Call for Papers

10th Workshop on Challenges in the Management of Large Corpora (CMLC-10)

Workshop in conjunction with LREC 2022
June 20-25, Marseille (France)

Workshop date: morning session of **Monday June 20, 2022**
Submission deadline: April 8, 2022
Workshop homepage: http://corpora.ids-mannheim.de/cmlc-2022.html

Workshop description

The upcoming CMLC meeting continues the successful series of “Challenges in the management of large corpora” events, previously hosted at *LREC* (since 2012) and at *Corpus Linguistics* conferences (since 2015). As in the previous meetings, we wish to explore common areas of interest across a range of issues in language resource management, corpus linguistics, natural language processing, and data science.

Large textual datasets require careful design, collection, cleaning, encoding, annotation, storage, retrieval, and curation to be of use for a wide range of research questions and to users across a number of disciplines. A growing number of national and other very large corpora are being made available, many historical archives are being digitized, numerous publishing houses are opening their textual assets for text mining, and many billions of words can be quickly sourced from the web and online social media.

A number of key themes and questions emerge which are of interest to the contributing research communities: (a) What can be done to deal with IPR and data protection issues? (b) What sampling techniques can we apply? (c) What quality issues should we be aware of? (d) What infrastructures and frameworks are being developed for the efficient storage, annotation, analysis and retrieval of large datasets? (e) What affordances do visualization techniques offer for the exploratory analysis approaches of corpora? (f) What kinds of APIs or other means of access would make the corpus data as widely usable as possible without interfering with legal restrictions? (g) How to guarantee that corpus data remain available and sustainably usable?

Motivation and topics of interest

This year’s event will cover the whole range of the standard CMLC themes, with some new additions including some of LREC 2022's focus topics:

Interoperability and accessibility

- Improved accessibility of large corpora
- Interoperable APIs for query and analysis software
- Provision of multiple levels of access for different tasks

Machine/Deep Learning

- Data preparation for machine learning input
- Creation, curation, maintenance and dissemination of language models based on machine learning (e.g. word embeddings and deep learning networks)
- Legal issues concerning language model distribution

Linguistic content challenges

- Dealing with the variety of language: multilinguality, historical texts, noisy OCR texts, user-generated content, etc.
- Diversity and inclusion in language resources
- Integration of human computation (crowdsourcing) and automatic annotation
- Quality management of annotations
- Integrating different linguistic data types (text, audio, video, facsimiles, experimental data, neuroimaging data, …)

Technical challenges

- Storage and retrieval solutions for big textual data corpora: primary data (potentially including facsimiles, etc.), metadata, and annotation data
- Scalable and efficient NLP tooling for annotating and analyzing large datasets: distributed and GPGPU computing; using big data analysis frameworks for language processing
- Dealing with streaming data (e.g. social media) and rapidly changing corpora
- Environmental impact of big language data computing
- Engineering and management of research software

Exploitation challenges

- Legal and privacy issues
- Query languages, data models, and standardization
- Licensing models of open and closed data, coping with intellectual property restrictions
- Innovative approaches for aggregation and visualization of text analytics

In the tradition of CMLC, we invite reports on national corpus initiatives whose submitters should be prepared to present a poster.

Current information is available at the workshop homepage:
http://corpora.ids-mannheim.de/cmlc-2022.html

Abstract submission

We invite extended abstracts for 15 to 20 minute presentations (4 pages maximum). All abstracts have to be submitted via the START Conference Manager (link tba. on the homepage http://corpora.ids-mannheim.de/cmlc-2022.html) .

Identify, Describe and Share your LRs!

Describing your LRs in the LRE Map is now a normal practice in the submission procedure of LREC (introduced in 2010 and adopted by other conferences). To continue the efforts initiated at LREC 2014 about "Sharing LRs" (data, tools, web-services, etc.), authors will have the possibility, when submitting a paper, to upload LRs in a special LREC repository. This effort of sharing LRs, linked to the LRE Map for their description, may become a new "regular" feature for conferences in our field, thus contributing to creating a common repository where everyone can deposit and share data.

As scientific work requires accurate citations of referenced work so as to allow the community to understand the whole context and also replicate the experiments conducted by other researchers, LREC 2022 endorses the need to uniquely Identify LRs through the use of the International Standard Language Resource Number (ISLRN, www.islrn.org), a Persistent Unique Identifier to be assigned to each Language Resource. The assignment of ISLRNs to LRs cited in LREC papers will be offered at submission time.

Important dates

- Deadline for abstract submission: **April 8, 2022**.
- Notification of acceptance: **May 3, 2022**.
- Deadline for the camera-ready papers: **May 23, 2022**.
- Meeting: **morning** session of **Monday June 20, 2022**

Organizing Committee

Piotr Bański (IDS Mannheim)
Adrien Barbaresi (BBAW Berlin)
Simon Clematide (University of Zurich, CH)
Marc Kupietz (IDS Mannheim)
Harald Lüngen (IDS Mannheim)

Programme committee:

tba. (see http://corpora.ids-mannheim.de/cmlc-2022.html for updates)