Title:

Research engineer position: Web architect – Semantic web technologies – Text and data mining

Information:

Employer:	University of Montpellier
Context:	ANR PractiKPharma project and VisaTM project
When:	November 2018
Duration	12-14 months minimum (extensible to 2 additional years on related projects)
Where:	LIRMM, Montpellier, France
Collaboration:	NCBO (Stanford, USA), LORIA & INSIT (Nancy, France), HEGP (Paris, France), INRA (Jouy-en-Josas, France), EMA (Alès, France)

Keywords:

Web development, ontologies & terminologies, semantic web, ontology repository, knowledge engineering, semantic interoperability, annotation, natural language processing, text and data mining.

Technologies:

Web development, Java/JEE, Ruby/Rails, RESTful web services, XML/JSON, Web technologies (HTML5, Bootstrap, JavaScript), Semantic web technologies (OWL, RDF, SPARQL, triplestore, Linked data), NCBO technology (BioPortal), OpenMinTeD technology.

Abstract:

A key aspect in addressing semantic interoperability is the use of terminologies, vocabularies and ontologies as a common denominator to structure data and make them interoperable. In partnership with Stanford University, LIRMM designs, develops and maintains two vocabulary and ontology repositories: (i) the SIFR BioPortal (http://bioportal.lirmm.fr) which targets the French biomedical community and offers an ontology-based annotation workflow to semantically index text biomedical and clinical data; (ii) AgroPortal (http://agroportal.lirmm.fr) a reference repository for semantic resources in agronomy, agriculture, food, plant sciences and biodiversity. These platforms allow us to tackle scientific problems in natural language processing, semantic annotation, ontology engineering, while being driven by concrete use cases with impacts in biomedicine and agronomy.

In the context of the ANR PractiKPharma and VisaTM projects, we are seeking a motivated, curious and interested research engineer and web developer to take hands on the platforms and prototypes developed. Your role will be both to support the current platforms and investigate technical decisions to enable the development of new features in relation to text and data mining. Within PractiKPharma, your role will consist in enhancing the SIFR Annotator to facilitate its use to annotate clinical data (in collaboration with HEGP and LORIA). Within the VisaTM project, your role will be to enable interoperation of both AgroProtal and BioPortal with the text and data mining infrastructure developed in the OpenMinTeD project (in collaboration with INRA and INIST). You will work with a small team (4 persons) at LIRMM in both a national and international context. Further extension of the proposed 12-month-contract are possible.

Context:

Pharmacogenomics (PGx) studies how individual gene variations cause variability in drug responses and constitutes a basis for implementing personalized medicine i.e., a medicine tailored to each patient by considering her/his genomic context. The goal of the PractiKPharma project (http://practikpharma.loria.fr) is to validate or moderate Pharmacogenomics state-of-the-art knowledge on the basis of practice-based evidences, i.e., knowledge extracted from Electronic Health Records. During this project, we extract state-of-the-art knowledge from (English) structured and unstructured descriptions in reference databases (e.g., PharmGKB) and literature (e.g., PubMed) as well as extract observational knowledge from (French) EHRs. Part of this multilingual knowledge extraction process is based on semantic annotation of plain-text data. We use and enhance tools developed in the context of

the NCBO (www.bioontology.org) [1, 2] and SIFR projects (www.lirmm.fr/sifr). Especially the SIFR Annotator, a web service allowing scientists to utilize available biomedical ontologies for annotating their French biomedical or clinical text automatically [3]. The SIFR Annotator service processes raw textual descriptions, tags them with relevant biomedical ontology concepts and returns the annotations to the users in several formats such as JSON-LD, RDF or BRAT. We have started to develop specific feature to process clinical text (e.g., detect negation, temporality and experiencer) [4], however they need to be improved and complemented by a general disambiguation module that will allow to increase the precision of the annotations. Another technical challenge is to enable the use of the system in data-sensitive environment (e.g., hospitals) for this a Docker packaging is available. The use cases will be discussed with our collaborators at HEGP hospital and in Nancy (LORIA and CHU).

The semantic Web relies on the construction of standard vocabularies and ontologies to formally capture the knowledge of a domain into semantic resources computers use to index, search or reason on the data [5]. In recent years, we have seen an explosion in the number of knowledge resources (thesaurus, terminologies, vocabularies and ontologies) being developed in life sciences, agronomy and biodiversity. However, those resources are spread out, in different formats, of different size, with different structures and from overlapping domains. Therefore, there is need for common reference platforms to receive and host them, align them, and enabling their use in external applications. More generally, ontologies and vocabularies are a key element to make data FAIR (Findable, Accessible, Interoperable and Reusable). For example, ontologies and terminologies are highly valuable in text and data mining workflows. To facilitate the development of such workflows, the EU OpenMinTeD project (http://openminted.eu) has developed a platform in which different text and data mining components can be used to process text corpora, possibly using ontologies and terminologies. Within the VisaTM project we develop the interoperation components for OpenMinTeD and AgroPortal (and by extension to any other ontology repository based on the NCBO technology) to rely on one another: OpenMinTed to consume AgroPortal's semantic resources [6] and AgroPortal to consume simplified/customized OpenMinTeD's workflows. More specifically, we are interested in using (and maybe contributing to) the OpenMinTeD technology to enable easy external access, via web service, to advanced text and data mining workflows.

