OpenMinTeD Interoperability Scenarios

OpenMinTeD Interoperability Team (T5.2)

Version 1.0.0

Contents

Introduction

The scenarios in this document aim to highlight TDM/NLP interoperabiliy aspects within particular use-cases. They approach the topic from the point of view of four different areas corresponding to interoperability working groups in the OpenMinTeD project:

- WG1 - Resource metadata
- WG2 - Language resources
- WG3 - IPR and licensing
- WG4 - Annotation and workflows

Structure

The scenario descriptions are structured as follows:

- Description
- Relevance to the WG
- Relevance other WGs
- Relevance in general
- Approaches
- Open questions

Interoperability scenarios

WG1 — Resource Metadata

Scenario 1 — Discover resources of various types at various locations

Description

A researcher is doing research on a certain topic, e.g. sentiment extraction in the domain of political conflicts. She wants to find all relevant literature and datasets that she could use for her research and/or construction of an application. Publications that might be of interest to her are stored at various points, e.g. OpenAire but also institutional/university repositories, various publishers' sites (to which she may have access through academic subscriptions but not necessarily), research repositories etc. Datasets, i.e. corpora (raw & annotated) as well as lexica, vocabularies, ontologies etc., are also scattered at various repos & portals. They are described with different metadata schemas, depending on the resource type & repository. She would ideally like to make one query at one point and get a list of all resources of relevance, a common description and a quick gist of the contents of the resource (e.g. abstract of the publication, sample of the dataset).

Relevance to the WG

It focuses on the core issue of the WG's interests, i.e. resource discovery & interoperability problems between different resource types and different metadata schemas used to describe them. Authentication/authorization depending on the user profile is also required.

Relevance to other WGs

- **Knowledge bases (WG2)**: Description of the contents and technical details of datasets, incl. lexical/conceptual resources is in order. Moreover, controlled vocabularies used as values for metadata elements (esp. for classification) and the way these can be made interoperable is an important aspect of the scenario.
- **IPR and licensing (WG3)**: Access to the resources may differ depending on the repo (e.g. publisher's site with academic subscription allowing downloading of a resource, open access repo from which all resources can be re-used), the resource itself and the user. The licensing conditions on each resource must be clearly visible to the researcher and trigger a mechanism that allows direct full access to the resource (if it's allowed) or a partial access to parts of it.
- **Annotation Workflows (WG4)**: Not as relevant

Relevance in general

They have come across similar situations and know how important description is for resource discovery, especially when it involves different resource types.

Approaches

By standardizing metadata schemas and mappings between the most popular ones, at least; addressing the way similar information (e.g. classification) is encoded in different resource types and communities is of great importance.

Open questions

External experts in this WG work in various areas where resources of different types are described with various metadata schemas. Their expertise will be of value for finding a "common/standard language" between them.

Scenario 2 — SME running research analytics for funders within the European Research Area

Description

The EC and other funders are looking into ways to discover correlations and trends in funded research in Europe. They want to go beyond the obvious relationships (project-publication-data-patents-software) and the related statistics (i.e., who produces what and co-funds). They need to identify the hidden relationships brought out via concepts and topics as these are discovered through advanced NLP and probabilistic algorithms. They need to find various side effects, e.g. structural effects stemming from collaboration from different institutions, countries and domains (cross-discipline)

As an addition, EC wants to compare their programmes with the ones in the US (NSF, NIH).

Relevance to the WG

To get some meaningful results the following tasks need to be completed:

- Discover all publications related (or not) to funding programmes from as many funders in Europe (and beyond). These publications may be OA or may be private/proprietary and can reside in institutional/thematic repositories, or be behind closed doors in publisher sites.
- Discover downloadable software tools (not content/data processing web services) that are interoperable when configured appropriately and combined together they will build a system/application. The IT people in the SME need to find the most appropriate tools (reliable, good performance, correct licences, …).

Relevance to other WGs

- **WG2** – Classification vocabularies like Eurovoc, MESH, DDC must be used.
- **WG3** – This is an SME building on OA and non-OA content using tools from various research labs, and they will build a profitable service. The SME does not own the scientific content but the funder has already paid subscription to metadata dbs (Scopus, Thomson Reuters). Still there is no access to the content itself (pdfs, xmls). The SME will use the tools as components of a system, and it’s important to see how the different tools are legally interoperable.
- **WG4** – The SME will probably be using components/tools from various providers (research labs) and needs to form workflows. Will also need to get results in a given format that will not be changing over time, or over different components. Operations may be run on a national data center or on dedicated servers on the EC/funder site (or on a paid cloud service like amazon).

Relevance in general

Relevant to our research analytics experts. Also practitioners that will deal more with structured data and face semantic interoperability problems with different semantic classification schemes.

Discovery of software solutions with appropriate software versioning and dependencies is a common problem to

many TDM users (and developers alike!).

Approaches

Promotion of standard metadata-based descriptions of software, standard (or standardisation of) vocabularies/authority lists and semantic interoperability of classification schemes & promotion (or standardisation) of licensing schemes for ensuring are among the main objectives.

