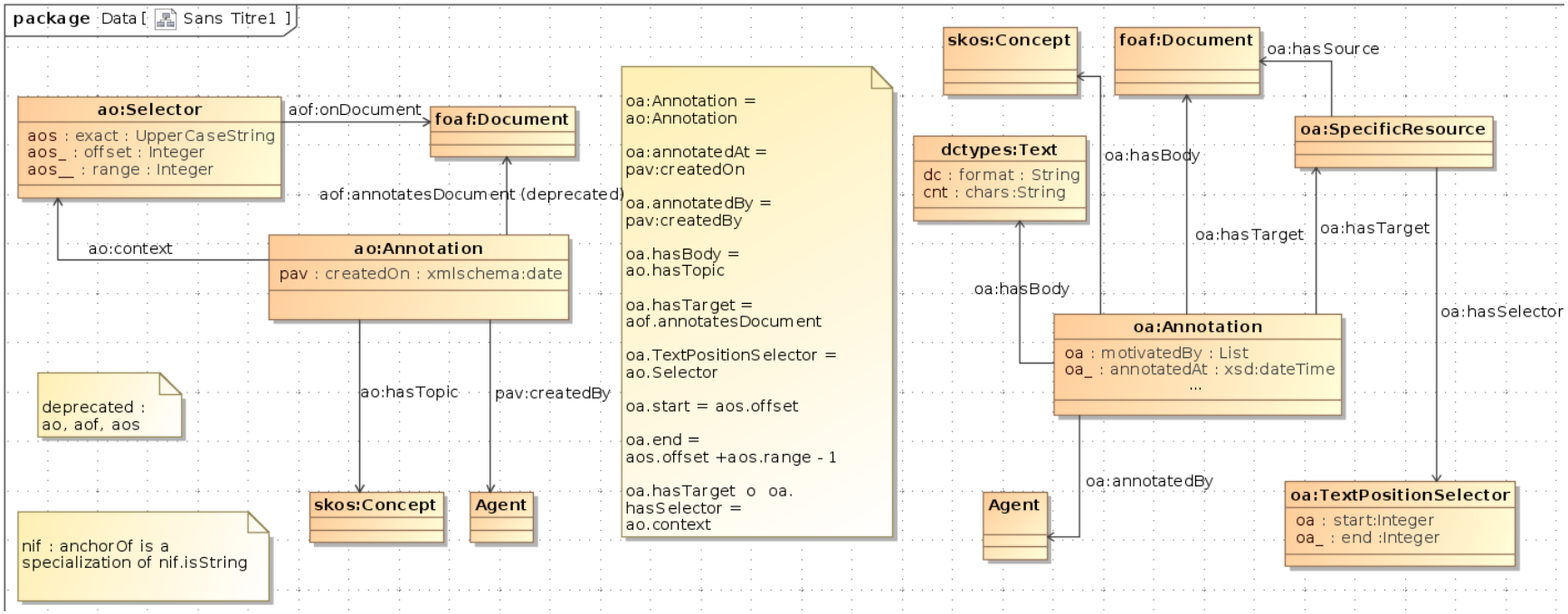
Text Annotation using FREME API



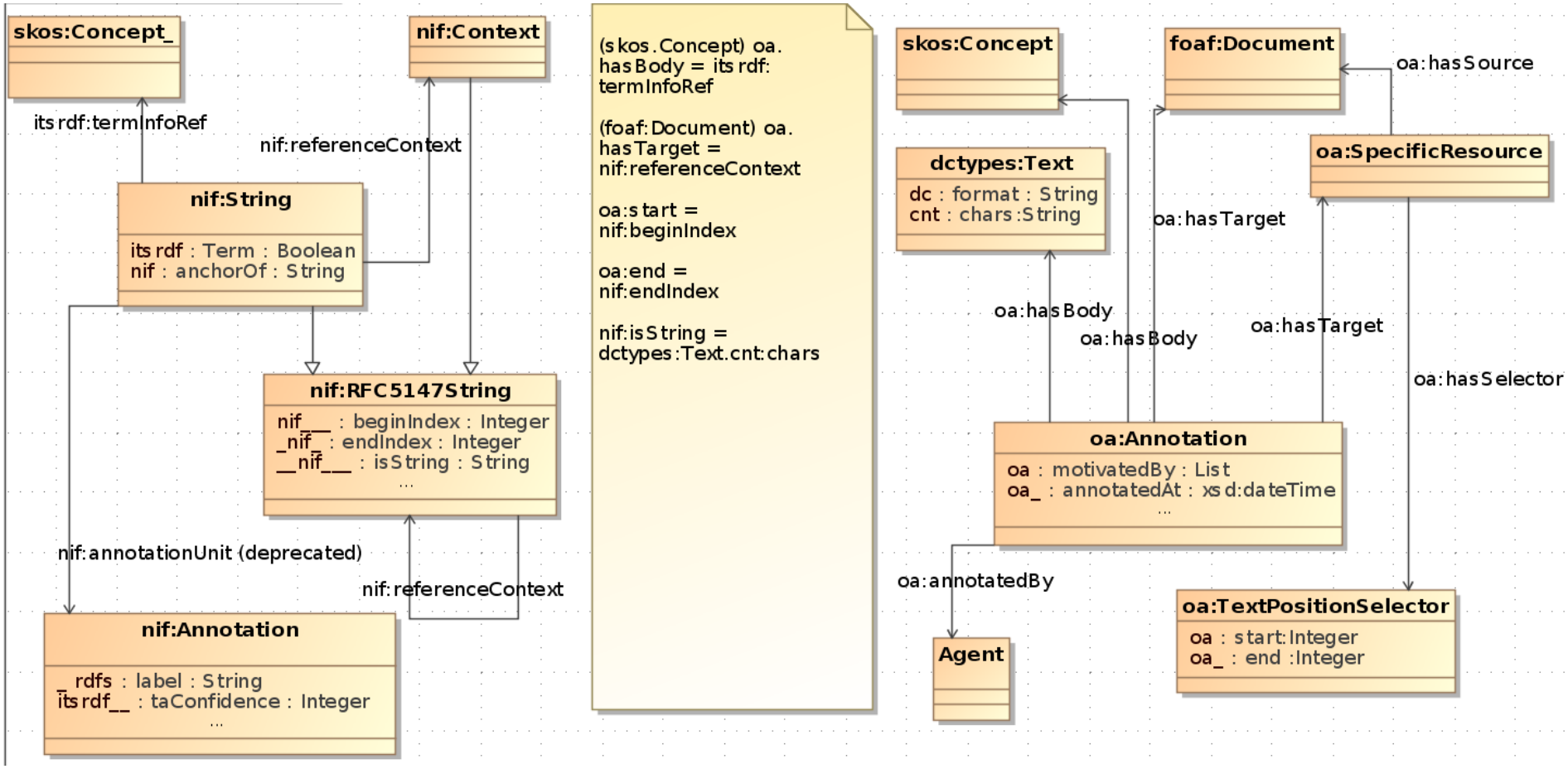
Mappings – AgroPortal with FREME



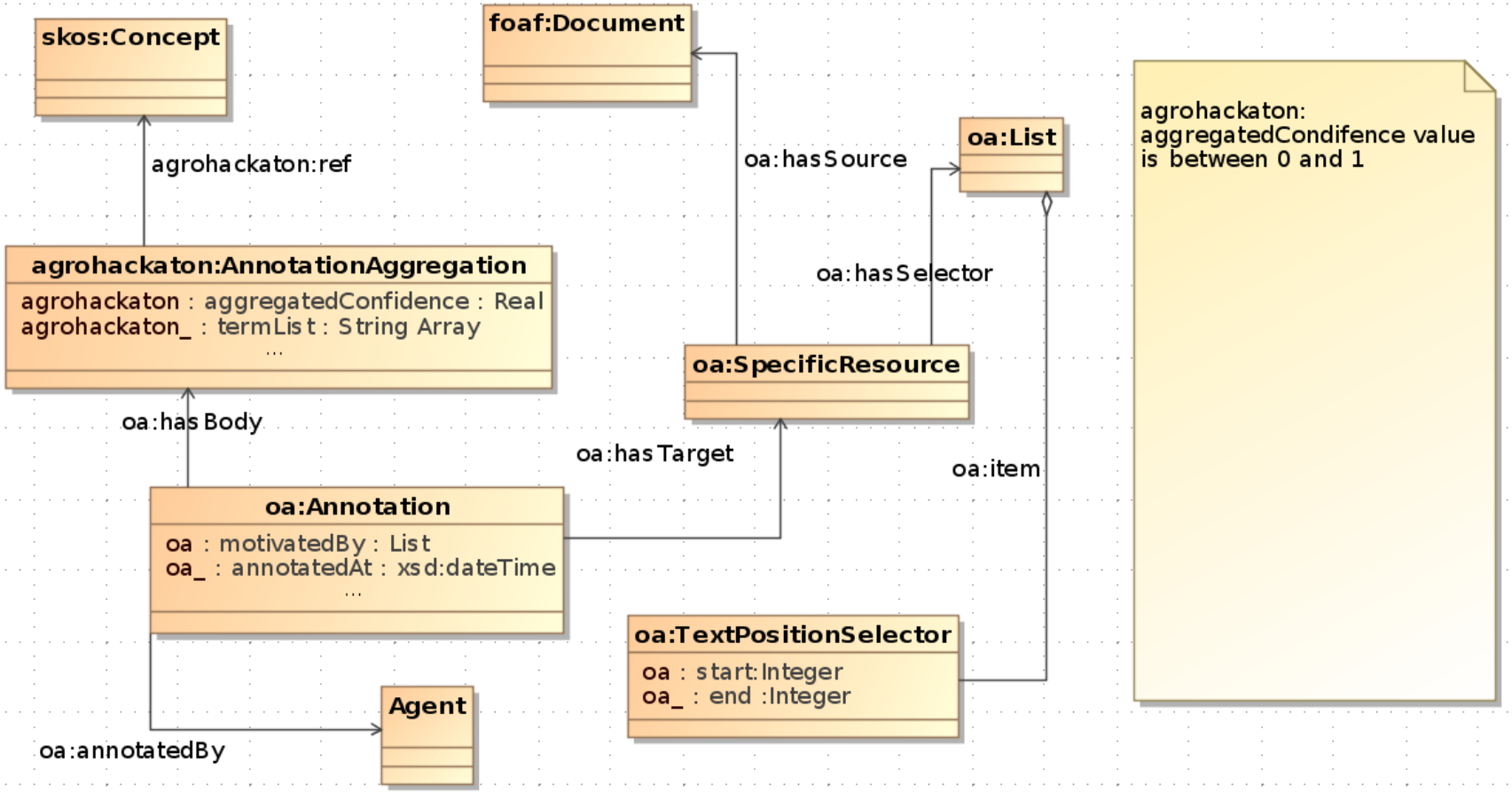
Mappings – AgroPortal with OA



Mappings – FREME with OA



Proposed Evolution of OA based on FREME



Future Work

- Integrate more controllers into the workflow.
- Design a cleansing component clearing out redundant information.
- Provide a richer API endpoint for the system.
- Benchmark various endpoints called by the controllers.
- Build a front-end web app on top to help end users.