UNIVERSIDAD POLITÉCNICA DE MADRID

DOCTORAL THESIS

SeMAntic RepresenTation for experimental Protocols

Author:

Olga Ximena Giraldo Pasmin

Supervisor:

Prof. Dr. Oscar Corcho

*A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy
in the*

Ontology Engineering Group
Department of Artificial Intelligence

April 23, 2019

Declaration of Authorship

I, Olga Ximena Giraldo Pasmin, declare that this thesis titled, “{SeMAntic Representation for Experimental Protocols” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

UNIVERSIDAD POLITÉCNICA DE MADRID

Abstract

Department of Artificial Intelligence
Escuela Técnica Superior de Ingenieros Informáticos

Doctor of Philosophy

SeMAntic RepresenTation for experimental Protocols

by Olga Ximena GIRALDO PASMIN

This research address the problem of semantically representing experimental protocols in life sciences and how to relate such information to data. The need for open interoperable data supporting research transparency, systematic reuse of existing data and, experimental reproducibility has been widely acknowledged. Several efforts are providing infrastructure for sharing and storing data. However, data per se does not imply reproducibility; there is the need to know how the data was produced -here is the data, where are the experimental protocols? Several efforts have studied the problem of "is this reproducible?" Fewer efforts have addressed the problem of semantically valid, machine-processable reporting structures. SMART Protocols (SP) makes use of Semantic Web technology, thus facilitating interoperability and machine processability; SP delivers an extendible infrastructure that allows researchers to search for similar protocols, or investigations with similar techniques, methods, instruments, variables and/or populations, etc. In order to achieve such degree of functionality, throughout this investigation a comprehensive vocabulary was gathered by annotating documents; the corresponding infrastructure, henceforth BioH, was specially developed to support this task. The evaluation of the vocabulary thus gathered made it possible to generate the SP gold standard; this is a gold standard corpus specifically engineered for experimental protocols. The tooling and methods applied when building this gold standard can be applied to other domains. Furthermore, this investigation also delivers a semantic publication platform for experimental protocols; Scientific publications aggregate data by encompassing it within a persuasive narrative. The SP approach addresses the problem of supporting such aggregation over a document that is to be born semantic, interoperable and conceived as an aggregator within a web-of-data publishing workflow.

Acknowledgements

First and foremost, thanks to my family. You are the foundation of all my strength. To my mother, thank you for your constant love and support, it is something that I have always depended on without thinking and I would be nowhere without it. To my husband, you have given more to me than I could ever ask, thank you for riding along with me through the storms and the doldrums of this journey and for reaching down and lifting me back up every time I started to drift beneath the surface. Most importantly, and from the bottom of my heart, thanks to my daughter in whom I have found my deepest happiness as well as my true inner strength. Since she was born, she has taught me more about myself than everything I taught I knew. To God, who blessed me with Alba. ...

Contents

Declaration of Authorship	iii
Abstract	vii
Acknowledgements	ix
1 Introduction	1
1.1 Introducing the problem	1
1.2 Motivation	2
1.3 Problem statement	4
1.4 Contributions of this thesis	5
1.4.1 Research Outcomes related to this Investigation	6
Awards	6
Journal Papers	6
Conferences and Workshops	6
1.5 Outline of this Thesis	7
Bibliography	11
2 A Guideline for Reporting Experimental Protocols in Life Sciences	13
2.1 Introduction	14
2.2 Materials and Methods	15
2.2.1 Materials	15
i) Instructions for authors from analyzed journals.	15
ii) Corpus of protocols.	16
iii) Minimum information standards and Ontologies.	16
2.2.2 Methods for developing this guideline	17
Analyzing guidelines for authors	17
Analyzing the protocols.	18
Analyzing Minimum Information Standards and ontologies	19
Generating the first draft	20
Evaluation of data elements by domain experts	21
2.3 Results	21
2.3.1 Bibliographic data elements	23
2.3.2 Data elements of the discourse	25
2.3.3 Data elements for materials	26
2.3.4 Data elements for the procedure	32
2.4 Data elements represented in the SMART Protocols Ontology	35
2.5 Discussion	36
2.6 Conclusion	38
Bibliography	41

3	Using Semantics for Representing Experimental Protocols	51
3.1	Background	52
3.2	Methods	53
3.2.1	The Kick-off, Scenarios and Competency Questions	53
3.2.2	Conceptualization and Formalization	53
	Domain Analysis and Knowledge Acquisition, DAKA	54
	Linguistic and Semantic Analysis, LISA	55
	Iterative ontology building and validation, IO	56
3.2.3	Ontology Evaluation	56
3.3	Results	57
3.3.1	The SMART Protocols ontology	57
	The Document Module	57
	The Workflow Module	57
3.3.2	Evaluation	59
	Syntax	59
	Conceptualization and Formalization	59
	Competency questions	62
3.4	Applying the SMART Protocols Ontology to the Definition of a Minimal Information Model	62
3.4.1	The Sample Instrument Reagent Objective (SIRO) Model	63
3.4.2	Evaluating the SIRO Model	64
3.5	Discussion	65
3.5.1	SMART Protocols Ontology	65
3.5.2	Modularization of the SP ontology	65
3.5.3	Limitations	66
3.5.4	The SIRO model, application of the ontology	66
3.6	Conclusions	66
	Bibliography	71
4	Laboratory Protocols in Bioschemas	77
4.1	Introduction	78
4.2	Why semantic structuring?	78
4.3	Bioschemas at a glance	78
4.3.1	Experimental Protocols and Bioschemas	80
4.4	Developing the LabProtocol profile	80
4.5	Results, The Labprotocol Profile	83
4.5.1	Mandatory properties	83
4.5.2	Recommended properties	83
4.6	Discussion	86
4.7	Conclusions and Future Work	87
	Bibliography	89
5	BioH, The Smart Protocols Annotation Tool	93
5.1	Introduction	94
5.2	The SIRO Curation Model	95
5.3	The Tool	96
5.3.1	Architecture	96
5.4	Discussion and Concluding Remarks	97
	Bibliography	99

6	Generating a Gold Standard Corpus for Experimental Protocols	101
6.1	Introduction	102
6.2	Materials and Methods	102
6.2.1	Materials	102
	Corpus of documents	102
	Annotators	103
	Annotation guidelines	103
6.3	Methods	104
6.4	Results	105
6.5	Discussion	108
6.6	Conclusions	108
	Bibliography	111
7	Semantics at Birth, the SMART Protocols Publication Platform	115
7.1	Introduction	116
7.2	Semantic Publishing for Experimental Protocols	117
7.2.1	Preserving the Resource Map for a Protocol	118
7.3	Results	119
7.3.1	Architecture and Data Workflow	119
7.4	Discussion	122
7.4.1	Granular preservation over Hyperledger	122
7.4.2	Nanopublications from SMART Protocols	123
7.5	Conclusions and Final Remarks	123
	Bibliography	125
8	Discussion and Conclusions	129
8.1	Summary	129
8.2	Reusable Data	130
8.2.1	Using the Semantic Layers	131
8.2.2	Concluding remarks	132
9	Future Work	135
	Appendix A User guide for the SMART Protocols Annotation Tool	137
	Appendix B Guidelines to annotate experimental protocols using the SIRO model	155

List of Figures

1.1	An overview of the structure of this thesis	9
2.1	Methodology Workflow.	19
2.2	Bibliographic data elements found in guidelines for authors. NC= Not Considered in guidelines; D= Desirable information if this is available.	23
2.3	Data elements related to the discourse as reported in the analyzed protocols	25
2.4	Data elements describing materials. NC= Not Considered in guidelines; D= Desirable information if this is available; R= Required information.	27
2.5	Data elements describing materials.	27
2.6	Data elements describing the process, as found in the guidelines for authors. NC= Not Considered in guidelines; O= Optional information; D= Desirable information if this is available; R= Required information.	32
2.7	Data elements describing the process, as found in analyzed protocols. .	33
2.8	Hierarchical organization of data elements in the SMART Protocols Ontology.	35
3.1	Developing the SMART Protocols ontology, methodology	54
3.2	SP-Document module. This diagram illustrates the metadata elements described in Table 2. The classes, properties and individuals are represented by their respective labels.	59
3.3	SP-Workflow module. This diagram illustrates the metadata elements described in Table 3. The classes, properties and individuals are represented by their respective labels.	61
3.4	Distribution of SIRO elements	63
3.5	The SIRO model	64
4.1	General overview of Bioschemas and the LabProtocol profile	79
4.2	A general overview of the development process	82
5.1	From general to specific, navigating an ontology	95
5.2	What and how to annotate using BioH	96
5.3	Architecture and components of the BioH annotation tool	97
6.1	An overview of the annotation process	104
6.2	Workflow summarizing annotation sections	105
6.3	Architecture for generating the gazetteers	106

6.4	Example illustrating 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.	107
6.5	Example illustrating a rule designed to find and annotate statements related to cell disruption	108
7.1	General view for an RMap represented as a Disco. IKn this figure, assets related to a protocol are presented. Small icons were taken from www.flaticon.com	118
7.2	General Architecture for SMART Protocols	120
7.3	A view of the publication process	121
7.4	Publishing a narrative as data	121
7.5	Nanopublications from a procedure	123
8.1	Reusable data	130

List of Tables

2.1	Guidelines for reporting experimental protocols.	16
2.2	Corpus of protocols analyzed.	16
2.3	Minimum Information Standards analyzed.	17
2.4	Ontologies analyzed.	18
2.5	Bibliographic data elements from guidelines for authors. Y= datum considered as “desirable information” if this is available, N= datum not considered in the guidelines.	18
2.6	Rhetorical/Discourse elements from guidelines for authors. R= Required information; NC= Not Considered in guidelines; D= Desirable information; O= Optional information.	20
2.7	Data elements for reporting protocols in life sciences	22
2.8	Examples illustrating two tittles. Issues in the ambiguous title: *Use of ambiguous terminology, ‡use of abbreviations.	24
2.9	Example illustrating the provenance of a protocol.	25
2.10	Examples of discursive data elements.	26
2.11	Example for the presentation of equipment.	29
2.12	Reporting consumables.	30
2.13	Reporting recipes for solutions.	30
2.14	Reporting reagents.	31
2.15	Examples of alert messages	34
3.1	Repositories and number of protocols analyzed	53
3.2	Metadata represented in SP-Document	58
3.3	Procedures and subprocedures from “Extraction of total RNA from fresh/frozen tissue (FT)”	60
3.4	Queries making use of external resources. Queries are available at https://smartprotocols.github.io/queries/	68
3.5	SIRO Elements	69
4.1	Mandatory properties proposed to represent the LabProtocol type	83
4.2	Thing properties from schema.org proposed as recommended properties	84
4.3	CreativeWork properties from schema.org proposed as recommended properties	85
4.4	Types from schema.org proposed as recommended properties	86
6.1	Corpus of annotated protocols	103
6.2	Number of annotators by institution	103
6.3	Protocols where the objective could not be annotated	105

To my daughter and husband with love...

Chapter 1

Introduction

1.1 Introducing the problem

Openness and reproducibility are not only related to data availability. When reproducing research, being able to follow the steps leading to the production of data is equally important. Reproducibility is related to *the degree of agreement between the results of experiments conducted by different individuals, at different locations, with different instruments. Put simply, it measures our ability to replicate the findings of others* [1]–[4]. Throughout this research, reproducibility can be thought of *as a different standard of validity* 149 *because it forgoes independent data collection and uses the methods and data collected by the original investigator.* Reproducibility is thus related to the ability of a researcher to reproduce an experiment and generate similar results; this practical definition is in agreement with Kitzes [4].

Experimental protocols are information structures that provide descriptions of the processes by means of which results, often data, are generated in experimental research [5]. Scientific experiments rely on several *in vivo*, *in vitro* and *in silico* methods and techniques; the protocols often include equipment, reagents, critical steps, troubleshooting, tips and all the information that facilitates reusability. Researchers write the protocols to standardize methods, to share these documents with colleagues and to facilitate the reproducibility of results. When reproducing research, experimental protocols are fundamental parts of the research record. This thesis addresses the problem of providing accurate, machine readable and configurable descriptions for the experimental protocols; this research also explores the use of semantic technology in the publication workflow for experimental protocols.

Being able to review the data makes it possible to evaluate whether the analysis and conclusions drawn are accurate. However, it does little to validate the quality and accuracy of the data itself. Evaluating research implies being able to obtain similar, if not identical, results. The data must be available, so does the experimental protocol detailing the methodology followed to derive the data. Journals and funders are now asking for datasets to be publicly available; there have been several efforts addressing the problem of data repositories e.g. Dryad [6], Figshare [7], DataCite [8]; if data must be public and available, shouldn't researchers be held to the same principle when it comes to methodologies? Researchers have studied the problem of reproducibility from various angles; however, fewer have proposed reporting structures for experimental protocols. Fewer have built their approaches upon exhaustive studies of published research using knowledge engineering methods. Freedman et al. [9] and Baker et al. [10] have studied and identified some of the sources for experimental irreproducibility, namely: I) poor study design and analytical procedures, II) Reagent variability, and variability in other materials used, III) Incomplete protocol reporting, and IV) Poor, or inexistent, access to the data and report of results. When reporting reagents and equipment, researchers sometimes include catalog numbers

and experimental parameters while in other occasions they refer to these items in a generic manner, e.g., “*Dextran sulfate, Sigma-Aldrich*” [11]. Having this information is important because reagents usually vary in terms of purity, yield, pH, hydration state, grade, and possibly additional biochemical or biophysical features. Similarly, experimental protocols often include ambiguities such as “*Store the samples at room temperature until sample digestion.*” [12]; but, how many Celsius degrees? What is the estimated time for digesting the sample? Having this information available not only saves time and effort, it also makes it easier for researchers to reproduce experimental results. Adequate and comprehensive reporting facilitates reproducibility [9], [10].

