openMINTED

Sophie Aubin - INRA

OpenMinTeD

Text mining services for e-infrastructures

CAPSELLA Open Data Workshop, Chania, 2 June 2017

European Commission

Amounts of scientific texts...

1 paper/sec

90% of papers **never** cited*

120,000 papers published on a single taxon "**Zea mays**"

50% of papers **never** read by anyone than its authors, referees, journal editors*

+ reports, patents, books, surveys, news, etc.

*Lokman I. Meho, The rise and rise of citation analysis, 2007

TDM services

Indexing documents and datasets

Entities **identification** and **normalisation** against reference data

Information **extraction**: from semi-structured to structured

Semantic/lexical resource **acquisition** from texts

However, …

Multitude of solutions catering for different

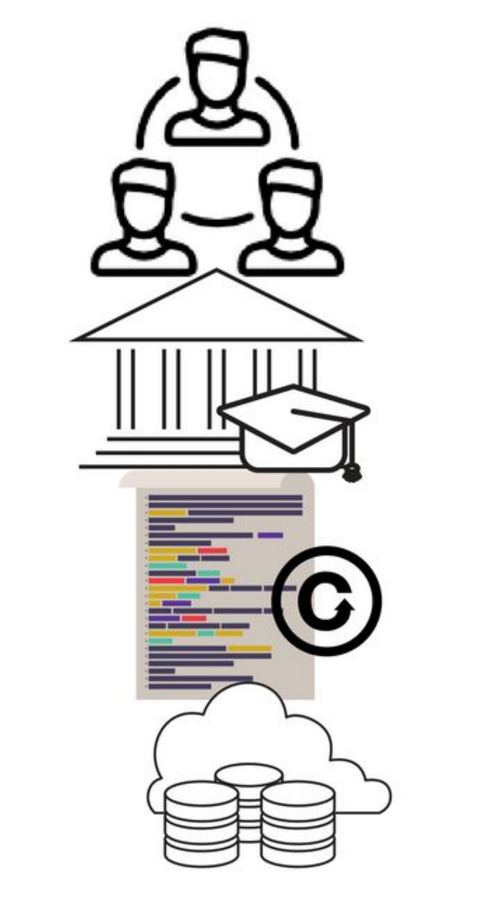
Text Types	Domains	Tasks	Languages
Newswire	Agriculture	Translation	English
Scientific Literature	Health	Knowledge acquisition	French
Tweets/blogs	Biology	Semantic Search	German
Patents	Social Sciences	Question Answering	Spanish
Clinical/medical records	Environment	Sentiment Analysis	Portuguese
Textbooks, monographs	….	Summarization	Italian
Online forums		Knowledge Discovery	Polish
		….	….

Creating a fragmented landscape

Then comes openMINTED

Duration: 3 years (2015-2018), **16 Partners**

- research groups in text-mining
- content providers
- a data center
- a library association
- legal experts
- community related partners
- SMEs



Partners of openMINTED

ATHENA
Research & Innovation
Information Technologies

LIBER LIBER

BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

University of Glasgow

The Open University

EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

INRA
SCIENCE & IMPACT

Agroknow

UNIVERSITY OF AMSTERDAM

NaCTeM
The National Centre for Text Mining

grnet

gesis
Leibniz Institute
for the Social Sciences

general architecture
GATE
for text engineering

EMBL-EBI

TECHNISCHE UNIVERSITÄT DARMSTADT

DKPro

frontiers

What is openMINTED?

openMINTED is a **platform** that works as an **infrastructural service** of the wider **research ecosystem**

