TEXT MINING: THE NEXT DATA FRONTIER

An infrastructural approach for mining scientific content

Natalia Manola
Athena Research & Innovation Center

OpenMinTeD

OUTLINE

Text mining on scientific content

What is involved?

Infrastructural approach – promoting Open Science

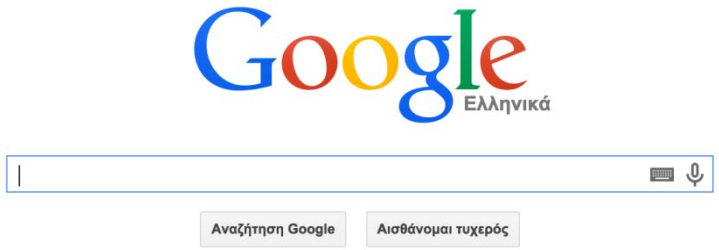
OpenMinTeD – putting it all together

TEXT MINING – TEXT ANALYTICS

Interchangeable terms

- What?
  - Text analytics applies **statistical**, **linguistic**, **machine learning**, and **data analysis** and **visualization** techniques to identify and extract salient information and insights.
  - **Not an end, in and of itself.** Text mining creates **new relationships and hypotheses** for domain experts to explore further to create **additional knowledge**.

- Where?
  - Text mining is everywhere, just … **behind the scenes**. Not in the front and center!



MINING SCIENTIFIC LITERATURE

Make sense of large volumes of data in a digital economy

European Commission

”

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

The STM report, 2009

”

It is a sobering fact that some 90% of papers that have been published in academic journals are never cited. Indeed, as many as 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

Scientific literature is most often the entry point to access and curate the research data.

TEXT ANALYTICS IN BUSINESS AND RESEARCH

Same processes, different perspectives

Business

- Text analytics describes software and transformational processes that uncover *business value* in “unstructured” text.
- The goal is to inform decision-making and support *business optimization*.

Research

- Text analytics describes software and transformational processes that uncover *new knowledge* in “unstructured” text.
- The goal is to improve *evidence base research* and support *research optimization* in process and quality.

DIGITAL ECONOMY

> Text mining and analytics of scholarly literature and other digitised text affords a real opportunity to support **innovation** and the development of **new knowledge**.

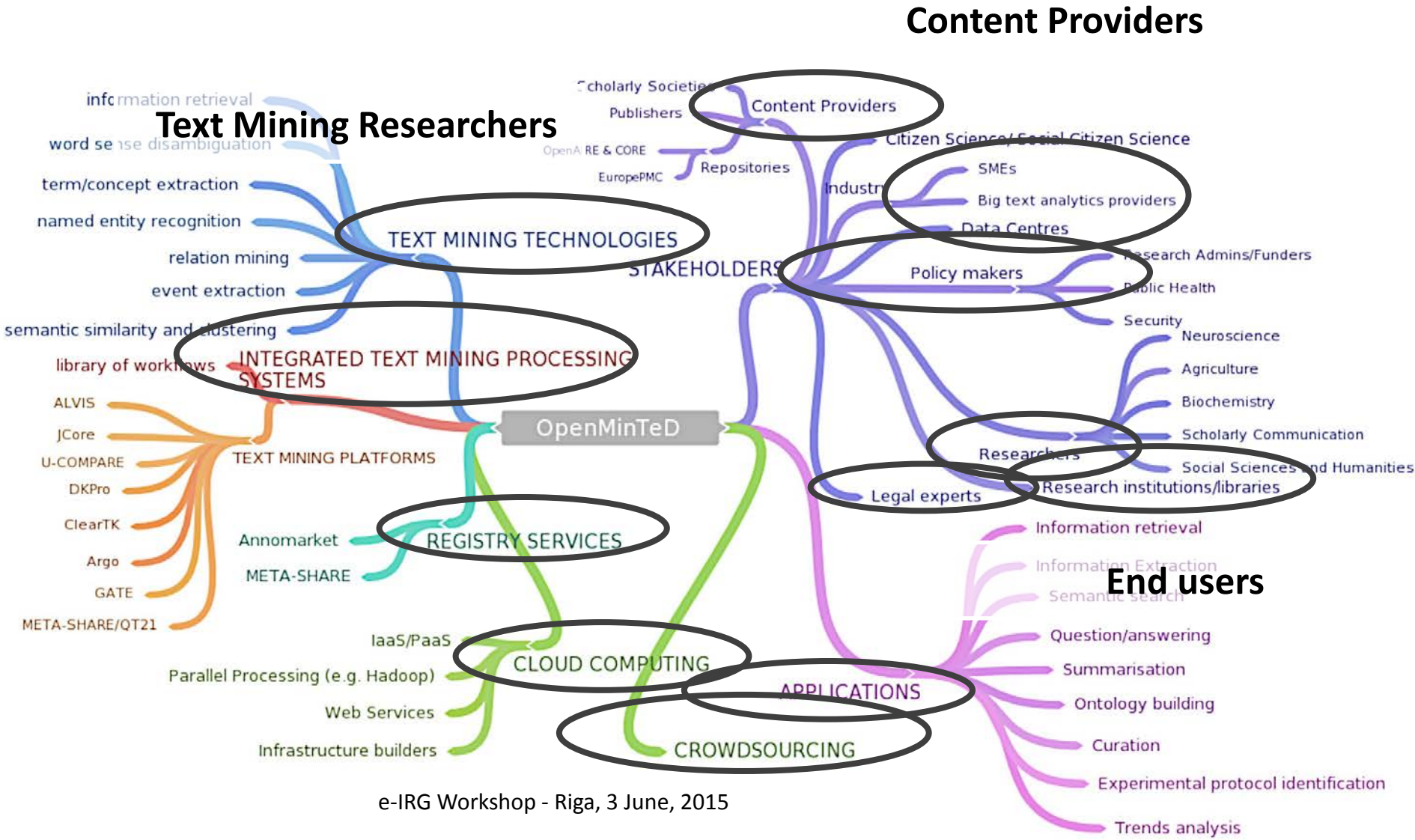
Hargreaves Report (Digital Opportunity, 2011)

