Using Semantics and NLP in Experimental Protocols

Olga Giraldo[1], Alexander Garcia[1], Jose Figueredo[1,2], and Oscar Corcho[1]

Ontology Engineering Group, Universidad Politécnica de Madrid, Spain,
ogiraldo@fi.upm.es, alexgarciac@gmail.com, ocorcho@fi.upm.es
Universidad Simón Bolívar, Venezuela
jfigueredofortes@gmail.com

Abstract. In this paper we present ***SMART Protocols***, a *semantic* and *NLP*-based infrastructure for processing and enacting experimental protocols. Our contribution is twofold; on the one hand, SMART Protocols delivers a semantic layer that represents the knowledge encoded in experimental protocols. On the other hand, it builds the groundwork for making use of such semantics within an NLP framework. We emphasize the semantic and NLP components, namely the *SMART Protocols (SP) Ontology*, the ***S**ample **I**nstrument **R**eagent **O**bjective* (SIRO) model and the text mining integrative architecture *GATE*. The SIRO model defines an extended layer of metadata for experimental protocols; SIRO is also a Minimal Information (MI) model conceived in the same realm as the **P**atient **I**ntervention **C**omparison **O**utcome (PICO) model that supports search, retrieval and classification purposes. By combining comprehensive vocabularies with NLP rules and gazetteers, we identify meaningful parts of speech in experimental protocols. Moreover, in cases for which SIRO is not available, our NLP automatically extracts it; also, searching for queries such as: *What bacteria have been used in protocols for persister cells isolation* is possible.

Keywords: semantic web, graph theory, biomedical ontologies, natural language processing, knowledge representation

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 [6]. Experimental protocols describe how the data was produced, the steps undertaken and conditions under which these steps were carried out. Biomedical experiments often rely on sophisticated laboratory protocols, comprising hundreds of individual steps; for instance, the protocol for chromatin immunoprecipitation on a microarray (Chip-chip) has 90 steps and uses over 30 reagents and 10 different devices [1]. Protocols are written in natural language; they are often presented in a "*recipe*" style; they are meant to make it possible for researchers to reproduce experiments.

In this paper, we present the semantic and Natural Language Processing (NLP) components for *SMART Protocols (SP)*; we want to facilitate the semantic representation and natural language processing for these documents. Our NLP layer makes use of various ontologies as well as of the ***S**ample **I**nstrument **R**eagent **O**bjective* (SIRO) model for minimal information (MI) that we have defined. Our work is based on an exhaustive analysis of over 400 published experimental protocols[1] (molecular biology, cell and developmental biology, biochemistry) and guidelines for authors from repositories and journals, e.g. Nature Protocols[2], Plant Methods (Methodology)[3], Cold Spring Harbor Protocols[4]. Moreover, the SP ontology and SIRO are built upon experiences such as those under the BioSharing umbrella, e.g. the MIBBI project [19], OBO

[1] https://github.com/oxgiraldo/SMART-Protocols/tree/master/corpus_of_protocols
[2] http://www.nature.com/nprot/info/gta.html
[3] http://plantmethods.biomedcentral.com/submission-guidelines/methodology
[4] http://cshlpress.com/cshprotocols/

foundry [15]. We have also considered the ISA-TAB because it is a general framework with which to collect and communicate complex metadata (i.e. sample characteristics, technologies used, type of measurements made) from *omics-based* experiments [12].

We have carefully considered and reused several ontologies, for instance: *i)* The Ontology for Biomedical Investigations (OBI) aims to provide a representation of biomedical investigations. OBI builds upon BFO and structures the ontology by using "*occurrences*" (processes) and "*continuants*" (materials, instruments, qualities, roles, functions) relevant to the biomedical domain [13]. *ii)* The Information Artifact Ontology (IAO)[5] is an ontology that represents information entities such as documents, file formats and specifications. *iii)* The ontology of experiments (EXPO) aims to formalize domain-independent knowledge about the planning, execution and analysis of scientific experiments. This ontology includes the class "*Experimental Protocol*" and defines some of its properties: "*has_applicability*", "*has_goal*", "*has_plan*" [17]. *iv)* The LABORS ontology (LABoratory Ontology for Robot Scientists) addresses the problem of representing the information required by robots to carry out experiments; LABORS is an extension of EXPO and defines concepts such as "*investigation*", "*study*", "*test*", "*trial*" and "*replicate*" [9, 10]. *v)* The ontology of experimental actions (EXACT) provides a terminology for the description of protocols in the biomedical domain. The core of this vocabulary is a hierarchical classification of verbs currently used in experimental protocols. These verbs are divided into three groups according to their goal (separation, transformation and combination) [16].