We are reusing the technologies developed by the National Center for Biomedical Ontologies at Stanford University: the BioPortal web application (http://bioportal.bioontology.org) made available via its virtual appliance (http://www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance). Please refer to our GitHub repository for more detail:

- https://github.com/ncbo
- https://github.com/sifrproject
- https://github.com/agroportal

The developer will have:

- to manage, administrate and modify the SIFR BioPortal and AgroPortal ontology repositories.
- to modify the SIFR Annotator with respect to the needs of the PractiKPharma project.
- to modify the AgroPortal (and generic technology) with respect to VisaTM's roadmap for interoperation with OpenMinTeD platform.

Expected profile:

We are seeking a motivated, curious and interested research engineer candidate with a computer science or bioinformatics training (master/engineer or PhD). Besides an important motivation for the technical challenges and excellent software development expertise, we are also looking for someone with some interest for the research topics presented. The candidate will demonstrate aptitudes or matches with some of the following aspects:

- Web developer with good knowledge of JEE technologies, Ruby/Ruby On rails, Bootsrap.
- Experience with semantic web technologies, especially JSON/RDF/SPARQL.
- Experience with text and data mining software (knowledge extraction, use of ontologies, etc.)
- Between 1 and 5 years of experience. Including experience in private companies.
- Excellent technical skills to push prototypes into production environment.

- Excellent remote working capabilities (emails, trackers, collaborative tools, etc.)
- Perfect English oral and writing skills.
- Few knowledge with French language with objective to learn the language during the contract.
- Excellent writing skills as reports, documentation, and technical notes will always be necessary.
- International trips accepted (collaboration with Stanford) and possibility to get a visa for the USA.
- Autonomy and initiative, take on technical decisions within the project and justify choices.
- Friendly person to join a small research team in Montpellier and to listen to their needs.
- Eventually interested in research valorization of the outcomes (everything will be published).
- Open source developer.

Application:

For more information about this position, please contact Clement Jonquet (jonquet@lirmm.fr). To apply, please send an email including links to (NO ATTACHED DOCUMENTS) the following:

- a curriculum vitae describing your training and experience;
- a motivation letter describing YOUR interest for the position and the matches with the expected profile;
- reference to already developed web application or project (URL, GitHub, technical documentation) clarifying your role;
- copies of diplomas and possibly other relevant certificates;
- names and contact details of referees.

Contract:

- The engineer will be hired under the “Ingénieur de recherche” or “ingénieur d’étude” status depending on qualification
- Social security and benefits are included.
- Salary will be between 1600 and 2000€ net per month depending on qualification and experience.

References:

1. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., Musen, M.A.: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 37, 170–173 (2009).
2. Whetzel, P.L., Team, N.: NCBO Technology: Powering semantically aware applications. Biomed. Semant. 4S1, 49 (2013).
3. Jonquet, C., Annane, A., Bouarech, K., Emonet, V., Melzi, S.: SIFR BioPortal : Un portail ouvert et générique d’ontologies et de terminologies biomédicales françaises au service de l’annotation sémantique. In: 16th Journées Francophones d’Informatique Médicale, JFIM’16. p. 16. , Genève, Suisse (2016).
4. Tchechmedjiev, A., Abdaoui, A., Emonet, V., Melzi, S., Jonnagaddala, J., Jonquet, C.: Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator+. Bioinformatics. 34, (2018).
5. Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzalé Yeumo, E., Emonet, V., Graybeal, J., Laporte, M.-A., Musen, M.A., Pesce, V., Larmande, P.: AgroPortal: A vocabulary and ontology repository for agronomy. Comput. Electron. Agric. 144, (2018).
6. Kettani, F., Schneider, S., Aubin, S., Bossy, R., François, C., Jonquet, C., Tchechmedjiev, A., Toulet, A., Nédellec, C.: Projet VisaTM : l’interconnexion OpenMinTeD – AgroPortal – ISTEX, un exemple de service de Text et Data Mining pour les scientifiques français. In: Rawnez, S. (ed.) 29èmes Journées Francophones d’Ingénierie des Connaissances, IC’18, Poster Session. pp. 247–249. , Nancy, France (2018).