INTEROP 2016 @ LREC

http://interop2016.github.io

Report

Monday, May 23, 2016

LREC 2016

Portoroz, Slovenia

Editors:

Richard Eckart de Castilho
Mouhamadou Ba
Penny Labropoulou
Giulia Dore
Wim Peters

Reviewers:

Thomas Margoni
Stelios Piperidis

Thanks for additional reviewing go to Dominique Estival!

Workshop organization

Motivation

Recent years have witnessed an upsurge in the quantity of available digital research data, offering new insights and opportunities for improved understanding. Following advances in Natural Language Processing (NLP), Text and data mining (TDM) is emerging as an invaluable tool for harnessing the power of structured and unstructured content and data. Hidden and new knowledge can be discovered by using TDM at multiple levels and in multiple dimensions. However, text mining and NLP solutions are not easy to discover and use, nor are they easy to combine for end users.

Multiple efforts are being undertaken world-wide to create TDM and NLP platforms. These platforms are targeted at specific research communities, typically researchers in a particular location, e.g. OpenMinTeD, CLARIN (Europe), Alveo (Australia), or LAPPS Grid (USA). All of these platforms face similar problems in the following areas: discovery of content and analytics capabilities, integration of knowledge resources, legal and licensing aspects, data representation, and analytics workflow specification and execution.

The goal of cross-platform interoperability needs to tackle many challenges. At the level of content access and discovery, metadata-based descriptions, language resources, and text annotations, we use different data representations and vocabularies. At the level of workflows, there is no uniform process model that allows platforms to smoothly interact. The licensing status of content, resources, analytics, and of the output created by a combination of such licenses is difficult to determine, and there is currently no way to reliably exchange such information between platforms. User identity management is often tightly coupled to the licensing requirements and likewise an impediment for cross-platform interoperability.

Workshop Goal

An important ingredient for interoperability is collaboration, which in turn is boosted by knowing the involved stakeholders and trusting them. The goal of the workshop was to provide a forum in which relevant stakeholders from the language resources and NLP communities meet, talk about ongoing efforts related to interoperability, and exchange experiences and ideas. Interoperability requires that stakeholders have trust in each other and that being interoperable is considered to be a win-win situation. Organizing the workshop as an official LREC workshop provided improved access to international experts also attending the main LREC conference, and in particular stakeholders from USA (LAPPS Grid) and Australia (Alveo).

With a focus on incentivizing communication and trust, the following mission statement was provided as:

- What are the present problems and open questions?
- What is currently being done?
- What are the solutions or promising approaches?
- What steps need to be taken next?

Topics of Interest

Workshop topics included but were not limited to:

- cross-repository discovery of content, language resources, and analytics
- uniform access to content repositories or heterogeneous data sources (content, knowledge)
- extraction of textual content from heterogeneous sources
- orchestration of analytics workflows composed from analytics from different sources
- orchestration of cross-platform analytics workflows
- linking knowledge sources and uniformly accessing them from analytics workflows
- annotation schema design best practices
- mapping and transformation between annotation schemata
- dynamic deployment of analytics to computing resources
- machine-interpretable representation of legal and licensing metadata
- policy making for TDM for an international open research environment and open access publishing

Target audience

NLP analytics providers, NLP researchers, NLP research infrastructure providers, interoperability experts, legal experts, language resource providers.

Workshop agenda

Monday, 23 May 2016	
9:00 - 9:10	Introduction
9:10 - 10:00	**Keynote: Interoperability - Can a model driven approach help to overcome organizational constraints?** *Alessandro Di Bari, IBM*
10:00 - 10:30	**Lightning talks** *5 minutes per talk, slides optional* Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not? *Petr Knoth and Nancy Pontika* Alveo: making data accessible through a unified interface – a pipe-dream? *Dominique Estival* The Language Application Grid *Nancy Ide, Keith Suderman, James Pustejovsky, Marc Verhagen, Christopher Cieri and Eric Nyberg* Interoperability of corpus processing work-flow engines: the case of AlvisNLP/ML in OpenMinTeD *Mouhamadou Ba and Robert Bossy* Integration of UIMA Text Mining Components into an Event-based Asynchronous Microservice Architecture *Sven Hodapp, Sumit Madan, Juliane Fluck and Marc Zimmermann* Interoperability = f(community, division of labour) *Richard Eckart de Castilho*
10:30 - 11:00	Coffee break
11:00 - 11:45	**Lightning talks (continued)** *5 minutes per talk, slides optional* Linked Data and Text Mining as an Enabler for Reproducible Research *John P. McCrae, Georgeta Bordea and Paul Buitelaar*