Unlike other approaches, the SP ontology is an application ontology designed to support NLP over experimental protocols, publish protocols as Linked Open Data (LOD) and annotate and classify these documents according to particularities in their workflows. The SP-document delivers an structured vocabulary for representing a specific type of document, the protocol. We extend the IAO to provide a structured vocabulary of concepts to represent the information that is necessary and sufficient for describing and experimental protocol as a document. In addition, the SP ontology also considers the protocol as an executable element to be carried out and maintained by humans it may also be transformed to workflow languages used by enactors such as robots or machines per se, the SP ontology is not a workflow language but a model.

For the representation of instructions in the SP-workflow module, we expand the class "*experiment action*" from EXACT and reuse classes from OBI, the BioAssay Ontology (BAO) [18], The Experimental Factor Ontology (EFO) [8], eagle-i resource ontology (ERO) [20], NCBI taxonomy [5] and Chemical Entities of Biological Interest (ChEBI) [7] that are related to "*instruments*", "*reageants/chemical compounds*", "*organisms*" and "*sample/specimen*". The property, `sp:has instruction`, is used to define the instructions involved in protocols; instructions have actions and are the units for the workflow in the SP-workflow module. The order in which these instructions should be executed is captured by the BFO[6] property "*is preceded by*" and "*precedes*".

SIRO has been conceived in a way similar to that of the **P**atient **I**ntervention **C**omparison **O**utcome (PICO) model; it supports information retrieval and provides an anchor for the records [2]. SIRO extends the document metadata and delivers the semantics for the registry of a protocol. SIRO facilitates classification and retrieval without exposing the content of the document. In this way, publishers and laboratories may keep the content private, exposing information that describes the sample, instruments, reagent and objective of the protocol. The combination between NLP and semantics in SMART Protocols makes it possible to answer 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?*.

[5] https://github.com/information-artifact-ontology/IAO/
[6] http://ifomis.uni-saarland.de/bfo/

For convenience, in this paper we use the protocol "*Extraction of total RNA from fresh/frozen tissue (FT)*" [11] as a running example. We model this protocol with the SP ontology and SIRO; we also use this protocol to illustrate how the NLP component is using the ontologies and facilitating information retrieval. We are using GATE [3, 4] as the NLP engine; the information extraction system is ANNIE (A Nearly-New Information Extraction), and extraction rules are coded in JAPE (Java Annotation Patterns Engine). This paper is organized as follows: We start by presenting the SMART Protocols ontology and the SIRO model for minimal information, section 2. We then introduce our NLP component, section 3. Discussion and conclusions are then presented.

2 Semantics in SMART Protocols

The development of the SP ontology, was the first step [6], then the SIRO model followed. Use cases making use of semantics, NLP and information retrieval guided the process. Both the ontology and SIRO benefited from the continuous use of NLP techniques in support of harvesting terminology and identifying meaningful parts of speech (PoS) such as actions in the narratives. NLP was also used to semantically enrich the protocols based on the identified terminology. The gazetteers and rules of extraction were developed iteratively; as terminology and PoS were identified and validated manually, rules were being defined, tested, and validated against the accuracy of extracted protocols.

2.1 The SP Ontology

The SMART Protocols (SP) ontology is an application ontology designed to facilitate the representation of experimental protocols in two ways, as a document and as a workflow [6]. Our ontology reuses the Basic Formal Ontology (BFO). We are also reusing the ontology of relations (RO) [14] to characterize concepts. In addition, each term from the SP ontology is represented by annotation properties imported from OBI Minimal metadata[7]. The classes, properties and individuals are represented by their respective labels to facilitate the readability. The prefix indicates the provenance for each term. The document module of SP (henceforth SP-document)[8] document) aims to provide a structured vocabulary of concepts to represent information for recording and reporting an experimental protocol. The class `iao:information content entity` and its subclasses `iao:document`, `iao:document part`, `iao:textual entity` and `iao:data set` were imported from IAO. This module also represents metadata such as `sp:title of the protocol`, `sp:purpose of the protocol`, `sp:application of the protocol`, `sp:reagent list`, `sp:equipment and supplies list`, `sp:manufacturer`, `sp:catalog number` and `sp:storage conditions`.