Our Services

1 Discover TDM Services and tools

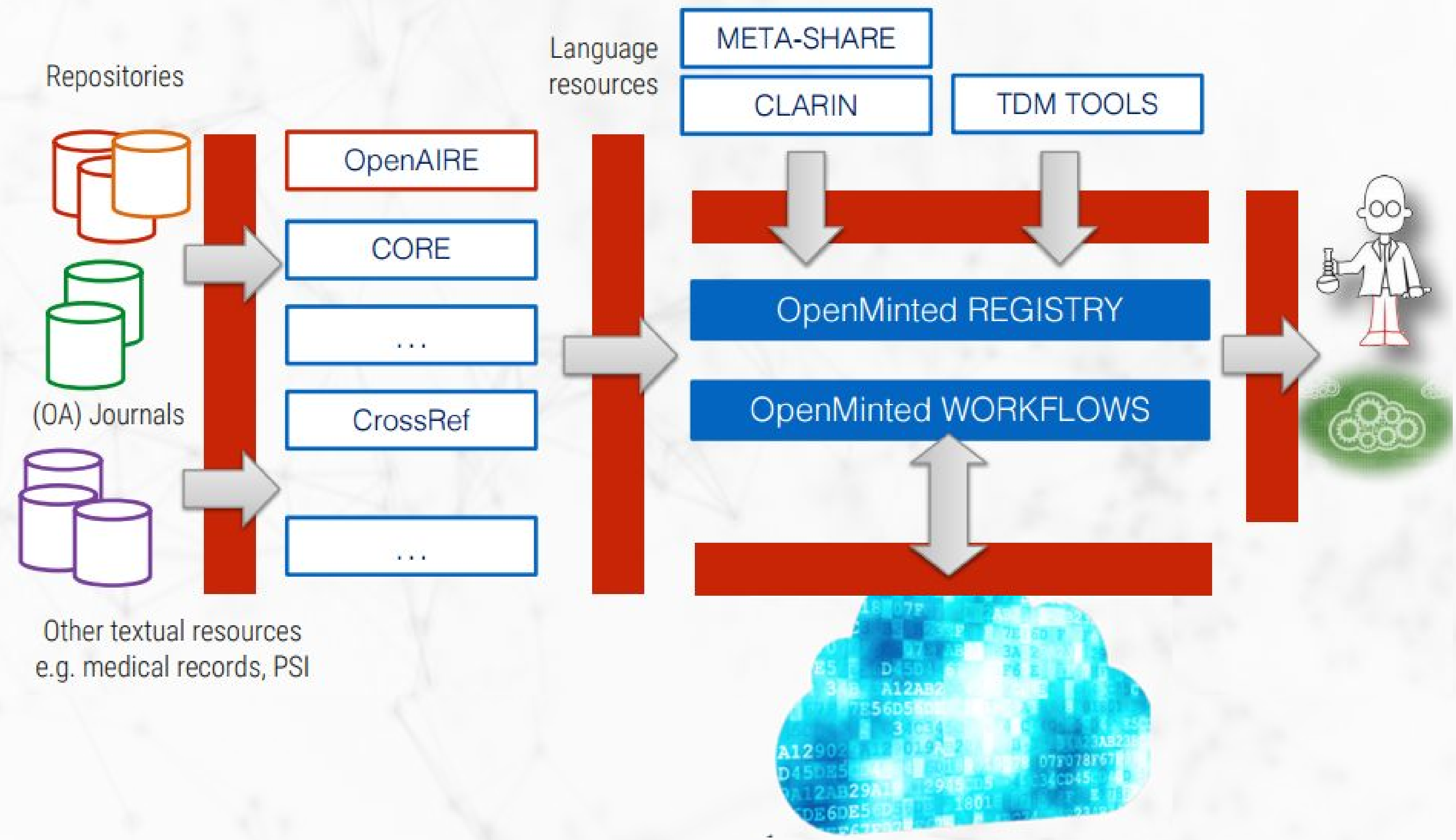
2 Feed with home content or easily get texts from (OA) content hubs

3 Pick adequate knowledge resources

4 Build your own service/applications – Combine components into a Workflow

5 Share and Re-Use

How does all this bind together?



Language resources

META-SHARE

CLARIN

TDM TOOLS

Repositories

OpenAIRE

CORE

. . .

CrossRef

. . .

