Publications Office
of the European Union

SUMMER SCHOOL LEX 2024

AI for OJ text conversion to AKN4EU

COLAVINCENZO Mauro, OP.A.3
HARDY Didier OP.D.4

7th September 2024

Content

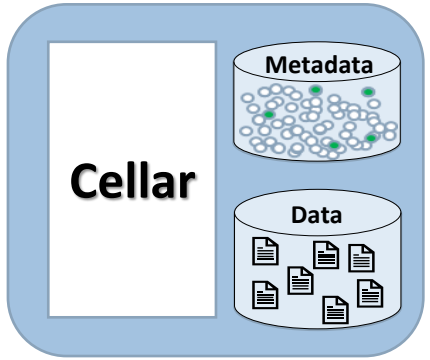
1. Introduction - global of the OJ publication process.
2. Objectives of the AI4XML project.
3. Experimentation.
4. Production vision integration.

1. Introduction - global of the OJ publication process.

Cellar:

The common data repository of the Publications Office of the European Union.

Digital publications and metadata are stored in and disseminated via Cellar, in order to be used by humans and machines.



Cellar domains:

- Official Journal and the acts with:
  - summaries,
  - consolidated legislation,
  - pre-legislative documents,
- Cases of the European Court of Justice and of national Court,
- General publications.
- …

More info on https://op.europa.eu/en/web/cellar

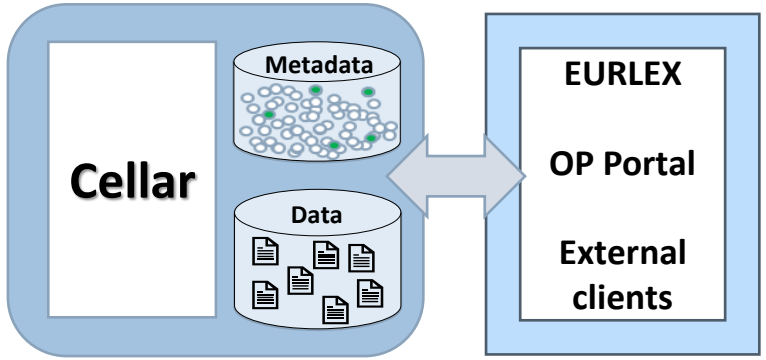
1. Introduction - global of the OJ publication process.

Ways to access the Cellar

- Content accessible to citizens via
  - internal (EurLex, OP Portal)
  - and external WEBSITE,

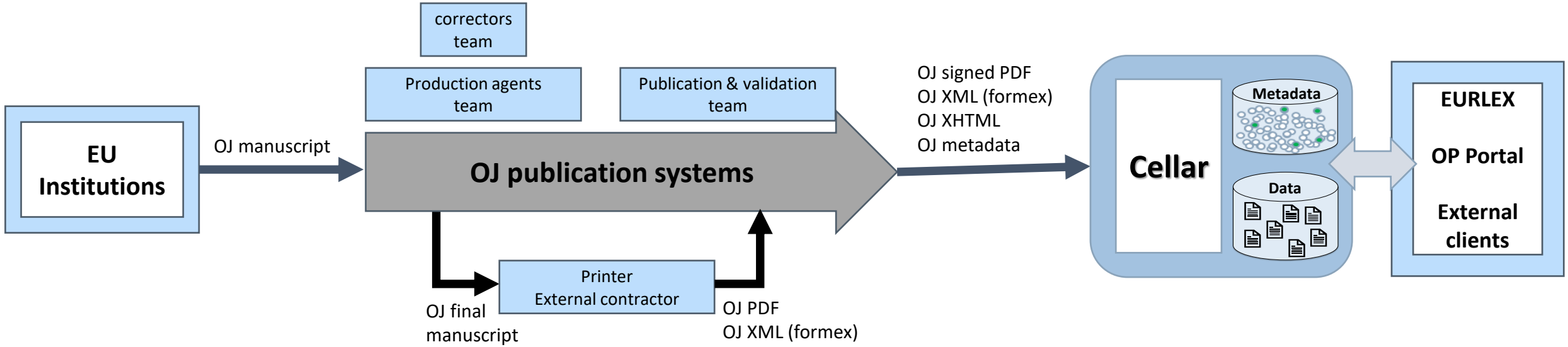
- Based on W3C semantic Web standards:
  - RDF, URIs (API with content negotiation), SPARQL, OWL ontology, etc.

- Metadata are formalised through an OWL ontology using RDF and linked data principles.

