openMINTED

NATALIA MANOLA
ATHENA RESEARCH AND INNOVATION CENTRE

A ONE-STOP SHOP COMPUTING PLATFORM FOR TEXT MINING OF SCIENTIFIC LITERATURE

FORCE2017
27 OCT, 2017 @ BERLIN

European Commission

PARTNERS

Agroknow

EMBL-EBI

EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

The University Of Sheffield

INRA
SCIENCE & IMPACT

ATHENA

frontiers

grnet
Networking Research and Education

KMi
Knowledge Media Institute

MANCHESTER 1824
The University of Manchester

BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

gesis
Leibniz Institute for the Social Sciences

University of Glasgow

LIBER LIBER

The Open University

NaCTeM
The National Centre for Text Mining

UBIQUITOUS KNOWLEDGE PROCESSING

TECHNISCHE UNIVERSITÄT DARMSTADT

PART I

THE PROBLEM

THE GLOBAL RESEARCH COMMUNITY GENERATES ~2.5 MILLION NEW SCHOLARLY ARTICLES PER YEAR (ENGLISH ONLY)

STM report (2015)

… one paper published every 12 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

PART II

HOW CAN WE MAKE SENSE OF THIS DATA?

TDM - AN EMERGING SOLUTION

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

HOWEVER, …

MULTITUDE OF SOLUTIONS CATERING FOR DIFFERENT

Text Types

Newswire
Scientific Literature
Tweets/blogs
Patents
Clinical/medical records
Textbooks, monographs
Online forums
….

Domains

Finance/Business
Health
Biology
Social Sciences
Humanities
….

Tasks

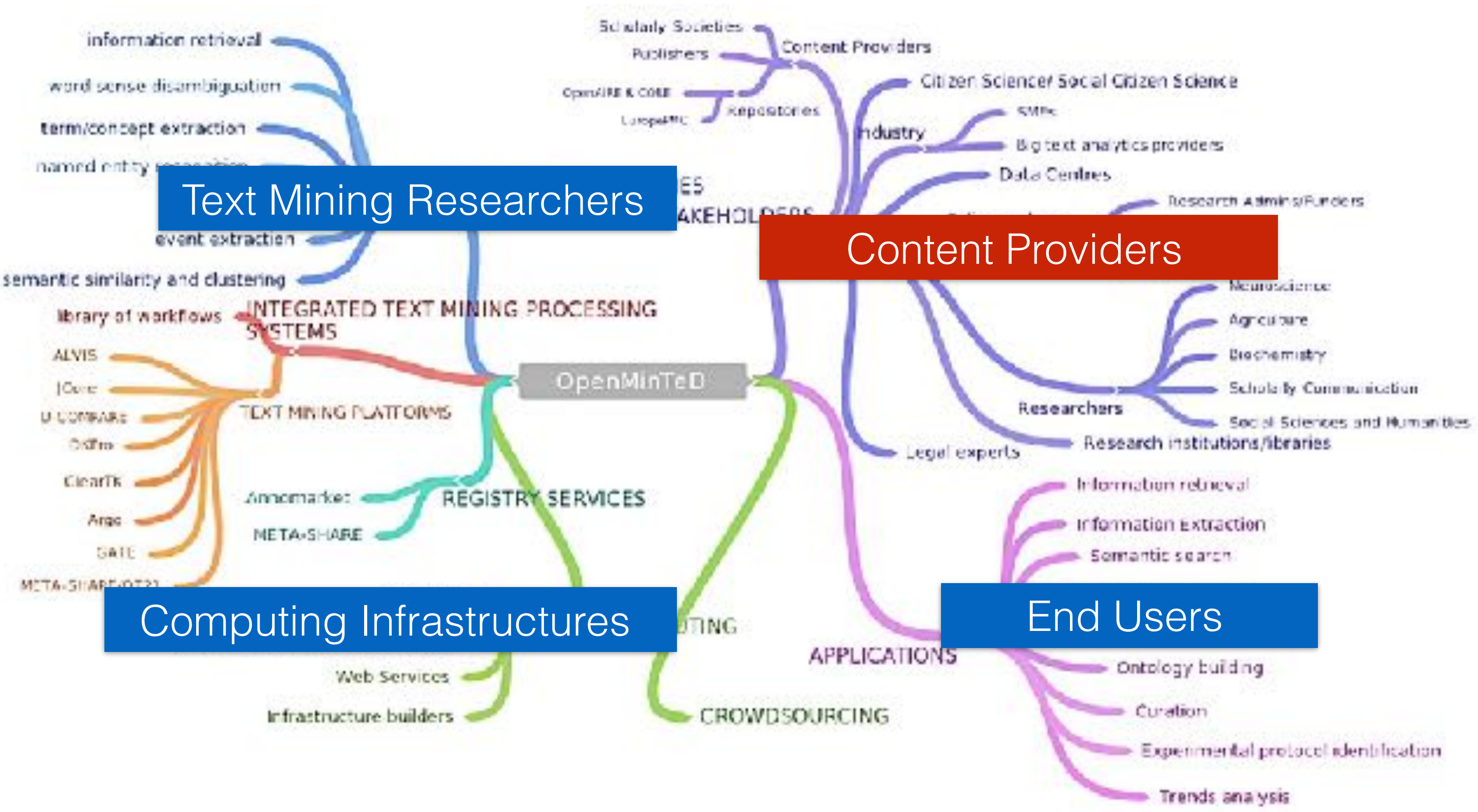
Translation
Information Extraction
Semantic Search
Question Answering
Sentiment Analysis
Summarization
Knowledge Discovery
….