This thesis focuses on the third cause of irreproducibility, a incomplete protocol reporting. An experimental protocol is a sequence of tasks and operations executed to perform experimental research. Protocols, as previously stated, often include references to critical steps, troubleshooting and tips, as well as a list of materials (samples, instruments, reagents, etc.), participating in the execution of steps. If the materials are not properly reported in the protocols, then, recreating the experiment becomes increasingly difficult and prone to error. In this sense, the second cause of irreproducibility, variability in materials used is also considered in this study.

This work investigates how to formally represent experimental protocols; understanding these as domain-specific workflows embedded within documents. By representing the knowledge embedded within these documents, this research facilitates the aggregation of the workflow and the data –the protocol describes how the data was produced; thus making it simpler to systematically reuse, evaluate, share and discover experimental protocols. By the same vein, the SMART Protocols approach, that taken throughout this thesis, makes data more reusable, as it provides important context that allows researchers to evaluate whether the approaches followed were methodologically sound.

Similarly, throughout this thesis the aggregative nature of scientific documents is studied; scientific publications aggregate data by encompassing it within a persuasive narrative. The aggregation is highly federated; authors reference external sources, analyze data elsewhere and summarize over the document, archive and publish methods, data and processes over heterogeneous resources and using a myriad of formats. Experimental protocols are part of this aggregative ecosystem; the workflows generate data that is supporting the narrative and making it possible to replicate experiments. This research investigates the use of semantic web technology to support the aggregation of meaningful parts within the context of experimental protocols. The approach conceived by the author is simple, instead of supporting *post-mortem* operations over published documents, why not making it possible to have a document that is to be born semantic, interoperable and, thought as an aggregator within a web-of-data publishing workflow?

1.2 Motivation

Reproducibility, although an elusive concept, helps researchers to verify results; it also allows others to build on previous experiments by making it possible to reuse, with a high degree of confidence, that by reproducing an experiment results will be similar -if not equal. It is at the core of experimental research; however, it is difficult to achieve; Freedman et al., [9] have reported that 50% of reported research is not reproducible.

As experiments become increasingly complex in the combination of technologies being used, reporting structures become less accurate in their descriptions. Also, the complex ecosystem of technologies make it difficult for existing publications to facilitate experimental reproducibility. Researchers often rely on the data as it is described in papers. But, sometimes the data description is incomplete; critical information to understand the workflow of an experiment is often excluded. For example, descriptions of column names in tabular data, libraries used in computational experiments, algorithms used in machine learning, proprietary software used to view files, information about the sample, etc. is very often missing or incomplete.

Funders, award-granting institutions, and peer-reviewed journals are taking notice of the general lack of reproducibility plaguing many scientific communities. Websites such as Retraction Watch (*Retraction Watch*) have sprung up to track which journal articles are being retracted. Very often these retractions are related to issues with reproducing the data based on the information provided by authors. These situations may be due to malpractice but they may also be the product of poor experimental reporting. One example that illustrates a case of malpractice involves Susana Gonzalez, a Spanish regenerative medicine scientist who lost a grant of 1.9 million of euros from the EU public funder ERC (European Research Council) and her position as group leader at the Centro Nacional de Investigaciones Cardiovasculares (CNIC) in Madrid. Her fifth publication in the scientific journal “Molecular and Cellular Biology” was retracted in 2017; this was due to digital manipulation of data (*fraude en ciencia española; For better science*) [13]. Another example of inconsistencies in published data involved a team of scientists that included Linda B. Buck, who shared the 2004 Nobel Prize in Physiology or Medicine. The researchers have retracted a scientific paper after other scientists could not reproduce the published findings. Fortunately, the paper is unrelated to her prize (*Nobel Winner Retracts Research Paper* [14]).

Experimental irreproducibility is a consequence of the inability to get the same or, statistically similar results. These differences can occur when there is variability across laboratories executing an experiment. There may be differences in methods, sample treatment, or reagents used; differences may also be due to the training of staff scientists. Independently from the causes of experimental irreproducibility, researchers should always be able to understand how data was produced, what sample treatments were there involved, what experimental methods were applied, what reagents, appliances and equipment were used. Files may go missing, protocols may be under reported, critical information such as sample or reagent data may be incomplete. These are situations that are usually related to inadequate reporting, a frequent cause of poor reproducibility. The focus has so far been on having data availability as a proxy for experimental reproducibility; being able to review the data makes it possible to evaluate whether the analysis and conclusions drawn are accurate. However, it does little to validate the quality and accuracy of the data itself. Evaluating research implies being able to obtain similar, if not identical, results. The data must be available, so does the experimental protocol detailing the methodology followed to derive the data. This research work aims to facilitate adequate reporting of experimental protocols and by doing so making it easier for researchers to specify the bundle data-protocol. Malpractice will always be possible; however, not having well defined reporting structures with the appropriate semantics should not be an excuse for experimental irreproducibility.

The experimental workflow, as well as details about materials and methods, are usually described in experimental protocols. An experimental protocol is a sequence of tasks and operations executed to perform experimental research in biological and

biomedical areas, e.g. biology, genetics, immunology, neurosciences, virology. Protocols often include references to critical steps, troubleshooting and tips, as well as a list of materials (samples, instruments, reagents, etc.), participating in the execution of the steps.

Protocols are part of the experimental record; they are widely used across laboratories around the world -big and small and with various degrees of infrastructure. Although central for the experimental record and widely used, reporting protocols remains highly idiosyncratic. Moreover, in spite of their workflow nature, the publication of experimental protocols remains largely based on a static narrative; for instance, the workflow does not have any machine processable components. Interestingly, although these documents are highly structured, have clearly identifiable entities with easy-to-establish- relations to the web of data, we continue to publish them using the same technology as any other document. Adequate reporting and semantic publishing of experimental protocols could help to improve reproducibility, bridge the gap between scientific documents and the web of data and, exemplify the production of executable documents.

Researchers execute workflows, these are represented in protocols and, by doing so data is produced. Again, there have been several efforts delivering infrastructure for data repositories. However, having data available does not imply having reproducible data. If data must be available, why not protocols?

1.3 Problem statement

This research work addresses the following challenges: i) incomplete description and variability in the content of protocols, ii) lack of machine readable protocols, ideally these should be equally intelligible for humans and machines, iii) limited support for the generation of semantic protocols. *“How to semantically represent experimental protocols?, How to generate semantic protocols?”*

In order to address these challenges and give an answer to the research question, the following objectives have been specified.

Objective 1: To design a guideline that formally represents bibliographic (e.g. title, author, version), and rhetorical components (e.g. purpose, materials, and procedure) from experimental protocols in life science.

Objective 2: To develop an ontology that represents the document and workflow aspects of the protocol.

Objective 3: To facilitate finding specific protocols based on common data elements in experimental protocols.

Objective 4: To publish experimental protocols as linked data so that the relation between reagents, samples and instruments with the larger web, e.g. pubchem, is possible.

Objective 5: To facilitate automatic entity recognition by using semantics and NLP techniques.

Objective 6: To facilitate the generation of semantic documents for experimental protocols.

1.4 Contributions of this thesis

The following are the contributions of this dissertation:

1 This thesis has delivered a comprehensive guideline for reporting experimental protocols, see chapter 2. Other guidelines focus on specific methods and techniques, e.g. Polymerase chain reaction (PCR); the SP guidelines may be specialized by these more particular guidelines. In this way the reporting structure for the experimental protocol results from the aggregation of a general non-method specific guideline, the SP, and that representing the particular method that was applied, e.g. PCR.

2 The SP ontology, see chapter 3, represents experimental protocols; it reuses existing ontologies and also specifies its own ontological structures. An interesting byproduct of this work is also presented in this chapter; the Sample Instrument Reagent Objective (SIRO) model, which represents the minimal common information shared across experimental protocols. The ontology was evaluated against competency questions so linked data was published in order to express the competency questions as SPARQL queries. Thus, also delivering a set of experimental protocols as linked data -to the best of my knowledge the first linked data set representing full text protocols.

3 The BioSchemas effort brings together the biomedical community in the definition of schema.org compliant vocabularies. In this fourth chapter the specification for laboratory protocols as well as the methodology that was followed is presented. Through the first chapters the semantics for experimental protocols was formalized; the proposed specification is an important byproduct of the initial chapters. It represents early the interest of the community and the adoption of this research.

3 The BioH annotation tooling, chapter 5, and the lessons learned deliver a reusable infrastructure that supports target specific annotation. It makes it possible to extend ontologies with specific terminology gathered by annotating documents. The tools and the lessons learned facilitate applying this method to other domains.

4 The SP gold standard, chapter 6, is the first and to the best of my knowledge the only gold standard for experimental protocols. It focuses on the identification of samples, instruments, reagents and experimental actions. Developing highly effective tools to automatically detect biological concepts depends on the availability of high quality annotated corpus

5 The SP publication platform, chapter 7. This contribution integrates all the previous ones; it delivers an end user semantic publication platform for experimental protocols. The SP approach facilitates the generation of the semantic document from the beginning of the publication workflow. Thus, making semantics at birth a reality for a scholarly document.

Throughout the development of this work special emphasis was placed in studying cases for which this work could have a direct impact. The search for and interest in real scenarios allowed me to extensively collaborate with other groups such as the EBI-ELIXIR (European Institute) Bioschemas working group, the Biotechnology group at the CIAT (Center for International Tropical Agriculture) and the Ontology Development Group at the Department of Medical Informatics and Clinical Epidemiology at Oregon Health and Science University.

1.4.1 Research Outcomes related to this Investigation

Awards

- Finalist in “actúaloop, Ideas Competition for Innovation in Research Social Networks”. June 23th of 2016 [15].
Title: Formalization of experimental protocols (SMART Protocols)
Description of the idea: SMART Protocols allow researchers to accurately generate and retrieve information from experimental protocols. It makes possible for publishers to expose ready-to-use data/content over the web as well as to deliver a content-based recommendation service for researchers.
- Best poster award in the International Conference on Biomedical Ontologies (ICBO 2015).
Title: Using semantics and NLP in the SMART Protocols.
Authors: Olga Giraldo, Alexander Garcia and Oscar Corcho.
- Internship sponsored by Elsevier – Oregon Health and Science University (OHSU).
Description: exploring products and standards/ontologies for experimental protocols.
- FORCE11, the Future of Research Communication and e-Scholarship (2013) [16].
Description: our work was selected as one of the fourteen best ideas about “Vision of the Future”
Title: Using nanopublications to model laboratory protocols.
Author: Olga Giraldo

Journal Papers

- Giraldo O, Garcia A, Corcho O. (2018) “A guideline for reporting experimental protocols in life sciences”. PeerJ 6:e4795 <https://doi.org/10.7717/peerj.4795>
- Giraldo, O., García, A., López, F., & Corcho, O. (2017). “Using semantics for representing experimental protocols”. Journal of biomedical semantics, 8 (1), 52. doi:10.1186/s13326-017-0160-y
- Garcia A, Lopez F, Garcia L, Giraldo O, Bucheli V, Dumontier M. 2018. Biotea: semantics for Pubmed Central. PeerJ 6:e4201 <https://doi.org/10.7717/peerj.4201>

Conferences and Workshops

- Leyla Jael García Castro, Olga X. Giraldo, Alexander Garcia and Dietrich Rebholz-Schuhmann. Biotea and Bioschemas knowledge graph. Submitted to the Biomedical Linked Annotation Hackathon. December, 13th/2018.
- Leyla Jael García Castro, Olga X. Giraldo, Alexander Garcia, Michel Dumontier, Bioschemas Community. Bioschemas: schema.org for the Life Sciences. Semantic Web Applications and Tools for Health Care and Life Sciences, SWAT4LS 2017. Rome, Italy, December 4-7, 2017.

- Olga Giraldo, Alexander Garcia, Tazro Ohta and Federico Lopez (2017). Annotating the SIRO model and discovering experimental protocols. Proposal at Biomedical Linked Annotation Hackathon 3, Tokyo, Japan, 16-20 January 2017.
- Olga Giraldo, Alexander García and Oscar Corcho (2016). Using Semantics and NLP in the SMART Protocols Repository. Poster accepted at FORCE11 (2016), Portland, Oregon, USA. April 17-19, 2016
- Olga Giraldo, Alexander Garcia, Jose Figueredo, and Oscar Corcho (2015). Using Semantics and NLP in Experimental Protocols. Paper accepted at Semantic Web Applications and Tools for Life Sciences 2015 (SWAT4LS 2015), Cambridge, England. December 7-10th, 2015.
- Olga Giraldo, Alexander García and Oscar Corcho (2015). Using Semantics and NLP in the SMART Protocols Repository. Poster accepted at International Conference on Biomedical Ontology 2015 (ICBO 2015), Lisbon, Portugal. July 27 - 30, 2015
- Olga Giraldo, Alexander Garcia and Oscar Corcho. (2014). SMART Protocols: SeMAnTic RepresenTation for Experimental Protocols. Paper accepted at the LISC, an International Semantic Web Conference (ISWC2014) Workshop, Riva del Garda, Trentino, Italy

1.5 Outline of this Thesis

This thesis is organized into a series of chapters addressing aspects related to the semantic representation of experimental protocols and the use of such semantics. This work begins by introducing the problem, motivation, and structure of the document, see Chapter 1. Chapter 2 "*A Guideline for Reporting Experimental Protocols in Life Sciences*" begins by addressing the problem of using a guideline to define and characterize important information elements in experimental protocols. A comprehensive reusable reporting structure and guideline was the main outcome.

