Application of machine learning and big data technologies in OpenAIRE system

Warsztaty Orange z cyklu
„Centrum Badawczo Rozwojowe zaprasza”

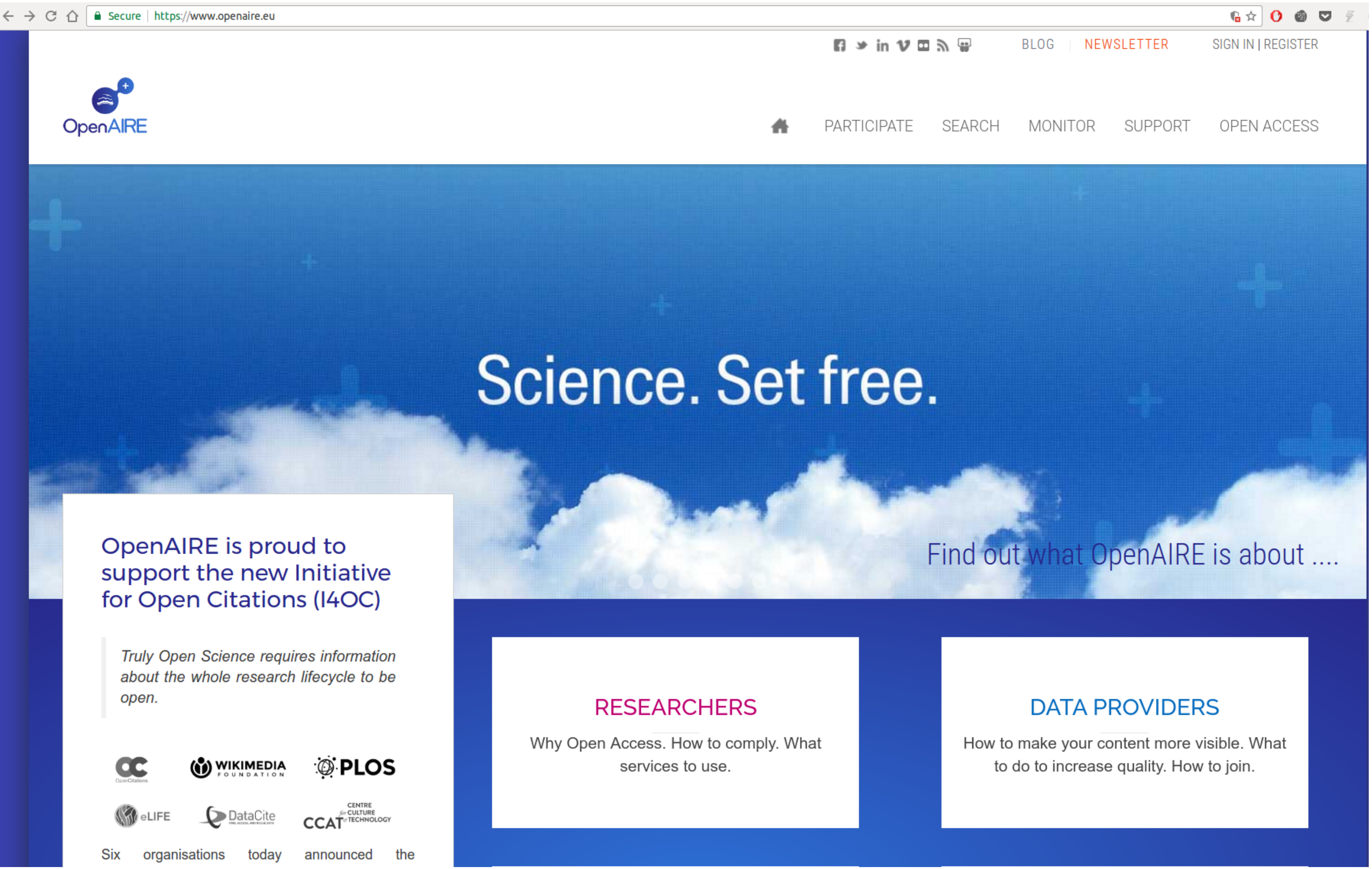
Mateusz Kobos, ICM, Univeristy of Warsaw

