TEXT & DATA MINING IN EUROPE IS
openMINTED
openMINTED
1 CONTENT
2 SERVICES
3 PROCESSING
meaningful insights
multiple views
MULTI DIMENSIONAL ANALYSIS
MACHINE BASED PREDICTIONS
relations
COOPERATE WITH RESEARCHERS
50% NEVER READ
"LAN
PAPE
IS A
1,5 MILLION NEW RESEARCH ARTICLES EVERY YEAR

Text Mining: the next data frontier

Repositories in the centre of new scientific knowledge

openMINTED

#openminted_eu, #or2016, #tdm

Natalia Manola
Athena Research &
Innovation Centre

Some facts About scientific literature

The global research community generates over 1.5 million new scholarly articles per annum.

The STM report (2009)

... some 90% of papers … are never cited.
… 50% of papers are never read by anyone other than their authors, referees and journal editors

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

... one paper published every 30 seconds

... 70,000 papers published on a single protein, the tumor suppressor p53

Spangler et al, Automated Hypothesis Generation based on Mining Scientific Literature, 2014

Emerging solution(S)

Machine reading

process textual sources, organise and classify in various dimensions, extract main (indexical) information items,

… and "understanding"

identify and extract entities and relations between entities, facilitate the transformation of unstructured textual sources into structured data

… and predicting

enable the multidimensional analysis of structured data to extract meaningful insights and improve the ability to predict