The workflow module[9] represents the "*steps/instructions*", "*actions*" and experimental inputs such as "*reagents*", "*instruments*", and "*samples/speciments*" for enacting the workflow. The representation of executable elements of a protocol (instructions or steps), is modeled by reusing and expanding "*experimental actions*" from EXACT; in addition, we reused and expanded terminology related to "*instruments*", "*reagents/chemical compounds*", "*organisms*" and "*sample/specimen*" from ontologies such as: OBI, BAO, EFO, ERO, NCBI Taxonomy and ChEBI. The representation of *steps/instructions* is modeled with the class `sp:protocol instruction`, The property `sp:has instruction`, is used to define instructions involved in protocols. The order in which these instructions should be executed is captured by the property "*is preceded by*" and "*precedes*" from BFO.

Our running example, "*Extraction of total RNA from fresh/frozen tissue (FT)*" [11], is illustrated in figures 1 and 2 as an SMART Protocol. The application of

[7] http://obi-ontology.org/page/OBI_Minimal_metadata
[8] http://vocab.linkeddata.es/SMARTProtocols/sp-documentV2.0.htm
[9] http://vocab.linkeddata.es/SMARTProtocols/sp-workflowV2.0.htm

the SP-Document module is presented in table 1 and Fig. 1; metadata elements are organized in SP-Document as "*textual entities*". Modeling the workflow aspects, SP-Workflow module is presented in table 2 and Fig. 2; the first column in table 2 includes frequently used instructions that should be executed in protocols for the extraction of nucleic acids. The second column includes instructions extracted from the protocol. Olga Giraldo domain expert and author of this papersemi automatically extracted the information for each metadata element and identified instructions frequently used in different types of protocolsmainly in molecular biology.

Table 1. Metadata represented in SP ontology

Bibliographic metadata			
sp:title of the protocol	Extraction of total RNA from fresh/frozen tissue (FT)		
sp:author name	Kim M. Linton, Yvonne Hey, Sian Dibben, Crispin J. Miller, Anthony J. Freemont, John A. Radford, and Stuart D. Pepper		
sp:protocol identifier	DOI:10.2144/000113260		
Descriptive metadata			
sp:application of the protocol	"Methods comparison for high-resolution transcriptional analysis of archival material on Affymetrix Plus 2.0 and Exon 1.0 microarrays"		
sp:provenance of the protocol	"The extraction method (steps 221) is taken from the method supplied with TRIzol reagent (Invitrogen, Paisley, UK)."		
Metadata about the materials used			
sp:specimen name	"tumor tissue"		
sp:reagent name	"TRIzol"	sp:manufacturer name	"Invitrogen"
sp:reagent name	"Chloroform"	sp:manufacturer name	"Sigma-Aldrich"
sp:reagent name	"Ethyl alcohol"	sp:manufacturer name	"Sigma-Aldrich"
sp:reagent name	"Isopropyl alcohol"	sp:manufacturer name	"Sigma-Aldrich"
sp:equipment or supplies name	"Tissue storage container", "Homogenizer blades", "Forceps", "Scalpel", "Scalpel holder"		

Table 2. Protocol instructions commonly used in nucleic acid extraction protocols

Protocol instruction	
sp:cell disruption	"Homogenize sample using tissue homogenizer."
sp:denaturation reaction	"Add 0.2 mL chloroform per 1 mL TRIzol and cap tube tightly."
sp:precipitation reaction	"Add 0.5 mL isopropyl alcohol per 1 mL TRIzol"
sp:washing nucleic acids	"Add 1 mL 75% ethanol per 1 mL TRIzol and vortex for 10 s."

2.2 The Sample Instrument Reagent Objective (SIRO) model

SIRO represents the minimal information for describing an experimental protocol. In doing so, it serves two purposes. Firstly, it extends and structures available metadata for experimental protocols, for instance, author, title, date, journal, abstract, and other properties that are available for published experimental protocols usually as PDFs. SIRO extends this layer of metadata by aggregating information about Sample,

