Using Semantics and NLP in the SMART Protocols Repository

Olga Giraldo[1,*] Alexander Garcia[1, 2] and Oscar Corcho[1]
[1] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[2]Linkingdata I/O LLC, Fort Collins, Colorado, USA

ABSTRACT

In this poster we present the semantic and NLP layers in the development of our repository for experimental protocols. We have studied existing repositories for experimental protocols as well the experimental protocols themselves. We have identified end-user features across existing repositories; we have also structured the semantics for these documents, defined by an ontology and a Minimal Information model for experimental protocols. In addition, we have built an NLP layer that makes extensive use of semantics. Our integrative approach focuses on facilitating search, retrieval and socialization of experimental protocols. We also focus on facilitating the generation of documents that are born semantics.

1 INTRODUCTION

Experimental protocols are fundamental information structures that support the description of the processes by means of which results are generated in experimental research. Well-structured and accurately described protocols (procesable by humans and machines) should facilitate experimental reproducibility. In this poster we present the semantic and NLP infrastructure that we are putting together for machine procesable protocols; we emphasize in the integration of key components of this infrastructure during the implementation of a repository for experimental protocols. Our components include: **i) The SMART Protocols (SP) Ontology**: this ontology results from the analysis of over 200 experimental protocols in various domains –molecular biology, cell and developmental biology and others. Domain experts also participated in the development of the SP ontology (Giraldo, García, & Corcho, 2014). Using the SP ontology allows us to annotate and generate Linked Open Data (LOD) for existing and *de novo* protocols –protocols to be born semantics. **ii) The Sample Instrument Reagent Objective (SIRO) model**. This is a twofold model; on the one hand it defines an extended layer of metadata for this kind of documents. On the other hand, SIRO is a Minimal Information (MI) model conceived in the same realm as PICO (Booth & Brice, 2004), supporting search, retrieval and classification purposes. SIRO is based on an exhaustive study of over 200 protocols in biochemistry, molecular biology, cell and developmental biology, health care as well as interviews with end users. SIRO includes information elements that were identified as central for describing, searching and sharing protocols. Furthermore, as SIRO is rooted in the content of the document, it defines a score of completeness and reproducibility for experimental protocols. iii) **The NLP engine**. The semantics defined by the SP ontology, SIRO, and several domain ontologies is used by our NLP engine, GATE[1]; thus, facilitating search, retrieval and socialization (SeReSo) over experimental protocols. We have generated rules based on the content of protocols; these rules allow us to identify meaningful parts of speech (PoS).

We have reviewed proposed standards for representing experimental protocols, investigations, experiments, scientific documents, rhetorical structures and annotations. In addition, we have analyzed existing repositories for protocols. Interestingly we have found that there are numerous similarities across these repositories –e.g. business model, end-user features, document management; by the same token, the lack of semantics for experimental protocols and the lack of specific features for this particular type of documents may be seen as a common deficiency in these repositories. This document is organized as follows; in section 2 the semantic components are presented; in this section we also inform on the use of semantics by our NLP engine. Some issues and final remarks are presented in section 3.

2 SEMANTICS PLUS NLP

The combination of semantics and NLP makes it possible to deliver a tool that facilitates the generation of experimental protocols that are to be born semantics –fully annotated, linked to the web of data, with fully identified PoS, procesable by machines as well as by humans. In the same vein, a similar process for existing experimental protocols in formats such as PDF is also supported. Furthermore, searching for queries such as: *"What* ***bacteria*** *have been used in protocols for* ***persister cells isolation****?", "What* ***imaging analysis software*** *is used for quantitative analysis of locomotor movements, buccal pumping and cardiac activity on* ***X. tropicalis****?", "How to prepare the stock solutions of the* ***H2DCF*** *and* ***DHE dyes****?",* is also possible.

We are using the SP ontology; SP aims to formalize the description of experimental protocols, which we understand as domain-specific workflows embedded within documents. SP delivers a structured workflow, document and domain knowledge representation written in OWL DL. For the representation of document aspects we are extending the

[*] To whom correspondence should be addressed: ogiraldo@fi.upm.es

[1] http://gate.ac.uk/

Information Artifact Ontology (IAO).[2] The representation of executable aspects of a protocol is captured with concepts from P-Plan Ontology (P-Plan) (Garijo & Gil, 2012); we are also reusing EXPO (Larisa N. Soldatova & D., 2006), EXACT (L. N. Soldatova, Aubrey, King, & Clare, 2008) and OBI (Courtot et al., 2008). For domain knowledge, we rely on existing biomedical ontologies. Our ontology-based representation for experimental protocols is composed of two modules, namely SP-document[3] and SP-workflow.[4] In this way, we represent the workflow, document and domain knowledge implicit in experimental protocols. By combining both modules we are delivering a born-semantics self- describing document.