Chapter 3 "*Using Semantics for Representing Experimental Protocols*" addresses the problem of having an ontology to represent experimental protocols. The resulting ontology represents the protocol as a workflow with domain specific knowledge embedded within a document. It also facilitates the production of linked data for full text protocols. In addition, in this chapter the Sample Instrument Reagent Objective minimal information model is also presented. Chapter 4, "*Laboratory Protocols in Bioschemas*" presents the contribution of this research to the Bioschemas effort. Chapters 2 through 4 present different layers of semantics, starting by a standardized checklist with data elements well defined, chapter 2, moving into an ontology, chapter 3, and finishing with a vocabulary for search engine optimization, chapter 4. These layers are interconnected and influenced each other. For instance the SIRO model, see chapter 3, is the basis for the LabProtocol profile developed for Bioschemas and presented in detail in chapter 4.

In order to gather terminology related to specifics within the protocol, e.g. samples, instruments, reagents and experimental actions, the BioH annotation tool was developed, see Chapter 5 "*BioH, The Smart Protocols Annotation Tool*". The annotation tool was used through chapter 6 The terminology thus gathered was organized in gazetteers; these were then used in the SP publication platform, see Chapter 6 "*Generating a Gold Standard Corpus for Experimental Protocols*"; in this chapter the rationale for developing such resource is explained. The gold standard made it possible to

build the semantic gazetteers and the rules for the automatic annotation of rules in the protocols.

Chapters 6 and 7, "*Semantics at Birth, the SMART Protocols Publication Platform*" are particularly important because they bring together the previous work and aim to deliver a general resource, e.g. the gold standard as well as an end user tool, e.g. the semantic publication platform. Chapter 6 "*Generating a Gold Standard Corpus of Experimental Protocols*" makes extensive use of the BioH annotation tool in order to build a gold standard for experimental protocols. Chapter 7 "*Semantics at Birth, the SMART Protocols Publication Platform*" makes extensive use of all the research presented in this work; it delivers a semantic publication infrastructure specially tailored for experimental protocols. As it relies on semantics, customizing this application for other types of documents does not represent a significant challenge.

Fig 1.1 illustrates the structure of this thesis.

FIGURE 1.1: An overview of the structure of this thesis

Bibliography

- [1] *What is the difference between repeatability and reproducibility?* labmate online, 2014. [Online]. Available: <https://www.labmate-online.com/news/news-and-views/5/breaking-news/what-is-the-difference-between-repeatability-and-reproducibility/30638>.
- [2] H. E. Plesser, "Reproducibility vs. replicability: A brief history of a confused terminology", *Frontiers Media S.A.*, vol. 11, p.76, 2018. DOI: <https://dx.doi.org/10.3389/fninf.2017.00076>.
- [3] J. P.A. I. Steven N. Goodman Daniele Fanelli, "What does research reproducibility mean?", *Science Translational Medicine*, vol. 8, p. 341, 2016. DOI: <http://doi.org/10.1126/scitranslmed.aaf5027>.
- [4] F. D. Justin Kitzes Daniel Turek, "The practice of reproducible research", *Science Translational Medicine*, p. 368, 2017.
- [5] L. Wissler, M. Almashraee, D. Monett, and A. Paschke, "The gold standard in corpus annotation", Jun. 2014. DOI: 10.13140/2.1.4316.3523.
- [6] Dryad, *Dryad*, Retrieved on 07/07/2017, 2017. [Online]. Available: <http://datadryad.org/>.
- [7] figshare, *Figshare*, Retrieved on 07/07/2017, 2017. [Online]. Available: <http://figshare.com>.
- [8] DataCite, *Datacite*, Retrieved on 07/07/2017 from <https://datacite.org/>, 2017. [Online]. Available: <https://datacite.org/>.
- [9] L. Freedman, G Venugopalan, and R Wisman, "Reproducibility2020: Progress and priorities [version 1; referees: 2 approved]", *F1000Research*, vol. 6, no. 604, 2017. DOI: 10.12688/f1000research.11334.1.
- [10] M Baker, "1,500 scientists lift the lid on reproducibility", *Nature*, vol. 53, no. 7604, 2016. DOI: 10.1038/533452a.
- [11] A. Karlgren, J. Carlsson, N. Gyllenstrand, U. Lagercrantz, and J. F. Sundström, "Non-radioactive in situ hybridization protocol applicable for norway spruce and a range of plant species", *Journal of Visualized Experiments : JoVE*, no. 26, p. 1205, 2009. DOI: 10.3791/1205. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3148633/>.
- [12] F Brandenburg, H Schoffman, N Keren, and M. Eisenhut, "Determination of mn concentrations in synechocystis sp. pcc6803 using icp-ms", *Bio-protocol*, vol. 7, no. 23, pp. 244–258, 2002. DOI: 10.21769/BioProtoc.2623. [Online]. Available: <https://bio-protocol.org/e2623>.
- [13] *Ciencia:el mayor fraude de la ciencia española sigue creciendo: Un nuevo estudio a la hoguera*, 2017. [Online]. Available: https://www.elconfidencial.com/tecnologia/ciencia/2017-09-18/mucho-mayor-escandalo-ciencia-espanola_1445736/.

- [14] *Nobel winner retracts research paper - the new york times*, 2008. [Online]. Available: <https://www.nytimes.com/2008/03/07/science/07retractw.html>.
- [15] *Changing research, one app at a time: Actúalooop awards – science research news | frontiers*, 2016. [Online]. Available: <https://blog.frontiersin.org/2016/06/07/changing-research-one-app-at-a-time-actualoop-awards/>.
- [16] *Visions for the future | force11*. [Online]. Available: <https://www.force11.org/Visions>.

Chapter 2

A Guideline for Reporting Experimental Protocols in Life Sciences

Experimental protocols are key when planning, doing and publishing research in many disciplines, especially in relation to the reporting of materials and methods. However, they vary in their content, structure and associated data elements. This article presents a guideline for describing key content for reporting experimental protocols in the domain of life sciences, together with the methodology followed in order to develop such guideline. As part of our work, we propose a checklist that contains 17 data elements that we consider fundamental to facilitate the execution of the protocol. These data elements are formally described in the SMART Protocols ontology. By providing guidance for the key content to be reported, we aim (1) to make it easier for authors to report experimental protocols with necessary and sufficient information that allow others to reproduce an experiment, (2) to promote consistency across laboratories by delivering an adaptable set of data elements and, (3) to make it easier for reviewers and editors to measure the quality of submitted manuscripts against an established criteria. Our checklist focuses on the content, what should be included. Rather than advocating a specific format for protocols in life sciences, the checklist includes a full description of the key data elements that facilitate the execution of the protocol.

2.1 Introduction

Experimental protocols are fundamental information structures that support the description of the processes by means of which results are generated in experimental research [1], [2]. Experimental protocols, often as part of “Materials and Methods” in scientific publications, are central for reproducibility; they should include all the necessary information for obtaining consistent results [3], [4]. Although protocols are an important component when reporting experimental activities, their descriptions are often incomplete and vary across publishers and laboratories. For instance, when reporting reagents and equipment, researchers sometimes include catalog numbers and experimental parameters; they may also refer to these items in a generic manner, e.g., “*Dextran sulfate, Sigma-Aldrich*” [5]. Having this information is important because reagents usually vary in terms of purity, yield, pH, hydration state, grade, and possibly additional biochemical or biophysical features. Similarly, experimental protocols often include ambiguities such as “*Store the samples at room temperature until sample digestion.*” [6]; but, how many Celsius degrees? What is the estimated time for digesting the sample? Having this information available not only saves time and effort, it also makes it easier for researchers to reproduce experimental results; adequate and comprehensive reporting facilitates reproducibility [2], [7].

Several efforts focus on building data storage infrastructures, e.g., 3TU. Datacenterum [8], CSIRO Data Access Portal [9], Dryad [10], figshare [11], Dataverse [12] and Zenodo [13]. These data repositories make it possible to review the data and evaluate whether the analysis and conclusions drawn are accurate. However, they do little to validate the quality and accuracy of the data itself. Evaluating research implies being able to obtain similar, if not identical results. Journals and funders are now asking for datasets to be publicly available for reuse and validation. Fully meeting this goal requires datasets to be endowed with auxiliary data providing contextual information e.g., methods used to derive such data [14], [15]. If data must be public and available, shouldn’t methods be equally public and available?

Illustrating the problem of adequate reporting, Morher et al. [16] have pointed out that fewer than 20% of highly-cited publications have adequate descriptions of study design and analytic methods. In a similar vein, Vasilevsky et al. [17] showed that 54% of biomedical research resources such as model organisms, antibodies, knockdown reagents (morpholinos or RNAi), constructs, and cell lines are not uniquely identifiable in the biomedical literature, regardless of journal Impact Factor. Accurate and comprehensive documentation for experimental activities is critical for patenting, as well as in cases of scientific misconduct. Having data available is important; knowing how the data were produced is just as important. Part of the problem lies in the heterogeneity of reporting structures; these may vary across laboratories in the same domain. Despite this variability, we want to know which data elements are common and uncommon across protocols; we use these elements as the basis for suggesting our guideline for reporting protocols. We have analyzed over 500 published and non-published experimental protocols, as well as guidelines for authors from journals publishing protocols. From this analysis we have derived a practical adaptable checklist for reporting experimental protocols.

Efforts such as the Structured, Transparent, Accessible Reporting (STAR) initiative [18], [19] address the problem of structure and standardization when reporting methods. In a similar manner, The Minimum Information about a Cellular Assay (MIACA) [20], The Minimum Information about a Flow Cytometry Experiment (MI-FlowCyt) [21] and many other “*minimal information*” efforts deliver minimal data elements describing specific types of experiments. Soldatova et al, [22], [23] proposes

the EXACT ontology for representing experimental actions in experimental protocols; similarly, Giraldo et al, [1] proposes the **SeM**Antic **R**epresen**T**ation of Protocols ontology (henceforth SMART Protocols Ontology) an ontology for reporting experimental protocols and the corresponding workflows. These approaches are not minimal, they aim to be comprehensive in the description of the workflow, parameters, sample, instruments, reagents, hints, troubleshooting, and all the data elements that help to reproduce an experiment and describe experimental actions.

There are also complementary efforts addressing the problem of identifiers for reagents and equipment; for instance, the Resource Identification Initiative (RII) [24], aims to help researchers sufficiently cite the key resources used to produce the scientific findings. In a similar vein, The Global Unique Device Identification Database (GUDID) [25] has key device identification information for medical devices that have Unique Device Identifiers (UDI); the Antibody Registry [26], gives researchers a way to universally identify antibodies used in their research and also the Addgene web-application [27], makes it easy for researchers to identify plasmids. Having identifiers make it possible for researchers to be more accurate in their reporting by unequivocally pointing to the resource used or produced. The Resource Identification Portal [28], makes it easier to navigate through available identifiers, researchers can search across all the sources from a single location.

In this paper, we present a guideline for reporting experimental protocols; we complement our guideline with a machine-processable checklist that helps researchers, reviewers and editors to measure the completeness of a protocol. Each data element in our guideline is represented in the SMART Protocols Ontology. This paper is organized as follows: we start by describing the materials and methods used to derive the resulting guidelines. In the "Results" section, we present examples indicating how to report each data element; a machine readable checklist in the JavaScript Object Notation (JSON) format is also presented in this section. We then discuss our work and present the conclusions.

2.2 Materials and Methods

2.2.1 Materials

We have analyzed: i) guidelines for authors from journals publishing protocols [29], ii) our corpus of protocols [30], iii) a set of reporting structures proposed by minimal information projects available in the FairSharing catalog [31] and, iv) relevant biomedical ontologies available in BioPortal [32] and Ontobee [33]. Our analysis was carried out by a domain expert, Olga Giraldo; she is an expert in text mining and biomedical ontologies with over ten years of experience in laboratory techniques. All the documents were read, and then data elements, subject areas, materials (e.g. sample, kits, solutions, reagents, etc), and workflow information were identified. Resulting from this activity we established a baseline terminology, common and non common data elements, as well as patterns in the description of the workflows (e.g. information describing the steps and the order for the execution of the workflow).

i) Instructions for authors from analyzed journals.

Publishers usually have instructions for prospective authors; these indications tell authors what to include, the information that should be provided, and how it should be reported in the manuscript. In Table 6.1 we present the list of guidelines that were analyzed.

Journal	Guidelines for authors
BioTechniques (BioTech)	[29]
CSH protocols (CSH)	[34]
Current Protocols (CP)	[35]
Journal of Visualized Experiments (JoVE)	[36]
Nature Protocols (NP)	[37]
Springer Protocols (SP)	[38]
MethodsX	[39]
Bio-protocols (BP)	[40]
Journal of Biological Methods (JBM)	[41]

TABLE 2.1: Guidelines for reporting experimental protocols.

ii) Corpus of protocols.

Our corpus includes 530 published and unpublished protocols. Unpublished protocols (75 in total) were collected from four laboratories located at the International Center for Tropical Agriculture (CIAT) [42]. The published protocols (455 in total) were gathered from the repository “Nature Protocol Exchange” [43] and from 11 journals, namely: BioTechniques, Cold Spring Harbor Protocols, Current Protocols, Genetics and Molecular Research [44], JoVE, Plant Methods [45], Plos One [46], Springer Protocols, MethodsX, Bio-Protocol and the Journal of Biological Methods. The analyzed protocols comprise areas such as cell biology, molecular biology, immunology, and virology. The number of protocols from each journal is presented in Table 6.2.

Source	Number of protocols
BioTechniques (BioTech)	16
CSH protocols (CSH)	267
Current Protocols (CP)	31
Genetics and Molecular Research (GMR)	5
Journal of Visualized Experiments (JoVE)	21
Nature Protocols Exchange (NPE)	39
Plant Methods (PM)	12
Plos One (PO)	5
Springer Protocols (SP)	5
MethodsX	7
Bio-protocols (BP)	40
Journal of Biological Methods (JBM)	7
non-published protocols from CIAT	75

TABLE 2.2: Corpus of protocols analyzed.

iii) Minimum information standards and Ontologies.

