Research ecosystem: From publications, through data to research artefacts

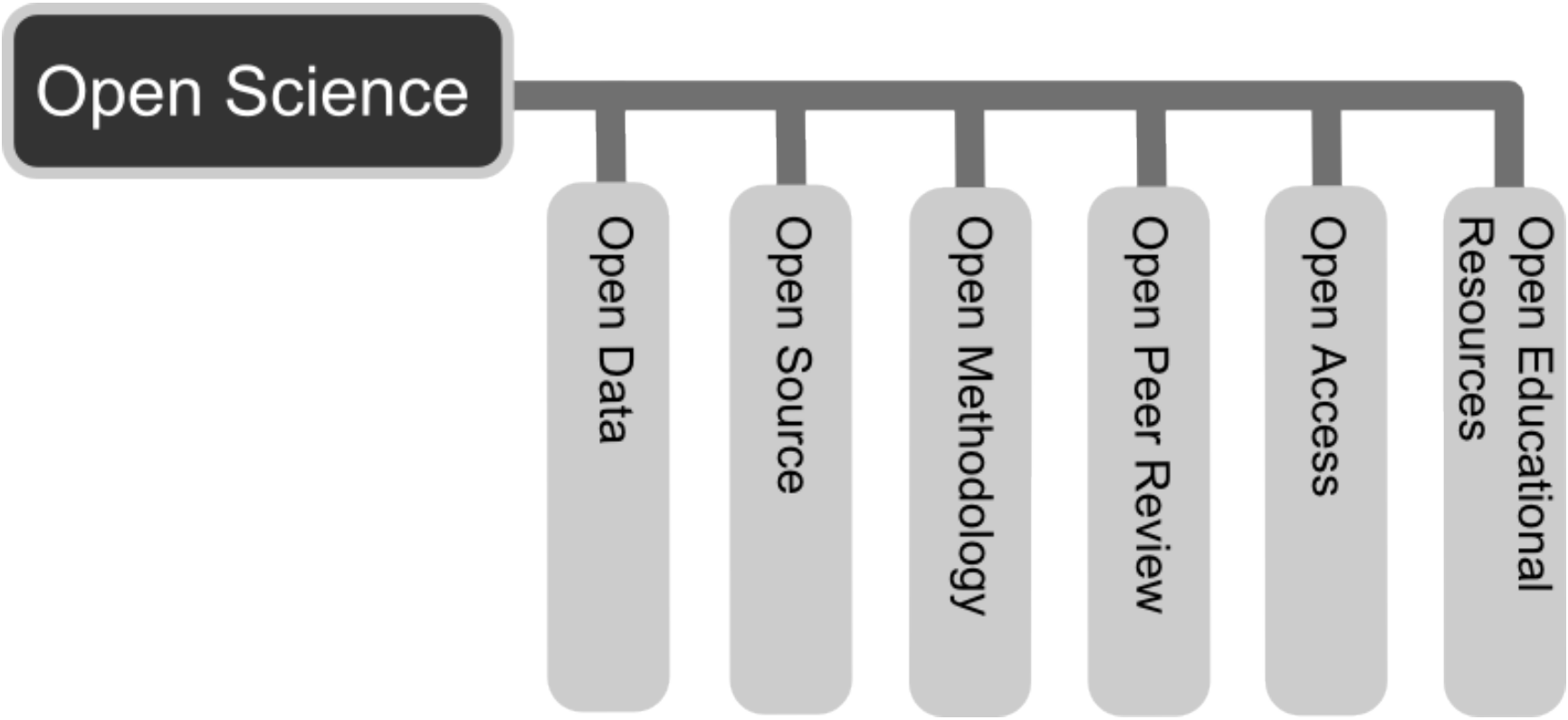
Wojtek Sylwestrzak
ICM, University of Warsaw

RDA meets researchers in Poland
1 February 2017, Warsaw, Poland

Open research data and the RDA

The RDA vision is researchers and innovators **openly sharing data** across technologies, disciplines, and countries to address the grand challenges of society.

The six principles of Open Science



Openly sharing research data

Purposes:

- Reproducibility
- Reuse

Requirement:

- Data have to be shared in a useful manner – together with their context

Data vs research articles

Historically, research datasets have been seen as an extension to traditional publications:

- Much of the OA activism has been driven by librarians
- Research evaluation model is still very conservative and primarily based on publications
- It supports the traditional publishers' publication model, even in data journals

Data vs research articles

Heavily interlinked: research publications can cite data, on the other hand collections of articles can be considered a form of data and text mined.

In most cases both data and articles are still considered to be static at the publication time.

The distinction between data and publications is blurred:

- Publications are (still) primarily intended to be consumed by humans
- Data are primarily consumed by machines

Broader perspective

If you take a step back, there are many more research artefacts than just papers and data.

Data and text articles can be seen as types or research artefacts.

publications, datasets, database tables, software, models and methods, research groups, labs, equipment, institutions etc.

Again, a collection of publications can be considered a dataset e.g. for text mining.

COMAC
The Common Map of Academia

A graph of objects and their relations extracted from aggregated:

- OA publications from repositories, journal databases, OA journals, blogs, Internet crawling, ORCID, DataCite, CrossRef

A public dataset of almost 2 billion RDF triples.