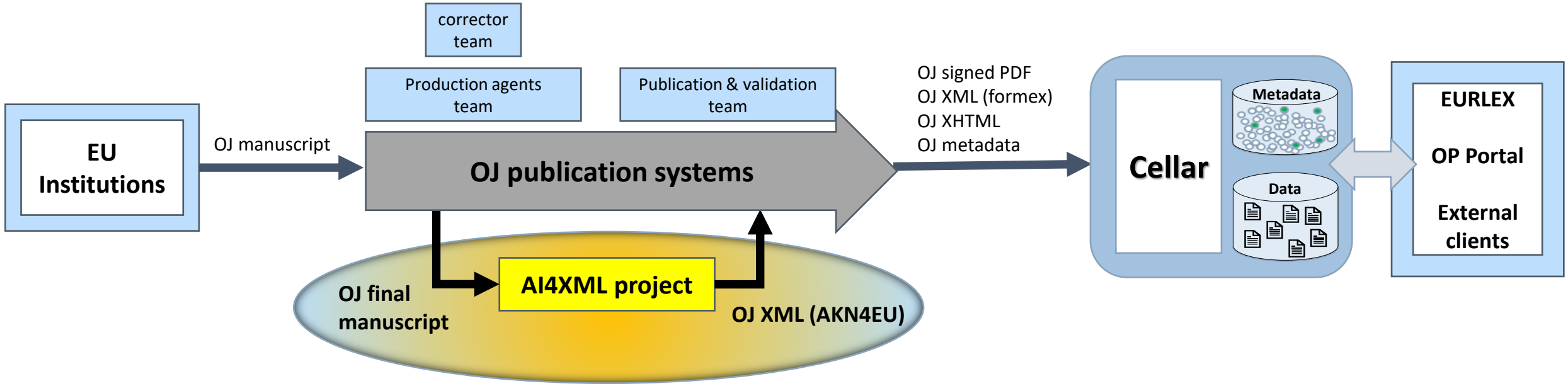


More info on https://op.europa.eu/en/web/cellar

1. Introduction - global of the OJ publication process.



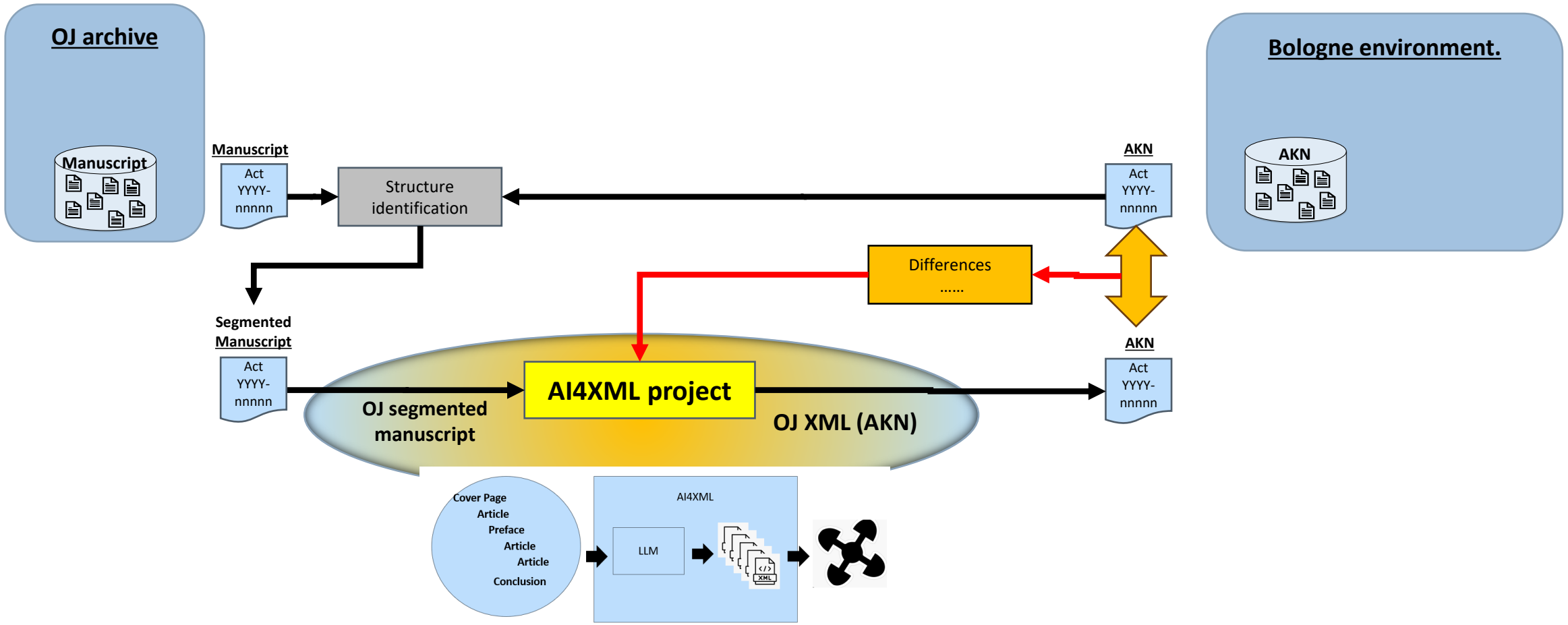
2. Objectives of the AI4XML project.



