Publications Office
of the European Union

AI-READY DATA IN THE PUBLICATIONS OFFICE

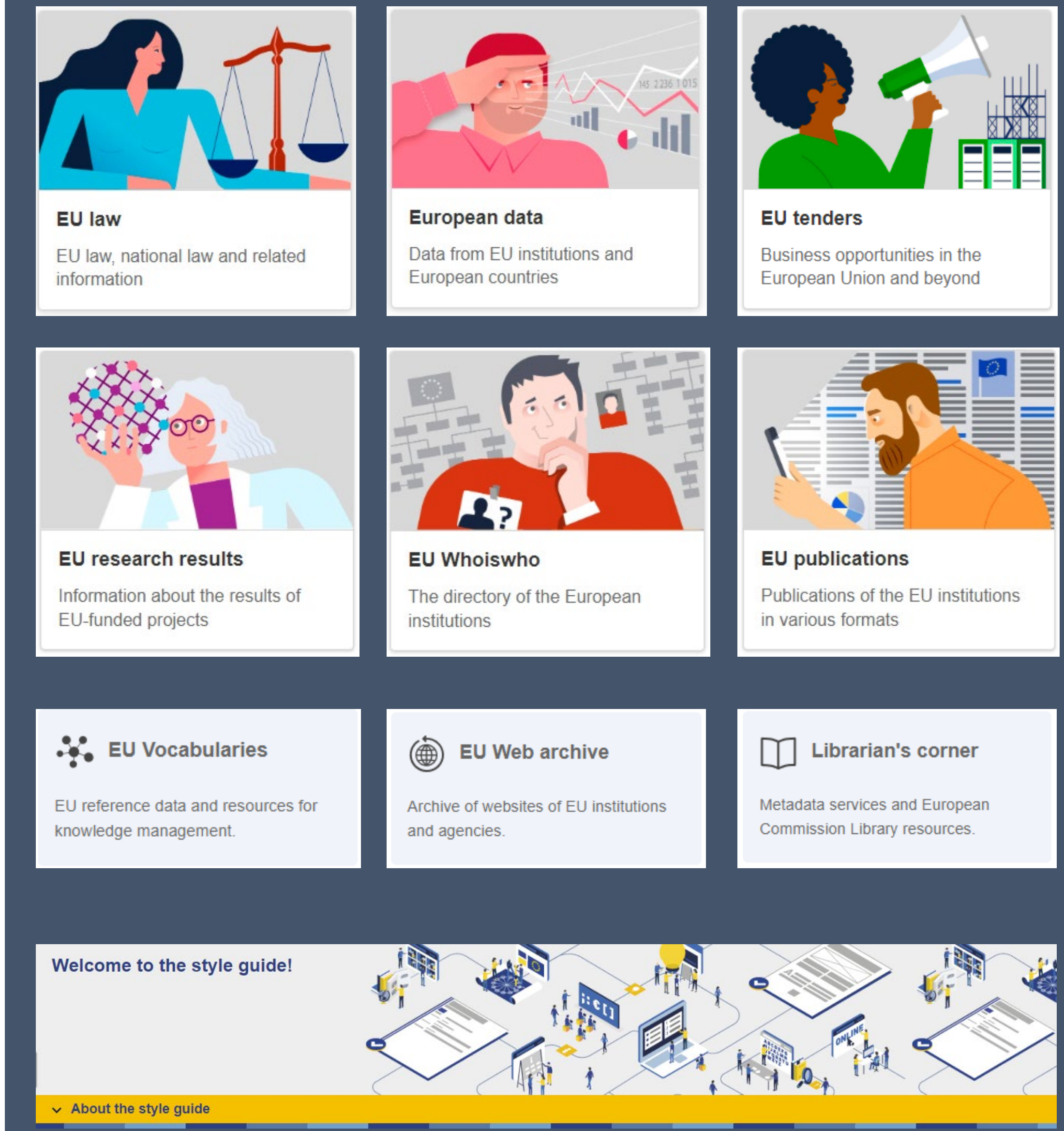
YOU HAVE THE AI FACTORIES, WE HAVE THE STANDARDS AND THE DATA

20 March 2026

vassilios.peristeras@ec.europa.eu

Publications Office of the European Union (OP): Who we are

- The official publisher for all EU institutions, bodies and agencies
- Collects, processes, annotates and curates huge volume of multilingual documents and data
- Makes available important open data via dedicated platforms e.g. data.europa.eu, EUR-Lex, TED, CORDIS
- We are deep in the data and metadata business



