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 on the semantic and NLP components, namely the SMART Protocols (SP) Ontology, the Sample Instrument Reagent Objective (SIRO) model and the text mining integrative architecture GATE. The SMART Protocols (SP) Ontology results from the analysis of over 300 experimental protocols in various domains –molecular biology, cell and developmental biology and others. The gathered terminology is then evaluated, rules are improved accordingly and then a new iteration starts. 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 Patient Intervention Comparison Outcome (PICO) model that supports search, retrieval and classification purposes. The SIRO ontology development process includes NLP as well as domain expertise in the extraction of the vocabulary; domain experts extract an initial seed of terminology, then the process is automated by using gazetteers and extraction rules in JAPE. Both SIRO and the SP ontology are then used by our NLP engine, GATE. 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

Several efforts have been proposed in the current state of the art for data sharing and preservation; these aim to facilitate reproducible science. There are from generalized data repositories (Open Science Framework[1] , DRYAD[2] and figshare[3]) to discipline specific initiatives (DNA Data Bank of Japan (DDBJ)[4] , Cancer Imaging Archive[5] , ChEMBL[6] , NOAA's National Centers for Environmental Information (NCEI)[7] and The NoMaD (Novel Materials Discovery) Repository).[8]

It is also important to know how the data was produced, and the steps undertaken when doing experiments. Fewer work has been done to formally represent such knowledge, which is usually encoded within experimental protocols. Experimental protocols are fundamental information structures that support the description of the processes by means of which results are generated in experimental research [6] . Biomedical experiments, for instance, often rely on sophisticated laboratory protocols, comprising hundreds of individual steps. For example, the protocol for chromatin immunoprecipitation on a microarray (Chip-chip) has 90 steps, uses over 30 reagents and 10 different devices [1] . Protocols are written in natural language; they are often presented in a "recipe" style and they provide a step-by-step description of procedures.

Efforts such as the ISA-TAB[9] and Biosharing[10] are proposing standards and reporting structures for biomedical investigations. Although the importance of experimental protocols is acknowledged, the workflow nature of experimental protocols and the minimal information describing such information artifacts is not yet part of available specifications. Also, the aggregative ever-changing nature of the process, we argue, is not entirely captured by previously proposed schemata; researchers plan a workflow, allocate resources for execution, and then execute. Although ideal, the execution usually imposes changes to the initial design; forks leading to entirely new workflows due to a myriad of factors are common in laboratory practices. Publishers have also addressed the issue of reporting the processes leading to the production of data; there is, however, no consensus about the specifics for reporting these kind of workflows.

In this paper we present the semantic and Natural Language Processing (NLP) infrastructure for SMART Protocols (SP); we aim to allow the generation and processing of experimental protocols. Our NLP layer makes use of various ontologies as well as of the Sample Instrument Reagent Objective (SIRO) model for minimal information (MI) that we have defined. The SP ontology as well as SIRO are built upon experiences like those reported in Biosharing, published ontologies and previously proposed standards; moreover, our work is based on an exhaustive analysis of over 300 real experimental protocols (molecular biology, cell and developmental biology, biochemistry) and guidelines for

[1] https://osf.io/

[2] http://datadryad.org/

[3] http://figshare.com/

[4] http://www.ddbj.nig.ac.jp/

[5] http://www.cancerimagingarchive.net/

[6] https://www.ebi.ac.uk/chembl/

[7] https://www.ncei.noaa.gov/

[8] http://nomad-repository.eu/cms/

[9] http://isatab.sourceforge.net/format.html

[10] https://www.biosharing.org/

authors from over 20 journals. The SP ontology has three main modules; it models the workflow, the document in which the workflow is communicated, as well as domain knowledge. SIRO has been conceived in a similar way to that of the Patient, Intervention, Comparison, and Outcome (PICO) model in support of information retrieval and providing an anchor for the records [2] . SIRO extends the document metadata; it delivers the semantics for the registry of a protocol, facilitating classification and retrieval without exposing the content of the document.

We are using GATE as our NLP engine; ANNIE (A Nearly-New Information Extraction) is our information extraction system, and extraction rules are coded in JAPE (Java Annotation Patterns Engine). 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?”. Central to our work it is to support sharing, discovering reusing and bridging the gap between data and experimental protocols.

This paper is organized as follows. we start by presenting the SMART Protocols ontology and the SIRO model for minimal information; we then introduce our NLP layer. Discussion and conclusions are then presented.

2 Semantics plus NLP in SMART Protocols

The development of the semantic layer of SMART Protocols, the SP ontology was the first step [6] . The definition of SIRO followed. 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 workflows. NLP, entity recognition, was also used to semantically enrich the protocols based on 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, validated against the accuracy of extracted protocols and then re-defined.