(OA) Journals

OpenMinted REGISTRY

OpenMinted WORKFLOWS

Other textual resources e.g. medical records, PSI

OpenMinTeD Scientific/Business Applications

Data & content repositories/registries

For Data analysts, Scientists

Curation tools

For Data providers

Analytical tools

For Scientists-policy makers

Decision tools

For Farmers, SMEs

Knowledge acquisition tools

For Scientists, Ontologists

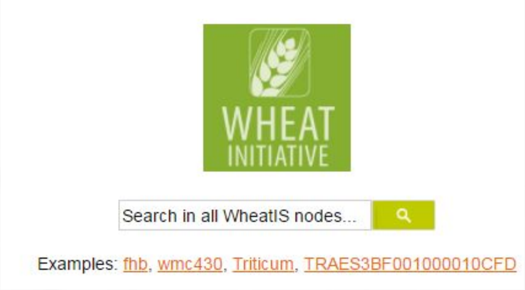
Example: WheatIS / gnpIS

Application: federated search of genomic and genetic data for wheat

Added value: direct access from data to related scientific articles

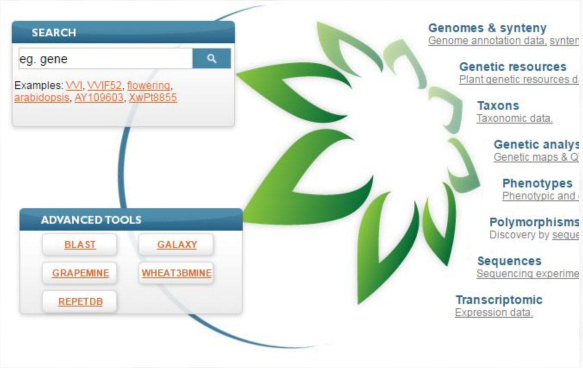
Objects: taxa, genes, markers, phenotypes and varieties

Challenges: naming heterogeneity & scale variety between textual and experimental data


WHEAT
INITIATIVE

Search in all WheatIS nodes...

Examples: fhb, wmc430, Triticum, TRAES3BF001000010CFD


SEARCH

eg. gene

Examples: VVI, VVIF52, flowering, arabidopsis, AY109603, XwPt8855

ADVANCED TOOLS

BLAST GALAXY
GRAPEMINE WHEAT3BMINE
REPETDB

Genomes & synteny
Genome annotation data, synten

Genetic resources
Plant genetic resources d

Taxons
Taxonomic data.