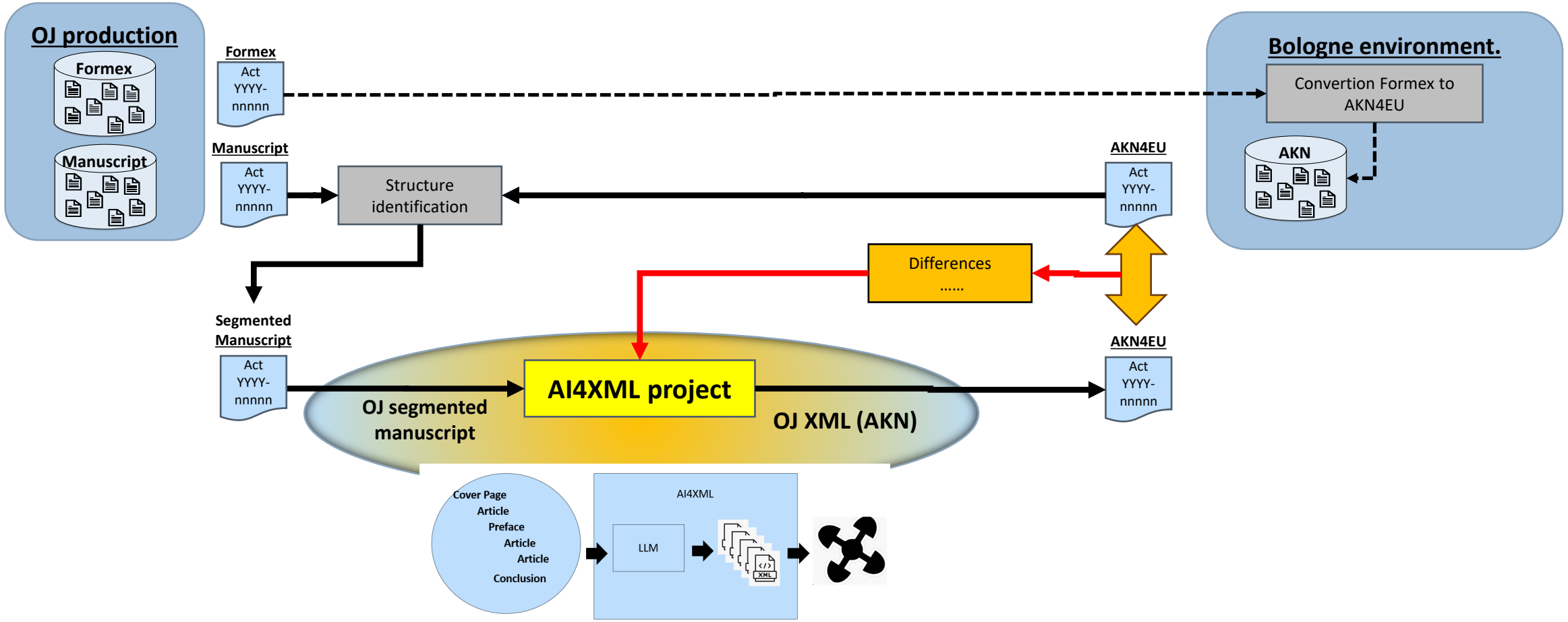
Objectives :

Usage of AI, particularly **Generative LLMs** models, to **automate** the XML (AKN4EU) generation.