	Tackling Resource Interoperability: Principles, Strategies and Models *Wim Peters* The DDINCBI Corpus — Towards a Larger Resource for Drug-Drug Interactions in PubMed *Lana Yeganova, Sun Kim, Grigory Balasanov, Kristin Bennett, Haibin Liu and W. John Wilbur* Multilingual Event Detection using the NewsReader pipelines *Rodrigo Agerri, Itziar Aldabe, Egoitz Laparra, German Rigau, Antske Fokkens, Paul Huijgen, Marieke van Erp, Ruben Izquierdo Bevia, Piek Vossen, Anne-Lyse Minard and Bernardo Magnini* Emotion Detection using Online Machine Learning Method and TLBO on Mixed Script *Shashank Sharma, PYKL Srinivas and Rakesh Balabantaray* Text mining for notability computation *Gil Francopoulo, Joseph Mariani and Patrick Paroubek* Why We Need a Text and Data Mining Exception (but it is not enough) *Thomas Margoni and Giulia Dore* Legal Interoperability Issues in the Framework of the OpenMinTeD Project: a Methodological Overview *Penny Labropoulou, Stelios Piperidis and Thomas Margoni* eInfrastructures: crossing boundaries, discovering common work, achieving common goals *Hege van Dijke and Stelios Piperidis*
11:45 - 12:00	Constitution of breakout groups
12:00 - 13:00	**Breakout Groups** Data Management and Metadata *Discovery, Access, Corpus Creation, and Managing Results* Processing Components, Data Exchange, Workflow Management *Composing, Deploying, Scaling* Language and Semantics Ambiguities, Semantic Resources, Underspecification, *Cross-/and Multi-Lingual Aspects*

	Legal Aspects and Policy *Aggregating, Deriving, and Transforming*
13:00 - 14:00	Lunch break
14:00 - 16:00	**Breakout Groups (continued)**
16:00 - 16:30	Coffee break
16:30 - 18:00	Presentation of the breakout groups results Plenary discussion

Methodology/Format

Peer reviewed short papers. Contributions to the workshop were submitted as short papers of up to 5 pages. The papers were peer-reviewed by an international committee of researchers from academia and industry. Fourteen of the submitted papers were accepted and included in the workshop proceedings published online along with the rest of the LREC 2016 proceedings.

Lightning talks. One author of each paper had the opportunity to present their position on interoperability during a five-minute lightning talk. Most authors used this time to closely follow the structure of their respective papers, others focussed more strongly on interoperability aspects during their talk.

Breakout groups. After the lightning talks, we formed four breakout groups around the topics of:

- Group 1: Data Management and Metadata
- Group 2: Processing Components, Data Exchange, Workflow Management
- Group 3: Language and Semantics
- Group 4: Legal Aspects and Policy

The discussion in the breakout groups was centered around topics/questions that were collected before the workshop through a poll on a website (http://www.well-sorted.org) and during the workshop by means of cards that attendants could fill in and assign to groups during the keynote and lightning talk sessions. During the breakout session, these topics were reviewed, prioritized and then discussed.

Plenary. Results from the breakout sessions were presented and discussed during a closing plenary session. Each group was allocated a 20 minute slot. A maximum of 10 minutes was permitted for presenting the results. The remaining time of the slot was used for discussion.