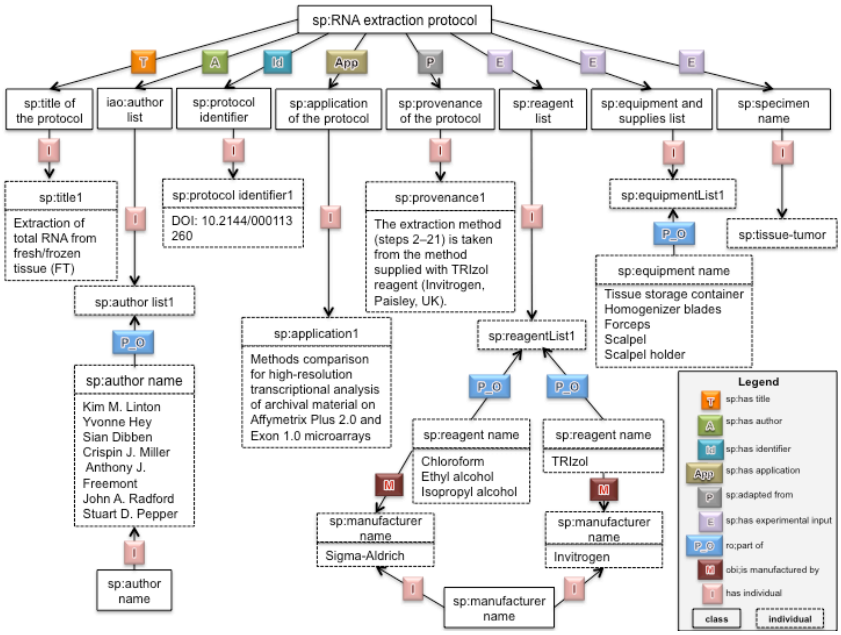

Fig. 1. SP-Document module. This diagram illustrates the metadata elements described in table 1. The classes, properties and individuals are represented by their respective labels.

Instrument, Reagent and Objective hence the name. If this information is part of the abstract or the full content, SIRO extracts and structures it as Linked Open Data (LOD). Secondly, SIRO, in combination with NLP and semantics, provides an anchor and structure for the minimal common data elements in experimental protocols. This makes it possible to find specific information about the protocol; if the owner of the protocol chooses not to expose the full content, as in the case of publishers and/or laboratories, SIRO may be exposed without compromising the full content of the document.

SIRO was developed after the SP ontology; Fig. 3 (step 2) illustrates the development process. The identification of common elements involved the following activities. Our "*kick-off*" phase started by redefining the use cases focusing on the identification of commonalities; it also entailed preparing the material to be used, e.g., ontologies, protocols and planning. Our main input was the SP Ontology and our corpus of documents. We then started to manually identify commonalities across protocols and, map these to the SP ontology as well as to ontologies in Bioportal[10], and OntoBee[11]. This **D**omain **A**nalysis and **K**nowledge **A**cquisition (DAKA) phase allowed us to gather common terminology with a raw classification. Our **Li**nguistic and **S**emantic **A**nalysis (LISA) was carried out in parallel with DAKA. LISA allowed us to automatically classify and identify the terminology we were gathering; LISA was extensively supported by GATE [3, 4]. The outcome allowed us to determine higher abstractions to which the terminology thus gathered could be mapped, e.g., "*sample*", "*reagent*" and "*instrument*". It also allowed us to recognize that, although the description of the objective was a common element, it was scattered throughout the narrative without an anchor.

3 Natural Language Processing for SMART Protocols

The SMART Protocols ontology models the execution of the workflow and relates reagents and instruments to steps in the workflow. Ontologies provide the gazetteers with the necessary knowledge for annotating beyond entity recognition. The gazetteers

[10] http://bioportal.bioontology.org/
[11] http://www.ontobee.org/index.php

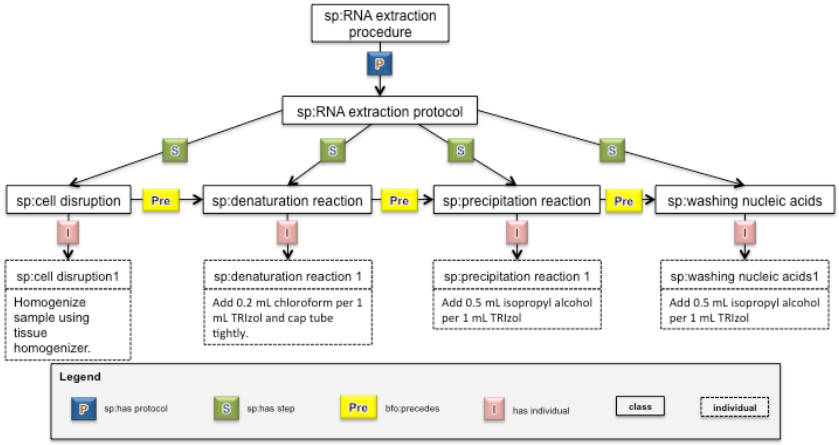

Fig. 2. SP-Workflow module. This diagram illustrates the metadata elements described in table 2. The classes, properties and individuals are represented by their respective labels.