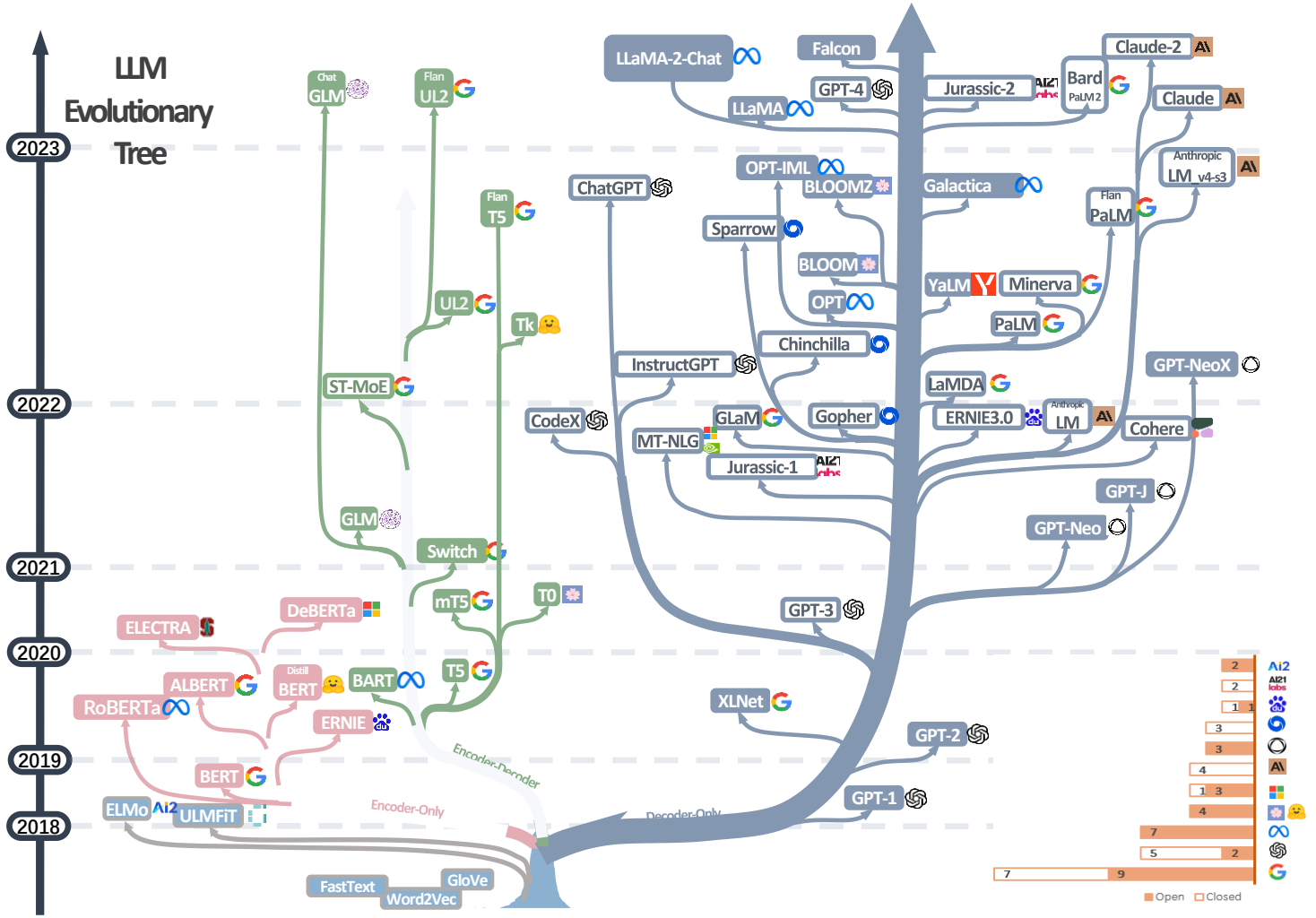
2. Objectives of the AI4XML project. – Phase 1



2. Objectives of the AI4XML project. – Phase 2



3. Experimentation.



Training an LLM from scratch is an extremely costly endeavor.

→ Consider using prompt engineering or fine-tuning as more efficient alternatives.