Open questions

Systems/methods/schemes used for metadata-based discovery. Vocabularies used.

Scenario 3 –- A content provider using text mining tools to enrich their content

Description

An open access journal or a repository wants to enrich the metadata of its content. Europe and national data services are funding infrastructures/aggregators like OpenAIRE and CORE to act as such information hubs/brokers. To be more specific, OpenAIRE operates a set of TDM tools on the aggregated content from 15mi publications and related data. These tools (either operated on the metadata+abstract or the full text) are able to extract all research related information (grants, authors and their affiliations, data/software/patent citations, classifications, topics, etc.) and clean/enrich the harvested metadata records. In addition, via de-duplication and disambiguation mechanisms OpenAIRE is able to link all results, organizations, authors, funders and data into an integrated research information system.

This scenario can be used in the following ways:

- OpenAIRE or any other service provider shares these tools (or services) with all participating (or not) content providers via OpenMinTeD. In this case, when a researcher self deposits a publication, a standard plugin installed at the journal or repository platform accesses such research information extraction services and applies them on the deposited publication to automatically extract all metadata. The researcher has only to verify the quality of metadata.
- When a content provider joins an e-Infrastructure or aggregator, they need to provide their full text/full data content to be mined at a central point. They need to describe the metadata and the push/pull protocols that will be used to access the full text.

Relevance to the WG

- Content providers need to register their content for TDM purposes. They need to describe the metadata using an appropriate schema, including formal licensing metadata.
- Plugin builders need to be aware of specific service descriptions. In the first sub case we need to return results in a standard format (annotated texts and extracted metadata information), usually via a repository platform plugin. It is important to include metadata related to the TDM operation (e.g., confidence level, performance, service version).

Relevance to other WGs

- **WG2** – All these text mining tools use a variety of ontologies.

- **WG3** – We need to fully describe the license of the content provided by the repository/journal, and the licences of the used services. Not sure about the licenses of the returned annotations (as the cleaning and enriching comes from the combination of content from different sources)
- **WG4** – Format and protocols for returned annotations.

Relevance in general

Automatic metadata extraction of content is a common requirement in many repositories. Mechanisms for documenting particular mining services discoverable and integratable/callable from within such e-infrastructures are required.

Approaches

Through promoting guidelines and schema(s) for formal metadata-based documentation and providing mechanisms for integrating mining services.

Open questions

What schemas do experts and practitioners use? Level of detail and completeness of metadata elements. Use of predefined vocabularies and value sets. Protocols for harvesting and exchanging research information.

Scenario 4 — Provide comprehensive statistical metadata for resources

Description

A user likes to make a detailed comparison of multiple versions of a language resource or a knowledge base. The user is interested in running a TDM experiment using a resource (e.g. Wikidata) and likes to evaluate the performance changes of the experiment with regards to using different versions of the same resource. To enable an analytical study, he requires detailed statistics related to revision changes (e.g. total number of concepts, total number of relations, number of added/updated/deleted concepts/relations). However, the problem is that this type of information is not typically shipped with resources.

Relevance to the WG

Knowledge encoded in a resource often has a great impact on results of an experiment; hence, they affect the reproducibility of results. However, statistical information regarding a resource or different revisions of a resource usually are not either well documented or published with the resource itself. This makes it difficult to trace changes made to a resource and also makes it difficult for researchers to analyze their results or compare with previous results.

Relevance to other WGs

- **Language Resources (WG2)**: It is important to store as much as possible information about a language resource. This will lead to a more transparent usage of resources.
- **IPR and licensing (WG3)**: Changes of usage license might be a rare event, but tracking these changes seems useful.
- **Annotation Workflows (WG4)**: Some TDM workflows should be implemented to extract and generate the required metadata.

Relevance in general

Providing detailed information about a resource makes it more transparent and easier to be adopted by the community. It also makes it easier to compare multiple resources prior to using them.

Approaches

Through designing metadata placeholder for statistical metadata of resources and implementing mechanism to calculate data that is not inherently announced with the resource.

Open questions

It is required to identify which metadata types are suited for describing each resource. The other requirement is to create components to generate statistical metadata information for each resource.

Scenario 5 — Domain specific researcher using a text mining tool or service to promote their research or use applied research results within their setting.

Description

A novice researcher, who may be in a research organization setting, an SME, or a different type of organization like a museum or a ministry that does not have access to a library subscription, wants to get some insights on his/her topic of interest. For example, an archaeologist in the Acropolis Museum trying to find scientific evidence on how to use storytelling mechanisms in a museum setting. This is a very trendy and state of the art technology and she would like specifically to know about which museums have applied the specific technologies on what types of settings, and moreover retrieve the ones that have somehow led to products. She will use this output within a research project which will ultimately help facilitate an iphone/android app for mobiles or tablets, which will be owned by the Museum and will be available for free download from the Apple/GPlay stores. The researcher does not know where to look and has no knowledge of the text mining algorithms or tools. She is acquainted to newer technologies as she has a MS. in cultural heritage digital technologies.