We are also working with the SIRO model; our model breaks down the protocol in key elements that are common to "*all*" laboratory protocols: i) Sample/Specimen (S), ii) Instruments (I), iii) Reagents (R) and iv) Objective (O). SIRO is motivated by minimal information models as well as by the **P**atient/Population/Problem **I**ntervention/Prognostic/Factor/Exposure **C**omparison **O**utcome (PICO) model. For the **sample** it is considered the strain, line or genotype, developmental stage, organism part, growth conditions, pre-treatment of the sample and, volume/mass of sample. For the **instruments** it is considered the commercial name, manufacturer and identification number. For the **reagents** it is considered the commercial name, manufacturer and identification number; it is also important to know the storage conditions for the reagents in the protocol. Identifying the **objective** or goal of the protocol, helps readers to make a decision about the suitability of the protocol for their experimental problem. The four elements are also automatically annotated with existing ontologies and exposed as LOD.

The NLP engine, GATE, uses the semantics defined by the SP ontology and SIRO. We have classified our corpus of protocols according to purpose/objective (e.g. extraction of nucleic acids, DNA amplification and visualization of nucleic acids) and then we transformed them to text. For each protocol, metadata available, reagents, instruments samples, actions and instructions were manually identified. We worked with full sentences to characterize PoS, relations, actions (verbs) and full instructions. Gazetteers and rules were thus generated. The results from our NLP workflow are very granular; for instance, we are able to identify DNA purification reagents, digest reaction reagents, cell disruption instruments, etc. Text like "*plant species*" is identified as sample, so are organisms and parts of organisms. The sentences and PoS where the vocabulary is located are also identified and characterized. For instance, PoS such as "***leaf tissue*** *finely* ***ground*** *using a* ***mortar and pestle****, then aliquoted (1 g) for each extraction*" are identified, characterized and annotated; in this example **sample**, **action**, **cell disruption instrument** are identified and characterized. We are using ANNIE (A Nearly-New Information Extraction) as our information extraction system and JAPE for coding rules.

3 FINAL REMARKS

We have presented the integration of three modules in the development of a repository for experimental protocols. Unlike existing repositories, the SP repository focuses on facilitating the production of semantic protocols, intelligent search and retrieval and social activity over experimental protocols. We have extensively studied existing experimental protocols; key functionalities from these will also been included in our repository. We have also presented the SP ontology, the SIRO model for MI and the use of GATE in our architecture. Our workflow addresses scenarios with PDFs and *de novo* protocols – those born semantics based on the SP ontology. For *de novo* documents we are using the ontology as a template; the resulting instantiated RDF is annotated and the conventional document metadata is extracted. For PDFs we are tuning the NLP workflow for extracting SIRO automatically. Extracting the ***O****bjective* has proven to be a challenging task. Actions e.g. *grind the sample*, usually have well defined grammatical structures; but, the ***O****bjective* of the experimental protocol is usually hidden in a complex prose. We are constantly improving the rules; new documents pertaining to other subdomains in biomedical sciences are added to the corpus; then, the rules are tested. Results are manually evaluated and the rules and gazetteers are consequently enriched.

REFERENCES

Booth, A., & Brice, A. (2004). Formulating answerable questions. In A. B. Booth, A (Eds) (Ed.), *Evidence Based Practice for Information Professionals: A Handbook* (pp. 61-70): London: Facet Publishing.

Courtot, Mélanie., Bug, William., Gibson, Frank., Lister, Allyson L., Malone, James., Schober, Daniel., . . . Ruttenberg, Alan. (2008). *The OWL of Biomedical Investigations* Paper presented at the OWLED workshop in the International Semantic Web Conference (ISWC), Karlsruhe, Germany.

Garijo, Daniel., & Gil, Yolanda. (2012). *Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data.* Paper presented at the The 2nd International Workshop on Linked Science 2012, Boston.

Giraldo, Olga., García, Alexander., & Corcho, Oscar. (2014). *SMART Protocols: SeMAntic RepresenTation for Experimental Protocols*. Paper presented at the 4th Workshop on Linked Science 2014 - Making Sense Out of Data (LISC2014), Riva del Garda, Trentino, Italy. http://ceur-ws.org/Vol-1282/lisc2014_submission_2.pdf

Soldatova, L. N., Aubrey, W., King, R. D., & Clare, A. (2008). The EXACT description of biomedical protocols. *Bioinformatics, 24*(13), i295-303. doi: btn156 [pii]10.1093/bioinformatics/btn156

Soldatova, Larisa N., & D., King Roos. (2006). An ontology of scientific experiments. *journal of the royal society interface, 3*(11), 795–803. doi: 10.1098/rsif.2006.0134

[2] https://code.google.com/p/information-artifact-ontology/

[3] http://vocab.linkeddata.es/SMARTProtocols/sp-documentV2.0.htm

[4] http://vocab.linkeddata.es/SMARTProtocols/sp-workflowV2.0.htm