Warszawa, 2017-05-10

OpenAIRE system

- What does the system do?
  - **Gathers** various scholarly information (publications, datasets, persons, fundings, and organizations) from publicly available sources (repositories, Current Research Information Systems).
  - **Presents** the aggregated and de-duplicated view to the user through www.openaire.eu portal.
- Who are its users?
  - **Scientists** – the portal provides tools for dissemination and discovery of research results.
  - **Funding bodies and organizations** – the portal provides tools for measuring and refining funding investments in terms of their research impact.
- Who develops it? A consortium of European research organizations and founded by European Union.

Screenshot of the OpenAIRE portal



Secure | https://www.openaire.eu

BLOG | NEWSLETTER SIGN IN | REGISTER

OpenAIRE

PARTICIPATE SEARCH MONITOR SUPPORT OPEN ACCESS

Science. Set free.

Find out what OpenAIRE is about ....

OpenAIRE is proud to support the new Initiative for Open Citations (I4OC)

> *Truly Open Science requires information about the whole research lifecycle to be open.*

OpenCitations WIKIMEDIA FOUNDATION PLOS eLIFE DataCite CCAT CENTRE for CULTURE & TECHNOLOGY

Six organisations today announced the

RESEARCHERS

Why Open Access. How to comply. What services to use.

DATA PROVIDERS

How to make your content more visible. What to do to increase quality. How to join.