Relevance to the WG

Explore which content to search into, which probably should include different content types: scientific publications, technical and project reports. She needs an easy way to discover what is around and make the appropriate choices. The researcher needs to find the TDM services (not tools, because they must be easy to use and the researcher has no own computing resources) that will do the job for her. She needs to identify them in a straightforward manner and an easy way to point to the content to be mined. It is also important to have the results in a human readable way. All content and services need to be run in the cloud as the museum doesn't have any supporting digital infrastructure. Authentication/authorization is also an important aspect that needs to be taken into consideration.

Relevance to other WGs

- **WG2** – Information probably exists in different collections around Europe (national projects) and language is an important facet. Entity resolution for museums/locations probably requires special vocabularies.
- **WG3** – As the researcher's institution has only access to OA content (no library subscription), we need to see what type of content is accessible. Also, as the services will be used to advance research towards a commercial project (non-profit organization?) the researcher needs advice on how to go about it.

- **WG4** – basic interoperability between data and tools is necessary. Services may need to be combined in a workflow. The researcher needs the results, i.e. the annotated content, in a human readable format.

Relevance in general

External experts will want to know how the specific content to be mined can be assembled into a virtual collection from different content types, and how text mining tools/workflows can be discovered.

Approaches

- creating the appropriate metadata infrastructure and enabling structured and unstructured data search
- adopting clear data licensing and services terms of use
- indicating which services can usefully operate on which types of content

Open questions

How are fairly advanced applications (e.g. storytelling) technologies formally described or potentially composed so that they are discoverable by non-expert users?

Reqiuirements for WG1 from other WGs' scenarios

From WG1

- discovery of various resource types (publications, datasets, tools/services)
- search at multiple places (repositories, archives, portals)
- licensing information per resource type
- specific formats for the output of queries/web services/workflows (e.g. statistical data, annotated corpora etc.)

From WG2

- mapping classification metadata on different resource types / from different sources (e.g. keywords, thesauri, technical vocabularies etc.) ⇒ interoperability between metadata elements
- resource discovery through classification metadata (e.g. that use different vocabularies for the same or similar elements) ⇒ interoperability between metadata elements & metadata values!
- resource access modes and points to be entered as metadata elements (e.g. through SPARQL endpoints) - cf. scenario 4

From WG3

- correct/proper labeling licence on source resource
- machine-processable metadata for licences
- identification of licences for services
- provenance information across TDM workflows
- Can legal interoperability be determined/facilitated through metadata?
- “derivatives” of TDM services based on analysis of source resources but not reproducing the source resource -

??

- metadata for resource parts, e.g. keywords ⇒ classification metadata required

From WG4

- standardisation of metadata as regards input & output of tools/services (mainly refers to format but also annotation schemes)
- standardisation of metadata as regards the description of components/modules of workflows
- metadata for versioning of components/modules
- metadata for annotated corpora and annotations
- for re-running experiments: detailed information about the resources used, their dependencies, their provenance, and particularly their versions is required. This entails a method of transitively resolving references from one resource to another (e.g. in a dependency or provenance graph)
- The ability to refer to stable versions of language resources which may be problematic e.g. for knowledge resources only available through web services. ⇒ same problem with virtual collections
- provenance, versioning, and licensing information
- automatic assignment of metadata for output of workflow
- metadata for user profiling, incl. skill of user

user authentication/authorisation

- funding information on metadata or extracted from contents of resources
- classification information regarding resources
- description & discovery of web services for extracting metadata information
- description of contents of resources & differences between versions
- description of structure of resources (in order to extract information)

General requirements

- metadata for
  - repositories/archives/portals/…
  - content resources/datasets
  - language/knowledge resources/databases (i.e. corpora, lexical/conceptual resources, language models etc.) ⇒ used as input but also used by text processing tools & services
  - publications
  - s/w tools & web services
  - workflows
  - users
- metadata for repositories/archives regarding
  - description, identification

    - licensing
    - access points
    - protocols & metadata schemas used for the resources
- metadata for content resources regarding
    - description, identification (can be different per resource type)
    - licensing (licence & conditions of use in a machine processable form)
    - access points
    - classification
    - funding
    - description of contents (e.g. size of entries, size and types of linguistic information encoded etc.)
- metadata for s/w tools & web services regarding
    - description, identification
    - access points
    - licensing (of the tool/service itself, but also of the input & output resources)
    - input resources: format, contents, classification, size?
    - output resources: format
- metadata for workflows regarding
    - description, identification
    - access points
    - licensing (of the workflow itself, but also of the input & output resources)
    - description & identification of components (cf. web services)
    - provenance
- metadata for users regarding
    - identification & authentication
    - level of expertise

WG2 — Language Resources

Scenario 1 — Combining heterogeneous resources for information extraction

Description

The user wants to build a knowledge base on animal diseases using information extracted from scientific or technical documents. The domain model is represented in an ontology describing relations between animals (hosts), diseases and pathogens (vectors). The lexical elements to identify the different objects are spread in disconnected resources (e.g. list of “livestock” from Agrovoc, “diseases” from a custom lightly hierarchical vocabulary and “pathogens” from the Uniprot taxonomy). Those resources provide information for scientific and common names, abbreviations, etc. There can be several relevant for a same type of object.

