EUROPEAN COMMISSION
DIRECTORATE-GENERAL HUMAN RESOURCES AND SECURITY
Directorate HR.AMC –Account management Centre
HR.DDG. AMC 8

2017-IPR-I-000-8664 **Multilingual Named Entity Recognition and Classification**

Position for: Trainee	**Short description of activity:** As the science and knowledge service of the Commission, the mission of DG Joint Research Centre is to support EU policies with independent evidence throughout the whole policy cycle. The JRC is located in 5 Member States (Belgium, Germany, Italy, the Netherlands and Spain). Further information is available at: http://www.jrc.ec.europa.eu **Short description of activity:** The JRC's *Europe Media Monitor* (EMM) team carries out research and development in the field of highly multilingual text mining (Language Technology; Computational Linguistics) for the purposes of media monitoring. EMM gathers an average of 300,000 online news articles per day in over 70 languages and analyses them to help its large international user community understand and use this enormous amount of media information. The *Europe Media Monitor* EMM is publicly accessible and widely used. The EMM team has produced over 200 international peer-reviewed publications. The team has also produced and distributes a number of highly multilingual Language Technology resources. The *Text and Data Mining Unit* (I3) of the European Commission's *Joint Research Centre* (JRC) in Ispra, Italy, is looking for a trainee to support the JRC's *Europe Media Monitor* (EMM) team in its effort to improve its Named Entity Recognition and Classification (NERC) tools, especially for multi-word entities such as organisation and event names. EMM gathers and analyses reports from traditional and social media in dozens of languages by clustering related news items; categorising them; extracting information such as entities (persons, organisations, locations), events (who did what to whom, where and when), quotations by and about people; identifying sentiment; as well as linking related news clusters over time and across languages. Methods used are mostly hybrid: machine learning tools are used to gather evidence, learn vocabulary and rules,

	but the results are usually controlled and optimised through human intervention. EMM is used by European Institutions, by national authorities in EU Member States, by international organisations and by the public. The public EMM applications NewsBrief, NewsExplorer and MedISys can be accessed freely by the general public. EMM is part of the JRC's Competence Centre on Text Mining and Analysis. As of now, the EMM team has accumulated several very large independent sets of multi-word entities and their monolingual and multilingual name variants. Some of the entities are classified according to an entity type hierarchy, while others are not. **The successful trainee will help to improve the current tools to recognise multi-word entities, classify entities, merge the various lists of entities and their variants into one single repository, and integrate the NERC tools with the EMM processing chain.** The trainee is also expected to contribute to writing a scientific publication on the work carried out. **Qualifications:** **Essential:** • a degree (or an almost completed degree) in computational linguistics, computer science or related areas;(Applications from students currently preparing a thesis for a University degree are eligible. The thesis should match with the subject of the project call). • Java programming skills; • good working knowledge of English. (B2 level) **Advantage:** • knowledge of further foreign languages; • proven advanced programming skills, especially in Java; • good knowledge of Language Technology-related tools and methods; • proven ability to work independently and as part of a team. In your application, please provide clear information on your skill set, by elaborating on the above-mentioned list of requirements and by listing your level of languages and your computer / programming skills.

	For general eligibility requirements, please read the rules governing the traineeship scheme of the JRC: https://ec.europa.eu/jrc/en/working-with-us/jobs/temporary-positions/jrc-trainees
Institute/Directorate Unit	Directorate *Competences* I03 – *Text and Data Mining Unit*
Indicative duration **Preferred starting date**	5 months As soon as possible
JRC Site **Country**	Ispra Italy
JRC contact details	**For any technical problems with your application, please contact:** JRC-ESRA@ec.europa.eu