3. Experimentation.

Prompt Engineering for structured documents generation

- **Does LLM perform well in xml generation ?**

❖Prepare small dataset based on AKN4EU documentation

  - Target section : cover page

❖Perform 0-Shot, 1-shot,2-shots prompting

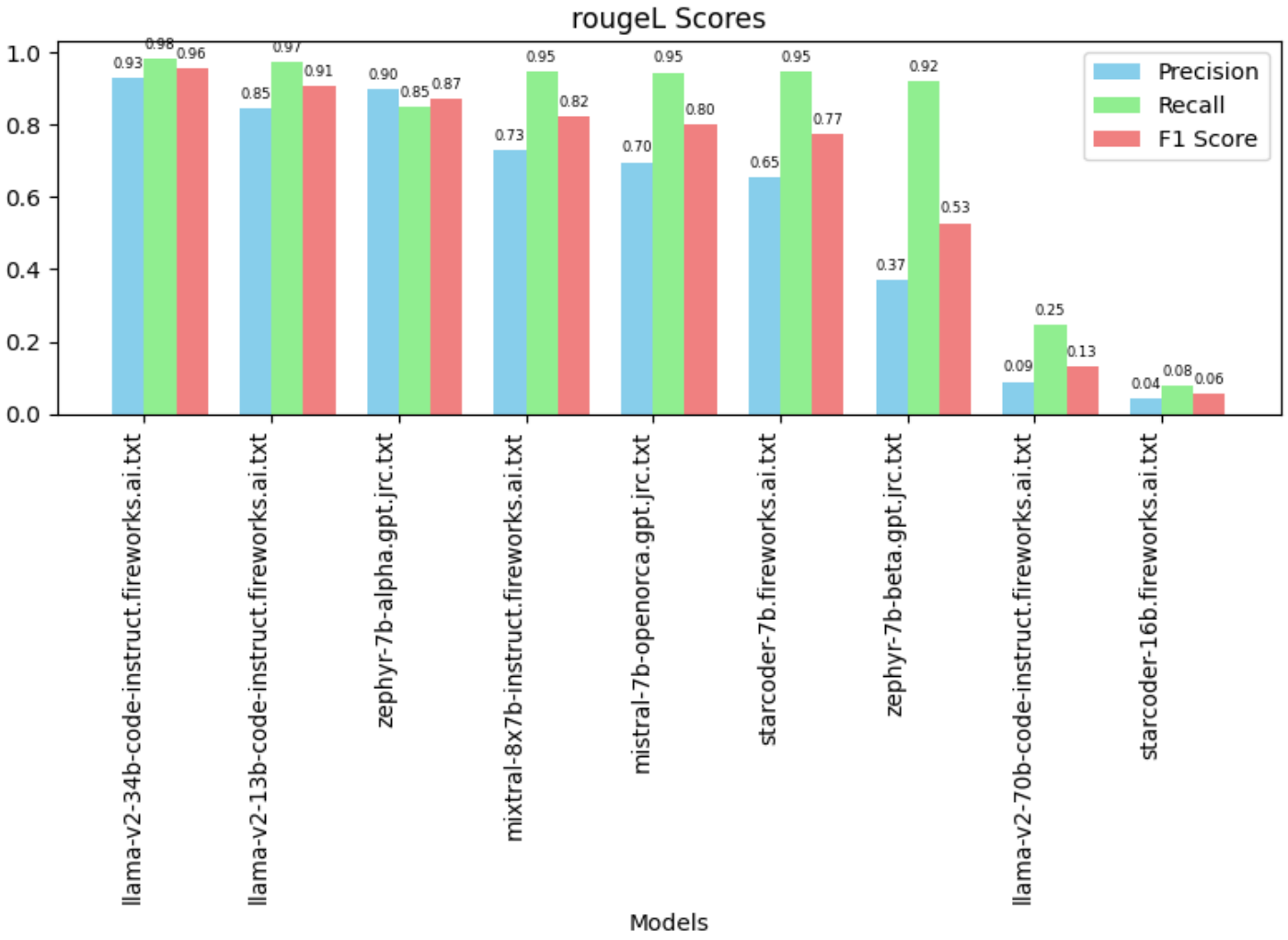
❖Test open and closed LLMs:

  - GPT
  - Mistral
  - Llama
  - …

3. Experimentation.

Prompt Engineering for structured documents generation

Preliminary results (2-shots)



3. Experimentation.
Example of 1-Shot Prompt with Cover page

Instruction

```
# OBJECTIVE #
Convert the plain text into AKN4EU XML format based on the provided
example.


ONLY return the XML format, DO NOT return ANYTHING else.
```