WG2 specific tasks:

- KB content harmonization including semantic mapping
- harmonization of I/O specifications for TDM workflows
- serialization of output annotations in OWL/RDF format

Relevance to the WG

This scenario requires uniform access to content from heterogeneous resources. In order to operationalise these resources, we need interoperability of their content enabling uniform access.

The user has to deal with vocabularies presenting different schemas and levels of information. This requires schema mapping. The vocabulary elements of these resources can be of many forms, e.g.

- resource specific
- conforming to standard schemas such as TBX, LMF, SKOS; and formats e.g. OWL, RDF

Detecting these elements in text may require different strategies or tools as they contain Named Entities (form and case sensitivity) or terms (fuzzy recognition, case insensitivity, etc.).

Relevance to other WGs

- WG1: discovery and technical metadata allow the user to identify relevant candidate vocabularies
- WG3: this scenario is an example of the combination of several vocabularies, and raises issues with respect to legal aspects of 1) resource access 2) the combination of resources with different licenses and 3) the publication of the annotated corpus.
- WG4: flexible insertion of resources into workflows based on interoperable content specification. modes of mapping vocabulary items onto text: consider the extraction strategy (e.g. strict, fuzzy, lemma to wordform).
- WG1/4: keep trace of the source and method used for producing annotation.

Relevance in general

Are vocabularies really reusable and combinable… and if yes, to what extent?

Approaches

- Provide easy access to resources and the possibility to test them by easy lookup on a corpus
- Allow the traceability of resource elements through metadata and identifiers
- Provide interfaces/ways to integrate vocabulary elements or subparts
- Clarify the legal status of extraction results with respect to the original vocabularies/resources used.

Open questions

- Identifications of problems and bottlenecks
- Discussion of possible solutions
- Help with defining preferred strategy

Scenario 2 — Including Custom Knowledge

Description:

A user wants to analyze mentions of pharmaceutical products based on the main agents used in these products. The user uses a publicly accessible dataset of products and their agents, but the dataset is incomplete, e.g. does not include products still in testing stage. The user wants to provide additional (confidential/unlicensed/proprietary) data about product/agents that should be taken into account for the analysis.

WG2 specific tasks:

- KB content harmonization including semantic mapping
- harmonization of I/O specifications for TDM workflows
- harmonization of output format

Relevance to the WG:

Use-case involves a public and a custom/private knowledge source

Relevance to other WGs:

- WG1: Not particularly relevant
- WG3: because the custom knowledge source may be confidential
- WG4: because the custom knowledge source may not be accessible

Relevance in general:

Including custom content and/or confidential data is a common problem in infrastructures

Approaches:

- WG 2
    - indicate if this is a real world problem

    - suggest how knowledge can be managed/packaged/harmonized.
- WG 3
    - indicate if there are legal/contractual devices to cater for confidentiality and/or use of an unlicensed resource transiently during processing (assuming that the user trusts the platform in the first place)
    - safe-harbor provisions?
- WG 4
    - allow deployment of TDM analytics in an environment trusted by the user
    - allow the TDM platform to access resources provided by a user
    - allow user to include excerpts (potentially anonymized, e.g. product/agent names replaces with IDs) of resources along with input data for TDM

Open questions:

- indicate if this is a real world problem
- suggest how knowledge can be managed/packaged
- report how such problems have been addressed (if at all)

Scenario 3 — The relation between documents and knowledge bases through keywords

Description:

The general idea is to treat keywords attached to a publication as a "mini lexicon" representing its topic and expand/enrich this lexicon by combining it with lexical/conceptual resources such as domain lexica, glossaries and ontologies. Possible applications of this idea:

- A user wants to find publications on a specific research topic. The query terms can be mapped to a resource consisting of keywords & their synonyms/hypernyms/semantically linked words/phrases as encoded in other lexical resources.
- The use of multilingual lexica in this scenario can also lead to the identification of publications on the same research topic in various languages; especially important when keywords in all these publications are not direct translation equivalents of each other, but linked through other semantically similar words/phrases/concepts
- Document clustering based on keywords can also be enhanced through the expansion of keywords with semantically related words from lexical resources and ontologies.

WG2 specific tasks: KB content harmonization including semantic mapping and multilingual keyword expansion.

Relevance to the WG:

It involves the interoperability between various language resources (publications, lexical resources, ontologies etc.).