EU law
EU law, national law and related information

European data
Data from EU institutions and European countries

EU tenders
Business opportunities in the European Union and beyond

EU research results
Information about the results of EU-funded projects

EU Whoiswho
The directory of the European institutions

EU publications
Publications of the EU institutions in various formats

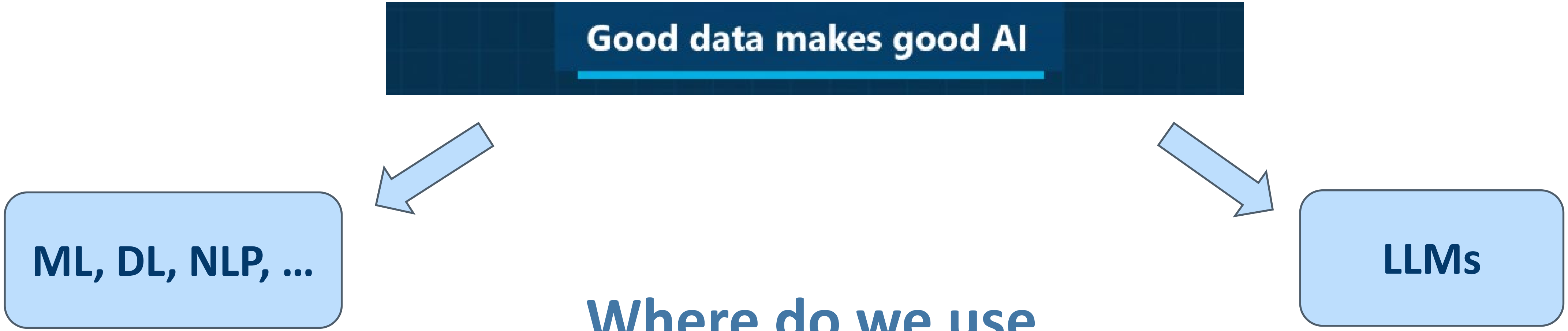
EU Vocabularies
EU reference data and resources for knowledge management.

EU Web archive
Archive of websites of EU institutions and agencies.

Librarian's corner
Metadata services and European Commission Library resources.

Welcome to the style guide!

AI-ready data



Good data makes good AI

ML, DL, NLP, …

LLMs

Where do we use “good” data in AI?