make it possible to identify SIRO elements in the narrative; they are structured with information such as definition, URIs, provenance, synonyms, etc. Gazetteers based on ontologies have context; rules making use of these gazetteers find meaningful parts of speech in the text. GATE uses the gazetteers and the rules for annotating the documents. In this way, it is possible to differentiate between "*centrifuge*" as an action and "*centrifuge*" as an instrument. Furthermore, the rule engine in GATE, in combination with the gazetteers, make it possible to find statements related to, for example, "*precipitation instructions*" that include words related to an action such as "*centrifuge*" and a reagent like "*isopropanol*" (used to facilitate the precipitation of DNA) see Fig. 6 for a more complete example.

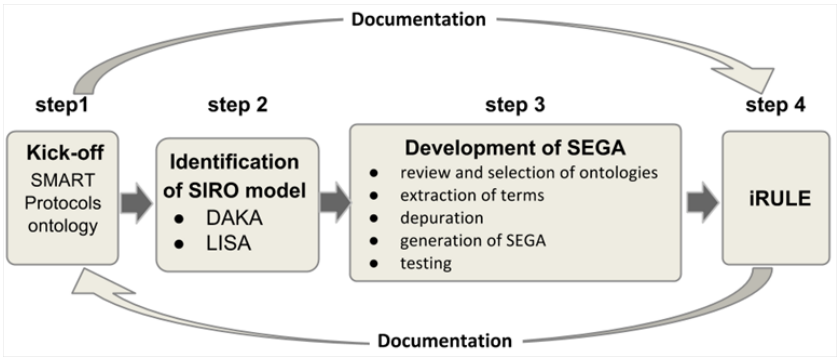

Fig. 3. Development process: identification of SIRO model, development of Gazetteers and rules for semantic annotation of experimental protocols.

3.1 Gazetteers and rules for NLP

The development of the **SE**mantic **GA**zetteers (SEGA) was a very complex and domain knowledge intensive activity. Developing the gazetteers entailed the identification of ontologies with terminology related to "*sample/specimen*" (including organisms), "*instruments*" and "*reagents*". Ontology repositories and their corresponding Application Programming Interfaces (APIs) were reviewed so that the process could be automated see Fig. 3, step 3 "*Review of Ontologies*". The ontologies identified during the development of SP ontology were then more carefully inspected; overlaps were identified, and availability of metadata for each term, object properties and complexity in the classification were addressed. For "*organisms*" (related to sample/specimen), the NCBI Taxonomy was chosen. For "*instrument*", the choices included EFO, ERO,

OBI and SP ontology. For "*reagents*" and "*chemical compounds*", ChEBI and SP were selected Fig. 3, step 3 "*Selection of ontologies*". During the stage "*Extraction of Terms*" see Fig. 3, step 3, we focused on enriching the terminology; depending on the limitations of the SPARQL endpoints for OntoBee and Bioportal, we were using either SPARQL queries or locally parsing the ontologies. The terminology was gathered with the corresponding annotation properties. Axioms and annotation properties were used to, for instance, discriminate if a term is synonymous with another term due to a case of acronyms or common name.

At this point, we had some gazetteers with over half a million terms. For quality control, we then started the depuration of the terminology. We removed the terms that had comments from the curators about the suitability of the terms in specific sub-domains. For instance, the class "*cell harvester*" (OBI_0001119) has a specific comment "*A device that is used to harvest cells from microplates and deposit samples on a filter mat. NOT AN INSTRUMENT*". We also removed terminology that was reused across ontologies. For instance, the OBI class "*thermal cicler*" is reused by SP and EFO. In this particular case, we use the term only once and from the original source -OBI. Classes with the same label represented in several ontologies with different axioms were conserved. For instance, SP reuses from the Sequence Ontology (SO) [8] the class "*forward primer*"; OBI also includes a class "*forward PCR primer*" (alternative term: forward primer). Once the terminology was cleaned, we then started the Generation of the Gazetteers; the gazetteers are used by GATE and together with the rules they support the NLP. GATE is based on a pipeline architecture, composed by Processing Resources (PR). Each PR has a specific function within the text processing (e.g. to create tokens and to tag PoS). We used ANNIE (A Nearly-New Information Extraction) as our information extraction system. We used the default ANNIE Gazetteer to build the gazetteers with less than 1 million terms per ontology and subdomain (Fig. 4); the gazetteers were configured as non case sensitive. For terms with synonyms, each synonym was added as an independent term, including features such as labels and URIs. To facilitate the recognition of terms varying from the corresponding roots, e.g. singular and plural, the gazetteers were nested into a Flexible Gazetteer (Fig. 4); this allows the extraction of the root for each token to be analyzed by a Morphological Analyzer. We also used a large KB Gazetteer to store sets of over 1 million terms related to organisms (Fig. 4). To facilitate data storage, we used a non-relational database and connected it to GATE.