Genetic analys
Genetic maps & Q

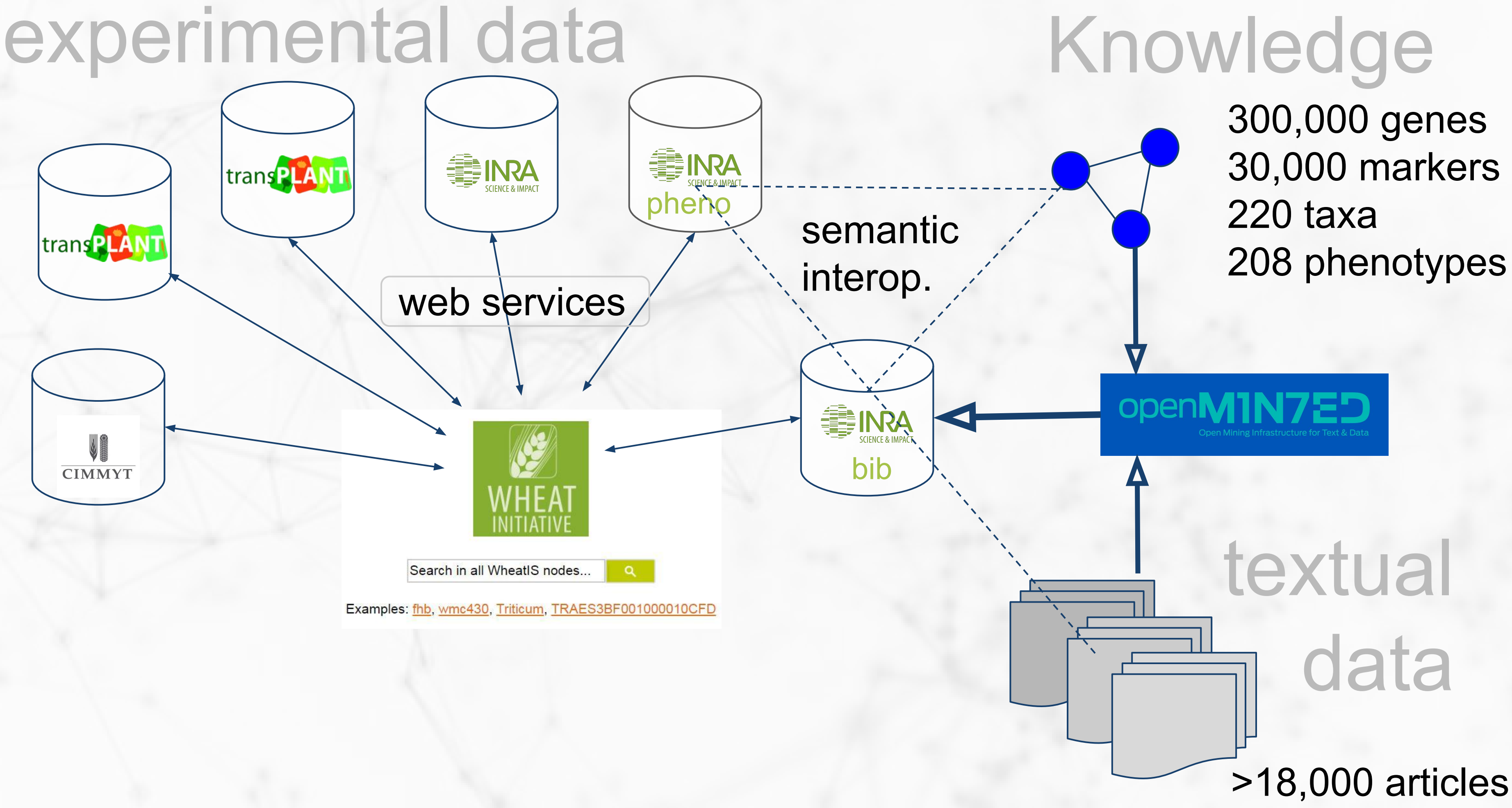
Phenotypes
Phenotypic and

Polymorphisms
Discovery by seque

Sequences
Sequencing experime

Transcriptomic
Expression data.

data - text interoperability



experimental data

Knowledge

300,000 genes
30,000 markers
220 taxa
208 phenotypes

web services

semantic interop.

textual data

>18,000 articles

transPLANT
transPLANT
INRA SCIENCE & IMPACT
INRA SCIENCE & IMPACT pheno
CIMMYT
WHEAT INITIATIVE
Search in all WheatIS nodes...
Examples: fhb, wmc430, Triticum, TRAES3BF001000010CFD
INRA SCIENCE & IMPACT bib
openMINTED
Open Mining Infrastructure for Text & Data

Example: living conditions of food-related micro-organisms

Application: study and characterize the microbial biodiversity of food ecosystems (dairy or meat products, fish, wine, etc.)