For model

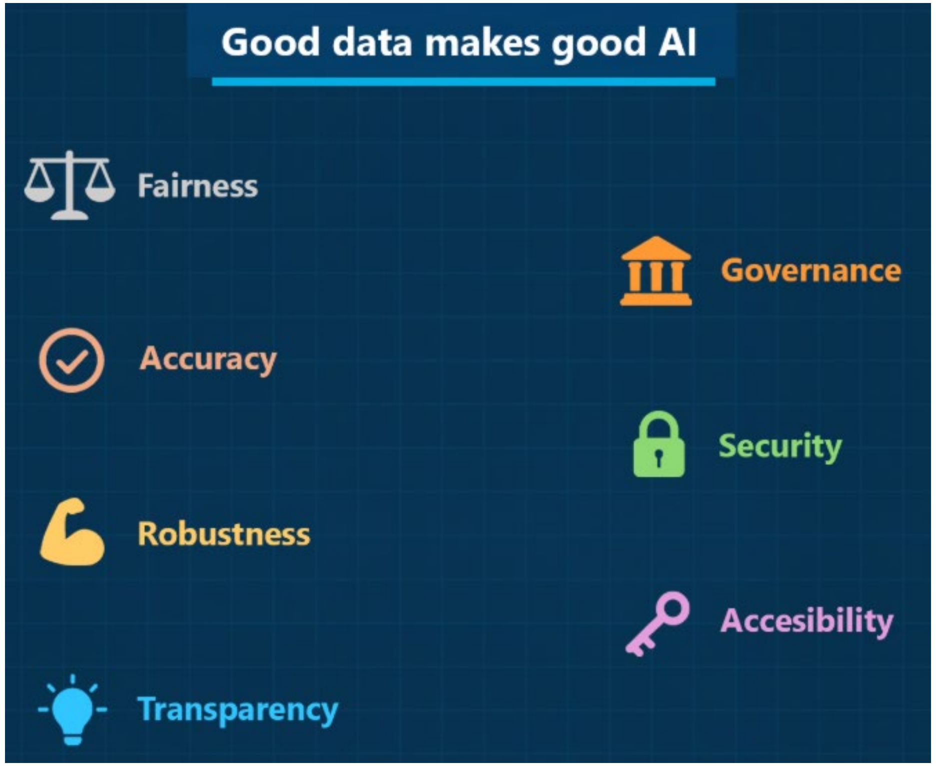
- Training
- Validation
- Testing

Maximize LLMs’ effectiveness with minimal noise

- Prompting
- Domain specific - RAG
- Evaluation/Benchmarking
- Fine-tuning

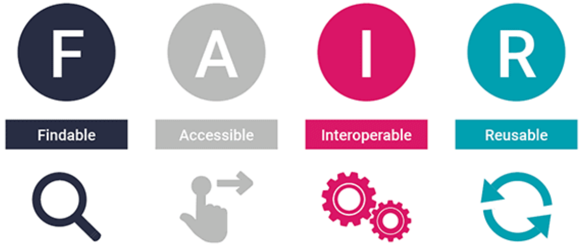
AI-Ready data

What is “good data” for AI?


Good data makes good AI

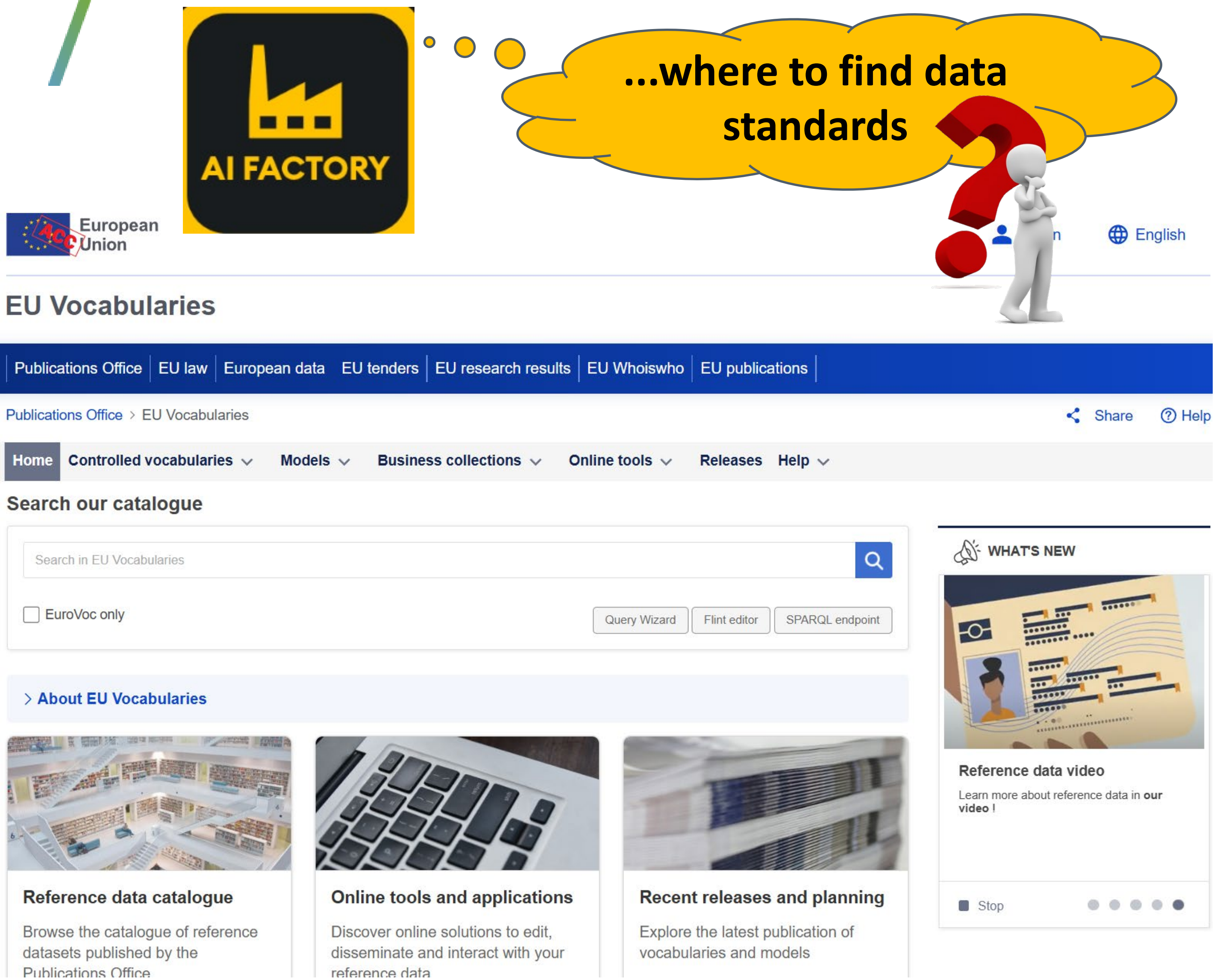
- Fairness
- Governance
- Accuracy
- Security
- Robustness
- Accesibility
- Transparency