Workshop results

Group 1: Data Management and Metadata

Issues discussed

Metadata enables discovery

The first topic of the discussion related to the easiness of discovering and accessing datasets, knowledge resources and processing tools. There was consensus that aggregators and catalogs are essential for searching and identifying such resources. However, the need for increasing their visibility was stressed, given that many end-users still resort to posting messages to dedicated community mailing lists. The need for getting access to full texts for TDM purposes through these aggregators, catalogs, or the actual data hosting repositories harvested by them, was considered a must. Unfortunately, this is also a major problem with most catalogs, as such access is usually not provided.

As pointed out by the group participants, one of the main benefits of using such catalogs is the availability of resource descriptions through formal metadata. The discussion went on to identify problematic issues in the current status of metadata and possible ways of addressing them. The following main criticisms were identified:

- the proliferation of metadata schemas and the consequent difficulty in their mappings (which hampers interoperability among the catalogs);
- the lack of strictness in metadata schemas and elements (especially as regards Dublin Core, which enables its use in a wide variety of applications and profiles across different communities, but lessens semantic accuracy and detail of descriptions);
- the admitted reluctance of resource providers such as researchers, to describe their resources manually and in detail. As pointed out, researchers just want to publish/upload their data and not describe them.

Creating metadata

One of the solutions discussed was the automatic extraction of metadata elements from various sources: e.g. detection of file formats from the uploaded data, use of NLP tools and techniques (especially unsupervised techniques) to extract metadata from the content itself (e.g. topic classification) but also from free-text metadata, extraction of user-generated/uploaded metadata (e.g. author names and affiliations) at the time of the submission of an article. This automatically generated metadata could then be presented to users for confirmation, thus ensuring the best combination of using manpower and computer techniques. User-generated metadata should be provided, of course, respecting privacy.

Metadata in the form of tag clouds was also discussed.

User tagging of other researchers' resources has also been discussed but doubts have been expressed; it entails a lot of work with unclear results. Still the idea of creating a collaborative platform for the collection of metadata and indexing of resources, where metadata elements automatically harvested from various sources are combined together into a single metadata record for each resource and are further edited/curated by the relevant resource providers/creators, seems promising.

As regards user-generated metadata, the need to create tools and documentation that will guide non-technical users (e.g. linguists) into using existing metadata schemas was pointed out.

The point of preservation of metadata elements and schemas and the appropriate ways of achieving this was raised in relation with legacy metadata that is often contained inside the data, as in the case of Alveo datasets. The metadata are important for language-related research, but the automatic extraction of these elements is extremely difficult. The question as to which are the best ways also of designing interfaces for displaying metadata information in generic yet user-adaptable ways (e.g. creating facets of datasets) was raised.

Too many metadata schemas

The existence of various metadata schemas was also discussed. It was agreed that there cannot be consensus for a single common metadata description schema and that although standards do exist, they are not widely adopted for various reasons. The most common metadata schema used mainly for exchange is DC but often criticized of being non-informative. The wide use of the OAI-PMH protocol for metadata exchange and the lack of a similarly popular protocol for the exchange of full-text data were mentioned - ResourceSync is not yet adopted.

Distinctions were made for metadata based on the intended end-usage, i.e. whether they are to be consumed by s/w tools or intended for human consumption; also the distinction between metadata extracted exclusively from the document and document structure vs. metadata generated by users was made. Different approaches in the application of metadata schemas were discussed, ranging from strict, rich in information and hierarchical metadata schemas such as META-SHARE, which require manual editing, vs. free structures of common elements to be taken from a common pool such as CLARIN relying on a flat ontology. Commercial concerns (e.g. of publishers) also seem to block the wide adoption of standards; however, the pressure of various communities is expected to change this attitude eventually.

The need for mappings and conversion tools between various metadata schemas and elements was accepted by all; the mappings need to be made at the level of the concept (metadata element and values) through relations such as "sameAs" and "similarTo". At the same time, the need for formally distinguishing and identifying criteria for selecting between similar metadata elements (e.g. author vs. creator) was also felt that it would help the clear usage of metadata schemas.

Consolidation of metadata schemas

As a general rule, it was pointed out that mappings should be used for existing schemas, standards should be promoted for new resources and their metadata descriptions, but for legacy/older schemas, the mappings should be very carefully designed in order to preserve the information that is contained in them without any loss.