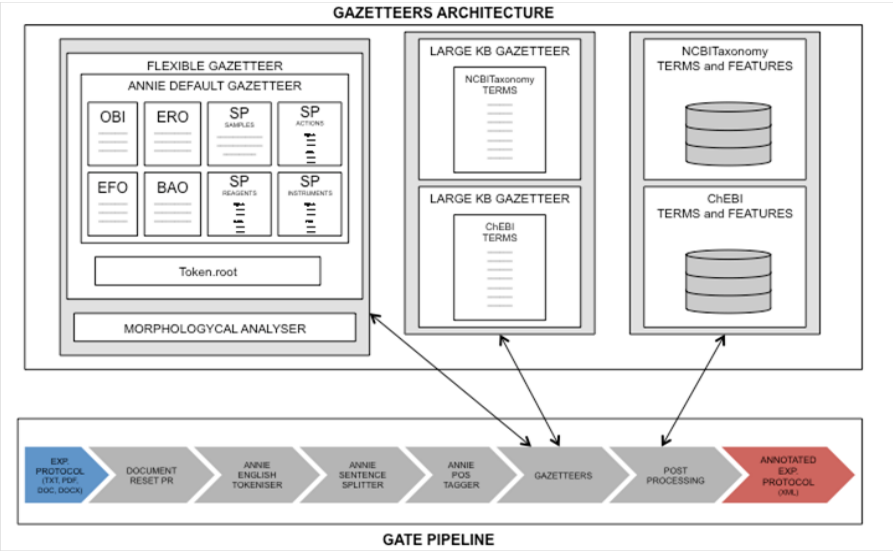

Fig. 4. Architecture for generating the gazetteers

For Testing the Gazetteers, we followed a manual process against our corpus of documents. Documents were loaded into GATE, and words related to SIRO elements were identified and annotated. We evaluated the following aspects, *i)* execution time, *ii)* correctness in the annotation of the terms and their synonyms, *iii)* failures in the recognition of terms in the texts, and *iv)* identification of terms incorrectly annotated, namely a word with different meaning For example, the word "*cat*" is a term from NCBI Taxonomy used to represent the common name of "*Felis catus*", but cat (or cat., Cat, CAT) also represents the short word for "*catalog*". From the gazetteers, linguistic patterns were identified so that The Iterative Rule Writing (see Fig. 5) step could commence.