WHAT IS INVOLVED?

Content – Services - Processing

European Commission

A COMPLEX LANDSCAPE



CHALLENGES

On multiple fronts

Content

- Barriers and obstacles due to non-availability or licensing issues
- No uniform way to retrieve content for TM

Services

- How to identify the most fitting one?
- Do I have permission to use it?
- Where to deploy? How to combine with MY content?

Processing

- TM needs to follow latest cloud trends - as they are evolving in Europe's cloud systems (and beyond)

BUILD SYNERGIES!

INFRASTRUCTURAL APPROACH

Promoting Open Science

European Commission

COSTS AND PROCESSES

On acquiring and accessing content

- Current high **transaction costs** due to the need to negotiate a maze of licensing agreements.

- In some cases the institutions may need to pay four different costs to enable the materials to be mined
  - traditional access (reading) costs
  - the right to copy
  - the right to digitise
  - the right to text mine

COSTS AND PROCESSES

On operating services

- Given the sophisticated technical nature of text mining, **entry costs** are high
  - No shared knowledge - TM remains a fragmented set of tools
  - Very specialized activity requiring significant technological and analytical skills as well as domain expertise
  - Lack of a central infrastructure may rule out the use of TM for small research groups

- Need to share **infrastructure cost**
  - No shared computing resources
  - RIs (CLARIN), e-Infrastuctures (OpenAIRE, EGI, EUDAT, AAI, …)

OPENMINTED

Putting it all together

Open Mining INfrastructure for TExt and Data

European Commission

Text mining to

- clean, disambiguate, enrich, link metadata based on OA content
- uncover hidden relationships

Issues with IPRs & licences on OA content

INSPIRED BY OpenAIRE

How to share our text mining tools with the community

Develop services/architectures that already existed

PROJECT SPECS

- **Started:** June 2015
- **Duration:** 3 years
- **Total budget:** 6,068,074 Euros
- **16 Partners**
  - 6 mining research groups
  - 3 content providers
  - 1 data center
  - 1 library association
  - 2 legal experts
  - 6 community related partners
  - 2 SMEs

PARTNERS

Athena RIC
Univ. of Manchester (NacTem)
Univ. of Darmstadt
INRA
EMBL-EBI
Agro-Know
LIBER
Univ. of Amsterdam
Open University UK
EPFL
CNIO
Univ. of Sheffield (GATE)
GESIS
GRNET
Frontiers
Univ. of Stirling

Users: researchers, curators, text-miners and new services developers

Platform services

Registry | Auth2 & Policy management | Workflow Management | Annotator | Accounting

Layer 1: Interoperability of text mining services (platforms or components)

Layer 2: Interoperability of language resources & corpora

Language resources and corpora registry service

Language resources | Publisher text corpus | Other text corpora | OpenAIRE/CORE text corpus | Other text corpora | PMC text corpus | Other text corpora | Language resources | Other types of text corpora

Layer 3: Interoperability to shared storage and computing resources

Data centre | Data centre | Data centre | Data centre in public cloud

Policies &interoperability on data AND services …across Europe, across scientific domains

WHO ARE OPENMINTED'S USERS

End users - *consume* TM services

- Researchers, data base curators, …
- *Novice:* use services to advance their science
- *Advanced:* service providers (e.g., SMEs) to create more complex research workflows for their clients.

Service providers - *provide* their TM services

- TM research communities
- SMEs

Content providers - *provide* their content

- Publishers, libraries, scientific dbs, …

OPENMINTED – MAIN ROUTES

An ambitious endeavour

ACCESSIBLE CONTENT	Via standardised programmatic interfaces and access rules
DISCOVERABLE SERVICES	Well-documented, easily discoverable text mining services and workflows which process, analyse and annotate text
EFFICIENT PROCESSING	Operate on public e-Infrastructures via standarized APIs
INVOLVE COMMUNITIES	Different scientific communities have different challenges
VALUE ADDED APPS	Community-driven applications to illustrate the value of the infastructure. Engage with industry.