Relevance to other WGs:

- WG1: identifying resources of different types on the same topic / of the same domain

- WG3: ensuring that the automatic access to the various resources is allowed; the combined use of resources makes the scenario more complicated as one needs to ensure that all resources allow their exploitation
- WG4: creating workflows for the use cases, taking care of the fact that different types of resources can be used at various stages of the process and thus need to be interoperable (ensuring appropriate input and output)

Relevance in general:

Interoperability issues between resources of the same or different type (e.g. lexica, keywords, term banks) is a central topic. Combining this with document search extends resource coverage

Approaches:

- WG1 & WG2 (each WG focusing on their area but interacting for similar/linked features):
  - annotating sections with relevant information inside the document/publication
  - relevant metadata included in the resource description (e.g. domain, keywords, …)
  - exploiting existing and finding appropriate ways of linking and mapping concepts as represented in publications, lexica, ontologies etc.
- WG3:
  - representing terms of use in a machine readable format
  - ensuring that interaction between multiple resources is allowed when required only if there are no conflicts between the terms of use
- WG4:
  - indicating appropriate input & output formats for resources involved in the workflow;
  - discussing the architecture of appropriate processes and workflows (e.g. how and when the various resources should be accessed)

Open questions:

- Ascertain that indeed it is worth exploiting
- Discuss interoperability issues between the various types of resources (texts, lexica, ontologies etc.), especially with regard to the architecture of relevant workflows, and suggest appropriate ways of tackling them.

Possible requirements/interoperability recommendations

Keywords should be in some way qualified with the source of the keyword if it belongs to a controlled vocabulary. E.g. a keyword could be a URI or there could be an additional metadata field indicating where keywords come from (latter related to WG1).

WG3 — IPR and Licensing

Scenario 1 — Legal status of aggregations: focus on content

Description

Researcher A is conducting research on a variety of sources. He needs to extract statistical information from 3 DB: X, Y and Z. He runs a TDM instance on the 3 databases and obtains insightful results. He wants to know how, if at all, he can use the resulting output. Database X is made publicly available on the Internet by the right holder but is not released under any specifically identified licence, term of use or any other policy. DB X is constituted by a text corpus (i.e. the individual elements constituting the database are not individually copyrightable). Database Y is available online under a CC-BY-SA 4.0 license and is constituted by a DB of academic articles (i.e. the elements constituting the database are individually copyrightable). Database Z is a "closed" database to which the researcher has access thanks to the "academic subscription" of its institution which allows to TDM the database but only for "academic and research activities and for non commercial purposes" and forbids the distribution or communication of results for any purpose. All these activities are performed using tools and or services belonging to the researcher (i.e. no licence or terms of use on tools and or services are relevant in this scenario).

Relevance to the WG

The scenario identifies the main elements of legal interoperability: absence of clear license or right statement, OA publishing, closed databases, Exceptions and Limitations to Copyright. The scenario also considers the type of results that TDM activities lead to (whether there is a redistribution of at least part of the original data).

Relevance to other WGs

- WG 1 - Are the licences or rights statements expressed in (machine-readable/processable) metadata?
- WG 2 - N/A
- WG 4 - License information and provenance information must be preserved throughout the workflow.

Relevance in general

External experts combine expertise in the legal and technical field. The combination of the three identified DB legal conditions is the starting point to determine legal interoperability.

Approaches

- Correct labeling of the resources with information about the licence, and with relevant licence metadata. With proper labeling, through the correct application of the licence and licence metadata, a first set of problems can be solved.
- Correct use of licences. The convergence towards a de facto standard (e.g. CCPL) will reduce incompatibility issues and transaction costs.
- Exceptions and Limitations to Copyright: proper exception to copyright and related rights (e.g. UK to some extent) can contribute to fix additional access and reuse issues.

Scenario 2 — Focus on TDM tools and TDM services

Description

The same scenario as scenario 1 above applies, however database (Y) is not directly accessible. On the contrary, the database owner allows the user (on the basis of the academic subscription) to run the rights holder dedicated TDM web service. Basically, the researcher does not have direct access to the database, but he is only able to place a query and obtain a result. The TDM web service is made available to the researcher for “non commercial uses” only. On the other hand, database Z is directly accessed but using the software (freely downloadable) of Company Alfa. The licence of software Alfa allows users to use the software only for non commercial purposes. For the purpose of this scenario, database Y and Z are not protected by any IP rights nor their use is contractually restricted, i.e. they are in the public domain (i.e. the scope of the question is to test the enforceability of services’ and tools’ agreements in relation to the results of the TDM activity).

The results of the TDM so obtained are all combined together in one new DB. Can the researcher publish a paper based on those results? Can he publish the paper under a CC-BY 4.0 license. Can he also publish the DB itself under a CC-BY 4.0 license? If yes, is he bound to apply a NC licence? Can the researcher develop a commercial product and sell it? Would the fact that the “results” reproduce substantial or insubstantial parts of the original DB, or on the contrary that those results do not reproduce parts of the original DB (for example the statistical recurrence of a given word in the DBs, but not the word itself) make a difference in the result?

Relevance to the WG

The scenario adds a main element of complexity: some TDM services restrict the use that you can do with the results obtained from their services. In this case researchers are not directly accessing a database but they are using a service offered by the data provider to analyse its own DB. Of particular relevance when the DB is not protected by copyright or sui generis database rights (SGDR) and therefore the SLA (service level agreement) is the only basis to restrict use of the results (not backed-up by property rights).