We analyzed minimum information standards from the FairSharing catalog, e.g., MIAPPE [47], MIARE [48] and MIQE [49]. See Table 6.3 for the complete list of minimum information models that we analyzed.

We paid special attention to the recommendations indicating how to describe specimens, reagents, instruments, software and other entities participating in different types of experiments. Ontologies available at Bioportal and Ontobee were

Standards	Description
Minimum Information about Plant Phenotyping Experiment (MIAPPE) CIMR: Plant Biology Context [50]	A reporting guideline for plant phenotyping experiments. A standard for reporting metabolomics experiments.
The Gel Electrophoresis Markup Language (GelML)	A standard for representing gel electrophoresis experiments performed in proteomics investigations.
Minimum Information about a Cellular Assay (MIACA)	A standardized description of cell-based functional assay projects.
Minimum Information About an RNAi Experiment (MIARE)	A checklist describing the information that should be reported for an RNA interference experiment.
The Minimum Information about a Flow Cytometry Experiment (MI-FlowCyt)	This guideline describes the minimum information required to report flow cytometry (FCM) experiments
Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE)	This guideline describes the minimum information necessary for evaluating qPCR experiments.
ARRIVE (Animal Research: Reporting of In Vivo Experiments) [51]	Initiative to improve the standard of reporting of research using animals.

TABLE 2.3: Minimum Information Standards analyzed.

also considered; we focused on ontologies modeling domains, e.g., bioassays (BAO), protocols (EXACT), experiments and investigations (OBI). We also focused on those modeling specific entities, e.g., organisms (NCBI Taxon), anatomical parts (UBERON), reagents or chemical compounds (ERO, ChEBI), instruments (OBI, BAO, EFO). The list of analyzed ontologies is presented in Table 2.4.

2.2.2 Methods for developing this guideline

Developing the guideline entailed a series of activities; these were organized in the following stages: i) analysis of guidelines for authors, ii) analysis of protocols, iii) analysis of Minimum Information (MI) standards and ontologies, and iv) evaluation of the data elements from our guideline. For a detailed representation of our workflow, see Figure 2.1

Analyzing guidelines for authors

We manually reviewed instructions for authors from nine journals as presented in Table 6.1. In this stage (step A in Figure 2.1), we identified bibliographic data elements classified as “desirable information” in the analyzed guidelines. See Table 2.5.

In addition, we identified the rhetorical elements. These have been categorized in the guidelines for authors as: i) required information (R), must be submitted with the manuscript; ii) desirable information (D), should be submitted if available, and; iii) optional (O) or extra information. See Table 2.6 for more details.

Ontology	Description
The Ontology for Biomedical Investigations (OBI) [52]	An ontology for the description of life-science and clinical investigations.
The Information Artifact Ontology (IAO) [53]	An ontology of information entities.
The ontology of experiments (EXPO) [54]	An ontology about scientific experiments.
The ontology of experimental actions (EXACT)	An ontology representing experimental actions.
The BioAssay Ontology (BAO) [55]	An ontology describing biological assays.
The Experimental Factor Ontology (EFO) [56]	The ontology includes aspects of disease, anatomy, cell type, cell lines, chemical compounds and assay information.
eagle-i resource ontology (ERO)	An ontology of research resources such as instruments, protocols, reagents, animal models and biospecimens.
NCBI taxonomy (NCBITaxon) [57]	An ontology representation of the NCBI organismal taxonomy.
Chemical Entities of Biological Interest (ChEBI) [58]	Classification of molecular entities of biological interest focusing on 'small' chemical compounds.
Uberon multi-species anatomy ontology (UBERON) [59]	A cross-species anatomy ontology covering animals and bridging multiple species-specific ontologies.
Cell Line Ontology (CLO) [60], [61]	The ontology was developed to standardize and integrate cell line information.

TABLE 2.4: Ontologies analyzed.

Bibliographic data elements	BioTech	NP	CP	JoVE	CSH	SP	BP	MethodsX	JBIM
title/name	Y	Y	Y	Y	Y	Y	Y	Y	Y
author name	Y	Y	Y	Y	Y	Y	Y	Y	Y
author identifier (e.g., orcid)	N	N	N	N	N	N	N	N	N
protocol identifier (DOI)	Y	Y	Y	Y	Y	Y	Y	Y	Y
protocol source (retrieved from, modified from)	N	Y	N	N	N	N	N	N	N
updates (corrections, retractions or other revisions)	N	N	N	N	N	N	N	N	N
references/related publications	Y	Y	Y	Y	Y	Y	Y	Y	Y
categories or keywords	Y	Y	Y	Y	Y	Y	Y	Y	Y

TABLE 2.5: Bibliographic data elements from guidelines for authors.

Y= datum considered as "desirable information" if this is available,

N= datum not considered in the guidelines.

Analyzing the protocols.

In 2014, we started by manually reviewing 175 published and unpublished protocols; these were from domains such as cell biology, biotechnology, virology, biochemistry and pathology. From this collection, 75 are unpublished protocols and

FIGURE 2.1: Methodology Workflow.

thus not available in the dataset for this paper. These unpublished protocols were collected from four laboratories located at the CIAT. In 2015, our corpus grew to 530; we included 355 published protocols gathered from one repository and eleven journals as listed in Table 6.2. Our corpus of published protocols is: i) identifiable, i.e. each document has a Digital Object Identifier (DOI) and ii) in disciplines and areas related to the expertise provided by our domain experts, e.g., virology, pathology, biochemistry, biotechnology, plant biotechnology, cell biology, molecular and developmental biology and microbiology. In this stage, step B in Figure 2.1, we analyzed the content of the protocols; theory vs. practice was our main concern. We manually verified if published protocols were following the guidelines; if not, *what was missing, what additional information was included?* We also reviewed common data elements in unpublished protocols.

Analyzing Minimum Information Standards and ontologies

Biomedical sciences have an extensive body of work related to minimum information standards and reporting structures, e.g., those from the FairSharing initiative. We were interested in determining whether there was any relation to these resources. Our checklist includes the data elements that are common across these resources. We manually analyzed standards such as MIQE, used to describe qPCR assays; we also looked into MIACA, it provides guidelines to report cellular assays; ARRIVE, which provides detailed descriptions of experiments on animal models and MIAPPE, addressing the descriptions of experiments for plant phenotyping. See Table 6.3 for a complete list of the standards that we analyzed. Metadata, data, and reporting structures in biomedical documents are frequently related to ontological concepts. We

Rhetorical/Discourse Elements	Bio-Tech	NP	CP	JoVE	CSH	SP	BP	Meth-odsX	JBM
Description of the protocol (objective, range of applications where the protocol can be used, advantages, limitations)	D	D	D	D	D	D	D	D	D
Description of the sample tested (name; ID; strain, line or eco-type; developmental stage; organism part; growth conditions; treatment type; size)	NC	NC	D	NC	NC	NC	NC	NC	NC
Reagents (name, vendor, catalog number)	R	D	D	D	R	D	R	NC	D
Equipment (name, vendor, catalog number)	R	D	D	D	R	D	R	NC	D
Recipes for solutions (name, final concentration, volume)	R	D	D	D	D	D	R	NC	D
Procedure description	R	R	R	D	R	R	R	R	D
Alternatives to performing specific steps	NC	NC	D	D	NC	D	NC	NC	NC
Critical steps	R	NC	D	NC	NC	NC	NC	NC	NC
Pause point	R	NC	NC	O	D	NC	NC	NC	NC
Troubleshooting	R	O	R	O	D	D	NC	NC	D
Caution/warnings	NC	NC	R	O	NC	D	NC	NC	D
Execution time	NC	O	D	NC	NC	D	NC	NC	NC
Storage conditions (reagents, recipes, samples)	R	NC	R	D	D	D	NC	NC	NC
Results (figure, tables)	R	NC	R	R	D	R	D	NC	D

TABLE 2.6: Rhetorical/Discourse elements from guidelines for authors. R= Required information; NC= Not Considered in guidelines; D= Desirable information; O= Optional information.

also looked into relations between data elements and biomedical ontologies available in BioPortal and Ontobee. We focused on ontologies representing materials that are often found in protocols; for instance, organisms, anatomical parts (e.g., CLO, UBERON, NCBI Taxon), reagents or chemical compounds (e.g., ChEBI, ERO), and equipment (e.g., OBI, BAO, EFO). The complete list of the ontologies that we analyzed is presented in Table 2.4.

Generating the first draft

The first draft is the main output from the initial analysis of instructions for authors, experimental protocols, MI standards and ontologies, see (step D in Figure 2.1). The data elements were organized into four categories: bibliographic data elements such as title, authors; descriptive data elements such as purpose, application; data elements for materials, e.g. sample, reagents, equipment; and data elements for procedures, e.g. critical steps, Troubleshooting. The role of the authors, provenance and properties describing the sample (e.g. organism part, amount of the sample, etc.) were considered in this first draft. In addition properties like "name", "manufacturer or vendor" and "identifier" were proposed to describe equipment, reagents and kits.

Evaluation of data elements by domain experts

This stage entailed three activities. The first activity was carried out at CIAT with the participation of 19 domain experts in areas such as virology, pathology, biochemistry, and plant biotechnology. The input of this activity was the checklist V. 0.1 (see step E in Figure 2.1). This evaluation focused on “*What information is necessary and sufficient for reporting an experimental protocol?*”; the discussion also addressed data elements that were not initially part of guidelines for authors -e.g., consumables. The result of this activity was the version 0.2 of the checklist; domain experts suggested to use an online survey for further validation. This survey was designed to enrich and validate the checklist V. 0.2. We used a Google survey that was circulated over mailing lists; participants did not have to disclose their identity (see step F in Figure 2.1). A final meeting was organized with those who participated in workshops, as well as in the survey (23 in total) to discuss the results of the online poll. The discussion focused on the question: *Should the checklist include data elements not considered by the majority of participants?* Participants were presented with use cases where infrequent data elements are relevant in their working areas. It was decided to include all infrequent data elements; domain experts concluded that this guideline was a comprehensive checklist as opposed to a minimal information. Also, after discussing infrequent data elements it was concluded that the importance of a data element should not bear a direct relation to its popularity. The analogy used was that of an editorial council; some data elements needed to be included regardless of the popularity as an editorial decision. The output of this activity was the checklist V. 1.0. The survey and its responses are available at [62]. This current version includes a new bibliographic element “license of the protocol”, as well as the property “equipment configuration” associated to the datum equipment. The properties: alternative, optional and parallel steps were added to describe the procedure. In addition, the datum “PCR primers” was removed from the checklist, it is specific and therefore should be the product of a community specialization as opposed to part of a generic guideline.

2.3 Results

Our results are summarized in table 2.7; it includes all the data elements resulting from the process illustrated in Figure 2.1. We have also implemented our checklist as an online tool that generates data in the JSON format and presents an indicator of completeness based on the checked data elements; the tool is available at <https://smartprotocols.github.io/checklist1.0> [63]. Below, we present a complete description of the data elements in our checklist. We have organized the data elements in four categories, namely: i) bibliographic data elements, ii) discourse data elements, iii) data elements for materials, and iv) data elements for the procedure. Ours is a comprehensive checklist, the data elements must be reported whenever applicable.

Data element	Property
Title of the protocol Author Version number License of the protocol Provenance of the protocol	Name Identifier
Overall objective or Purpose Application of the protocol Advantage(s) of the protocol Limitation(s) of the protocol	
Organism	Whole organism / Organism part Sample/organism identifier Strain, genotype or line Amount of Bio-Source Developmental stage Bio-source supplier Growth substrates Growth environment Growth time Sample pre-treatment or sample preparation
Laboratory equipment	Name Manufacturer or vendor (including homepage) Identifier (catalog number or model) Equipment configuration
Laboratory consumable	Name Manufacturer or vendor (including homepage) Identifier (catalog number)
Reagent	Name Manufacturer or vendor (including homepage) Identifier (catalog number)
Kit	Name Manufacturer or vendor (including homepage) Identifier (catalog number)
Recipe for solution	Name Reagent or chemical compound name Initial concentration of a chemical compound Final concentration of chemical compound Storage conditions Cautions Hints
Software	Name Version number Homepage
Procedure	List of steps in numerical order Alternative / Optional / Parallel steps Critical steps Pause point Timing Hints Troubleshooting

TABLE 2.7: Data elements for reporting protocols in life sciences

2.3.1 Bibliographic data elements

From the guidelines for authors, the datum “author identifier” was not considered, nor was this data element found in the analyzed protocols. The “provenance” is proposed as “desirable information” in only two of the guidelines (Nature Protocols and Bio-protocols), as well as “updates of the protocol” (Cold Spring Harbor Protocols and Bio-protocols). 72.5% (29) of the protocols available in our Bio-protocols collection and 61.5% (24) of the protocols available in our Nature Protocols Exchange collection reported the provenance (Figure 2.2). None of the protocols collected from Cold Spring Harbor Protocols or Bio-protocols had been updated –last checked December 2017.

FIGURE 2.2: Bibliographic data elements found in guidelines for authors. NC= Not Considered in guidelines; D= Desirable information if this is available.

As a result of the workshops, domain experts exposed the importance of including these three data elements in our checklist. For instance, readers sometimes need to contact the authors to ask about specific information (quantity of the sample used, the storage conditions of a solution prepared in the lab, etc.); occasionally, the correspondent author does not respond because he/she has changed his/her email address, and searching for the full name could retrieve multiple results. By using author IDs, this situation could be resolved. The experts asserted that well-documented provenance helps them to know where the protocol comes from and whether it has changed. For example, domain experts expressed their interest in knowing where a particular protocol was published for the first time, who has reused it, how many research papers have used it, how many people have modified it, etc. In a similar way, domain experts also expressed the need for a version control system that could help them to know and understand how, where and why the protocol has changed. For example, researchers are interested in tracking changes in quantities, reagents, instruments, hints, etc. For a complete description of the bibliographic data elements proposed in our checklist, see below.