ACCESSIBLE CONTENT

IPR and licensing

- Study IPR restrictions for reuse of sources
  - Exceptions?
  - What about non commercial research?
- Translate the legal & policy aspects into authentication & authorization specifications (OAuth2, GÉANT's EduGain, …)

OUTPUTS
Guidelines
Content retrieval tools & services
Legal recommendations

Metadata & transfer standards

- Document literature content, language resources, data categories taxonomies, provenance information
  - Generic and domain-specific metadata descriptions
- Identify standards for metadata harvesting and federated search in distributed repositories (OpenAIRE)

DISCOVERABLE SERVICES

IPR and licensing

- ated textual sources and

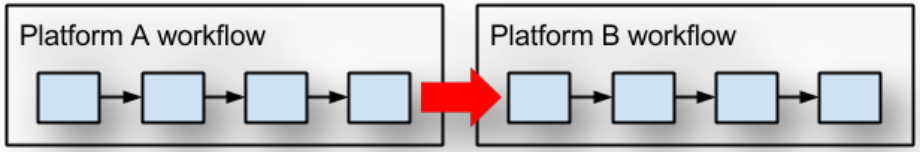
S

- how to define interoperable input/output format?

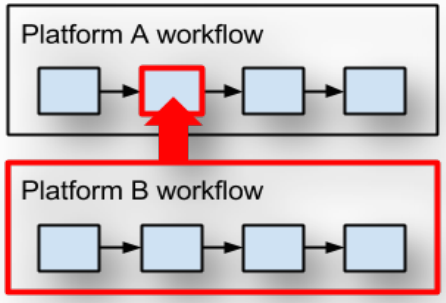
MIXING AND MATCHING SERVICES

Workflows – various levels of complexity

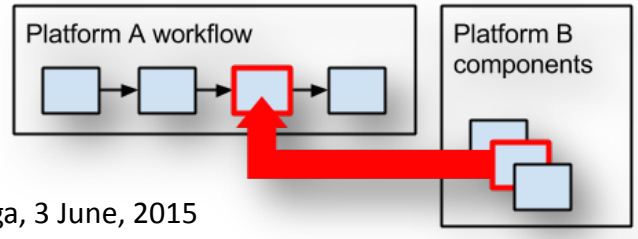
- Remote workflow's output as input



- Remote workflow as component



- Portable components processing



EFFICIENT PROCESSING

- How to use and optimize resources related to the physical layer (storage and processing units)
  - Efficient distribution and parallelization
  - Interlinked European cloud and national environments

- Build on existing expertise and e-Infras
  - Shop around
  - Require QoS

INVOLVE COMMUNITIES – THE TM EXPERTS

Working groups

1. Resource Metadata: content, services, language resources
2. Text, lexica, terminologies and ontologies representation and access
3. IPR and licensing
4. Text annotation and text-mining services workflows

- Start from existing practices/standards
- Synergies with other initiatives
  - RDA and beyond

INVOLVE COMMUNITIES – THE SCIENTISTS

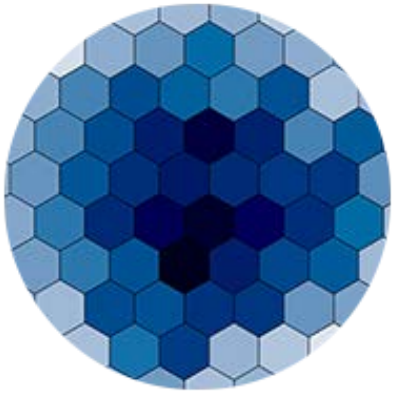
Through focussed applications

From the very beginning

Requirements, content, barriers, expected outcomes.

To the very end

Create applications, evaluate the results.



RESEARCH ANALYTICS

LIFE SCIENCES

AGRICULTURE

SOCIAL SCIENCES

VALUE ADDED APPLICATIONS

- Scholarly communication & research analytics

  OpenAIRE, CORE, Frontiers

  - Semantic search and discovery of open scientific outcomes
  - Map of academia – scholarly communication network
  - Research monitoring and analytics

- Life sciences

  EBI and Human Brain Project

  - Text mining assisted curation of the EMBL-EBI chemical databases
  - Curation of the neurosciences resources KnowledgeBase and Neurolex

VALUE ADDED APPLICATIONS

- Agriculture and Biodiversity

  INRA, Agro-Know, EFSA

  - Enrich agricultural databases to assist food- and water-borne disease outbreak alerts and product recalls
  - Image, figure and dataset discovery in the AGRIS FAO online service

- Social sciences

  GESIS

  - Develop and evaluate methods for the automatic detection and linking of *named entities, citation traces* and *intentions* in social science scientific publications.

THANK YOU

www.openminted.eu

natalia@di.uoa.gr

European Commission