Classification information is a good example as to the problems of metadata: schemas use various elements (genres, text types, domains, topics etc.) and values to record them even when the same vocabulary is used, the way this appears may vary from schema to schema (e.g. DC codes vs. full names). Moreover, keywords remain one of the main classification items provided by users; the main problem is that keywords are too detailed, not the high-level tags required by s/w programs. The need for normalization was stressed. Approaches for automatic topic classification were discussed; highly successful approaches to classification are not necessarily based on the content, but rely on information about who reads what (e.g. in Mendeley). The fact that the academic community is losing these data by not maintaining these platforms is unfortunate ...

Levels at which to add metadata

The level at which metadata should be added was discussed but not to a great length. Metadata can also be added at the level of the annotation (e.g. who has annotated which sentence, the source of the annotation etc.), but at this point this was considered too detailed. For the time being, the focus should be on at least adding at the level of the document information about the annotations and structural encoding of documents. It is felt that we should first overcome problems such as conversion of PDF files (especially of scanned files) and extraction of divisions (citations, figures, chapters, sections etc.) from texts.

Copyright

Legal issues were only briefly mentioned, as this was the topic of another dedicated breakout group. Everyone agreed that metadata records should be "open access" while resources themselves should include a metadata element recording their licensing.

Sustainability and Quality Assurance

One of the main problems that were noted is the lack of sustainable sources of funding for research infrastructures. The need to have permanent funding to ensure continuous archiving and preservation of data, and curation of data and metadata was brought up again and again during the discussion. Voluntary or user-provided data and metadata-related activities cannot ensure quality and commercial platforms may hinder accessibility. One cannot rely on aggregators that provide metadata based on inaccurate methods extracting information from documents themselves; a more formal source is required.

Trustability and security of resources were raised as questions: how far should providers be trusted, and what security measures can be undertaken to ensure that uploaded tools do not infringe copyright, or to block devious s/w tools? The detection of duplicates was considered important; it's already usually done through hashing; however, this cannot be performed for metadata without data. It was felt, though, that for archiving reasons it is desirable to have data duplicates at various sites, but applications should include a facility for checking duplicates and generating links among them.

Personalized data collections

Discussion evolved also around the topic of "virtual research datasets". As a rule, users use APIs to create datasets they want for their research. This may suffice for doing the research but to ensure reproducibility of the experiment, the data itself must be downloaded and stored.

Taking as a starting point the proposal for identifying datasets via double hashing, the topics of data storage, archiving and deduplication were addressed. Double hashing[1] is a project that has just started, so there isn't yet much to report; its goal is the sustainability of datasets.

Possible actions and plans

Most of the actions and plans have already been hinted at in the previous section. In summary, the main actions should promote the following points:

- Get publishers and resource providers for making data available and discoverable
- Need for a TDM exception for the data that is being processed but also for the outputs of the processing procedure; the problem is not getting the data (for personal research/use) but redistributing it
- Collaboration between providers & infrastructures in order to have a method of uniquely identifying resources across multiple catalogs. A DOI may work, but not all works have a DOI. So the idea of content-based hashing was raised and will be further explored. Content-based hashing would permit tracking resources across multiple registries and repositories.
- Open academic infrastructures are important in a number of ways and should be sustained through public funding; they are instrumental in retaining control over user generated metadata against commercial entities, in ensuring that future research trends are not directed through TDM operations in the hands of private interests, as well as in promoting standards, cultivating community spirit and sharing mentality.

Pointers to related activities and work

This list is an indicative and non-exhaustive list of the activities and resources mentioned during the workshop.