Title. The title should be informative, explicit, and concise (50 words or fewer). The use of ambiguous terminology and trivial adjectives or adverbs (e.g., novel, rapid, efficient, inexpensive, or their synonyms) should be avoided. The use of numerical values, abbreviations, acronyms, and trademarked or copyrighted product names is discouraged. This definition was adapted from BioTechniques [29]. In Table 2.8, we present examples illustrating how to define the title.

Author name and author identifier. The full name(s) of the author(s) is required together with an author ID, e.g., ORCID [66] or research ID [67]. The role of each author is also required; depending on the domain, there may be several roles. It is important to use a simple word that describes who did what. Publishers, laboratories, and authors should enforce the use of an “author contribution section” to

am- biguous title	A single * protocol for extraction of gDNA ‡ from bacteria and yeast.	Protocol available at [64]
compre- hensible title	Extraction of nucleic acids from yeast cells and plant tissues using ethanol as medium for sample preservation and cell disruption.	Protocol available at [65]

TABLE 2.8: Examples illustrating two titles. Issues in the ambiguous title: *Use of ambiguous terminology, ‡use of abbreviations.

identify the role of each author. We have identified two roles that are common across our corpus of documents.

- **Creator of the protocol:** This is the person or team responsible for the development or adaptation of a protocol.
- **Laboratory-validation scientist:** Protocols should be validated in order to certify that the processes are clearly described; it must be possible for others to follow the described processes. If applicable, statistical validation should also be addressed. The validation may be procedural (related to the process) or statistical (related to the statistics). According to the Food and Drug Administration (FDA) [68], validation is “*establishing documented evidence which provides a high degree of assurance that a specific process will consistently produce a product meeting its predetermined specifications and quality attributes*” [69].

Updating the protocol. The peer-reviewed and non peer-reviewed repositories of protocols should encourage authors to submit updated versions of their protocols; these may be corrections, retractions, or other revisions. Extensive modifications to existing protocols could be published as adapted versions and should be linked to the original protocol. We recommended to promote the use of a version control system; in this paper we suggest to use the version control guidelines proposed by the National Institute of Health (NIH) [70].

- **Document dates:** Suitable for unpublished protocols. The date indicating when the protocol was generated should be in the first page and, whenever possible, incorporated into the header or footer of each page in the document.
- **Version numbers:** Suitable for unpublished protocols. The current version number of the protocol is identified in the first page and, when possible, incorporated into the header or footer of each page of the document.
 - **Draft document version number:** Suitable for unpublished protocols. The first draft of a document will be Version 0.1. Subsequent drafts will have an increase of “0.1” in the version number, e.g., 0.2, 0.3, 0.4, ... 0.9, 0.10, 0.11.
 - **Final document version number and date:** Suitable for unpublished and published protocols. The author (or investigator) will deem a protocol final after all reviewers have provided final comments and these have been addressed. The first final version of a document will be Version 1.0; the date when the document becomes final should also be included. Subsequent final documents will have an increase of “1.0” in the version number (1.0, 2.0, etc.).

- **Documenting substantive changes:** Suitable for unpublished and published protocols. A list of changes from the previous drafts or final documents will be kept. The list will be cumulative and identify the changes from the preceding document versions so that the evolution of the document can be seen. The list of changes and consent/assent documents should be kept with the final protocol.

Provenance of the protocol. The provenance is used to indicate whether or not the protocol results from modifying a previous one. The provenance also indicates whether the protocol comes from a repository, e.g., Nature Protocols Exchange, protocols.io [71], or a journal like JoVE, MethodsX, or Bio-Protocols. The former refers to adaptations of the protocol. The latter indicates where the protocol comes from. See Table 2.9.

example	<i>"This protocol was adapted from "How to Study Gene Expression," Chapter 7, in Arabidopsis: A Laboratory Manual (eds. Weigel and Glazebrook). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA, 2002."</i>	Protocol available at [72]
----------------	--	----------------------------

TABLE 2.9: Example illustrating the provenance of a protocol.

License of the protocol. The protocols should include a license. Whether as part of a publication or, just as an internal document, researchers share, adapt and reuse protocols. The terms of the license should facilitate and make clear the legal framework for these activities.

2.3.2 Data elements of the discourse

Here, we present the elements considered necessary to understand the suitability of a protocol. They are the “overall objective or purpose”, “applications”, “advantages,” and “limitations”. 100% of the analyzed guidelines for author suggest the inclusion of these four elements in the abstract or introduction section. However, one or more of these four elements were not reported. For example, “limitations” was reported in only 20% of the protocols from Genetic and Molecular Research and PLOS One, and in 40% of the protocols from Springer. See Figure 2.3.

FIGURE 2.3: Data elements related to the discourse as reported in the analyzed protocols

Interestingly, 83% of the respondents considered the “limitations” to be a data element that is necessary when reporting a protocol. In the last meeting, participants

considered that “limitations” represents an opportunity to make suggestions for further improvements. Another data element discussed was “advantages”; 43% of the respondents considered the “advantages” as a data element that is necessary to be reported in a protocol. In the last meeting, all participants agreed that “advantages” (where applicable) could help us to compare a protocol with other alternatives commonly used to achieve the same result. For a complete description of the discourse data elements proposed in our checklist, see below.

Overall objective or Purpose. The description of the objective should make it possible for readers to decide on the suitability of the protocol for their experimental problem. See Table 2.10.

Discourse data element	Example	Source
Overall objective/Purpose	<i>“Development of a method to isolate small RNAs from different plant species (...) that no need of first total RNA extraction and is not based on the commercially available TRIzol® Reagent or columns.”</i>	Protocol available at [73]
Application	<i>“DNA from this experiment can be used for all kinds of genetics studies, including genotyping and mapping.”</i>	Protocol available at [74]
Advantage(s)	<i>“We describe a fast, efficient and economic in-house protocol for plasmid preparation using glass syringe filters. Plasmid yield and quality as determined by enzyme digestion and transfection efficiency were equivalent to the expensive commercial kits. Importantly, the time required for purification was much less than that required using a commercial kit.”</i>	Protocol available at [75]
Limitation(s)	<i>“A major problem faced both in this and other safflower transformation studies is the hyperhydration of transgenic shoots which result in the loss of a large proportion of transgenic shoots.”</i>	Protocol available at [76]

TABLE 2.10: Examples of discursive data elements.

Application of the protocol. This information should indicate the range of techniques where the protocol could be applied. See Table 2.10.

Advantage(s) of the protocol. Here, the advantages of a protocol compared to other alternatives should be discussed. See Table 2.10. Where applicable, references should be made to alternative methods that are commonly used to achieve the same result.

Limitation(s) of the protocol. This datum includes a discussion of the limitations of the protocol. This should also indicate the situations in which the protocol could be unreliable or unsuccessful. See Table 2.10.

2.3.3 Data elements for materials

From the analyzed guidelines for authors, the datum “sample description” was considered only in the Current Protocols guidelines. The “laboratory consumables or supplies” datum was not included in any of the analyzed guidelines. See Figure 2.4.

FIGURE 2.4: Data elements describing materials. NC= Not Considered in guidelines; D= Desirable information if this is available; R= Required information.

Our Current Protocols collection includes documents about toxicology, microbiology, magnetic resonance imaging, cytometry, chemistry, cell biology, human genetics, neuroscience, immunology, pharmacology, protein, and biochemistry; for these protocols the input is a biological or biochemical sample. This collection also includes protocols in bioinformatics with data as the input. 100% of the protocols from our Current Protocols collection includes information about the input of the protocol (biological/biochemical sample or data). In addition, 87% of protocols from this collection include a list of materials or resources (reagents, equipment, consumables, software, etc.).

We also analyzed the protocols from our MethodsX collection. We found that despite the exclusion of the sample description in guidelines for authors, the authors included this information in their protocols. Unfortunately, these protocols do not include a list of materials. Only 29% of the protocols reported a partial list of materials. For example, the protocol published by Vinayagamoorthy et al.[64], includes a list of recommended equipment but does not list any of the reagents, consumables, or other resources mentioned in the protocol instructions. See Figure 2.5.

FIGURE 2.5: Data elements describing materials.

Domain experts considered that the input of the protocol (biological/biochemical sample or data) needs an accurate description; the granularity of the description varies depending on the domain. If such description is not available then the reproducibility could be affected. In addition, domain experts strongly suggested to include consumables in the checklist. It was a general surprise not to find these data elements in the guidelines for authors that we analyzed. Domain experts shared with us bad experiences caused by the lack of information about the type of consumables. Some of the incidents that may arise from the lack of this information include: i) cross contamination, when no information suggesting the use of filtered pipet tips is available; ii) misuse of containers, when no information about the use of

containers resistant to extreme temperatures and/or impacts is available; iii) misuse of containers, when a container made of a specific material should be used, e.g., glass vs. plastic vs. metal. This is critical information; researchers need to know if reagents or solutions prepared in the laboratory require some specific type of containers in order to avoid unnecessary reactions altering the result of the assay. Presented below is the set of data elements related to materials or resources used for carrying out the execution of a protocol.

Sample. This is the role played by a biological substance; the sample is an experimental input to a protocol. The information required depends on the type of sample being described and the requirements from different communities. Here, we present the data elements for samples commonly used across the protocols and guidelines that we analyzed.

- **Bio-source properties:**

Strain, genotype or line: This datum is about subspecies such as ecotype, cultivar, accession, or line. In the case of crosses or breeding results, pedigree information should also be provided.

Starting material: This datum is about the physical biological specimen from which your experimental data are derived. The starting material could be a whole organism, or a part of this.

- **whole organism** Typical examples are multicellular animals, plants, and fungi; or unicellular microorganisms such as protists, bacteria, and archaea.
- **organism part** Typical examples of an organism part include a cell line, a tissue, an organ, corporal bodily fluids protoplasts, nucleic acids, proteins, etc.
- **organism/sample identifier** This is the unique identifier assigned to an organism. The NCBI taxonomy id, also known as “taxid”, is commonly used to identify an organism; the Taxonomy Database is a curated classification and nomenclature for all organisms in the public sequence databases. Public identification systems, e.g. the Taxonomy Database, should be used when ever possible. Identifiers may be internal; for instance, laboratories often have their own coding system for generating identifiers. When reporting internal identifiers it is important to also state the source and the nature (private or public) of the identifier, e.g. A0928873874, barcode (CIAT-DAPA internal identifier) of a specimen or sample.

Amount of Bio-Source: This datum is about mass (mg fresh weight or mg dry weight), number of cells, or other measurable bulk numbers (e.g., protein content).

Developmental stage: This datum includes age and gender (if applicable) of the organism.

Bio-source Supplier: This datum is defined as a person, company, laboratory or entity that offers a variety of biosamples or biospecimens.

- **Growth conditions:**

Growth substrates: This datum refers to an hydroponic system (type, supplier, nutrients, concentrations), soil (type, supplier), agar (type, supplier), and cell culture (media, volume, cell number per volume).

Growth environment: This datum includes, but is not limited to, controlled environments such as greenhouse (details on accuracy of control of light, humidity, and temperature), housing conditions (light/dark cycle), and non-controlled environments such as the location of the field trial.

Growth time: This datum refers to the growth time of the sample prior to the treatment.

- **Sample pre-treatment or sample preparation:** This datum refers to collection, transport, storage, preparation (e.g., drying, sieving, grinding, etc.), and preservation of the sample.

Laboratory equipment. The laboratory equipment includes apparatus and instruments that are used in diagnostic, surgical, therapeutic, and experimental procedures. In this subsection, all necessary equipment should be listed; manufacturer name or vendor (including the homepage), catalog number (or model), and configuration of the equipment should be part of this data element. See Table 2.11.

example	Name / manufacturer / model: <i>"Inverted confocal microscope, PC and image acquisition software / Zeiss / LSM 780."</i>	equipment configuration: <i>"Configure a four-channel microscope with appropriate excitation light sources and emission filters: FITC-488 excitation, 490–560-nm emission; ..."</i>	Protocol available at [77]
----------------	---	--	----------------------------

TABLE 2.11: Example for the presentation of equipment.

- **Laboratory equipment name:** This datum refers to the name of the equipment as it is given by the manufacturer (e.g., FocalCheck fluorescence microscope test slide).
- **Manufacturer name:** This datum is defined as a person, company, or entity that produces finished goods (e.g., Life Technologies, Zeiss).
- **Laboratory equipment ID (model or catalog number):** This datum refers to an identifier provided by the manufacturer or vendor (e.g., F36909 –catalog number for FocalCheck fluorescence microscope test slide from Life Technologies).
- **Equipment configuration:** This datum should explain the configuration of the equipment and the parameters that make it possible to carry out an operation, procedure, or task (e.g., the configuration of an inverted confocal microscope).

Laboratory consumables or supplies. The laboratory consumables include, amongst others, disposable pipettes, beakers, funnels, test tubes for accurate and precise measurement, disposable gloves, and face masks for safety in the laboratory. In this subsection, a list with all the consumables necessary to carry out the protocol should be presented with manufacturer name (including the homepage) and catalog number. See Table 2.12.

ambiguous example	Filter paper	Protocol available at [78]
descriptive example	Filter paper (GE, catalog number: 10311611)	Protocol available at [79]

TABLE 2.12: Reporting consumables.