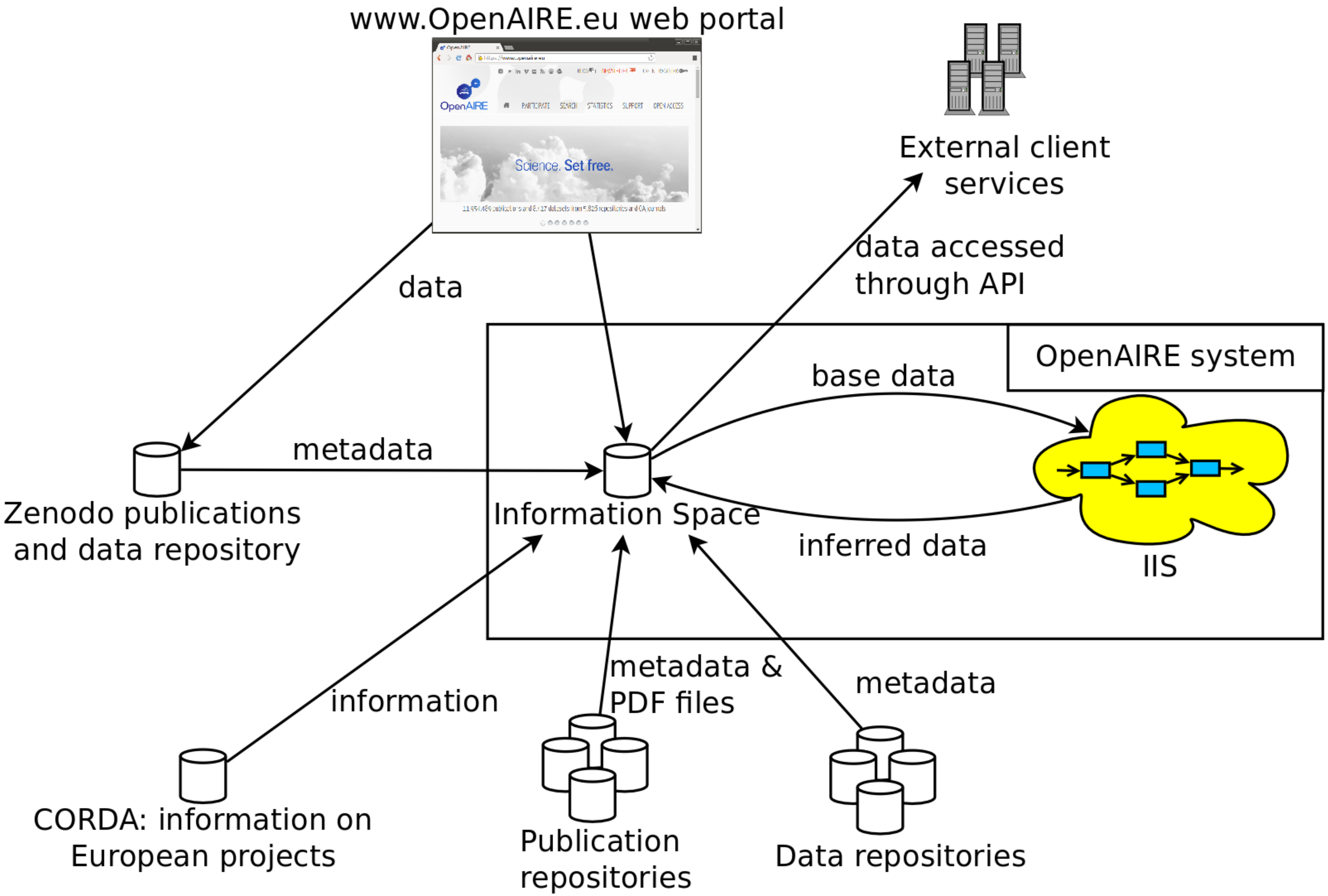
Data processed by the OpenAIRE system

- Where does the data come from?
  - Data providers that OpenAIRE continuously collects information from: ~800
    - They have to provide OpenAIRE-compatible API
    - Types of repositories: institutional, thematic, data, journals, aggregators, etc.
  - CORDA: a system with information about EU projects
  - Others
- What data is stored in the system?
  - Publication metadata: 17 millions
  - Authors: 16 millions
  - Full-text documents: 4 millions
  - Dataset metadata: 3 millions
  - Projects: 700 thousands
  - Organizations: 60 thousands

Information Inference Service (IIS)

- Information Inference Service (IIS) is a part of the OpenAIRE system. It is developed by ICM in cooperation with partners from other countries from the consortium.
- Its goal is to do **data/text mining** of the gathered data available in OpenAIRE system.
- It's **open source**. The code is available at https://github.com/openaire/iis
- The development of IIS **started in 2012**.

IIS as a part of the OpenAIRE system



www.OpenAIRE.eu web portal

External client services

data

data accessed through API

OpenAIRE system

base data

metadata

Zenodo publications and data repository

Information Space

inferred data

IIS

information

metadata & PDF files

metadata

CORDA: information on European projects

Publication repositories

Data repositories

What functionality does IIS provide?

- IIS consists of a few data processing workflows. The workflows contain various data mining modules that work mostly on the content of documents. Their functionalities:
  - Extract: references to datasets, references to projects, references to research communities, software links, protein database references, citation links,
  - Infer metadata from the content of the PDF document (uses CERMINE)
  - Classify documents
  - Find similar documents
  - Match citation links extracted from document content with actual documents
  - Match author affiliations extracted from document content with actual organizations

How does IIS work?

- IIS is a Hadoop **cluster “application”**.
  - Based on Apache Hadoop technologies: Oozie, MapReduce, Spark, Pig, Avro, Hive (for analytics).
- Using IIS is **like calling a function** with subsequent stages**:**
  - The client **starts IIS** passing it **parameters** that define:
    - what modules will be run,
    - what data sets they will be run on,
    - parameters of the modules.
  - IIS **execution**:
    - imports required data,
    - processes the data using selected modules (this takes a few hours),
    - exports produced data.
  - IIS **shuts down**.
- IIS is **stateless** – no information is kept between subsequent runs of IIS
  - (apart from cache used internally)

A few numbers about IIS

- Hadoop cluster specification:
  - Distribution: Cloudera CDH5 (v.5.9.0)
  - 16 slave nodes, each one with identical specification. This sums up to:
    - CPU: 384 cores, 768 threads
    - RAM: 2048GB
    - HDD: 384TB (HDFS)
- Duration of the longest processing workflow: 15 hours

Thank you for your attention!