2.1 The SP Ontology

The SMART Protocols approach follows the OBO Foundry principles [14] . Our modules reuse the Basic Formal Ontology (BFO). Also, we are reusing the ontology of relations (RO) [13] to characterize concepts. In addition, each term from SP is represented by annotation properties imported from OBI Minimal metadata. [11] The classes, properties and individuals are represented by their respective labels to facilitate the readability. The prefix indicates the provenance of each term. The class `iao:information content entity` and its subclasses `iao:document`, `iao:document part`, `iao:textual entity` and `iao:data set` were imported from The Information Artifact Ontology

[11] http://obi-ontology.org/page/OBI_Minimal_metadata

(IAO) to represent the document aspects in the protocol. The representation of executable aspects of a protocol is modeled with the classes `p-plan:Plan`, `p-plan:Step` and `p-plan:Variable` from the P-Plan Ontology (P-Plan).

The document module of SMART Protocols[12] reuses classes from CHEBI [9] , EXACT [15] , MGED [8] , SO [10] , OBI [3] and SNPO. [13] Also, SMART Protocols- document (henceforth SP-document) extends the class `iao:information content entity` proposed by the Information Artifact Ontology (IAO) to represent the experimental protocol as an `iao:document` that has parts, `ro:has_part` , such as `iao:document part` (iao:author list, `sp:introduction section`, `sp:materials section` and `sp:methods section`). SP-document represents information such as, the protocol type, `sp:DNA extraction protocol`; it has a tittle, identified by the property `sp:has title`, it is instantiated by genomic DNA isolation. Also, the author entry, `iao:author identification`, is instantiated by CIMMYT [4] . This protocol is derived, `sp:provenance of the protocol`, from the protocol published by [12] (sp:PNAS 81:8014-8019) and its purpose is instantiated by plant DNA extraction of high quality.

The workflow module[14] extends the P-Plan Ontology (P-Plan) [5] . This ontology was developed to describe scientific processes as plans and link them to their previous executions. In the workflow module of SMART Protocols (henceforth SP-workflow), the experimental protocol, `p-plan:Plan`, is a description of a sequence of operations, `p-plan:Step`, that includes an input and an output `p-plan:Variable`. In this sense, a protocol is a type of workflow. SP-workflow also reuses classes from CHEBI, MGED, SO, OBI and NPO [17] . The use case illustrates DNA extraction, this is a procedure frequently used to collect DNA for subsequent molecular or forensic analysis, see Fig 1. DNA extraction includes 3 basic p-plan:Steps: i) cell disruption or cell lysis, ii) Digestion reaction (in this step, contaminants such as lipid membrane, proteins and RNA are removed from the DNA solution), and iii) DNA purification. Each one of these steps may include different protocols (or p-plan:Plans) to be executed. For example, the step `sp:cell disruption` or cell lysis may be achieved by chemical and physical methods - blending, grinding or sonicating the sample. Also, the ontology considers that each step is executed following a predetermined order. For instance, according to the protocol published by CIMMYT, the cell disruption by lyophilization and grinding has an input variable, `p-plan:hasInputVar`, as well as `sp:plant tissue`; it also has an output, `p-plan:hasOutputVar`, and `sp:powdered tissue`. The next step, `sp:digestion reaction`, has as input the output of the immediately previous step, `sp:powdered tissue`, and as output `sp:digested contaminant`. The last one, `sp:DNA purification` has as input `sp:digested contaminant`, and as output `obi:DNA extract` .