- **Laboratory consumable name:** This datum refers to the name of the laboratory consumable as it is given by the manufacturer e.g., Cryogenic Tube, sterile, 1.2 ml.
- **Manufacturer name:** This datum is defined as a person, enterprise, or entity that produces finished goods (e.g., Nalgene, Thermo-scientific, Eppendorf, Falcon)
- **Laboratory consumable ID (catalog number):** This datum refers to an identifier provided by the manufacturer or vendor; for instance, 5000-0012 (catalog number for Cryogenic Tube, sterile, 1.2 mL from Nalgene).

Recipe for solutions. A recipe for solutions is a set of instructions for preparing a particular solution, media, buffer, etc. The recipe for solutions should include the list of all necessary ingredients (chemical compounds, substance, etc.), initial and final concentrations, pH, storage conditions, cautions, and hints. Ready-to-use reagents do not need to be listed in this category; all purchased reagents that require modification (e.g., a dilution or addition of β -mercaptoethanol) should be listed. See Table 2.13 for more information.

ambiguous example	See in the section recipes, the recipe 1 (PBS)	Protocol available at [79]
descriptive example	Phosphate-buffered saline (PBS) recipe	Protocol available at [80]

TABLE 2.13: Reporting recipes for solutions.

- **Solution name:** This is the name of the preparation that has at least 2 chemical substances, one of them playing the role of solvent and the other playing the role of solute. If applicable, the name should include the following information: concentration of the solution, final volume and final pH. For instance, Ammonium bicarbonate (NH_4HCO_3), 50 mM, 10 ml, pH 7.8.
- **Chemical compound name or reagent name:** This is the name of a drug, solvent, chemical, etc.; for instance, agarose, dimethyl sulfoxide (DMSO), phenol, sodium hydroxide. If applicable, a measurable property, e.g. concentration, should be included.
- **Initial concentration of a chemical compound:** This is the first measured concentration of a compound in a substance.
- **Final concentration of chemical compound:** This is the last measured concentration of a compound in a substance.
- **Storage conditions:** This datum includes, among others, shelf life (maximum storage time) and storage temperature for the solutions e.g., “Store the solution

at room temperature", "maximum storage time, 6 months". Specify whether or not the solutions must be prepared fresh.

- **Cautions:** Toxic or harmful chemical compounds should be identified by the word 'CAUTION' followed by a brief explanation of the hazard and the precautions that should be taken when handling e.g., "CAUTION: NaOH is a very strong base. Can seriously burn skin and eyes. Wear protective clothing when handling. Make in fume hood".
- **Hints:** The "hints" are commentaries or "tips" that help the researcher to correctly prepare the recipe e.g., "Add NaOH to water to avoid splashing".

Reagents. A reagent is a substance used in a chemical reaction to detect, measure, examine, or produce other substances. List all the reagents used when performing the protocol, the vendor name (including homepage), and catalog number. Reagents that are purchased ready-to-use should be listed in this section. See Table 2.14.

ambiguous example	Dextran sulfate, Sigma-Aldrich	Protocol available at [5]
descriptive example	Dextran sulfate sodium salt from <i>Leuconostoc</i> spp., Sigma-Aldrich, D8906-5G	Protocol available at [81]

TABLE 2.14: Reporting reagents.

- **Reagent name:** This datum refers to the name of the reagent or chemical compound. For instance, "Taq DNA Polymerase from *Thermus aquaticus* with 10X reaction buffer without MgCl₂".
- **Reagent vendor or manufacturer:** This is the person, enterprise, or entity that produces chemical reagents e.g., Sigma-Aldrich.
- **Reagent ID (catalog number):** This is an identifier provided by the manufacturer or vendor. For instance, D4545-250UN (catalog number for Taq DNA Polymerase from *Thermus aquaticus* with 10X reaction buffer without MgCl₂ from Sigma-Aldrich).

Kits. A kit is a gear consisting of a set of articles or tools for a specific purpose. List all the kits used when carrying out the protocol, the vendor name (including homepage), and catalog number.

- **Kit name:** This datum refers to the name of the kit as it is given by the manufacturer e.g., SpectrumTM Plant Total RNA Kit, sufficient for 50 purifications.
- **Kit vendor or manufacturer:** This is the person, enterprise, or entity that produces the kit e.g., Sigma-Aldrich.
- **Kit ID (catalog number):** This is an identifier provided by the manufacturer or vendor e.g., STRN50, catalog number for SpectrumTM Plant Total RNA Kit, sufficient for 50 purifications.

Software. Software is composed of a series of instructions that can be interpreted or directly executed by a processing unit. In this subsection, please list software used in the experiment including the version, as well as where to obtain it.

- **Software name:** This datum refers to the name of the software. For instance, “LightCycler 480 Software”.
- **Software version:** A software version number is an attribute that represents the version of software e.g., Version 1.5.
- **Software availability:** This datum should indicate where the software can be downloaded from. If possible, license information should also be included; for instance, <https://github.com/MRCIE-U/ariesmql>, GPL3.0.

2.3.4 Data elements for the procedure

All the analyzed guidelines include recommendations about how to document the instructions; for example, list the steps in numerical order, use active tense, organize the procedures in major stages, etc. However, information about documentation of alternative, optional, or parallel steps (where applicable) and alert messages such as critical steps, pause point, and execution time was infrequent (available in less than 40% of the guidelines). See Figure 2.6.

FIGURE 2.6: Data elements describing the process, as found in the guidelines for authors. NC= Not Considered in guidelines; O= Optional information; D= Desirable information if this is available; R= Required information.

We chose a subset of protocols (12 from our Plant Methods collection, 7 from our Biotechniques collection, and 5 unpublished protocols from CIAT) to review which data elements about the procedure were documented. 100% of the protocols have steps organized in major stages. 100% of the unpublished protocols list the steps in numerical order, and nearly 60% of the protocols from Plant Methods and Biotechniques followed this recommendation. Alert messages were included in 67% of the Plant Methods protocols and in 14% of the Biotechniques protocols. Neither of the 5 unpublished protocols included alert messages. Troubleshooting was reported in just a few protocols; this datum was available in 8% of the Plant Methods protocols and in 14% of the Biotechniques protocols. See Figure 2.7.

In this stage, the discussion with domain experts started with the description of steps. In some protocols, the steps are poorly described; for instance, some of them include working temperatures, e.g., cold room, on ice, room temperature; but, *what exactly do they mean?* Steps involving centrifugation, incubation, washing, etc., should specify conditions, e.g., time, temperature, speed (rpm or g), number of washes, etc. For experts, alert messages and troubleshooting (where applicable) complement the description of steps and facilitate a correct execution. This opinion coincides with the results of the survey, where troubleshooting and alert messages such as critical steps, pause points, and timing were considered relevant by 83% -

FIGURE 2.7: Data elements describing the process, as found in analyzed protocols.

87% of the respondents. The set of data elements related to the procedure is presented below.

- Recommendation 1. Whenever possible, list the steps in numerical order; use active tense. For example: "Pipette 20 ml of buffer A into the flask," as opposed to "20 ml of buffer A are/were pipetted into the flask" [37].
- Recommendation 2. Whenever there are two or more alternatives, these should be numbered as sets of consecutive steps [35]. For example: "Choose procedure A (steps 1-10) or procedure B (steps 11-20); then continue with step 21 . . .". Optional steps or steps to be executed in parallel should also be included.
- Recommendation 3. For techniques comprising a number of individual procedures, organize these in the exact order in which they should be executed [37].
- Recommendation 4. Description of steps. Those steps that include working temperatures, e.g., cold room, on ice, room temperature, should be clearly specified. From the European Pharmacopoeia (Pharm.Eur.) [82], World Health Organization resource guidance (WHO guidance) [83], and the U.S. Pharmacopoeia (USP) [84], the most common storage conditions were extracted, see below:
 - Frozen/deep-freeze temperature (-20°C to -15°C)
 - Refrigerator, cold room or cold temperature (2°C to 8°C)
 - Cool temperature (8°C to 15°C)
 - Room/Ambient temperature (15°C to 25°C)
 - Warm/Lukewarm temperature (30°C to 40°C)

For centrifugation steps, specify time, temperature, and speed (rpm or g). Always state whether to discard/keep the supernatant/pellet. For incubations, specify time, temperature, and type of incubator. For washes, specify conditions e.g., temperature, washing solution and volume, specific number of washes, etc.

Useful auxiliary information should be included in the form of "alert messages". The goal is to remind or alert the user of a protocol with respect to issues that may

arise when executing a step. These messages may cover special tips or hints for performing a step successfully, alternate ways to perform the step, warnings regarding hazardous materials or other safety conditions, time considerations. For instance, pause points, speed at which the step must be performed and storage information (temperature, maximum duration) [35].

- **Critical steps:** Highlight critical steps in the protocol and give indications that help to carry these out in a precise manner. For instance, time and temperature information if these are deemed crucial. Or, whether the use of RNase free solutions is required. Information should be provided in order to indicate how these steps are critical and how to overcome the issues. "Critical Steps" should help the user to maximize the likelihood of success; use the heading CRITICAL STEP followed by a brief explanation. See Table 2.15.

Alert message	Step	Note	Source
Critical step	<i>"Remove dirt from the surface of the specimen with a tissue. If necessary, moisten the tissue with ..."</i>	<i>"Dirt may introduce a variety of inhibitory substances (...); these substances may interfere or even completely block subsequent enzymatic manipulations of the DNA extracts."</i>	Protocol available at [85]
Pause point	<i>"Weigh out no more than 500 mg of sample powder and transfer it to a 15 ml tube."</i>	<i>"The sample powder can be stored at room temperature, but should be subjected to the extraction as soon as possible."</i>	Protocol available at [85]
Timing	<i>"Preparation of the bone or tooth sample"</i>	<i>"15–30 min per sample"</i>	Protocol available at [85]
Hint	<i>"Add the following components to a nuclease-free microcentrifuge tube:..."</i>	<i>"We tested several commercial thermostable DNA polymerases (...), the most consistent results were obtained using Advantage 2 PCR Polymerase Mix ..."</i>	Protocol available at [86]

TABLE 2.15: Examples of alert messages

- **Pause point:** This datum is appropriate after steps in the protocol where the procedure can be stopped. i.e., when the experiment can be stopped and resumed at a later point in time. Any PAUSE POINTS should be indicated with a brief description of the options available. See Table 2.15.
- **Timing:** This datum is used to include the approximate time of execution of a step or set of steps. Timing could also be indicated at the beginning of the protocol. See Table 2.15.
- **Hints:** Provide any commentary, note, or hints that will help the researcher to correctly perform the protocol. See Table 2.15.

- **Troubleshooting:** This datum is used to list common problems, possible causes, and solutions/methods of correction. This can be submitted as a 3-column table or listed in the text. An example is presented in "Table 1.Troubleshooting table", available at [85].

2.4 Data elements represented in the SMART Protocols Ontology

The data elements proposed in our guideline are represented in the SMART Protocols Ontology. This ontology was developed to facilitate the semantic representation of experimental protocols. Our ontology reuses the Basic Formal Ontology (BFO) [87] and the Relation Ontology (RO) [88] to characterize concepts. In addition, each term in the SMART Protocols ontology is represented with annotation properties imported from the OBI Minimal metadata. The classes and properties are represented by their respective labels to facilitate the readability; the prefix indicates the provenance for each term. Our ontology is organized in two modules. The document module represents the metadata necessary and sufficient for reporting a protocol. The workflow module represents the executable elements of a protocol to be carried out and maintained by humans. Figure 2.8 presents the hierarchical organization of data elements into the SMART Protocols Ontology.

FIGURE 2.8: Hierarchical organization of data elements in the SMART Protocols Ontology.

2.5 Discussion

In this paper, we have described 17 data elements that can be used to improve the reporting structure of protocols. Our work is based on the analysis of 530 published and non-published protocols, guidelines for authors, and suggested reporting structures. We examined guidelines for authors from journals that specialize in publishing experimental protocols, e.g., *Bio-protocols*, *Cold Spring Harbor Protocols*, *MethodsX*, *Nature Protocols*, and *Plant Methods (Methodology)*. Although *JoVE* [89] is a video methods journal, its guidelines for authors were also considered. Online repositories were also studied; these resources deliver an innovative approach for the publication of protocols by offering platforms tailored for this kind of document. For instance, *protocols.io* [90] structures the protocol by using specific data elements and treats the protocol as a social object, thus facilitating sharing. It also makes it possible to have version control over the document. Protocol Exchange from *Nature Protocols* is an open repository where users upload, organize, comment, and share their protocols. Our guideline has also benefited from the input from a group of researchers whose primary interest is having reproducible protocols. By analyzing reporting structures and guidelines for authors, we are contributing to the homogenization of data elements that should be reported as part of experimental protocols. Improving the reporting structure of experimental protocols will add the necessary layer of information that should accompany the data that is currently being deposited into data repositories.

Ours was an iterative development process; drafts were reviewed and analyzed, and then improved versions were produced. This made it easier for us to make effective use of the time that domain experts had available. Working with experimental protocols that were known by our group of domain experts helped us to engage them in the iterations. Also, for the domain experts who worked with us during the workshops, there was a pre-existing interest in standardizing their reporting structures. Reporting guidelines are not an accepted norm in biology [91]; however, experimental protocols are part of the daily activities for most biologists. They are familiar with these documents, the benefits of standardization are easy for them to understand. From our experience at CIAT, once researchers were presented with a standardized format that they could extend and manage with minimal overhead, they adopted it. The early engagement with domain experts in the development process eased the initial adoption; they were familiar with the outcome and aware of the advantages of implementing this practice. However, maintaining the use of the guideline requires more than just availability of the guideline; the long-term use of these instruments requires an institutional policy in data stewardship. Our approach builds upon previous experiences; in our case, the guidelines presented in this paper are a tool that was conceived by researchers as part of their reporting workflow, thus adding a minimal burden on their workload. As domain experts were working with the guideline, they were also gaining familiarity with the Minimum Information for Biological and Biomedical Investigations (MIBBI) [91] that were applicable to their experiments. This made it possible for us to also discuss the relation between MIBBIs and the content in the experimental protocols.