- Compliance to the AI Act
- FAIR principles
- Data governed transparently
- Consistent/standard format
- Enriched with authoritative metadata
- Known provenance
- Validated for quality
- Human-in-the-loop (annotation, validation)



CONTROL QUALITY CONTROL

We have the standards…



European Union

English

EU Vocabularies

Publications Office | EU law | European data | EU tenders | EU research results | EU Whoiswho | EU publications

Publications Office › EU Vocabularies

Share | Help

Home | Controlled vocabularies | Models | Business collections | Online tools | Releases | Help

Search our catalogue

Search in EU Vocabularies

□ EuroVoc only

Query Wizard | Flint editor | SPARQL endpoint

› About EU Vocabularies

WHAT'S NEW

Reference data video

Learn more about reference data in **our video** !

Stop

Reference data catalogue

Browse the catalogue of reference datasets published by the Publications Office

Online tools and applications

Discover online solutions to edit, disseminate and interact with your reference data

Recent releases and planning

Explore the latest publication of vocabularies and models

- *Open standards, metadata and reference data*
- *Need for seamless integration and Interoperable across factories*

➢ 300 semantic assets
  - ❑ Reference data
  - ❑ Data standards
  - ❑ Thesauri
  - ❑ Taxonomies
  - ❑ Ontologies

➢ Collaborative editing
  - ❑ VocBench multilingual collaborative platform for managing controlled vocabularies

➢ Common governance with EU Institutions

➢ Authoritative status

We have the standards…

Authority lists

- Harmonised codes and labels used to support metadata exchange between EU institutions and beyond
- *Corporate bodies, Countries and territories, Currencies, Languages, Places*

Taxonomies

- *Eurostat statistical classifications, European Learning Model taxonomies, EuroSciVoc*

Thesauri

- *EuroVoc, Digital Europa Thesaurus, DE-BIAS Vocabulary*

Ontologies

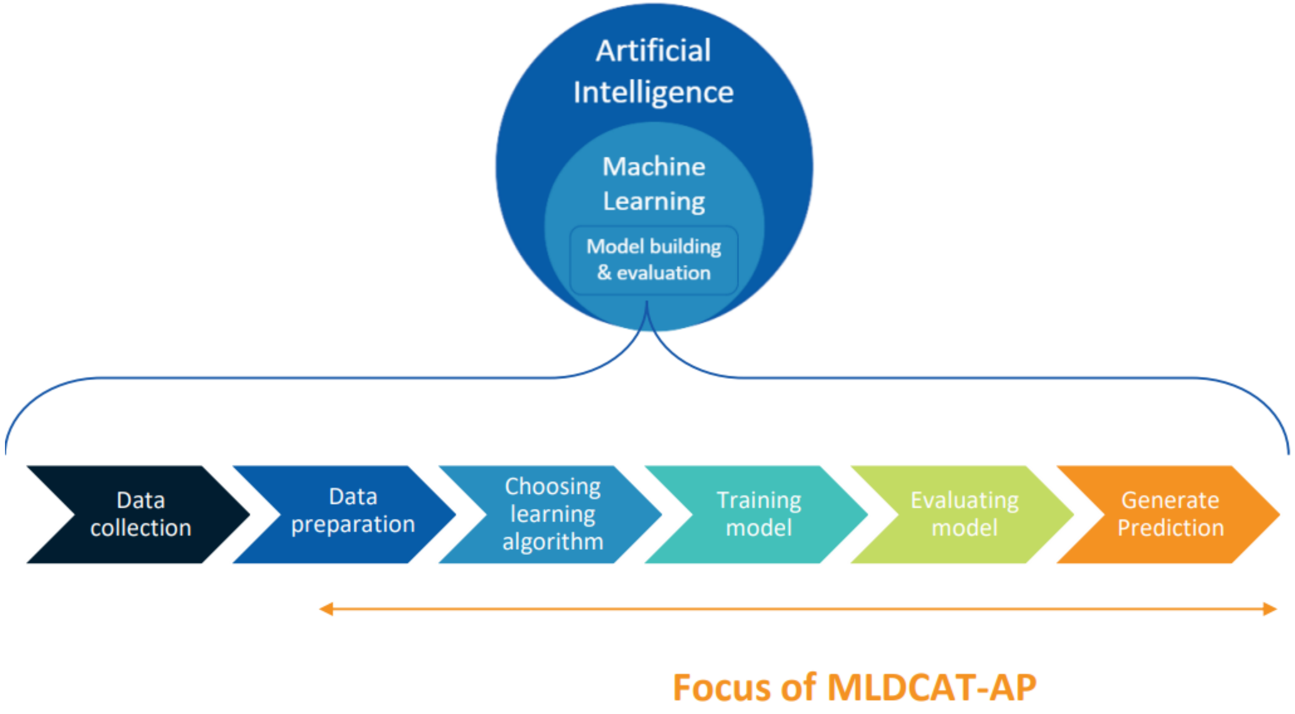
- *ELI ontology, ERA Vocabulary, AKN4EU schema, DE-BIAS ontology, eProcurement ontology*