- risk management
- food quality improvement

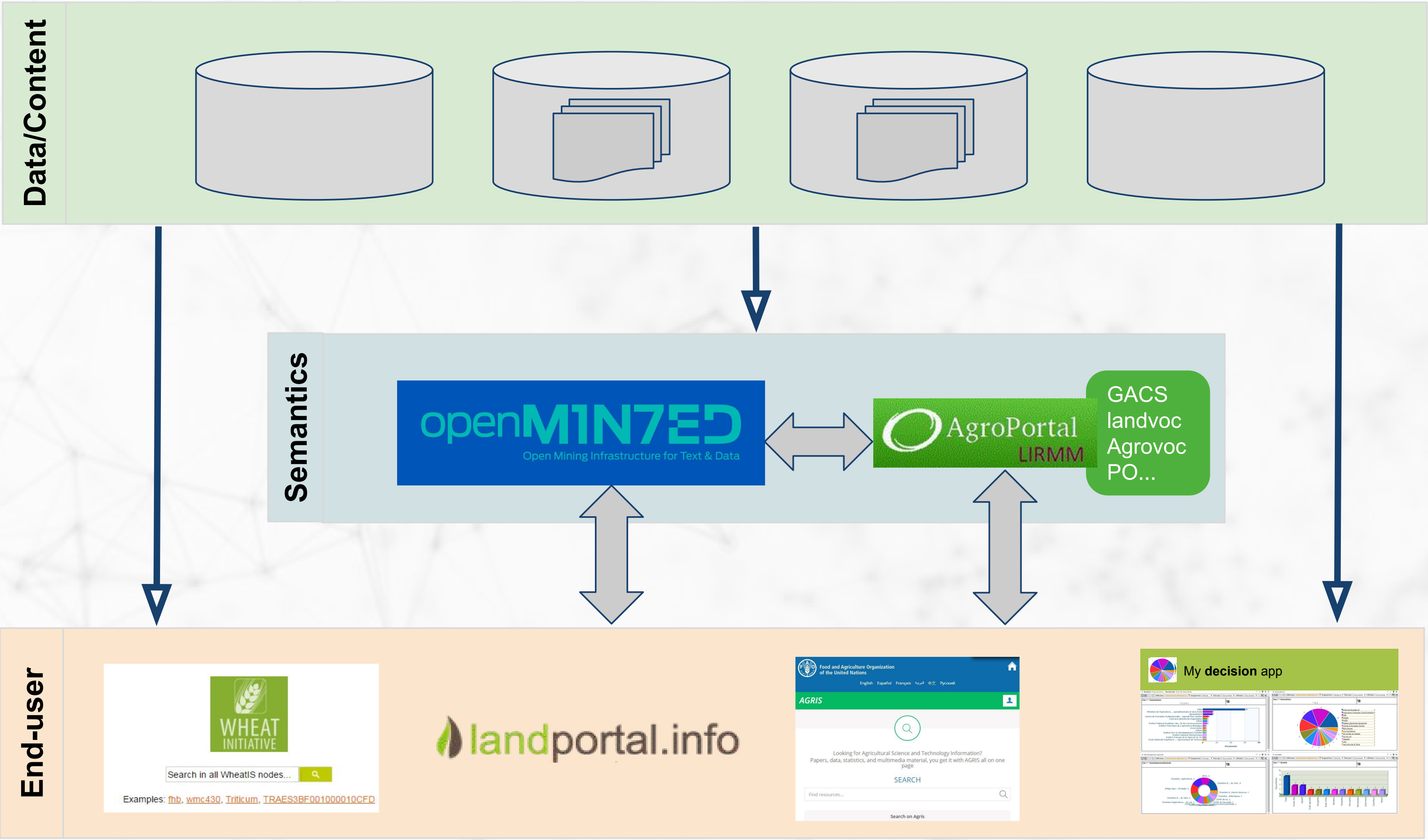
Added value: completion of knowledge databases with info from the literature

Objects: bacteria, habitats and phenotypes

Challenges: heterogeneous data integration, object identity

CAPSELLA Open Data Workshop, Chania, 2 June 2017

OpenMinTeD for agriculture



Data/Content
Semantics
openMINTED
Open Mining Infrastructure for Text & Data
AgroPortal
LIRMM
GACS
landvoc
Agrovoc
PO…
End-user
WHEAT INITIATIVE
Search in all WheatIS nodes...
landportal.info
Food and Agriculture Organization of the United Nations
English Español Français العربية 中文 Русский
AGRIS
SEARCH
Find resources...
Search on Agris
My decision app

Examples: fhb, wmc430, Triticum, TRAES3BF001000010CFD

Looking for Agricultural Science and Technology Information?
Papers, data, statistics, and multimedia material, you get it with AGRIS all on one page

Text mining

- creates value from texts
- creates even more value if the results are linked to other data

This relies on shared semantics, standards & protocols

This requires less tech competencies & resources thanks to the common e-infra openMINTED

openMINTED

THANK YOU!

Sophie Aubin
sophie.aubin@inra.fr

twitter.com/openminted_eu

facebook.com/openminted

bit.do/openmintedlinkedin

vimeo.com/openminted

bit.do/openmintedplus