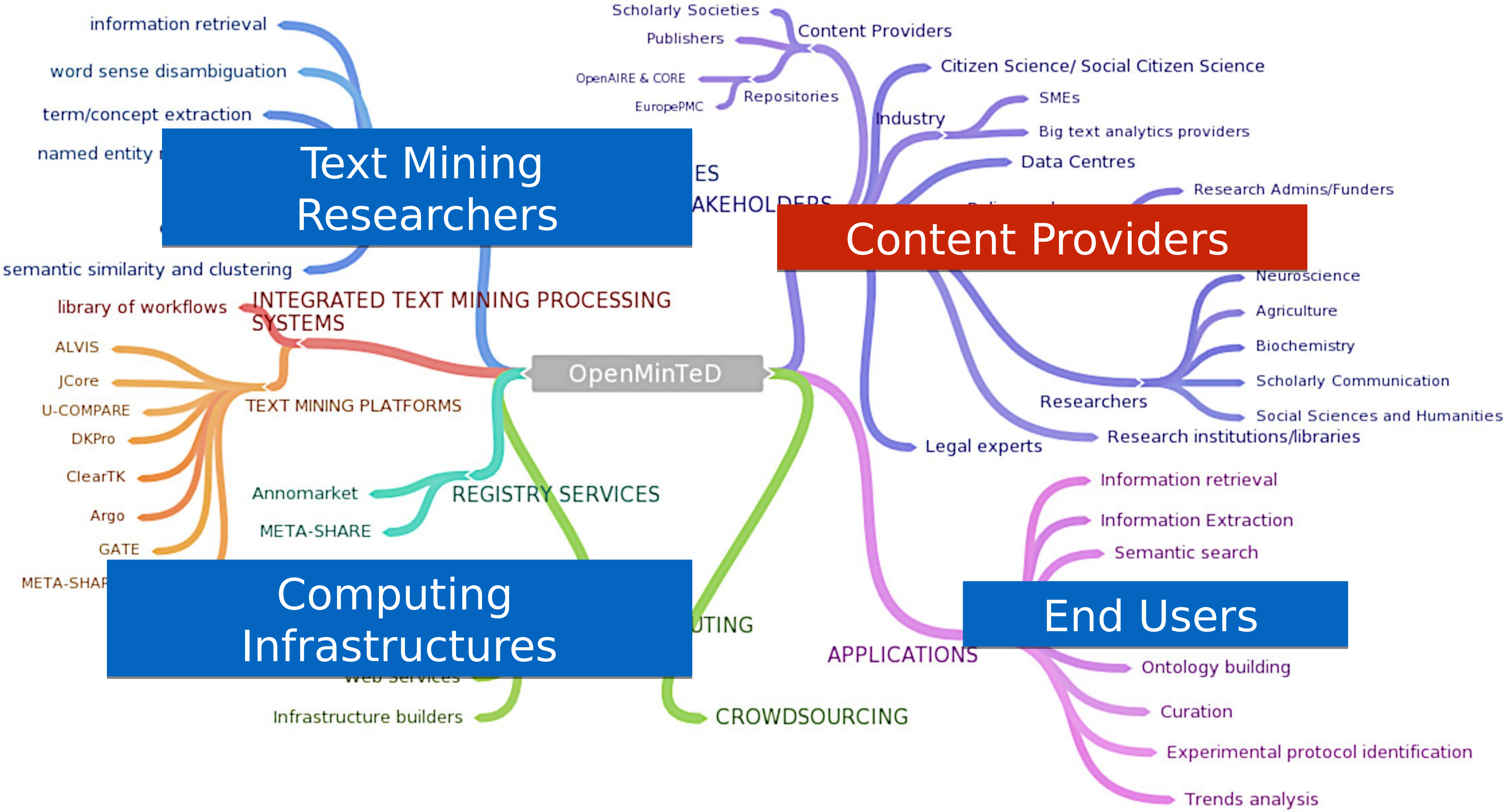
What OpenMinted is About

MAIN Objectives

Establish an **open** and **sustainable** Text and Data Mining (TDM) **platform** and **infrastructure** where researchers can discover, collaboratively create, share and re-use knowledge from a wide range of text based **scientific and scholarly**