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

[13] http://www.loria.fr/ coulet/snpontology1.4_description.php

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

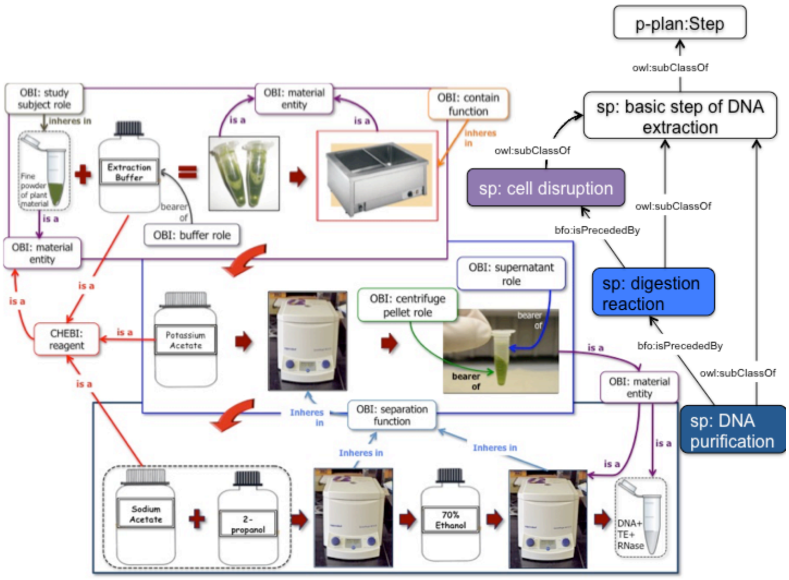

Fig. 1. SMART Protocols ontology development process

The SP ontology is in constantly enrichment. Our owl model is periodically evaluated according to the criteria for evaluation proposed by Suárez-Figueroa [16] . The OntOlogy Pitfall Scanner (OOPS), was useful to detect and correct anomalies or pitfalls in our ontologies [11] . here, is also evaluated the precision, error rate and recall of the annotated protocols.

2.2 The Sample Instrument Reagent Objective (SIRO) model

The protocols don't always include an explicit statement detailing the Objective; by the same token information about the Sample, Instruments and Reagents is usually scattered all over the narrative in the document. SIRO delivers a simple, intuitive and rigorous structure that facilitates retrieval and classification. 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 some other properties are available for published experimental protocols. SIRO extends this layer of metadata by aggregating information about Sample, Instrument, Reagent and Objective. If this information is part of the abstract or the full content, SIRO extracts and structures it as Linked Open Data (LOD) and could expose it over a SPARQL endpoint. 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. 1 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 loaded with domain specific terminology, e.g. CHEBI, to be used as a seed for subsequent NLP tasks. We then started to manually identify commonalities across protocols, and mapping these to the SP ontology as well as to ontologies in Bioportal[15] , and OntoBee.[16] This Domain Analysis and Knowledge Acquisition (DAKA) phase allowed us to gather common terminology with a raw classification. Our Linguistic and Semantic Analysis (LISA) was carried out in parallel to DAKA. LISA allowed us to automatically classify and identify the terminology we were gathering; LISA was extensively supported by GATE. The outcome allowed us to determine to which higher abstractions could the terminology thus gathered be mapped -.e.g sample, reagent, instrument. It also allowed us to identify that, although the description of the objective was a common element, it was scattered throughout the narrative.

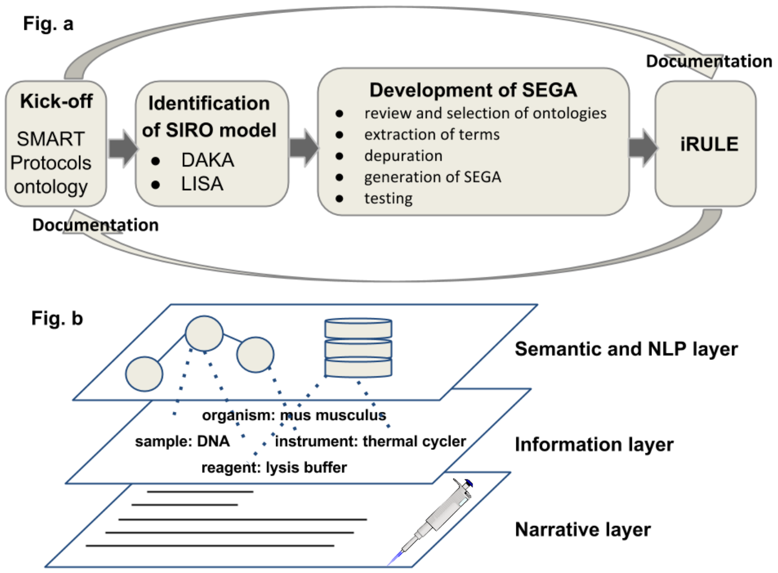

Fig. 2. SMART Protocols ontology development process

2.3 Gazetteers and rules for NLP

The development of the semantic gazetteers (SEGA) followed; this was a very complex and domain knowledge intensive activity. The gazetteers, in combina-

[15] http://bioportal.bioontology.org/

[16] http://www.ontobee.org/index.php

tion with the rules for extraction, make it possible to identify SIRO elements in the narrative. Actions may also be identified in this way. Developing the gazetteers entailed the identification of ontologies with terminology related to sample/specimen, instruments and reagents. Ontology repositories and their corresponding APIs were also reviewed so that the process could be automated -step Review of Ontology repositories, Fig 2. From this step OntoBee and Bioportal were selected. The identified ontologies were then more carefully inspected -step Selecting Ontologies; overlappings were identified, availability of metadata for each term, object properties and complexity in the classification were addressed. For organisms (related to sample/specimen), the NCBITaxon was chosen. For instruments, the choice was EFO [7] , ERO[17] , OBI and SP ontologies. For reagents and chemical compounds, CHEBI and SP were selected. During the stage Extraction of Terms, we focused on enriching the terminology; depending on the limitations of the endpoints for OntoBee and Bioportal we were using either SPARQL queries or locally parsing the ontologies. The terminology was gathered with the corresponding associated 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 -at least one of them had over 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 cycler 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) 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, 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; 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; this allows the extraction of the root for each token to be analyzed by a Morphological Analyzer. We also used large KB Gazetteer to store large sets (over 1 million) of