Example

```
########################################
# PLAIN TEXT EXAMPLE #


EUROPEAN COMMISSION
Brussels, 21.12.2016
2016/0411 (COD)
Proposal for a
REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE
COUNCIL
amending Regulation (EC)
No 1008/2008 on common rules for the operation of air services in the
Community
EN

########################################
# XML OF TEXT EXAMPLE #
 <coverPage>
  <container name="logo">
    <p><img src="EC.png" alt="EUROPEAN COMMISSION"/></p>
  </container>
  <container name="actingEntity">
    <p><organization refersTo="~_COM">EUROPEAN COMMISSION</organization></p>
  </container>
  <container name="mainDoc">
    <block name="placeAndDate">
      <location refersTo="~_BEL_BRU">Brussels</location>, <date date="2016-12-21">21.12.2016</date>
    </block>
  </container>
  <container name="procedureIdentifier">
    <p><docketNumber refersTo="~_procedure_2016_411">2016/0411 (COD)</docketNumber></p>
  </container>
  <longTitle>
    <p><docStage>Proposal for a</docStage>
      <docType refersTo="~_REG">REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE
        COUNCIL</docType>
      <docPurpose>amending <ref href="http://data.europa.eu/eli/reg/2008/1008">Regulation (EC)
          No 1008/2008 on common rules for the operation of air services in the
        Community</ref></docPurpose></p>
  </longTitle>
  <container name="mainDocLanguage">
    <p><inline name="language" refersTo="~_FRBRlanguage">EN</inline></p>
  </container>
</coverPage>
```

Target Section of the document

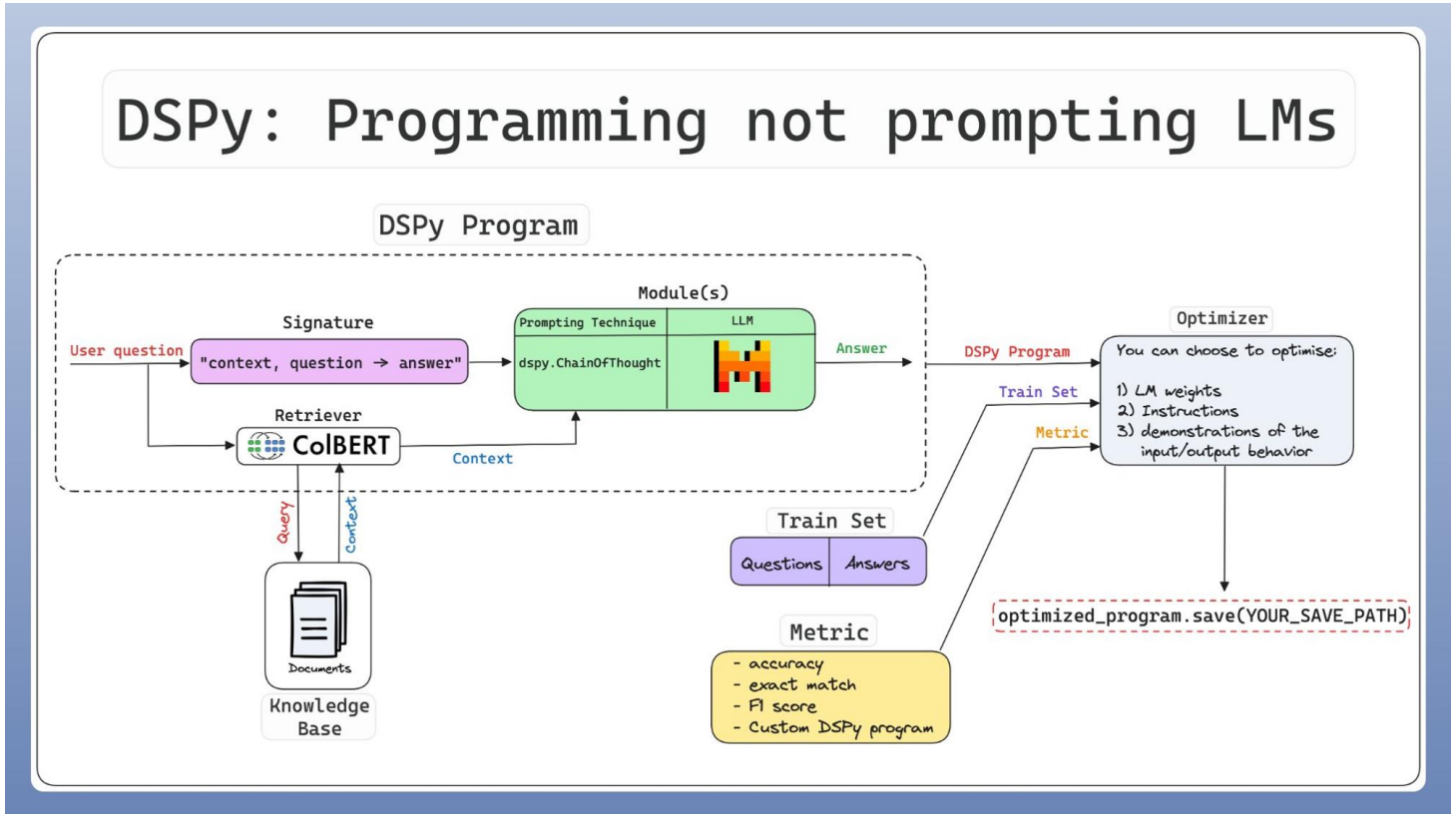
```
##############
# TEXT TO CONVERT #
EUROPEAN COMMISSION
Brussels, 21.12.2017
2012/0412 (COD)
Proposal for a
REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending
Regulation (EC)
No 1009/2009 on common rules for the operation of air services in the Community
EN
```