Catalogs of resources

- Linguist List (http://linguistlist.org),
- PubMed (http://www.ncbi.nlm.nih.gov/pubmed)
- Google Scholar (http://scholar.google.com)

[1] Cf. http://interop2016.github.io/pdf/INTEROP-7.pdf

- Semantic Scholar (https://www.semanticscholar.org/)

Harvesting protocols

- ResourceSync - for full texts (http://www.openarchives.org/rs)
- OAI-MPH (https://www.openarchives.org/pmh/)

Metadata schemas and vocabularies

- CrossRef (http://www.crossref.org): for citation of scholarly publications
- VoID (https://www.w3.org/TR/void/): vocabulary for metadata between RDF datasets; mainly used in the Semantic Web community, not necessarily known by the NLP community
- Linked Open Vocabularies (http://lov.okfn.org/dataset/lov/): various linked data ontologies and vocabularies
- ISOcat: vocabularies put together by various experts; in the process of being validated; taken up by CLARIN Concept Registry (https://openskos.meertens.knaw.nl/ccr/browser/)
- META-SHARE (http://www.meta-share.org/knowledgebase/homePage): metadata schema for language resources
- CLARIN metadata (https://catalog.clarin.eu/ds/ComponentRegistry): community-created profiles and components
- A review of ontologies for describing scholarly and scientific document structure (http://ceur-ws.org/Vol-1155/paper-07.pdf)

Preservation and archiving

- LOCKSS (https://www.lockss.org/): Lots of Copies Keep Stuff Safe; tools for data storage & preservation

Identification

- ORCID id (http://orcid.org/): Unique identifiers for researchers

Structural annotation of documents

- JATS (https://jats.nlm.nih.gov/publishing/0.4/xsd.html): Tags for journal articles, also for structural divisions (e.g. title, abstract, acknowledgements, citations etc.)
- TEI (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/): schema for digital texts; widely used in linguistic communities; includes tags for division and annotation of sections

Group 2: Processing Components, Data Exchange, Workflow Management

Issues discussed

The discussion focused mainly on:

- workflow management and related tools (with Galaxy being a prime example)
- the profile of the users who define the workflows
- data provenance
- documentation and tutorials about the interoperability solutions
- scaling of the workflows.

Workflow management

The discussion pointed out that the NLP/TDM community can benefit in many aspects from existing scientific workflow systems. For example, Galaxy is a platform originating from the bioinformatics community to define workflows for data analysis, particularly in bioinformatics. It enables users to create and execute workflows but also to manage the data and workflow life cycles. Galaxy manages different types of data and services. It offers accessibility, reproducibility and shareability of the resources. In Galaxy, one can access data from external platforms (e.g., UCSC Genome Browser[2], Ensembl platform[3]) process them with workflows and save the results. The data and the workflows can be shared and re-used through communities. Furthermore, Galaxy is already used by several language-oriented projects, for example LAPPS Grid, Alveo, or CLARINO.

Similar systems to Galaxy are mentioned during the discussion: for example Taverna[4] that supports web services and some complex constructs in workflows, VisTrails[5] that is an open-source scientific workflow and provenance management for simulation, data exploration and visualization, Kepler[6] that specially offers wrappers to integrate the MapReduce components (e.g., Hadoop, Storm), and Alveo[7] that uses the Galaxy Workflow Engine to run a range of text, audio and video analysis tools. The mentioned systems globally seem to use normalized workflow descriptions, simplified interfaces to define workflows, and they offer wrappers for local services (and web services in some systems).

[2] https://genome.ucsc.edu/index.html

[3] http://www.ensembl.org/index.html

[4] https://taverna.incubator.apache.org/

[5] http://www.vistrails.org/

[6] https://kepler-project.org/

[7] http://alveo.edu.au/

Heterogeneous workflows

Next, the discussion turned to the issues of heterogeneous workflows in which the workflow components use different input and output formats. This is in contrast to the NLP/IE communities more prominent homogeneous workflows where a common format is used for data exchange at the workflow level, e.g. as used in GATE and UIMA. A shimming approach was suggested (and is being used in LAPPS Grid) for managing interoperability between the services/tools. The use of shims (format converters) permits workflows to use multiple formats but requires the maintenance of these converters. While a star-like approach using a pivot format for data exchange was generally deemed more maintainable, the problem of not being able to agree on a single pivot format makes the shimming approach more realistic. At least in a specific environment, such as LAPPS Grid, the number of format converters can be kept low by incorporating whole sets components from existing ecosystems (e.g. GATE and UIMA).