[17] https://open.med.harvard.edu/wiki/display/eaglei/Ontology

terms related to organisms. To facilitate data storage we used a non-relational database and connected it to GATE. The development process for the gazetteer is illustrated in Fig. 2.

For Testing the Gazetteers we followed a manual process against our corpus of documents. Documents were loaded into GATE, then annotated and then SIRO elements were identified. 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 NCBItaxon used to represent the common name of Felis catus, but 'cat' (or cat., Cat, CAT) also represent the short word for 'catalog'. From the gazetteers, linguistic patterns were identified so that The Iterative Rule Writing step could start. 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 will also be mapped to the corresponding vocabularies.

3 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 [15] are important descriptors for biomedical protocols; however, in order for actions to be meaningful, attributes such as measurement units, 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, idem actions, could be specialized based on the activities carried out by a 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 all laboratory 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 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, NCBITaxon 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 the ontologies e.g. NCBITaxon 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, personnel executing actions, 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 wont 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, sample, roles, actions, reagents, instruments, etc are aggregated. Representing the execution requires the confluence of metadata that allows to track down 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 workflow 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.

4 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 to Edna Ruckhaus for her 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 A Booth and A (Eds) Brice. 2004.

[3] RR. Brinkman et al. “Modeling biomedical experimental processes with OBI”. In: J Biomed Semantics 22.1 (2010), S1–S7.

[4] CIMMYT. Laboratory Protocols: CIMMYT Applied Molecular Genetics Laboratory. Third Edition. Mexico, D.F.: CIMMYT, 2005.

[5] D. Garijo and Y Gil. “Augmenting PROV with Plans in P-PLAN: Scientific Processes as Linked Data”. In: in 2nd international Workshop on Linked Sicence 2012 - Tackling Big Data (LISC2012), in conjunction with 11th International Semantic Web Conference (ISWC2012). 2012.

[6] Olga Giraldo, Alexander Garcia, and Oscar Corcho. “SMART Protocols: SeMAntic RepresenTation for Experimental Protocols”. In: 4th Workshop on Linked Science 2014- Making Sense Out of Data (LISC2014), in conjunction with the International Semantic Web Conference (ISWC2014). 2014.

[7] Malone James et al. “Modeling sample variables with an Experimental Factor Ontology”. In: Bioinformatics 26.8 (2010), pp. 1112–1118.

[8] Stoeckert Jr, Christian J, and Helen Parkinson. “The MGED ontology: a framework for describing functional genomics experiments”. In: Comparative and Functional Genomics 4 (2003), pp. 127–132.

[9] Paula de Matos et al. “Chemical Entities of Biological Interest: an update”. In: Nucleic Acids Research 38.suppl 1 (2010), pp. 1D249–D254.

[10] C. J Mungall, C Batchelor, and K Eilbeck. “Evolution of the Sequence Ontology terms and relationships”. In: J Biomed Inform 44.1 (2011), pp. 87–93.

[11] Mariá Poveda-Villaloń, M.C Suaŕez-Figueroa, and Asuncioń Goḿez-Peŕez. Validating Ontologies with OOPS! Ed. by Annette ten Teije et al. 2012.

[12] M. A. Saghai-Maroof et al. “Ribosomal DNA spacer-length polymorphisms in barley: mendelian inheritance, chromosomal location, and population dynamics”. In: Proc Natl Acad Sci U S A 81.24 (1984), pp. 8014–8018.

[13] Barry Smith et al. “Relations in biomedical ontologies”. In: Genome Biology 6.5 (2005), R46.

[14] Barry Smith et al. “The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration”. In: Nature Biotechnology 25.11 (2007), pp. 1251–1255.

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

[16] M.C Suaŕez-Figueroa, Asuncioń Goḿez-Peŕez, and Mariano Fernańdez-Loṕez. “The NeOn Methodology framework: A scenario-based methodology for ontology development”. In: Applied Ontology (2015), pp. 1–39.

[17] D. G Thomas, R. V Pappu, and N. A. Baker. “NanoParticle Ontology for cancer nanotechnology research”. In: J Biomed Inform 44.1 (2011), pp. 59–74.