Application profiles

- *DCAT-AP and its specifications: Mobility DCAT-AP, Health DCAT-AP, MLDCAT-AP*

We have the standards…

Domain of MLDCAT-AP



Data preparation — Classes describing the **Dataset** and **Quality Measurement** (cf. AI Act)

Choosing learning algorithm — **Algorithm** class (cf. AI Act)

Training model — **Machine Learning Model, Computer Infrastructure, Energy Consumption, Hardware,** and **Library** classes (cf. AI Act, AI CoP)

Evaluating model — **Benchmarks** and **Risks** of the model (cf. AI Act)

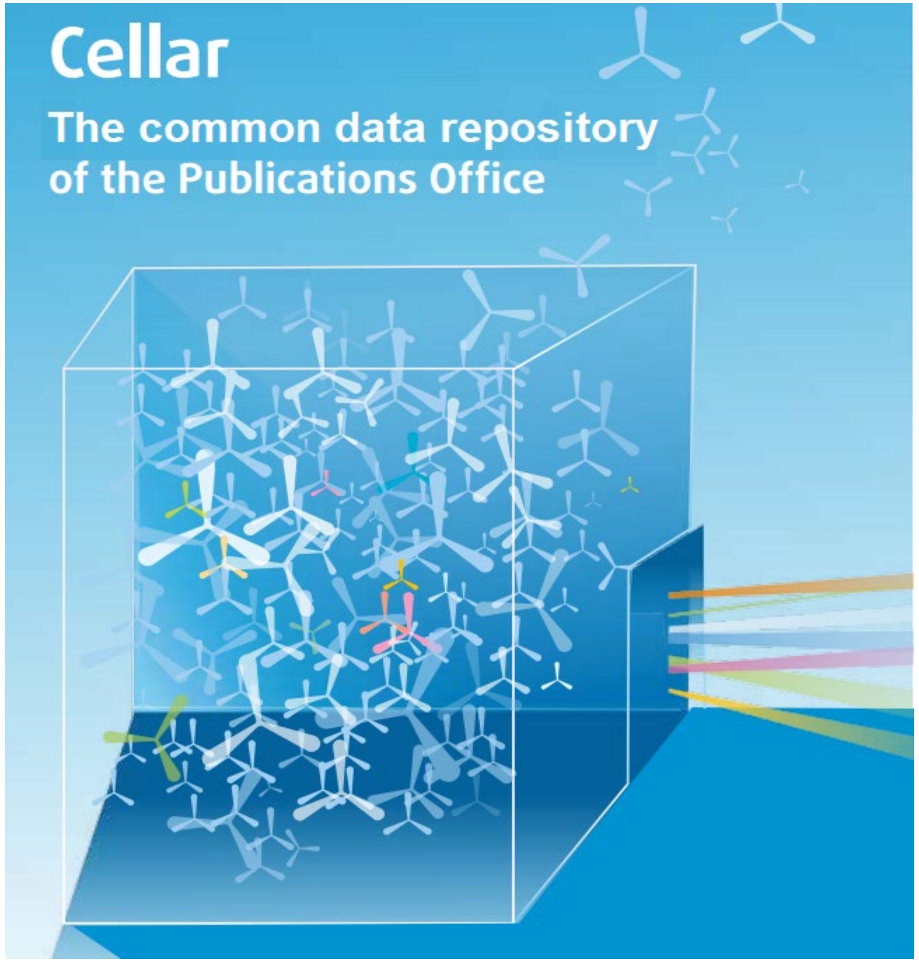
Generate Prediction — **Dataset** class as output of a **Machine Learning Model**

We have the data…

AI FACTORY

...where to find good data

1. ✓ Standards
2. ✓ Structure
3. ✓ Findability/Discovery
4. ✓ Access
5. ✓ Interoperability
6. ✓ Reusability
7. ✓ Veracity
8. ✓ Quality
9. ✓ Quantity
10. ✓ Multilingualism
11. ✓ Documentation
12. ✓ Community
13. ✓ Monitoring and updates