The quality of the information reported in experimental protocols and methods is a general cause for concern. Poorly described methods generate poorly reproducible research. In a study conducted by [92] in *Trypanosoma* experiments, they report that none of the investigated articles met all the criteria that should be reported in these kinds of experiments. The study reported by [93] has similar results leading to similar conclusions; key metadata elements are not always reported by researchers. The

widespread availability of key metadata elements in ontologies, guidelines, minimal information models, and reporting structures was discussed. These were, from the onset, understood as reusable sources of information. Domain experts understood that they were building on previous experiences; having examples of use was helpful in understanding how to adapt or reuse from existing resources. This helps them to understand the rationale of each data element within the context of their own practice. For us, being able to consult previous experiences was also an advantage. Sharing protocols is a common practice amongst researchers from within the same laboratories or collaborating in the same experiments or projects. However, there are limitations in sharing protocols, not necessarily related to the lack of reporting standards. They are, for instance, related to patenting and intellectual property issues, as well as to giving away competitive advantages implicit in the method.

During our development process, we considered the SMART Protocols ontology [1]; it reuses terminology from OBI, IAO, EXACT, ChEBI, NCBI taxonomy, and other ontologies. Our metadata elements have been mapped to the SMART Protocols ontology; the metadata elements in our guideline could also be mapped to resources on the web such as PubChem [94] [95] and the Taxonomy database from UniProt [96]. Our implementation of the checklist illustrates how it could be used as an online tool to generate a complement to the metadata that is usually available with published protocols. The content of the protocol does not need to be displayed; key metadata elements are made available together with the standard bibliographic metadata. Laboratories could adapt the online tool to their specific reporting structures. Having a checklist made it easier for the domain experts to validate their protocols. Machine validation is preferable, but such mechanisms require documents to be machine-processable beyond that which our domain experts were able to generate. Domain experts were using the guideline to implement simple Microsoft Word reporting templates. Our checklist does not include aspects inherent to each possible type of experiment such as those available in the MIBBIs; these are based on the minimal common denominator for specific experiments. Both approaches complement each other; where MIBBIs offer specificity, our guideline provides a context that is general enough for facilitating reproducibility and adequate reporting without interfering with records such as those commonly managed by Laboratory Information Management Systems.

In laboratories, experimental protocols are released and periodically undergo revisions until they are released again. These documents follow the publication model put forward by Carole Goble, *"Don't publish, release"* with strict versioning, changes, and forks [97]. Experimental protocols are essentially executable workflows for which identifiers for equipment, reagents, and samples need to be resolved against the Web. The use of unique identifiers can't be underestimated when supporting adequate reporting; identifiers remove ambiguity for key resources and make it possible for software agents to resolve and enrich these entities. The workflows in protocols are mostly followed by humans, but in the future, robots may be executing experiments [98]; it makes sense to investigate other publication paradigms for these documents. The workflow nature of these documents is more suitable for a fully machine-processable or -actionable document. The workflows should be intelligible for humans and processable by machines; thus, facilitating the transition to fully automated laboratory paradigms. Entities and executable elements should be declared and characterized from the onset. The document should be "born semantic" and thus inter-operable with the larger web of data. In this way post-publication and linguistic processing activities, such as Named Entity Recognition and annotation, could be more focused.

Currently, when protocols are published, they are treated like any other scientific publication. Little attention is paid to the workflow nature implicit in this kind of document, or to the chain of provenance indicating where it comes from and how it has changed. The protocol is understood as a text-based narrative instead of a self-descriptive Findable Accessible Interoperable and Reusable (FAIR) [99] compliant document. There are differences across the examined publications, e.g., JoVE builds the narrative around video, whereas Bio-protocols, MethodsX, Nature Protocols, and Plant Methods primarily rely on a text-based narrative. The protocol is, however, a particular type of publication; it is slightly different from other scientific articles. An experimental protocol is a document that is kept “alive” after it has been published. The protocols are routinely used in laboratory activities, and researchers often improve and adapt them, for instance, by extending the type of samples that can be tested, reducing timing, minimizing the quantity of certain reagents without altering the results, adding new recipes, etc. The issues found in reporting methods probably stem, at least in part, from the current structure of scientific publishing, which is not adequate to effectively communicate complex experimental methods [92].

2.6 Conclusion

Experimental research should be reproducible whenever possible. Having precise descriptions of the protocols is a step in that direction. Our work addresses the problem of adequate reporting for experimental protocols. It builds upon previous work, as well as over an exhaustive analysis of published and unpublished protocols and guidelines for authors. There is value in guidelines because they indicate how to report; having examples of use facilitate how to adapt them. The guideline we present in this paper can be adapted to address the needs of specific communities. Improving reporting structures requires collective efforts from authors, peer reviewers, editors, and funding bodies. There is no “one size that fits all.” The improvement will be incremental; as guidelines and minimal information models are presented, they will be evaluated, adapted, and re-deployed.

Authors should be aware of the importance of experimental protocols in the research life-cycle. Experimental protocols ought to be reused and modified, and derivative works are to be expected. This should be considered by authors before publishing their protocols; the terms of use and licenses are the choice of the publisher, but where to publish is the choice of the author. Terms of use and licenses forbidding “reuse”, “reproduce”, “modify”, or “make derivative works based upon” should be avoided. Such restrictions are an impediment to the ability of researchers to use the protocols in their most natural way, which is adapting and reusing them for different purposes –not to mention sharing, which is a common practice among researchers. Protocols represent concrete “know-how” in the biomedical domain. Similarly, publishers should adhere to the principle of encouraging authors to make protocols available, for instance, as preprints or in repositories for protocols or journals. Publishers should enforce the use of repository or journal publishing protocols. Publishers require or encourage data to be available; the same principle should be applied to protocols. Experimental protocols are essential when reproducing or replicating an experiment; data is not contextualized unless the protocols used to derive the data are available.

This work is related to the SMART Protocols project. Ultimately we want, (1) to enable authors to report experimental protocols with necessary and sufficient information that allows others to reproduce an experiment, (2) to ensure that every data item is resolvable against resources in the web of data, and (3) to make the protocols available in RDF, JSON, and HTML as web native objects. We are currently working on a publication platform based on linked data for experimental protocols. Our approach is simple, we consider that protocols should be born semantics and FAIR.

Bibliography

- [1] O. Giraldo, A. García, F. López, and O. Corcho, “Using semantics for representing experimental protocols”, *Journal of Biomedical Semantics*, vol. 8, no. 1, p. 52, 2017, ISSN: 2041-1480. DOI: 10.1186/s13326-017-0160-y. [Online]. Available: <https://doi.org/10.1186/s13326-017-0160-y>.
- [2] L. P. Freedman, G. Venugopalan, and R. Wisman, “Reproducibility 2020: Progress and priorities”, *F1000Research*, vol. 6, p. 604, 2017. DOI: 10.12688/f1000research.11334.1. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5461896/>.
- [3] A. Casadevall and F. C. Fang, “Reproducible science”, *Infection and Immunity*, vol. 78, no. 12, pp. 4972–4975, Dec. 2010. DOI: 10.1128/IAI.00908-10. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2981311/>.
- [4] M. F. W. Festing and D. G. Altman, “Guidelines for the design and statistical analysis of experiments using laboratory animals”, *ILAR Journal*, vol. 43, no. 4, pp. 244–258, 2002. DOI: 10.1093/ilar.43.4.244. eprint: /oup/backfile/content_public/journal/ilarjournal/43/4/10.1093_ilar.43.4.244/3/ilar-43-4-244.pdf. [Online]. Available: <http://dx.doi.org/10.1093/ilar.43.4.244>.
- [5] A. Karlgren, J. Carlsson, N. Gyllenstrand, U. Lagercrantz, and J. F. Sundström, “Non-radioactive in situ hybridization protocol applicable for norway spruce and a range of plant species”, *Journal of Visualized Experiments : JoVE*, no. 26, p. 1205, 2009. DOI: 10.3791/1205. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3148633/>.
- [6] F. Brandenburg, H. Schoffman, N. Keren, and M. Eisenhut, “Determination of Mn concentrations in *Synechocystis* sp. pcc6803 using ICP-MS”, *Bio-protocol*, vol. 7, no. 23, pp. 244–258, 2002. DOI: 10.21769/BioProtoc.2623. [Online]. Available: <https://bio-protocol.org/e2623>.
- [7] M. Baker, “1,500 scientists lift the lid on reproducibility”, *Nature*, vol. 533, no. 7604, p. 452, 2016. DOI: 10.1038/533452a. [Online]. Available: <https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970>.
- [8] 4TU, *4tu, centre for research data*, Retrieved on 07/07/2017 from <http://researchdata.4tu.nl/en/home/>, 2017. [Online]. Available: <http://researchdata.4tu.nl/en/home/>.
- [9] CSIRO, *The commonwealth scientific and industrial research organisation data access portal*, Retrieved on 07/07/2017, 2017. [Online]. Available: <https://data.csiro.au>.
- [10] Dryad, *Dryad*, Retrieved on 07/07/2017, 2017. [Online]. Available: <http://datadryad.org/>.
- [11] figshare, *Figshare*, Retrieved on 07/07/2017, 2017. [Online]. Available: <http://figshare.com>.

- [12] G. King, "An introduction to the dataverse network as an infrastructure for data sharing", *Sociological Methods and Research*, vol. 36, 173–199, 2007.
- [13] Zenodo, *Zenodo*, Retrieved on 07/07/2017, 2017. [Online]. Available: <https://zenodo.org/>.
- [14] M. Assante, L. Candela, D. Castelli, and A. Tani, "Are Scientific Data Repositories Coping with Research Data Publishing?", *Data Science Journal*, no. 15, p.6, 2016. DOI: <http://doi.org/10.5334/dsj-2016-006>.
- [15] Y. L. Simmhan, B. Plale, and D. Gannon, "A survey of data provenance in e-science", *SIGMOD Rec.*, vol. 34, no. 3, pp. 31–36, Sep. 2005, ISSN: 0163-5808. DOI: 10.1145/1084805.1084812. [Online]. Available: <http://doi.acm.org/10.1145/1084805.1084812>.
- [16] D. Moher, M. Avey, G. Antes, and D. G. Altman, "The national institutes of health and guidance for reporting preclinical research", *BMC Medicine*, vol. 13, p. 34, 2015. DOI: 10.1186/s12916-015-0284-9. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4332445/>.
- [17] N. A. Vasilevsky, M. H. Brush, H. Paddock, L. Ponting, S. J. Tripathy, G. M. LaRocca, and M. A. Haendel, "On the reproducibility of science: Unique identification of research resources in the biomedical literature", *PeerJ*, vol. 1, e148, Sep. 2013, ISSN: 2167-8359. DOI: 10.7717/peerj.148. [Online]. Available: <https://doi.org/10.7717/peerj.148>.
- [18] E. Marcus, "A STAR Is Born", *Cell*, vol. 166, no. 5, pp. 1059–1060, 2016, ISSN: 00928674.
- [19] STAR, *Structured, transparent, accessible reporting, the cell star methods guide for authors*, Retrieved on 07/07/2017, 2017. [Online]. Available: <http://www.cell.com/star-authors-guide>.
- [20] MIACA. (). Minimum Information About a Cellular Assay (MIACA). Accessed 12 march 2018 from <http://miaca.sourceforge.net/>.
- [21] J. A. Lee, J. Spidlen, K. Boyce, J. Cai, N. Crosbie, M. Dalphin, J. Furlong, M. Gasparetto, M. Goldberg, E. M. Goralczyk, B. Hyun, K. Jansen, T. Kollmann, M. Kong, R. Leif, S. McWeeney, T. D. Moloshok, W. Moore, G. Nolan, J. Nolan, J. Nikolich-Zugich, D. Parrish, B. Purcell, Y. Qian, B. Selvaraj, C. Smith, O. Tchuvatkina, A. Wertheimer, P. Wilkinson, C. Wilson, J. Wood, R. Zigon, R. H. Scheuermann, and R. R. Brinkman, "Miflowcyt: The minimum information about a flow cytometry experiment", *Cytometry Part A*, vol. 73A, no. 10, pp. 926–930, 2008, ISSN: 1552-4930. DOI: 10.1002/cyto.a.20623. [Online]. Available: <http://dx.doi.org/10.1002/cyto.a.20623>.
- [22] L. N. Soldatova, W. Aubrey, R. D. King, and A. Clare, "The EXACT description of biomedical protocols", *Bioinformatics*, vol. 24, no. 13, pp. i295–i303, 2008, ISSN: 1367-4803. [Online]. Available: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn156>.
- [23] L. N. Soldatova, D. Nadis, R. D. King, P. S. Basu, E. Haddi, V. Baumlé, N. J. Saunders, W. Marwan, and B. B. Rudkin, "EXACT2: the semantics of biomedical protocols", *BMC Bioinformatics*, vol. 15, no. Suppl 14, S5, 2014, ISSN: 1471-2105. [Online]. Available: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S14-S5>.
- [24] (). Resource identification initiative. Accessed 12 march 2018, [Online]. Available: <https://www.force11.org/group/resource-identification-initiative>.