Relevance to other WGs

- WG 1 - Are the licences or rights statement expressed in DB metadata? How to identify this license in case of services?
- WG 2 - N/A
- WG 4 - License information and provenance information must be preserved throughout the workflow.

Relevance in general

External experts combine expertise in the legal and technical field. The combination of the DB legal conditions and the add-on of the service element an additional step toward the determination of legal interoperability.

Approaches

- Correct labeling of the resources with information about the licence, and with relevant licence metadata. With proper labeling, through the correct application of the licence and licence metadata, a first set of problems can be solved.
- Correct use of licences. The convergence towards a de facto standard (e.g. CC variants) will reduce incompatibility issues and transaction costs.
- ELC: proper exception to copyright and related rights (e.g. UK to some extent) can contribute to fix additional

access and reuse issues.

Scenario 3 — The type and nature of TDM results (or How far do copyright and SGDR reach)?

Description

The same scenario as scenario 1 above applies. Would it make any difference if the resulting DB, although being based on the data of the original databases, does not reproduce any substantial or insubstantial parts of those databases? In other words, the results of TDM are directly based on the analysis of the original DBs but do not materially reproduce any part of them? For example the results of the TDM activity is "in the analysed corpora the word "and" occurs 235 times".

Relevance to the WG

The scenario looks into the scope of copyright and SGDR protection. To what extent a DB that "derives" from another one, without however reproducing it, can be considered a derivative of the originals, under both copyright rules and SGDR rules?

Relevance to other WGs

- WG 1 - This is fundamental aspect of TDM that should be relevant for all WG.
- WG 2 - N/A
- WG 4 - Workflows can produce abstract representations that are a long way from the original text (relates to "copyright protects the expression of a fact, not the idea, not the fact itself"). They can also produce associations that were never explicitly mentioned anywhere in any of the original data, i.e., produce new knowledge previously unknown by anyone. That knowledge is derived, it could not have been found without the original data, but does it constitute a derivative work?

Relevance in general

External experts combine expertise in the legal and technical field. Try to identify the limits in the scope of the applicable legal tools is a fundamental element.

Approaches

- Correct labeling of the resources. With proper labeling, through the correct application of the licence and licence metadata, a first set of problems can be solved.
- Correct use of licences. The convergence towards a de facto standard (e.g. CC variants) will reduce incompatibility issues and transaction costs.
- ELC: proper exception to copyright and related rights (e.g. UK to some extent) can contribute to fix additional access and reuse issues.

Reqiuirements for WG3 from other WGs' scenarios

From WG1

- Interoperability of different tools (scenario 1)

- Conditions of access to different resources (scenario 1)
- Discovery of legally interoperable tools on the basis of licences (scenario 2)
- Standardisation of licensing schemes at the metadata level (scenario 2)
- Combination of licences from repository/journals and licences (aka terms of use) of services used and how these impact the returned annotations (scenario 3)

From WG2

- combination of several vocabularies: resource access; combination of resources with different licences; publication of annotated corpus (scenario 1)
- Issue of confidential information such as NDAs (scenario 2)
- Use of non licensed resources (scenario 2)
- Safe harbour provisions? (scenario 2)
- Conditions for access to resources and need to represent those conditions in machine readable format (scenario 3)

From WG4

- Licensing issues (ToS) which may affect which services can be combined (scenario 1)
- Use of TDM for internal uses only, without redistribution or licensing (scenario 3)
- Legal status and licence conditions of input data (scenario 4)
- Possibility and risks associated with distributing or linking a workflow, and in distributing data processed by the workflow (scenario 5)

WG4 — Annotations and Workflows

Type of scenarios

- Typical problems that may occur during the process of building workflow (Scenario 1, …)
- Typical usage of workflows (Scenario 2, 4)
- Typical usage of text annotations (Scenario 3, …)

Scenario 1 — Transferability of components between ecosystems

Description

I am building a text mining workflow in a specific workflow editor (e.g. Argo). There is a really useful component, but it is in a repository for a different workflow editor (say GATE). I cannot use this really useful Gate workflow (which may be made of one or more modules) in my Argo workflow. I must either reimplement the really useful component in Argo or convert the work I have already done into Gate.

Relevance to the WG

This scenario exposes the core issue in workflow interoperability. Because different workflow editors use differing annotation schema, a component written for one workflow ties the user into that specific ecosystem. This has a direct negative impact on the scientific goals of openness and reproducibility.

Relevance to other WGs

- Resource MetaData (WG1): A key form of metadata is the input and output of each tool. Standardisation of this metadata would lead to greater interoperability between tools.
- Language Resources (WG2): Annotations produced by a specific component may be inaccessible to alternate components due to the annotation scheme in use. Also, workflows which are built in a specific workflow editor will not be usable within other workflow editors.
- IPR and licensing (WG3): Less relevant. Licensing issues may affect which services can be combined. May also affect the usage of the workflow's results if an incorporated component has a stricter licensing policy.