Cellar

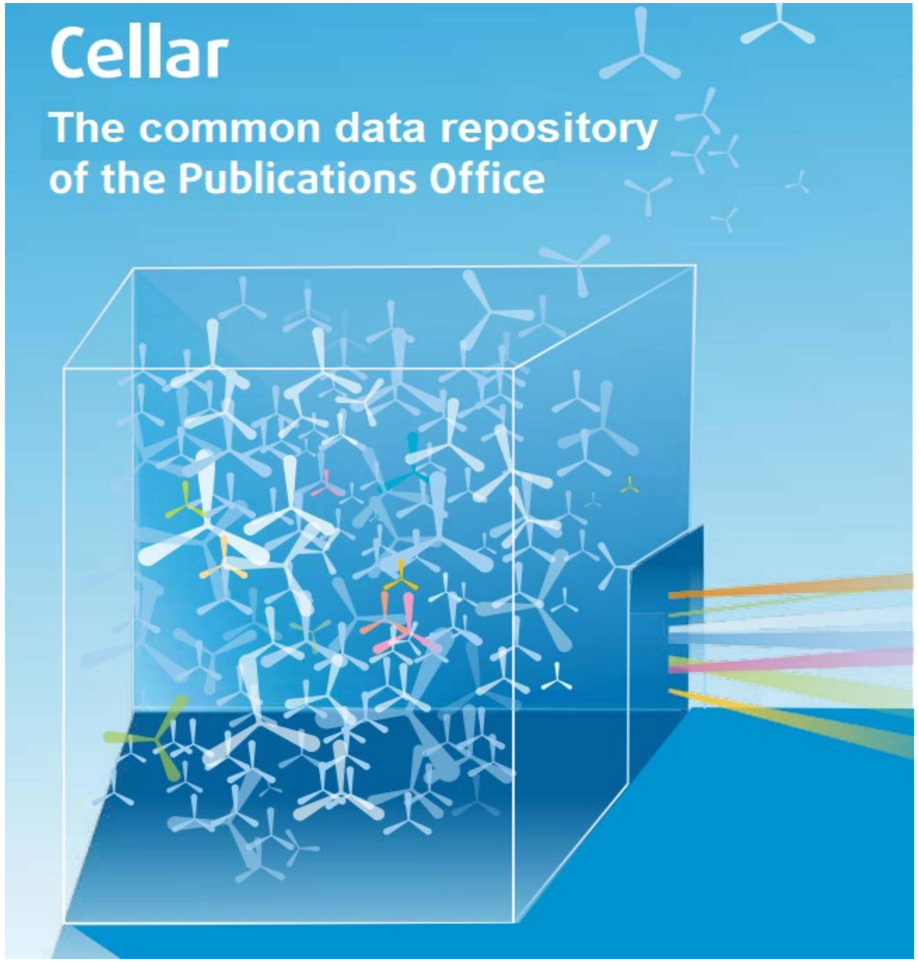
The common data repository of the Publications Office

We have the data…

AI FACTORY

...where to find good data

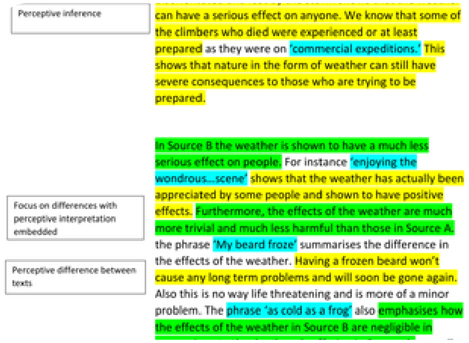
1. ✓ **Standards** => EU Vocabularies
2. ✓ **Structure** => Documents’ annotation
3. ✓ **Findability/Discovery** => Metadata & APIs
4. ✓ **Access =>** APIs & Bulk Downloads
5. ✓ **Interoperability =>** data standards and Knowledge Graphs
6. ✓ **Reusability =>** Clear licensing for reuse
7. ✓ **Veracity** => Authoritative documents and data
8. ✓ **Quality** => Curated, annotation, validation, cleaning, provenance
9. ✓ **Quantity** => over 44M documents
10. ✓ **Multilingualism** => 24 languages
11. ✓ **Documentation** => Machine-Readable
12. ✓ **Community** => Shared governance, feedback & Collaboration
13. ✓ **Monitoring and updates** => Analytics, monitor data quality, update schemas



We have the data…

Content

- What: Education material, reports, studies, legal document, Official Journal, Case Law
- By whom: Data providers are all EU institutions
- How: PDF, HTML, XML, RDF



Perceptive inference

can have a serious effect on anyone. We know that some of the climbers who died were experienced or at least prepared as they were on 'commercial expeditions.' This shows that nature in the form of weather can still have severe consequences to those who are trying to be prepared.