COMAC
The Common Map of Academia

Building and updating process:

- Import phase: aggregating and storing data in Hbase.
- Processing phase: a workflow of Spark based machine-learning modules to extract metadata (where missing) from fulltext, extract, deduplicate and disambiguate objects, extract and match citations, keywords, classify, build relations, enrich metadata etc.
- Publishing phase: storing the results in a triplestore (RDF) and sequence files, make available for download and through an online navigator UI.

COMAC Navigator

Search | Info | Help

Browsing graph

Double click on a node to expand it, i.e. to add all of its neighobours.

Single click on a node to view additional informations about a node.

Node types:

- Journal
- Dataset
- Project
- Paper
- Organisation
- Author
- Blog

OpenAIRE – the European OA Research e-infrastructure

Collecting information from publications in OA repositories (links to datasets, affiliations, projects, funders).

Running its own repository for data and publications.

All the text-mining modules are run on Hadoop and Spark clusters in Poland.

Currently extending its data model to support a wider range of artefacts and provide them in research context (e.g. how to use the data, how were they created, legal context etc.)

OpenAIRE – the European OA Research e-infrastructure

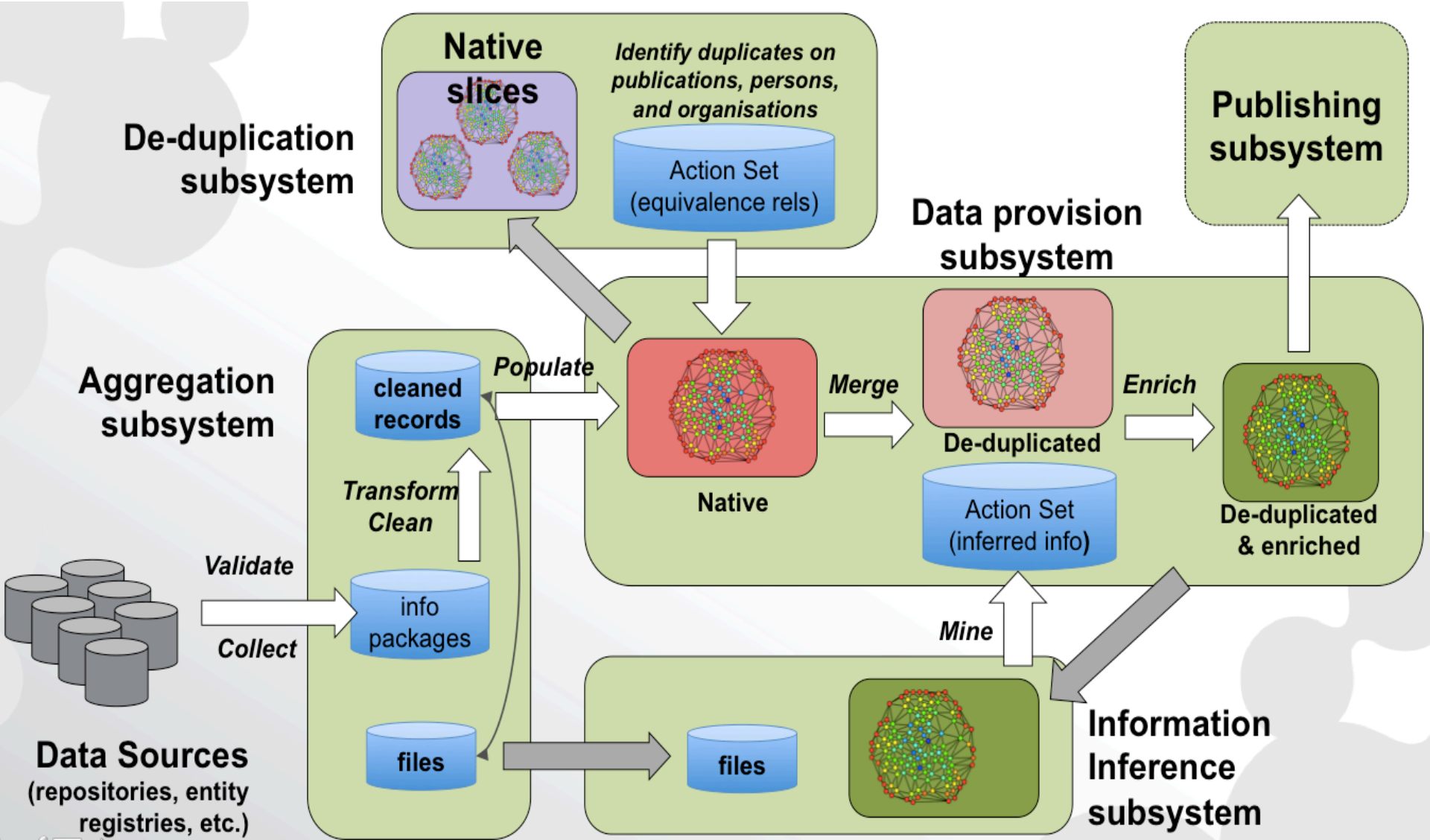
Collecting information from publications in OA repositories (links to datasets, affiliations, projects, funders).

Running its own repository for data and publications.

All the text-mining modules are run on Hadoop and Spark clusters in Poland.

Currently extending its data model to support a wider range of artefacts and provide them in research context (e.g. how to use the data, how were they created, legal context etc.)

OpenAIRE workflow



RepOD – the Polish repository do open research data

Thank you

COMAC http://comac.ceon.pl/

OpenAIRE http://www.openaire.eu/

RepOD http://repod.pon.edu.pl/

Wojtek Sylwestrzak W.Sylwestrzak@icm.edu.pl