Contribution ID: **29** Type: **Regular Presentation**

Metacurate-ML: Metadata Extraction from CAI

Tuesday, 3 December 2024 10:45 (20 minutes)

Extending the results of our work on pre-trained language models with recent developments in text-layout models and zero-shot techniques. Since relying solely on textual information makes it difficult to accurately classify and extract metadata, a combination of textual content and visual logic that incorporates vision transformers with optimisation techniques will be explored.
This will allow us to extract the specific items with questionnaires such as question texts, responses and routing to create a rich source of metadata which provenances'data collection methodology to the resultant data which can be transformed into DDI-Lifecycle. We will investigate the feasibility of document understanding multimodal models that employ masked language techniques and present the resulting challenges.

Primary authors: DE, Suparna (University of Surrey); JOHNSON, Jon (CLOSER, UCL)

Co-authors: Mr WANG, Zeqiang (University of Surrey); Dr PRAVIN, Chandresh (University of Surrey); LUNGLEY, Deirdre; Mr BRADSHAW, Paul (Scottish Centre for Social Research (ScotCen))

Presenter: DE, Suparna (University of Surrey)

Session Classification: Metacurate-ML