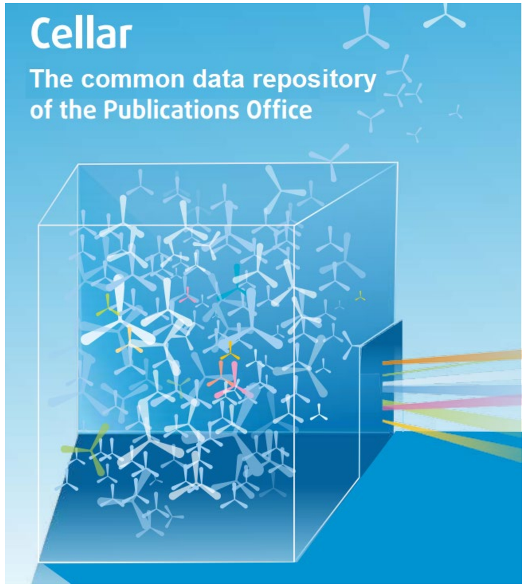
Focus on differences with perceptive interpretation embedded

Perceptive difference between texts

In Source B the weather is shown to have a much less serious effect on people. For instance 'enjoying the wondrous...scene' shows that the weather has actually been appreciated by some people and shown to have positive effects. Furthermore, the effects of the weather are much more trivial and much less harmful than those in Source A. the phrase 'My beard froze' summarises the difference in the effects of the weather. Having a frozen beard won't cause any long term problems and will soon be gone again. Also this is no way life threatening and is more of a minor problem. The phrase 'as cold as a frog' also emphasises how the effects of the weather in Source B are negligible in

Annotation

- Human annotation by professional cataloguers
- With structured metadata e.g. controlled vocabularies, thesauri