3. Experimentation.
LLM programs for XML generation

❖Use DSPy

- Compile small LLM programs
- Each program handles specific document parts
- Experiment open/closed LLMs
- Use AKN dataset

- Prompting with hints
- Chain of thoughts , ReAct



- Output is an optimized Program that based on the signature and the trainset DSPY will do several calls to the model in order to find the Optimal program to respond to questions similar to the singnature

3. Experimentation.
Example program for Prefaces generation

Signature:
Input: Plain text of the Preface,
Context: Publication date, doc number
Output: AKN's representation of the preface

Module
- ChainOfThoughts
- LLM: GPT 4o

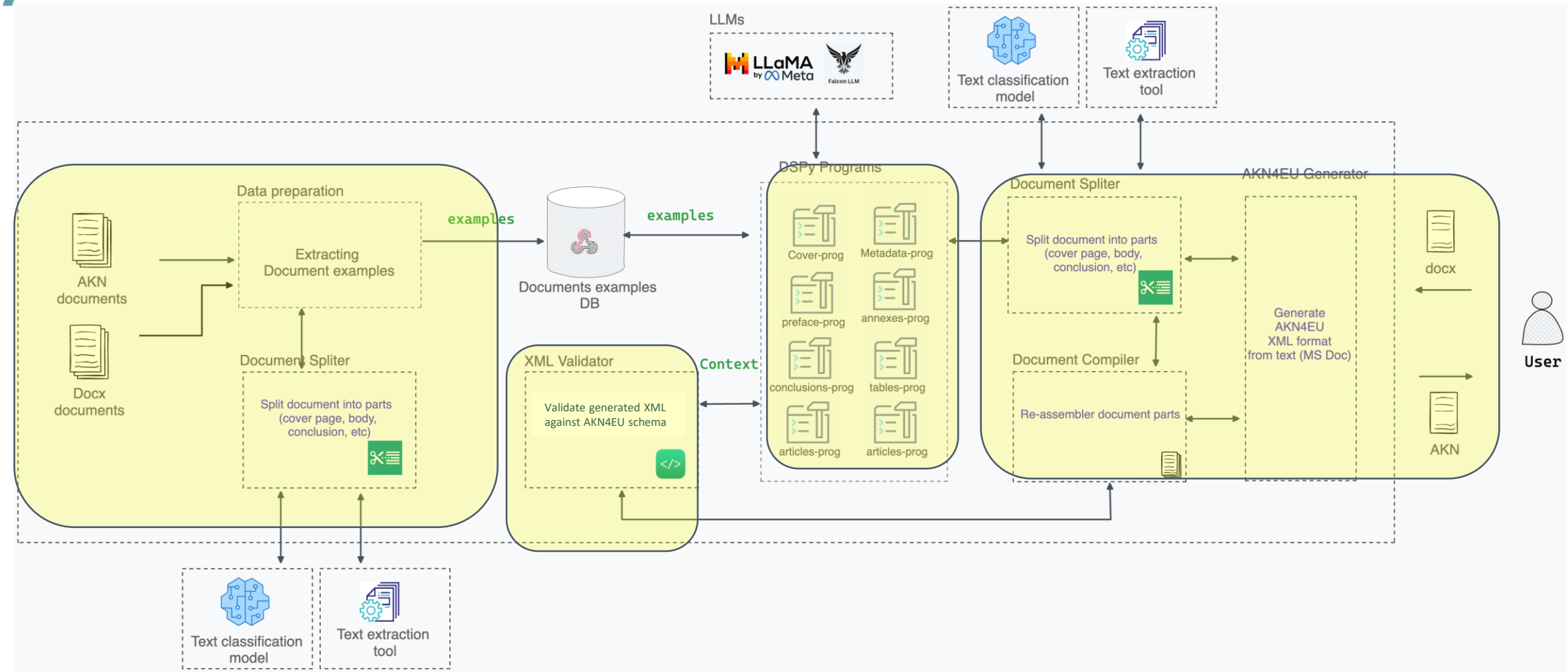
Program Optimizer:
BootstrapFewShot

Trainset:
Prefaces from EUR-Lex AKN Dataset (plain text + XML)