Relevance in general

External experts will want a smooth experience when using a TDM framework. Having to reimplement components is frustrating and costs time and money.

Approaches

- Common annotation formats and annotation schemes between components.
- Workflows as web services with the ability to incorporate web services into workflows.
- Building a grand unified text annotation workflow editor (and getting everybody to use it) or facilitating the embedding of components from one ecosystem in the workflow of another one, cf. GATE-UIMA bridge such that users of ecosystems that already exist for the respective frameworks can continue to work in their known environment while still profiting from components previously out of reach.

Open questions

- Experiences of being constrained to a specific ecosystem
- Requirements for the integration of a set of incompatible components.
- Estimation of the cost to their business of having to reimplement a component / workflow.
- Experiences of actual benefits of reimplementing a component / workflow.

Scenario 2 — Comparison of competing components or parameters

Description

I want to compare competing components for TDM (e.g. which is the best parser/tokeniser/event extraction component for my text?). I have a workflow which is made up of several components (e.g. A tokeniser, A POS tagger and an event extractor). I have several options for each component. I would like to know if changing the tokenizer / POS tagger / event extractor will improve or degrade my results. I would like to know if changing a parameter of the tokenizer / POS tagger, etc. (e.g. which model to use, which kinds of abbreviation lists to use, etc.) affects the results. I would like to know if a specific version of a component which I use will improve my results. This not only be the version “1.2.3” but be the “latest”, or “latest stable”, or “latest snapshot” and the respective version is automatically looked up in a repository.

Relevance to the WG

Workflows provide a seamless way to replace components with similar functionality. In the easiest case, competing components will have exactly the same inputs and outputs as each other and the user will be able to directly replace one component with another. In a more complicated scenario components may require extra resources or annotations. These can still be incorporated, as long as the prerequisite components are also available.

Relevance to other WGs

- Resource MetaData (WG1): Less relevant. Competing components can only be discovered if they are stored with appropriate metadata.
- Language Resources (WG2): Almost directly correspondent. WG2 are interested in both LR and annotation interoperability, which are very important for this scenario.
- IPR and licensing (WG3): Less relevant. Competing components should be available for testing.

Relevance in general

External experts will be interested to know which option from many components is the best for their specific task. If it is difficult to interface components (say, a wrapper must be written for each new component), then the expert is likely to only evaluate one or two components. The more components they can compare for a specific sub-task, the likelier it becomes that they will find the correct component for their purposes.

Approaches

By creating specification for interoperable components, OpenMinTeD will provide a framework for the easy comparison of components within a workflow setting.

Open questions

Types of components where they would like to compare competing workflows. Experiences of building workflows, but not knowing which tagger / tokenizer / etc. to use or how to parametrize it.

Scenario 3 — Non-expert provider of TDM resource

Description

I want to use text mining techniques to analyse literature in an area that has not previously been examined with TDM (e.g. paleontology). Unfortunately, although I am an expert in the subject area, I know very little about language processing or TDM. I want to start with annotating a set of documents. How do I retrieve the articles? Where do I find the necessary tools? Do I need dedicated hardware or infrastructure? Where should I start?

Relevance to the WG

This scenario, taking a non-expert perspective, shows how hard it is to start TDM research, even by building the simplest workflow. Most of them require a user to prepare infrastructure, install and configure the tools, and convert the data into a specified format. Only a small fraction of TDM tools is available as simple-to-use web tools.

Relevance to other WGs

- Resource MetaData (WG1): In the scenario the initial annotated corpus is created by an expert from different field. A clear metadata standard is necessary to describe it and ensure that others (probably TDM experts) could access and process it.
- Language Resources (WG2): In some cases the user may want to refer to some external resource (e.g. ontology containing extinct species) in the created annotations — he would need then to have a quick overview of available resources of that kind.
- IPR and licensing (WG3): The non-expert user creates a resource, but probably doesn't have knowledge about copyright law. He needs to be provided a basic information about relevant licensing schemes to choose from. The user may not even want to publish or license the data, but rather apply TDM only for internal use, potentially without the data leaving the house at all.

Relevance in general

External experts frequently have little knowledge about TDM but possess domain expertise necessary to create annotated resources. They know how the TDM field looks "from the outside".

Approaches

OpenMinTeD could build (or at least help to build by providing interoperability standards) a single, easy to use interface that would let inexperienced users to employ different tools in their work, regardless of their data formats and infrastructure requirements. What is more, OpenMinTeD could provide a simple guide that will help them to choose the right text representation standard, metadata format and license.

Open questions

Experiences of trying to build resources without expertise in the field: which was the hardest step? What kind of information they needed?

Scenario 4 — Reproducibility of TDM-related research

Description

I want to give readers of my paper the possibility to re-run my experiments involving text processing with modified settings or data and check the results. Unfortunately, contrary to statistical data analysis, it is not enough to provide the input files and program script — instead, the reader would have to install all the tools I used in the project in his computing environment, configure them, take care for data format conversion, etc. As he is very unlikely to be able to do it, my results are virtually irreproducible.