Workflow structure

The necessary expressiveness of workflow descriptions was discussed, e.g. whether it was necessary to support conditions or loops. Here, a conflict between usability and expressiveness was detected. For example, Galaxy offers a user-friendly editor to create workflows but the workflows are presently limited to directed acyclic graphs. Galaxy does not offer complex constructs such as conditions and loops in the workflows. However, one participant mentioned that they actually had removed such advanced functionality from their proprietary workflow system because users found it too complex. Instead, their system was switched to a dynamic routing strategy where each component determines to which next component data will be sent. In this context, also the difference between synchronous and asynchronous processing was discussed. Asynchronous processing was seen not only as beneficial for workflows using a dynamic routing strategy, but also for workflows involving a human in the loop (e.g. an annotation task). To manage asynchronous communication, the use of message queues like ApolloMQ or ActiveMQ was suggested.

Deployment

Attending users of Galaxy were using it on computing resources with pre-installed software or to drive statically deployed services. The ability to dynamically allocate computing resources and to deploy and run tools/services on demand was considered generally desirable to make better use, e.g of cloud resources. Also other workflow and scaleout mechanisms mentioned were largely based on statically deployed resources (e.g. UIMA AS, WebLicht, etc.).

Scalability

The issue of scaling was raised during the discussion. Scaling seems to depend on the data to transfer, nature of the services (remote, unsynchronized or synchronized services) and infrastructures. Some technos (HPC, Cloud computing) are mentioned to tackle these aspects.

User profiles and user friendliness

It was discussed that there is a range of users with different profiles interested in NLP/TDM workflows and applications. This again raises the issue of workflow expressiveness and complexity vs. user friendliness, briefly already touched upon while discussing workflow structure. The discussion seems to suggest providing simple workflows that domain end users (e.g., biologists, documentalists) can manage.

However, complex workflows that are not easy to manage for common users are also required for specific tasks. Thus it was discussed to separate the complex workflow constructs and let workflow experts manage them. The expected solution should rely on a simplified and standardized workflow model. It should consider facilities (e.g., question/answering, forms) common users are able to deal with and the requirements (e.g., experimental configurations) for experimenting and testing through a workflow. The issue of interactive workflows is also discussed. Humans need to be integrated in the loop (e.g., annotation task) and the human work itself needs be organized in a kind of workflow (e.g., annotation campaign).

Provenance

The data provenance that consists of tracking the steps by which data are consumed, handled and produced, was identified as an important issue to consider in scientific workflows. It is needed as a documentation to enable the reproducibility of experiments. It may be used to handle licensing aspects or to check errors. Provenance was considered not only important in automatically executed workflows, but also for manual annotation tasks, i.e. who annotated what, which annotation were updated, etc. In a mixed automatic/manual workflow, being able to identify manually created annotations vs. automatically created annotations is essential in order to compare automatic annotations against the manually created gold standard. A typical scenario would be machine-assisted annotation where the annotator is supported through automatically generated annotations.

However, specifically in the NLP/TDM community, the recording of provenance data from processing workflows is still underdeveloped. Thus, the question is how to deal with provenance in that context. A part of the response is to deal with data provenance at workflow level, that means capturing information before and after the execution of an individual service by the workflow engine. It seems to be a solution for services that are black boxes that do not reveal their internal steps. For services that internally deal with data provenance, it is discussed to capture and use the information.

Documentation

During the discussion it is noted that project funding generally does not sufficiently cover the issues of documentation and tutorials. These nevertheless remain significant for the partners and for the solutions to be used. Tutorials and documentation can also help users/developers save time to adapt and extends the solutions. To facilitate adoption and thus to reduce fragmentation, clear documentation and guidelines are needed.

Possible actions and plans

- Make it easier to write code in a high level language (Domain Specific Language) to coordinate tasks
- Learn from the bioinformatics community (cf. Galaxy)
- Deal on-demand dynamic deployment of services
- Somehow “compile” workflows such that adjacent components speaking the same format do not require format conversions/shims
- Deal with data provenance
- Deal with interactive workflows
- Take into account the importance of guidelines, tutorials and documentation for partners and future members
- Foster a user-base and a user-community, because users make the best tutorials

Pointers to related activities and work

This list is an indicative and non-exhaustive list of the activities and resources mentioned during the workshop.