Cellar

The common data repository of the Publications Office

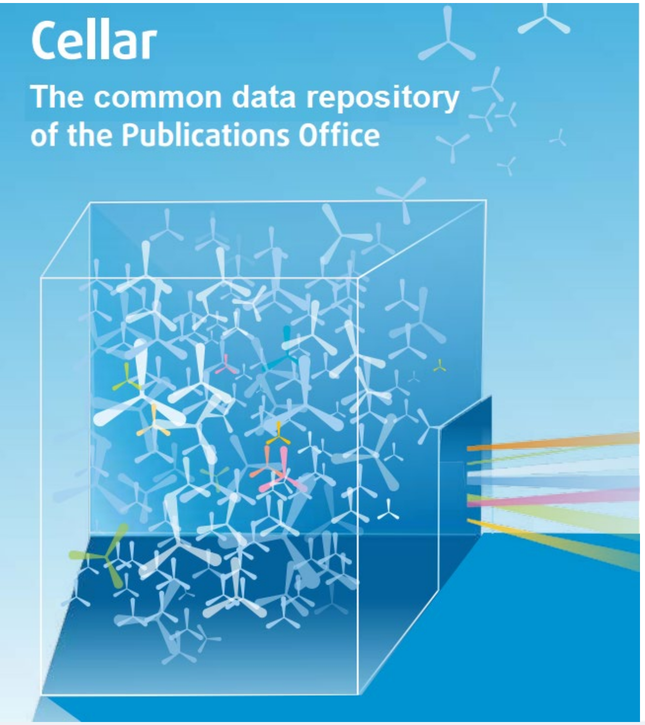
Cellar

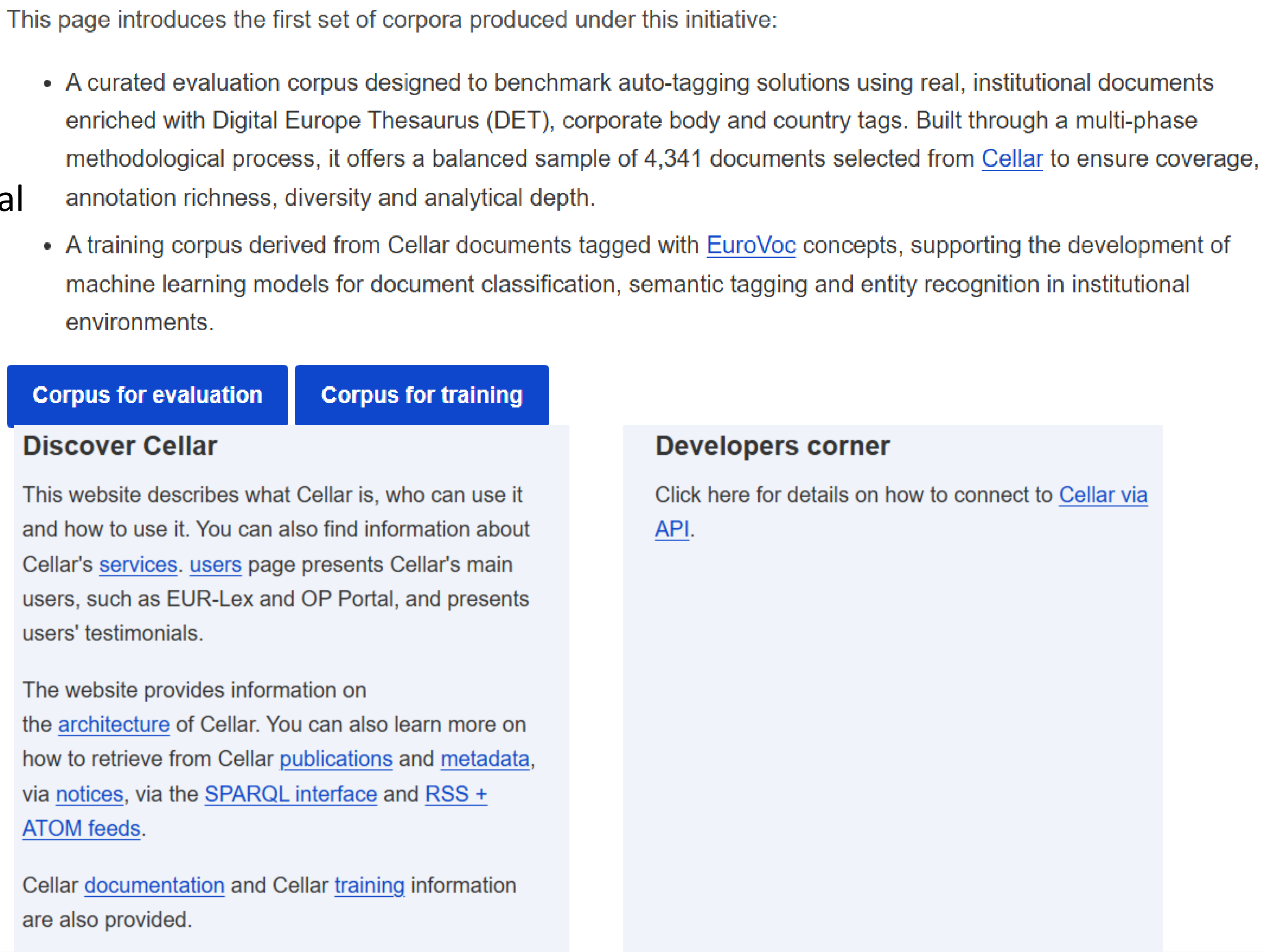
- 44 million of files, 40 TB
- 1B triples, 95M resources stored on knowledge graph
- Size increases every day
- Average 20 million requests/day
- Available via API
- Authoritative data
- Expert-validated annotations can be used for training AI models

We have the data…

Appetizer

- Training data for document classification, semantic tagging and information retrieval
- Datasets for benchmarking auto-tagging solutions





This page introduces the first set of corpora produced under this initiative:

- A curated evaluation corpus designed to benchmark auto-tagging solutions using real, institutional documents enriched with Digital Europe Thesaurus (DET), corporate body and country tags. Built through a multi-phase methodological process, it offers a balanced sample of 4,341 documents selected from Cellar to ensure coverage, annotation richness, diversity and analytical depth.
- A training corpus derived from Cellar documents tagged with EuroVoc concepts, supporting the development of machine learning models for document classification, semantic tagging and entity recognition in institutional environments.

Discover Cellar

This website describes what Cellar is, who can use it and how to use it. You can also find information about Cellar's services. users page presents Cellar's main users, such as EUR-Lex and OP Portal, and presents users' testimonials.

The website provides information on the architecture of Cellar. You can also learn more on how to retrieve from Cellar publications and metadata, via notices, via the SPARQL interface and RSS + ATOM feeds.

Cellar documentation and Cellar training information are also provided.

Developers corner

Click here for details on how to connect to Cellar via API.

Coming soon…

Webinar

MLDCAT-AP

for AI factories

Thursday, 7 May 2026

10:30h – 12:00h CEST

interoperable
europe
European
Commission

Publications Office
of the European Union

AI-READY DATA IN THE PUBLICATIONS OFFICE

YOU HAVE THE AI FACTORIES, WE HAVE THE STANDARDS AND THE DATA

20 March 2026

vassilios.peristeras@ec.europa.eu