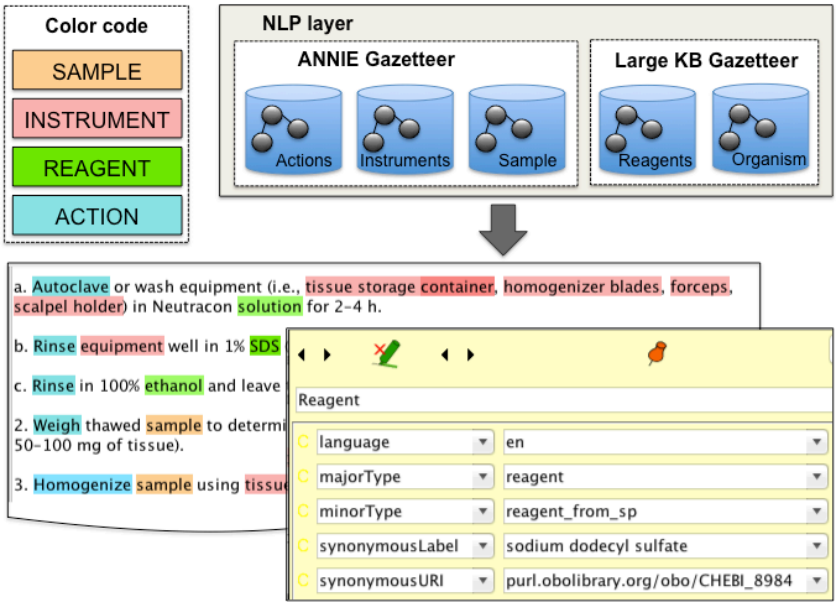

Fig. 5. This example illustrates a protocol annotated with terms related to sample/specimen, instruments, reagents and actions. Each annotated word is enriched with information related to: provenance (e.g. SDS is a concept reused by the SP ontology from ChEBI) and synonyms (sodium dodecyl sulfate). This term, reused from ChEBI, does not include a definition.

We are using JAPE (Java Annotation Patterns Engine) to code the rules. In this stage, we are designing rules to automate the identification of meaningful elements in the narrative. This step runs iteratively with previous stages; as linguistic structures and meaningful PoS, e.g. instructions, are characterized, then rules are written, tested and improved. Ontologies and domain terminology is mapped to the corresponding vocabularies.

4 Discussion and Conclusions

We have presented our approach to the Semantics for representing experimental protocols, the SP ontology and the SIRO model. The SP ontology is composed of two modules, namely SP-document and SP-workflow. In this way, we represent the workflow, document and domain knowledge implicit in experimental protocols. Actions, as presented by [7], are important descriptors for biomedical protocols; however, in order for actions to be meaningful, attributes such as measurement units and material entities (e.g., sample, instrument, reagents, personnel involved, etc.) are also necessary. Modularization, as it has been implemented in SP, facilitates specializing the ontology with more specific formalisms; this makes it easier for laboratories to adapt the ontology to their needs. For instance, reagents, instruments and experimental steps (actions), could be specialized based on the activities carried out by a

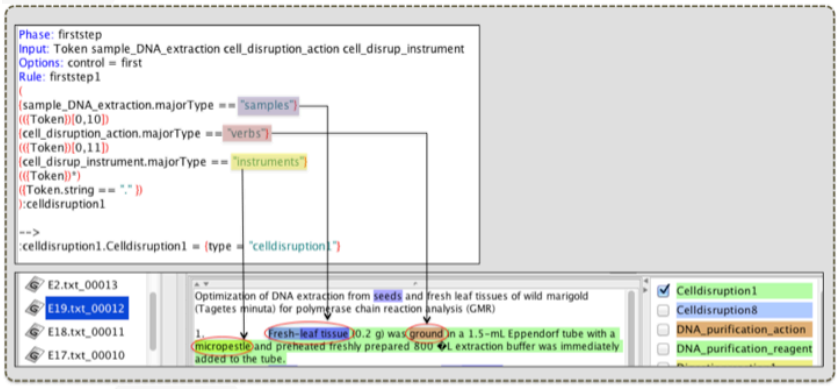

Fig. 6. Example illustrating a rule designed to find and annotate statements related to cell disruption.

particular laboratory. The document module facilitates archiving; the structure also allows to have fully identified reusable components.

The SIRO model for minimal information breaks down the protocol in key elements that we have found to be common to our corpus of experimental protocols: *i)* Sample/ Specimen (**S**), *ii)* Instruments (**I**), *iii)* Reagents (**R**) and *iv)* Objective (**O**). 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. SIRO and the SP Ontology facilitate the generation of a self-describing document with structured annotation.

Our NLP layer makes use of the semantics we have defined. We currently have six gazetteers with over 1.400.000 terms in all; these terms will be further refined and then added to the SP ontology. The gazetteers are currently reusing terminology from EFO, ERO, OBI, NCBI Taxonomy and ChEBI; we will continue adding terminology from other ontologies and also adding more documents to our corpus. We are making use of existing infrastructure provided by BioPortal and OntoBee. For managing large ontologies, we are not using their respective SPARQL endpoints but locally parsing them, e.g. NCBI Taxonomy and ChEBI. Our Semantics plus NLP infrastructure makes it possible to retrieve information where specifics from the protocols are used to construct the query. Our NLP layer is able to extract SIRO automatically; we have encountered issues with the free narrative often used for describing the objectives.