Relevance to the WG

The problem in this scenario is that although an interested paper reader may be a TDM expert, he/she probably uses different workflow and environment. To make him able to analyse a given data set, we need the possibility to run a given program without the user manually installing and configuring all the tools on a target environment (e.g. a local PC, a local server, a cloud server, etc…). The user may either choose to re-run the experiment with slightly different parameters or input data, or he may edit the workflow (e.g. replace some of the components) and check whether the conclusions remain valid. That clearly requires a lot of work in the area of annotations and workflows interoperability.

Relevance to other WGs

- Resource MetaData (WG1): In order to reproduce the experiment, detailed information about the resources used, their dependencies, their provenance, and particularly their versions is required. This entails a method of transitively resolving references from one resource to another (e.g. in a dependency or provenance graph).
- Language Resources (WG2): The ability to refer to stable versions of language resources which may be problematic e.g. for knowledge resources only available through web services.
- IPR and licensing (WG3): This problem applies to the researcher sharing his input data: he needs to have necessary knowledge about licensing to ensure that he is not violating any laws by doing so.

Relevance in general

Here, experts with experiences in TDM would be very helpful, because they know what prevents them from sharing all their input data and procedures or replicating workflows of others.

Approaches

OpenMinTeD could build (or at least help to build by providing interoperability standards) a single, easy to use interface that would let curious readers repeat the experiments without cost and time needed to fully install the whole environment. In case it is not possible, OpenMinTed could alternatively provide standards, best practices, and tooling to facilitate the installation of the environment. The latter is also an important step towards dynamically deploying TDM workflows to computing resources in the age of scalable cloud computing.

Open questions

They could share knowledge about what prevents them from sharing input and procedures of their research projects. Is it licensing? Or maybe amount of work necessary to fully describe the computing environment?

Scenario 5 — Integration of a TDM workflow in a service/embedding in an application

Description

I, project manager, need a TDM workflow in order to provide TM/NLP functionality to the service we are developing. Examples:

- A smart search function to a web site.
- The recommendation system of a vendor makes use of product descriptions and user reviews.
- A bioinformatics workflow for the annotation of a large genome includes the extraction of information from relevant papers.

Relevance to the WG

The developers need to know how to invoke the workflow when new input is provided and how to fetch the result of the TDM analysis. They might require continuous processing of incoming text.

They will also require some level of robustness, the TDM should then be able to provide failure/recovery information. Also they might want to get as many results as possible as soon as possible even if the whole workflow has not finished.

From the workflow design PoV, we assume in this scenario that the team includes one TDM professional. However they might want to optimize the components according to specific criteria (storage, CPU usage, robustness, scalability).

Relevance to other WGs

- Resource MetaData (WG1): provenance, versioning, and licensing information might be critical in this scenario.
- Language Resources (WG2): low relevance.
- IPR and licensing (WG3): in this scenario the TDM workflow is part of a service that gives important context for WG3.

Relevance in general

Experience of deploying enterprise or end-user services that includes some level of TDM would be welcome.

Approaches

OpenMinTeD could specifiy a standard API to access worklows (parameters, LR), output (annotations, clusters, classifiers), and status (running, fail).

Open questions

From experts:

- What kind of services are known to use TDM analyses?
- Which issues arise when deploying a workflow, what hinders the use and adoption of a workflow engine?

From WGs:

- From WG1: formal description of provenance, versioning, and licensing.
- From WG3: IPR issues in distributing or linking a workflow, and in distributing data processed by the workflow.

Scenario 6 — Development of TDM resources

Description

A TDM specialist or a domain expert is building a linguistic/semantic resource, it could be:

- an annotated corpus for training NER or IE components;
- a trained classifier
- a terminology or a lexicalized ontology;
- hand crafted rules.

These resources are typically used in other workflows. However the development of resources itself requires a TDM workflow. On one hand it could feed the user with suggestions (pre-annotation for corpus annotation, IE or term extraction for terminology design). On the other hand the evaluation of the resource can be partially achieved through a TDM workflow (inter-annotator agreement, testing with the resource).

Relevance to the WG

The TDM workflow could have protocols compatible with the main annotation and ontology editors on the market.

Moreover the manual intervention in the middle of the workflow poses a challenge because it cannot be designed exclusively as a batch process. The pace of human produced output is at the same time slower and more valuable than the result of fully automated output. The TDM workflow could proceed either continuously as the experts modify the resource, or with partial input.

Relevance to other WGs

- Resource MetaData (WG1): in this scenario the resource is one of the outputs of the workflow, they should conform to the standards provided by WG1.
- Language Resources (WG2): They need to be enabled to precisely define the information types they want as output from the customized workflows. These information types should correlate with WG2 data interoperability.
- IPR and licensing (WG3): IP status of derived products, IP status of collaborative artefacts.

Relevance in general

Experience and insights on not strictly batch workflows.

Approaches

Allow interactive steps within the TDM workflow. Propose solutions compatible with editors.

Open questions

Knowledge about how they have been solving the problem so far and how could OpenMinTeD improve the process.