A next step from Open Access to Open Science

A complex Landscape



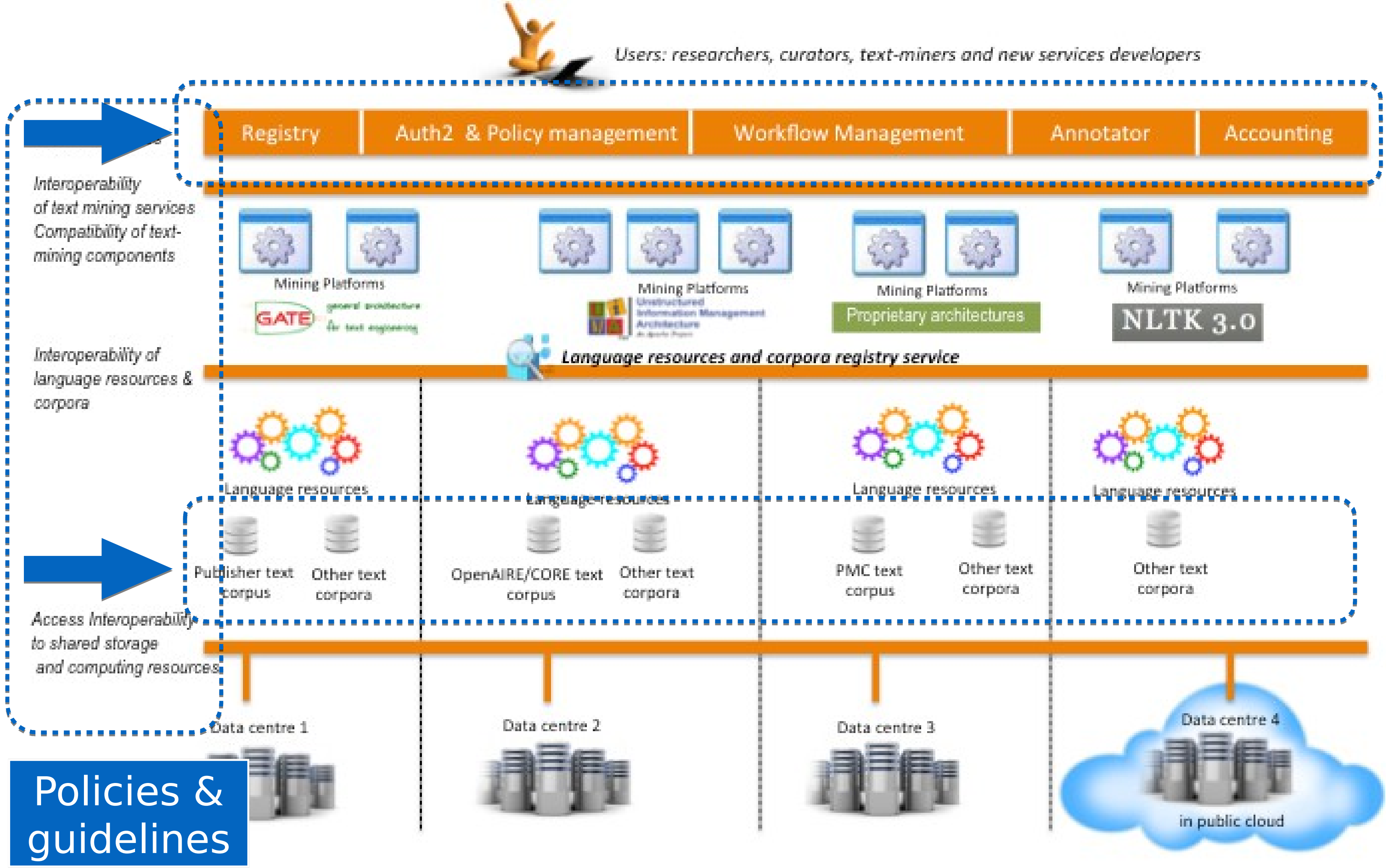
Text Mining Researchers

Content Providers

Computing Infrastructures

End Users

HIGH LEVEL ARCHITECTURE


Users: researchers, curators, text-miners and new services developers
Registry
Auth2 & Policy management
Workflow Management
Annotator
Accounting
Interoperability of text mining services
Compatibility of text-mining components
Mining Platforms
GATE
general architecture for text engineering
Mining Platforms
Unstructured Information Management Architecture
Mining Platforms
Proprietary architectures
Mining Platforms
NLTK 3.0
Language resources and corpora registry service
Interoperability of language resources & corpora
Language resources
Language resources
Language resources
Language resources
Publisher text corpus
Other text corpora
OpenAIRE/CORE text corpus
Other text corpora
PMC text corpus
Other text corpora
Other text corpora
Access Interoperability to shared storage and computing resources
Data centre 1
Data centre 2
Data centre 3
Data centre 4
in public cloud

Policies & guidelines

Key Characteristics

1 service oriented – discovery, re-use of content and tools

2 build on existing TDM tools - no focus on new algorithms

3 infrastructure – focus on interoperability

4 community driven - user centric requirements

5 open science - openness at all levels

Challenges

Discoverable & accessible content & services

- Document literature cont
  data categories taxonomies, provenance information
- Document langu
  workflows
- Generic and domain specific metadata descriptions

Starting with repositories and OA publishers
via OpenAIRE and CORE

Building on existing language resources repositorie
and infras (meta-share, clarin)

Interoperabl

- Combine services i
- Combine content and language resources with services and workflows
- Combine automat
  services

Promoting existing standards and best practice
AND technologies

In close collaboration with the FUTURETDM proje
http://project.futuretdm.eu/

IPR and licensing

- Study IPR restrictions for reuse of sources as well as possible

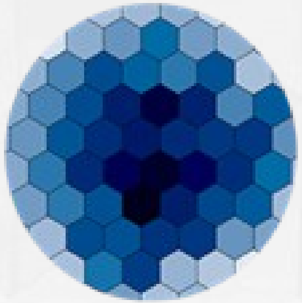
Community Driven

om the very beginning…

equirements, content, barriers, expected outcomes.

to the very end

eate applications, **validate and evaluate** the results.



Scholarly Comm.

Feature extraction

Data citation

Research analytics

Life Sciences

Curation of databases and lexica in Chembolomics & neuroinformatics

Agriculture

Extracting information from tables for food safety alerts

Social Sciences

Data citation

openMINTED

THANK YOU!

Natalia Manola
natalia@di.uoa.gr

twitter.com/openminted_eu

facebook.com/openminted

bit.do/openmintedlinkedin

vimeo.com/openminted

bit.do/openmintedplus