Workflow descriptions

- The Language Application Grid provides LSD (Lapps Services DSL) for scripting workflows.
  - https://github.com/lappsgrid-incubator/org.anc.lapps.dsl
  - https://github.com/ksuderman/lsd-scripts/wiki
- Common Workflow Language https://github.com/common-workflow-language

Workflow Management Engines

- https://galaxyproject.org/
- https://taverna.incubator.apache.org/
- http://www.vistrails.org/
- https://kepler-project.org/

Human in the loop

- http://belief.scai.fraunhofer.de/BeliefDashboard/
- In case of questions feel free to ask marc.zimmermann@scai.fraunhofer.de
- GATE TeamWare https://gate.ac.uk/teamware/
- GATE crowdsourcing https://gate.ac.uk/wiki/crowdsourcing.htm

Group 3: Language and Semantics

Issues discussed

The group worked, in accordance with the setup of the workshop, with issues submitted by the workshop participants.

Linking knowledge to text

The low level requirements posited were generally accepted: content derived from knowledge resources (KR) needs to be added as text annotations. This enables the TDM modules to work in a uniform manner with the informational content the KRs provide. Further, it was generally agreed that KR elements should always have URIs, in order to be resolvable.

Overall, there was a preference for stand-off annotation because this allows maximum flexibility with respect to overlapping annotation spans where there are multiple annotation layers. Stand off annotation is considered to be standard practice with notable systems making use of it such as UIMA and GATE.

Error propagation

It was recognised in the discussion that the application domain is important in order to establish the scope of applicability for modules intended to be incorporated into a particular TDM workflow. Semantic interoperability does not guarantee performance/quality within a particular domain if any of the modules has been tuned to a different domain. Accumulated error rates from modules - when concatenated within a TDM workflow - will unerringly produce low quality output.

This highlights an important aspect of module exchangeability that goes beyond pure interoperability in the sense that the workflow modules successfully fit together in a workflow and produce output.

In order to build a workflow that is functional in terms of output quality it is necessary to evaluate and choose the best performing module; each component needs to be tested individually. For this purpose, it is worthwhile to adopt existing methodologies such as Open Advancement, a methodology originally formulated to enable progress in question answering, which aims to "combine formal metrics and rigorous module and end-to-end system evaluation with a collaborative research process that allows our field, as a research community, to achieve monotonically increasing performance levels across multiple instances of QA problems, while managing overall research and development cost effectively."[8]. This entails problem typology,

[8] http://domino.watson.ibm.com/library/CyberDig.nsf/papers/D12791EAA13BB952852575A1004A055C/$File/rc24789.pdf

and common evaluation metrics in order to accurately measure the contribution of each technology to end to end performance.

Mappings

With respect to annotation classes and the mapping of classes between KRs for interoperability purposes, it was considered an important issue how to make the content of these classes clear to other users. For this purpose, a meta-model language enhances communication, and a high level schema guarantees uniformity and perdurance.

Various schema mapping strategies were considered:

- manual
- the use of converters
- (semi-)automatic mapping as exemplified by Predicate Matrix (http://adimen.si.ehu.es/web/PredicateMatrix)

Differences in KR granularity cause information loss when they are aligned, especially where mapping from a more specific to a less specific KR is involved. Pairwise mappings constitute a lot of work. Using one singular hub model for mapping mediation is considered a better solution. A highly granular model as the reference model for expressing KR knowledge is ideal for mapping purposes.

The OLIA reference model (http://purl.org/olia) was suggested as a hub model for linguistic and terminological information. This model allows decomposition of KR content into fine-grained ontological descriptions, which then can be used for e.g. rule based inference with SPARQL.

We further briefly discussed EMF, which was considered very useful for instantiation, i.e. transformation into lower level processes.

Possible actions and plans

- The organisation of a workshop on automatic schema mapping for NLP, for instance under the header of ACL SIGANN : Linguistic Annotation Workshop (the LAW).