Languages

English
French
German
Spanish
Portuguese
Italian
Polish
….

CREATING A FRAGMENTED LANDSCAPE

A COMPLEX AND FRAGMENTED LANDSCAPE



Text Mining Researchers

Content Providers

Computing Infrastructures

End Users

PART III

THE COMPONENTS

1. SHARE CONTENT

- Document literature content
- Share in a meaningful way: what does Open Access really mean?

IPR AND LICENSING

- Study IPR restrictions for reuse of sources as well as possible **exceptions**
- Promote clarity and standardisation of legal rights and obligations

CHALLENGES

- Rights statement vs. Open licenses (for repositories)
- No access to full text. We live in a metadata world
- No standard protocols, formats and APIs for access and retrieval
- No capacity to handle extra traffic

2. SHARE TDM SERVICES

- Document language processing/text mining services and workflows in a meaningful way for domain discipline researchers
- Document language/knowledge resources, data categories taxonomies, provenance information

INTEROPERABLE SERVICES

- Common way of presenting annotated results
- Combine services into workflows
- Combine content and language resources with services and workflows
- Combine automatic and manual/crowdsourcing annotation services

IPR AND LICENSING

- Translate the legal & policy aspects into specifications for lawful user-to-service and service-to-service interactions

CHALLENGES

- Bring text miners close to the researcher problems and needs
- Semantic interoperability (not just technical)

3. USE/SHARE COMPUTING RESOURCES

- Capacities and capabilities

INTEROPERABLE SERVICES AT THE LOWER LEVEL

- Common way of deploying operations/jobs
- Authentication and Authorisation services: Single Sign On (SSO)
- Accounting

CHALLENGES

- Legal, organisational, …

PART III

THE OPENMINTED PLATFORM

OUR SERVICES

1 REGISTER AND DISCOVER TDM SERVICES AND TOOLS

2 LINK TO CONTENT HUBS - SHARE CORPORA

3 RUN A TDM JOB

4 BUILD YOUR OWN SERVICE – COMBINE COMPONENTS INTO A WORKFLOW AND SHARE

5 STORE, DOCUMENT, PUBLISH AND SHARE RESULTS (ANNOTATED CORPORA)

PART IV

WHO IS OPENMINTED FOR

END USERS AS CONSUMERS

DOMAIN SPECIFIC RESEARCHERS & RESEARCH COMMUNITIES

Rather novice users and who want to find services (end to end) that fill their needs in an off the shelf type of situation. **(>100.000)**

APPLICATION DEVELOPERS / RI DATA SCIENTISTS

Understand basic usage of NLP and TDM services, but not the details. They know how to connect components, which content they must work on to get the required results. They need to develop end to end applications. **(>10.000)**

INFRASTRUCTURE OPERATORS

agnostic to the internal specifics of TDM, but they need to integrate and operate TDM services into daily workflows. **(<100)**

