

Enhancing the real impact of scholarly publications through text and data mining

Nancy Pontica (1), Natalia Manola (2), Androniki Pavlidou (2), Petr Knoth (3), Martine Oudenhoven (4), Mappet Walker (5), Giulia Dore (6), and Sophia Ananiadou (7)

(1) CORE, KMi, The Open University, UK, Orcid 0000-0002-2091-0402, (2) Athens Research Institute, ARC, (3) CORE, KMi, The Open University, (4) LIBER, (5) FRONTIERS, (6) University of Glasgow, (7) University of Manchester

OpenMinTeD is an European Union funded project that aims at creating an open, service oriented e-Infrastructure that facilitates the use of text and data mining technologies focused on scientific and scholarly publications. Researchers can collaboratively create, discover, share and re-use knowledge from a wide range of text based scientific related resources in a seamless way. The project's main goals are to collect requirements from various subject based scientific communities, to build an interoperability framework and to develop a service oriented platform.

In FORCE2017, we propose a two hours workshop that is primarily aimed mainly at publishers but also all text and data mining stakeholders. We plan to present the work we are doing with regards to interoperability on machine accessibility of open access publications, the compatibility matrix, and the creation of trusted hubs of information.

Three areas will be explored in more depth:

- The OpenMinTeD project has investigated how accessible the open access content from publisher platforms is, and has created a publisher connector that helps text and data miners to access scientific publications. In this session we will share our experiences with regards to this work and propose best interoperability practices. We will also show how the data are utilised by the OpenMinTeD platform to implement applications for the scholarly communication domain. Specifically we will present blueprints of an application that tracks research performance (citations, downloads, altmetrics, semantometrics) of papers, individuals, organisations, etc. and delivers this in the form of a dashboard. The application needs access to full texts and metadata to be able to slice and dice the data and extract full text features that can indicate performance (for semantometrics). The application then displays graphs and allows to perform benchmarking, comparison and trend analysis. We will also present how the data can be used to create a scholarly oriented recommender application. This application offers content-based recommendations for scholarly materials, such as related research papers. The recommender relies on metadata, full text features and other features that can be mined from the full text articles or can be obtained from external sources.

- OpenMinTeD builds upon existing tools and text mining platforms, rendering them to be discoverable through appropriate registries, and interoperable through a standards-based interoperability layer. The interoperability layer is largely based on a combination of existing standards. OpenMinTeD supports awareness of the benefits of this approach and will work on training for text mining users and developers. The merits of the approach are demonstrated through a number of use cases identified by scholars and experts from different scientific areas, ranging from life sciences (bioinformatics, biochemistry, etc.) to food and agriculture and social sciences and humanities related literature.

- Another important ambition of OpenMinTeD is to promote license standardisation and interoperability against the very fragmented and uncertain legal framework that features the field of text and data mining. Only few licensing terms and conditions are standardised, while most of them are unclear and often incompatible. OpenMinTeD urges boosting a rather interoperable and efficient use of the resources to mine.

With the support of its legal team, the project provides guidance for users who wish to undertake TDM activities by developing a tool that will help to overcome the ambiguity of the current setting. To pursue this crucial goal, OpenMinTeD has developed specific compatibility tools, including a matrix that will help users and miners to choose the most appropriate licence for their resources.

The matrix is articulated in three main layers: (a) content (b) software and (c) services. It assesses the possibility

that the resources licensed under the licences therein considered may be combined. Their combination can feasibly result in a derivative work. The matrix also assesses whether it is likely that there is a conflict between the licensing terms examined. In doing this, it considers some specific clauses, such as a non-commercial, no-derivatives or share-alike clause. The graphic layout of the matrix pictures precisely this type of compatibility: it signals a non conditional “green” yes when the licensing terms considered allow derivative works and no further restrictions apply, or when no conflict whatsoever applies looking at their terms; a “yellow” yes specifies compatibility under some conditions, as for derivative works that are allowed some restrictions apply or some specific terms may be conflicting; a warning “red” no will be given when no compatibility whatsoever is foreseen, that is when derivative works are not allowed given the terms of the licences considered, or when the terms of the licences conflict.

By opening up scholarly publications for text and data mining we can increase the impact of the scholarly outputs and as a consequence advance science. By demonstrating our work in the OpenMinTeD project we wish to bring one of the key stakeholders of text and data mining, the publishers, closer to our work. We hope that in this way we can increase our collaboration with publishers.