Experimental protocols are meant to capture a complex and nested set of roles actions, derivations of original plans, actions executed by personnel, robots taking care of some specific steps in the workflow, computational workflows often used in support of laboratory work, data being produced at every step of the workflow, etc. Representing and enacting all of these is not a simple task; laboratories require flexibility in their conceptual models so that parameterizing their own workflows won't become an overwhelming task. The laboratories only carry out a limited set of actions over a limited set of samples; high-level abstractions for general process models are needed. These could be made more concrete as workflow constructs, samples, roles, actions, reagents, instruments, etc. are aggregated. Representing the execution requires the confluence of metadata that allows tracking of everything that has occurred, who has done it, how, where, etc. Our ontology model may easily be extended and adapted to these realities. The metadata schemata to represent laboratory protocols should be kept independent from the workflow enactors; robots will surely have their own procedural languages. The descriptive schemata should interoperate with the work-

flow enactors. The SP ontology was conceived considering all of these; our use cases are incrementally becoming more complex as we are moving from protocols published in journals to those registered in laboratory notebooks needless to say, that gaining access to laboratory notebooks is not easy.

5 Acknowledgments

OG acknowledges the support from the EU project DrInventor FP7-ICT-2013.8.1, AG acknowledges the KOPAR project, H2020-MSCA-IF-2014, Grant Agreement nr: 655009. We thank Edna Ruckhaus and Melissa Carrion for their useful discussions.

References

[1] LG. Acevedo et al. "Genome-scale ChIP-chip analysis using 10,000 human cells". In: *Biotechniques* 43.6 (2007), pp. 791–797.

[2] A Booth and A. Brice. *Formulating answerable questions*. Ed. by Editor. A.B. Booth A (Eds). 2004.

[3] H. Cunningham et al. "GATE - a General Architecture for Text Engineering". In: Computers and the Humanities. 2002.

[4] H. Cunningham et al. "Getting more out of biomedical documents with GATE's full lifecycle open source text analytics". In: *PLoS Comput Biol* 9.2 (2013), e1002854.

[5] S. Federhen. "Type material in the NCBI Taxonomy Database". In: *Nucleic Acids Res* 43 (2015), pp. D1086–98.

[6] O. Giraldo, A. Garcia, and C. Oscar. "SMART Protocols: SeMAntic RepresenTation for Experimental Protocols". In: 4th Workshop on Linked Science 2014- Making Sense Out of Data (LISC2014). 2014.

[7] J. Hastings et al. "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013". In: *Nucleic Acids Res* 41 (2013), pp. D456–63.

[8] Malone J. et al. "Modeling sample variables with an Experimental Factor Ontology". In: *Bioinformatics* 26.8 (2010), pp. 1112–1118.

[9] R.D. King et al. "On the formalization and reuse of scientific research". In: *J R Soc Interface* 8.63 (2011), p. 1440.

[10] R.D. King et al. "The automation of science". In: *Science* 324.5923 (2009), p. 85.

[11] K.M. Linton et al. "Extraction of total RNA from fresh/frozen tissue (FT)". In: *The International Journal of Life Science Methods* (2010).

[12] P. Rocca-Serra et al. *Release candidate 1, ISA-TAB v1.0 specification document, version 24th*. 2008, p. 36.

[13] Brinkman R.R. et al. "Modeling biomedical experimental processes with OBI". In: *Journal of Biomedical Semantics* 1.1 (2010), p. 11.

[14] B. Smith et al. "Relations in biomedical ontologies". In: *Genome Biology* 6.5 (2005).

[15] B.and others Smith. "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration". In: *Nature Biotechnology* 25.11 (2007), pp. 1251–1255.

[16] L. N. Soldatova et al. "The EXACT description of biomedical protocols". In: *Bioinformatics* 24.13 (2008), pp. i295–303.

[17] L.N. Soldatova and K.R. D. "An ontology of scientific experiments". In: *Journal of the royal society interface* 3.11 (2006), p. 795.

[18] L.N. Soldatova et al. "Evolving BioAssay Ontology (BAO): modularization, integration and applications." In: *J Biomed Semantics* 5.Suppl 1 Proceedings of the Bio-Ontologies Spec Interest G (2014), S5.

[19] C.F. Taylor et al. "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project". In: *Nat Biotechnol* 26.8 (2008), pp. 889–96.

[20] C. Torniai et al. "Developing an Application Ontology for Biomedical Resource Annotation and Retrieval: Challenges and Lessons Learned". In: ICBO: International Conference on Biomedical Ontology. Buffalo, NY, USA., pp. 101–108.