Machine learning for large-scale legacy reference extraction at BMJ

Background
The ability to extract, parse and link citations to their source documents (‘reference mining’) is a fundamental task in creating citation networks and building digital libraries. Over recent years, a number of tools have been developed to address this; however, they are often limited by input format requirements, availability, infrastructure requirements and runtime performance[1]. Recent developments have focused on reference mining PDFs from the arts and humanities literature[2] and German social science papers[3]. Yet, there is still a need for fast, accurate tool, available via API, that can extract and parse references from a wide range of documents and formats across the research literature, and that allows vendor-specific models to be trained and deployed.

Objective
At the BMJ, we have an archive of hundreds of thousands of articles that exist only in PDF format. We wish to mine these PDFs for references, and automatically structure them in CrossRef XML format[4], to make them available to the community as part of the Open Citations Initiative[5].

Methods
Scholarcy[6] provides a set of APIs that use machine learning and automated linguistic analysis to extract key facts, references and data from research papers. Scholarcy has built models to identify, extract and parse references in any style or format with at a rate of up to 1000 PDFs per minute and an accuracy of up to 95%.

Challenges
- Many PDFs are OCR scans, so the text accuracy will be variable
- Older articles do not have a defined References section
- Many PDFs contain multiple articles

References