Pointers to related activities and work

This list is an indicative and non-exhaustive list of the activities and resources mentioned during the workshop.

- Open Advancement (methodology)
- OPenRefine for mapping (http://openrefine.org/)
- OAEI: ontology alignment and evaluation initiative (http://oaei.ontologymatching.org/2016/)
- OLIA reference model (http://purl.org/olia)

Group 4: Legal Aspects and Policy

Issues discussed

Participants showed interest in a broad range of topics related to copyright and sui generis database right. In particular, they asked and discussed:

- why copyright law prevents TDM in the first place;
- what are the shortcomings and pitfalls of current copyright (and SGDR) framework;
- what are the real inefficacies of current copyright framework;
- whether there is a best license for NLP researchers and
- whether a best practice is recommended.

In addition, further issues and questions emerged during the discussion, which were then addressed by the working group.

TDM exception

Overall, emerged the need to emphasise that, as a fundamental principle of current EU and international copyright law, making a copy of a protected work, even if temporary and partial, such as the one required for most if not all TDM activities, has to be authorised by the right holder unless the law provides for an exception.Failure to obtain authorisation where needed will constitute an act of infringement of copyright and / or *sui generis* database right.

In Europe at the EU level, although there is not yet a specific TDM exception, there could be one in the near future, according to the very recent approach of the EU Commission in this regard. One EU country (UK) has recently implemented a dedicated TDM exception, although limited to non commercial and research activities.

Considering the current obstacles that copyright and sui generis database right (SGDR) represent for TDM, the working group has considered in detail the problematic issues of defining a derivative work as well as actually defining a database, which clearly add further complexity to the current legal framework where TDM takes place. To such an extent, it emerged that a case by case analysis should always be made, with reference to the specific jurisdiction in which the researcher-miner is willing to conduct her activities. The need for a fully harmonised approach, at least at the EU level, emerged with particular strength.

The group reached the conclusion that the existing legal framework for exceptions does not provide a uniform and consistent answer that researchers are asking for.

Inefficacies that have so far emerged include the highly fragmented legal setting that features copyright exceptions and limitations within the European Union and in particular their different and inconsistent implementation given by Member States. It appears promising to develop a truly uniform legal framework for TDM. This could be feasibly reached through a broader and fair use like exception that would also not be limited to non-commercial purposes.

Licensing

Furthermore, participants demonstrated particular interest in discussing the topic of licensing, considering the example of NLP as a basis for a broader discussion on the best applicable license and recommended best practices. With regard to the latter, there emerged the need of being informed and cautious when choosing the license, although participants agreed that there is still much to be done to provide clear guidelines for researchers and thus avoiding incompatibility.

Finally, other related topics were addressed and few concrete examples were analysed, such as the instance in which the person who provides and uploads data does not have the right to do so; the case of international research bodies or consortia that include researchers from different countries and aim at conducting TDM activities; the issue of removing data or metadata upon request; the specific consequences of processing the data and making them available.

In all these instances, it was made clear that copyright exceptions and any specific analysis should have a case by case approach. At the same time, it was suggested to refer to the current draft of the OpenMinTeD WG3 case scenario 3[9], which also offers a prospective test for infringement. Indeed, this made clear the need of always considering not only the broader statutory (and often case law) framework, but also the terms of the contract where applicable.

Overall, the complexity that arises from the variety of approaches caused by a fragmented legal framework (in 28 different legal systems) could arguably be overcome if a broader open-ended and flexible type of exception would be designed, particularly at the EU level.

Possible actions and plans

- There emerged the necessity of further investigating the issue of applicable law/s in a transnational research environment, as one of the topic that participants find more difficult to understand.
- It also became clear that there is the need to address in more depth the issue of what are the outer limits of derivative works within the framework of copyright and sui generis database right.

Pointers to related activities and work

This list is an indicative and non-exhaustive list of the activities and resources mentioned during the workshop.

Following the outcomes of the discussion, the OpenMinTed WG3 case scenarios were accordingly addressed and reviewed.

[9] https://github.com/openminted/interoperability-spec