Metrics:
RougeL

DSPY will produce an Optimized program which can be called for inference as a tuned model

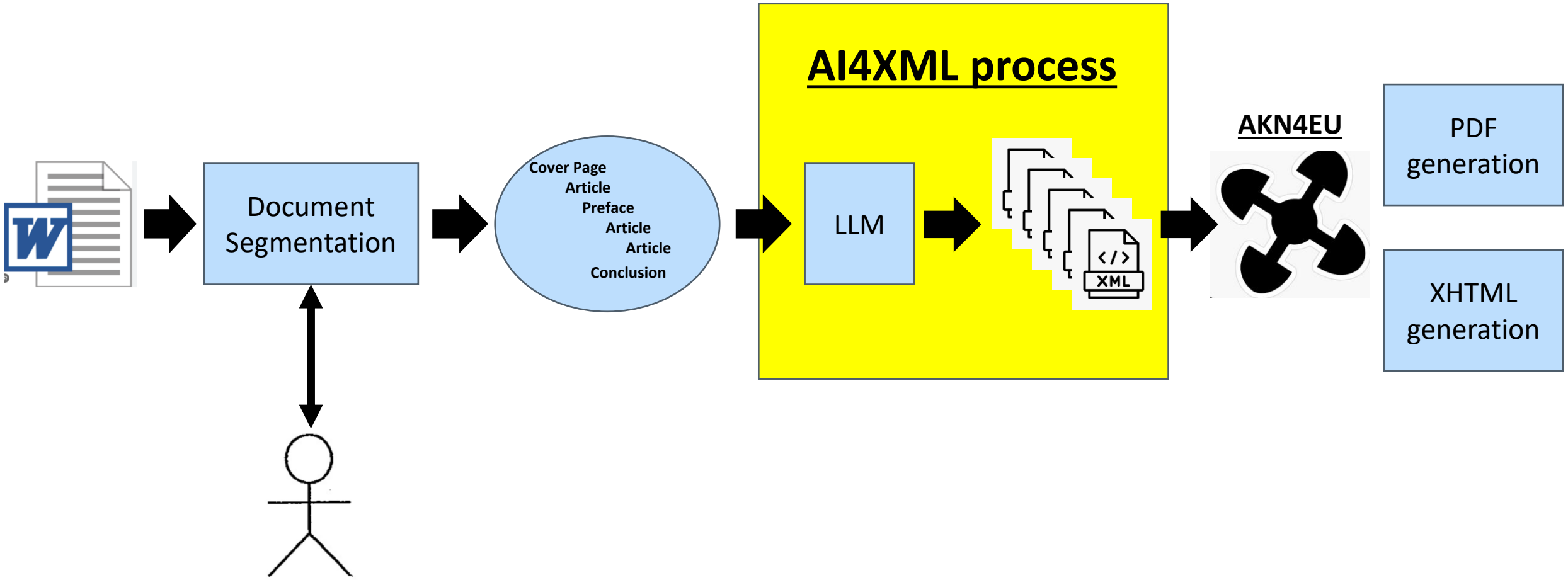
3. Experimentation.
LLM programs for XML generation



3. Experimentation.
LLM programs for XML generation :: next steps

- **Data preparation :**
  - Complete the preparation of the dataset :
    - A dataset per document part
    - Incorporate more context from AKN documentation
- **Define more complex LLM programs**
  - DSPy is a powerful tool to define complex prompting strategies
  - Evaluate each program separately, with its dataset and context
  - Involve XML schema, XML validation messages in the program compilation process
- **Improve validation methodology**
  - Currently we're using ROUGE and Bleu metrics for evaluation,
  - We intend to use METEOR, or other XML similarity measures
- **Fine-tune models and deploy the first version of the model (programs)**
  - Fine-tune (open) LLMs that provided best results
  - Deploy fine-tuned model or LLM programs

4. Product vision integration.



THANK YOU

Keep in touch via:

Dr KUSTER Marc (OP) Marc.KUSTER@ec.europa.eu

Head of Unit – OP.D.4 – "Interinstitutional Relations, Innovation, Programming and Compliance"