CONTENT AND SERVICES CONTRIBUTORS

FOR CONTENT

Publishers and repository managers (research libraries). **(<1000)**

FOR SERVICES

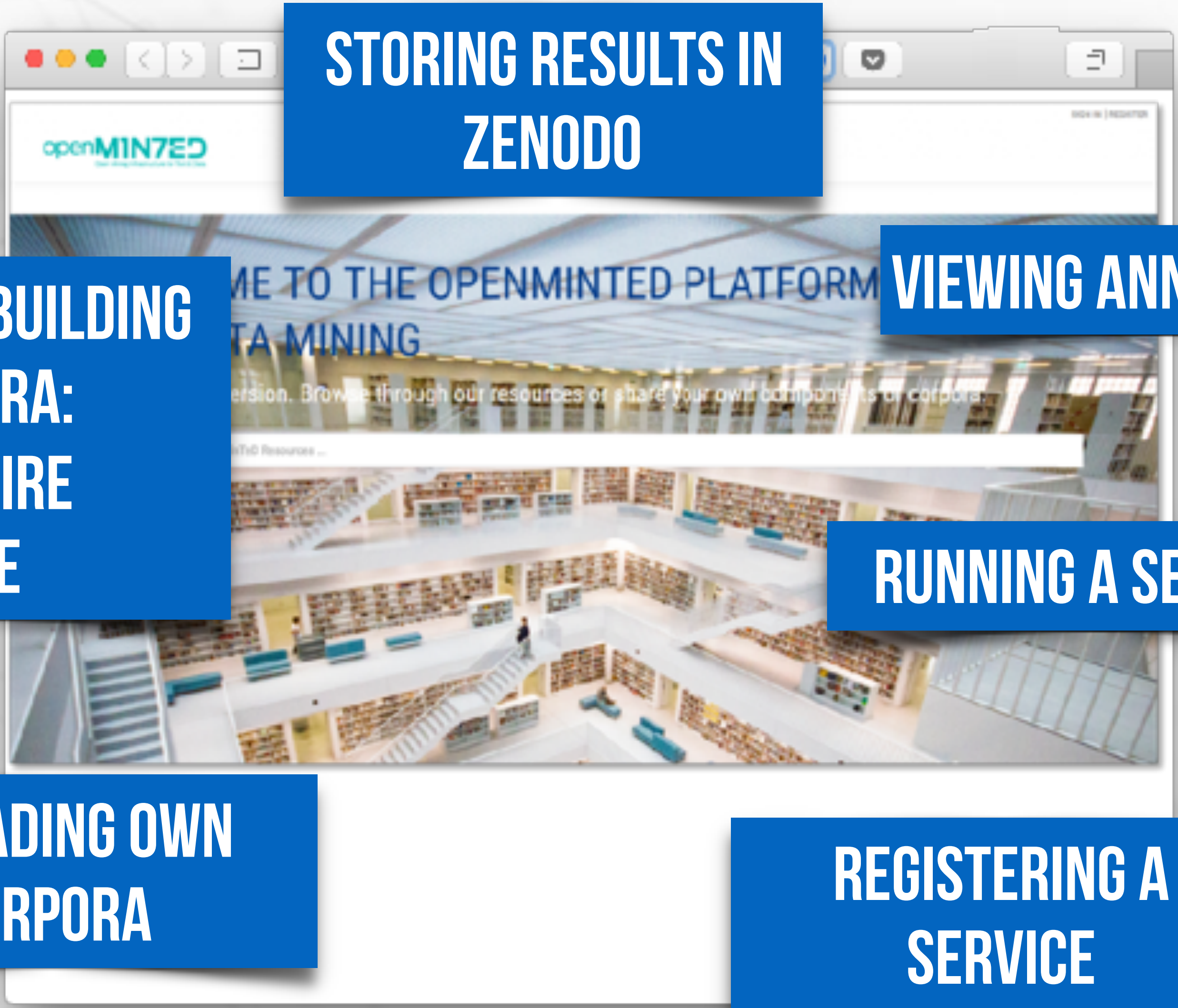
Expert language technology oriented people, who are using specific technologies and frameworks to develop and enhance their services. **(< 500)**

Non NLP expert developers, creating TDM modules based on off the shelf libraries and tools (e.g. Python, Jupyter). Not familiar with NLP frameworks and terminology but are eager to publish their small services. **(<5.000)**

PART V

WHERE WE ARE NOW

BETA RELEASE



openMINTED

...ME TO THE OPENMINTED PLATFORM

...TA MINING

...ersion. Browse through our resources or share your own components or corpora

STORING RESULTS IN ZENODO

REAL TIME BUILDING CORPORA: OPENAIRE CORE

VIEWING ANNOTATIONS

RUNNING A SERVICE

UPLOADING OWN CORPORA

REGISTERING A SERVICE

openMINTED

THANK YOU!

QUESTIONS?

NATALIA MANOLA
NATALIA@DI.UOA.GR