- [25] (). The Global Unique Device Identification Database. Accessed 13 march 2018, [Online]. Available: <https://accessgudid.nlm.nih.gov/>.
- [26] (). Antibody registry. Accessed 13 march 2018, [Online]. Available: <http://antibodyregistry.org/>.
- [27] (). Addgene. the nonprofit plasmid repository. Accessed 13 march 2018, [Online]. Available: <http://www.addgene.org/>.
- [28] (). Resource identification portal. Accessed 13 march 2018, [Online]. Available: <https://scicrunch.org/resources>.
- [29] O. Giraldo, A. Garcia, and O. Corcho, *Guidelines for reporting experimental protocols [data set]*. zenodo. <http://doi.org/10.5281/zenodo.1204887>, Mar. 2018. DOI: 10.5281/zenodo.1204887. [Online]. Available: <https://doi.org/10.5281/zenodo.1204887>.
- [30] —, *Corpus of protocols [data set]*. zenodo. <http://doi.org/10.5281/zenodo.1204838>, Mar. 2018. DOI: 10.5281/zenodo.1204838. [Online]. Available: <https://doi.org/10.5281/zenodo.1204838>.
- [31] P. McQuilton, A. Gonzalez-Beltran, P. Rocca-Serra, M. Thurston, A. Lister, E. Maguire, and S.-A. Sansone, "Biosharing: Curated and crowd-sourced meta-data standards, databases and data policies in the life sciences", *Database: The Journal of Biological Databases and Curation*, vol. 2016, baw075, 2016. DOI: 10.1093/database/baw075. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4869797/>.
- [32] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.", *Nucleic acids research*, vol. 39, no. Web Server issue, W541–5, 2011, ISSN: 1362-4962.
- [33] Z. Xiang, C. Mungall, A. Ruttenberg, and Y. He, "Ontobee: A Linked Data Server and Browser for Ontology Terms", *Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO), July 28-30, 2011, Buffalo, NY, USA*, pp. 279–281, URL: <http://ceur-ws.org/Vol-833/paper48.pdf>. Accessed 26 Oct 2016.
- [34] CSH-Protocols, *Cold Spring Harbor Protocols, Instructions for Authors*, Retrieved on 07/07/2013, 2013. [Online]. Available: <http://cshlpress.com/cshprotocols/>.
- [35] Current-Protocols, *Current Protocols, The Fine Art of Experimentation. Instructions for Authors*, Retrieved on 07/07/2012, 2012. [Online]. Available: <http://www.currentprotocols.com/WileyCDA/Section/id-810273.html>.
- [36] JoVE, *Journal of Visualized Experiments, Instructions for Authors*, Retrieved on 07/07/2012, 2012. [Online]. Available: https://www.jove.com/files/Instructions_for_Authors.pdf.
- [37] Nature-Protocol, *Nature Protocol, Instructions for Authors*, 2012. [Online]. Available: <https://www.nature.com/nprot/for-authors/preparing-your-submission>.
- [38] Springer-Protocols, *Springer Protocols, Instructions for Authors*, 2013. [Online]. Available: [http://www.springer.com/cda/content/document/cda/_downloaddocument/Springer+\\\$Protocols+Manuscript+Instructions+-+MIMB.pdf?SGWID=0-0-45-1331963-p173723003](http://www.springer.com/cda/content/document/cda/_downloaddocument/Springer+\$Protocols+Manuscript+Instructions+-+MIMB.pdf?SGWID=0-0-45-1331963-p173723003).

- [39] MethodsX, *MethodsX, Instructions for Authors*, 2014. [Online]. Available: <https://www.elsevier.com/journals/methodsx/2215-0161/guide-for-authors>.
- [40] Bio-protocol, *Bio-protocol LLC, Instructions for Authors*, 2012. [Online]. Available: https://bio-protocol.org/Protocol_Preparation_Guidelines.aspx.
- [41] JBM, *Journal of Biological Methods, Instructions for Authors*, 2013. [Online]. Available: <http://www.jbmmethods.org/jbm/about/submissions/#onlineSubmissions>.
- [42] CIAT, *International Center for Tropical Agriculture (CIAT)*, 2017. [Online]. Available: <https://ciat.cgiar.org/>.
- [43] NPE, *Nature Protocol Exchange*, 2017. [Online]. Available: <http://www.nature.com/protocolexchange/>.
- [44] GMR, *Genetics and Molecular Research*, 2017. [Online]. Available: <http://www.geneticsmr.com/>.
- [45] Plant-Methods, *Plant Methods*, 2017. [Online]. Available: <http://plantmethods.biomedcentral.com/>.
- [46] Plos-One, *Plos One*, 2017. [Online]. Available: <http://journals.plos.org/plosone/>.
- [47] MIAPPE, *Minimum Information about Plant Phenotyping Experiment*, 2017. [Online]. Available: http://cropnet.pl/phenotypes/?page_id=15.
- [48] MIARE, *Minimum Information About an RNAi Experiment*, 2017. [Online]. Available: <http://miare.sourceforge.net/HomePage>.
- [49] S. A. Bustin, V. Benes, J. A. Garson, J. Helleman, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, J. Vandesompele, and C. T. Wittwer, "The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments", *Clinical Chemistry*, vol. 55, no. 4, pp. 611–622, 2009, ISSN: 0009-9147.
- [50] B. Nikolau, O. Fiehn, the participants of the 2006 Plant, M. Conference, S. Rhee, J. Dickerson, M. Lange, G. Lane, U. Roessner, J. Ward, R. Last, and C. Chapple, *CIMR: Plant Biology Context Metabolomics Standards Initiative (MSI)*, Retrieved on 07/07/2017 from <http://cosmos-fp7.eu/system/files/presentation/plant.pdf>, 2006. [Online]. Available: <http://cosmos-fp7.eu/system/files/presentation/plant.pdf>.
- [51] C. Kilkenny, W. J. Browne, I. C. Cuthill, M. Emerson, and D. G. Altman, "Improving bioscience research reporting: The arrive guidelines for reporting animal research", *PLoS Biology*, vol. 8, no. 6, e1000412, Jun. 2010. DOI: 10.1371/journal.pbio.1000412. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2893951/>.
- [52] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, L. Fan, J. Fostel, G. Fragoso, F. Gibson, A. Gonzalez-Beltran, M. A. Haendel, Y. He, M. Heiskanen, T. Hernandez-Boussard, M. Jensen, Y. Lin, A. L. Lister, P. Lord, J. Malone, E. Manduchi, M. McGee, N. Morrison, J. A. Overton, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, D. Schober, B. Smith, L. N. Soldatova, C. J. Stoeckert, C. F. Taylor, C. Torniai, J. A. Turner, R. Vita, P. L. Whetzel, and J. Zheng, "The Ontology for Biomedical Investigations", *PLOS ONE*, vol. 11, no. 4, Y. Xue, Ed., e0154556, 2016, ISSN: 1932-6203.

- [53] (). Information Artifact Ontology (IAO). <https://github.com/information-artifact-ontology/IAO/>. Accessed 7 May 2016.
- [54] L. Soldatova and R. King, "An ontology of scientific experiments", *Journal of the Royal Society Interface*, vol. 3, no. 11, pp. 795–803, 2006.
- [55] S. Abeyruwan, U. D. Vempati, H. Küçük-McGinty, U. Visser, A. Koletti, A. Mir, K. Sakurai, C. Chung, J. A. Bittker, P. A. Clemons, S. Brudz, A. Siripala, A. J. Morales, M. Romacker, D. Twomey, S. Bureeva, V. Lemmon, and S. C. Schürer, "Evolving BioAssay ontology (BAO): Modularization, integration and applications.", *Journal of Biomedical Semantics*, vol. 5, no. Suppl 1 Proceedings of the Bio-Ontologies Spec Interest G, S5, 2014.
- [56] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, "Modeling sample variables with an Experimental Factor Ontology", *Bioinformatics*, vol. 26, no. 8, pp. 1112–1118, 2010.
- [57] S. Federhen, "Type material in the NCBI Taxonomy Database", *Nucleic Acids Res*, vol. 43, pp. D1086–98, 2015.
- [58] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, "The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013", *Nucleic Acids Res*, vol. 41, pp. D456–63, 2013.
- [59] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel, "Uberon, an integrative multi-species anatomy ontology", *Genome Biology*, vol. 13, no. 1, R5, 2012, ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-1-r5. [Online]. Available: <http://dx.doi.org/10.1186/gb-2012-13-1-r5>.
- [60] S. Sarntinvijai, Y. Lin, Z. Xiang, T. F. Meehan, A. D. Diehl, U. D. Vempati, S. C. Schürer, C. Pang, J. Malone, H. Parkinson, Y. Liu, T. Takatsuki, K. Saijo, H. Masuya, Y. Nakamura, M. H. Brush, M. A. Haendel, J. Zheng, C. J. Stoeckert, B. Peters, C. J. Mungall, T. E. Carey, D. J. States, B. D. Athey, and Y. He, "Clo: The cell line ontology", *Journal of Biomedical Semantics*, vol. 5, pp. 37–37, 2014. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4387853/>.
- [61] S. Sarntinvijai, Z. Xiang, T. Meehan, A. Diehl, U. Vempati, S. Schurer, C. Pang, J. Malone, H. Parkinson, B. Athey, and Y. He, "Cell line ontology: Redesigning the cell line knowledgebase to aid integrative translational informatics", *Neoplasia*, vol. 833, pp. 25–32, 2011, ISSN: 1522-8002.
- [62] O. Giraldo, A. Garcia, and O. Corcho, *Survey - reporting an experimental protocol [data set]*. zenodo. <http://doi.org/10.5281/zenodo.1204916>, Mar. 2018. DOI: 10.5281/zenodo.1204916. [Online]. Available: <https://doi.org/10.5281/zenodo.1204916>.
- [63] F. L. Gómez, Alexander, and O. Giraldo, *SMARTProtocols/SMARTProtocols.github.io: First release of SMARTProtocols.github.io (Version v1.0.0)*. Zenodo. <http://doi.org/10.5281/zenodo.1207846>, Mar. 2018. DOI: 10.5281/zenodo.1207846. [Online]. Available: <https://doi.org/10.5281/zenodo.1207846>.
- [64] F. E. Vingataramin L, "A single protocol for extraction of gdna from bacteria and yeast.", *BioTechniques*, vol. 58, no. 3, 120–125, 2015. DOI: 10.2144/000114263.

- [65] B. Linke, K. Schröder, J. Arter, T. Gasperazzo, H. Woehlecke, and R. Ehwald, "Extraction of nucleic acids from yeast cells and plant tissues using ethanol as medium for sample preservation and cell disruption.", *BioTechniques*, vol. 49, no. 3, 655–657, 2010. DOI: 10.2144/000113476.
- [66] ORCID | *Connecting Research and Researchers*. [Online]. Available: <https://orcid.org/> (visited on 04/11/2017).
- [67] ResearcherID, *ResearcherID*. Retrieved on 07/07/2017 from <http://www.researcherid.com/>, 2017. [Online]. Available: <http://www.researcherid.com/>.
- [68] FDA, *Food and Drug Administration, White paper: FDA Guidance for Industry Update – Process Validation*. Retrieved on 07/07/2017 from <http://community.learnaboutgmp.com/uploads/db7093/original/1X/2ee62ef6c571868afa1a9fd1f35cca1b3ab00def.pdf>, 2017. [Online]. Available: <http://community.learnaboutgmp.com/uploads/db7093/original/1X/2ee62ef6c571868afa1a9fd1f35cca1b3ab00def.pdf>.
- [69] B. Das, "Validation protocol: First step of a lean-total quality management principle in a new laboratory set-up in a tertiary care hospital in india.", *Ind J Clin Biochem*, vol. 26, no. 3, 235–243, 2011. DOI: 10.1007/s12291-011-0110-x.
- [70] NIH, *National Institute of Dental and Craniofacial Research from NHI, Version Control Guidelines*. Retrieved on 07/07/2017 from <http://www.nidcr.nih.gov/Research/ToolsforResearchers/Toolkit/VersionControlGuidelines.htm>, 2017. [Online]. Available: <http://www.nidcr.nih.gov/Research/ToolsforResearchers/Toolkit/VersionControlGuidelines.htm>.
- [71] L. Teytelman, A. Stoliartchouk, L. Kindler, and B. L. Hurwitz, "Protocols.io: Virtual communities for protocol development and discussion", *PLOS Biology*, vol. 14, no. 8, pp. 1–6, Aug. 2016. DOI: 10.1371/journal.pbio.1002538. [Online]. Available: <https://doi.org/10.1371/journal.pbio.1002538>.
- [72] M. Blazquez, "Quantitative gus activity assay in intact plant tissue", *Cold Spring Harbor Protocols*, vol. 2007, no. 2, pdb.prot4688, 2007. DOI: 10.1101/pdb.prot4688. [Online]. Available: <http://cshprotocols.cshlp.org/content/2007/2/pdb.prot4688.abstract>.
- [73] F. Rosas-Cárdenas, N. Durán-Figueroa, J.-P. Vielle-Calzada, A. Cruz-Hernández, N. Marsch-Martínez, and S. de Folter, "A simple and efficient method for isolating small rnas from different plant species", *Plant Methods*, vol. 7, no. 1, p. 4, 2011, ISSN: 1746-4811. DOI: 10.1186/1746-4811-7-4. [Online]. Available: <http://dx.doi.org/10.1186/1746-4811-7-4>.
- [74] Y. Lu, "Extract genomic dna from arabidopsis leaves (can be used for other tissues as well)", *Bio-protocol*, 2011. DOI: 10.21769/BioProtoc.90. [Online]. Available: <http://www.bio-protocol.org/e90>.
- [75] Y.-C. Kim and S. L. Morrison, "A rapid and economic in-house dna purification method using glass syringe filters", *PLoS ONE*, vol. 4, no. 11, C. Lalueza-Fox, Ed., e7750, 2009. DOI: 10.1371/journal.pone.0007750. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2773934/>.
- [76] S. Belide, L. Hac, S. P. Singh, A. G. Green, and C. C. Wood, "Agrobacterium-mediated transformation of safflower and the efficient recovery of transgenic plants via grafting", *Plant Methods*, vol. 7, no. 1, p. 12, 2011, ISSN: 1746-4811. DOI: 10.1186/1746-4811-7-12. [Online]. Available: <http://dx.doi.org/10.1186/1746-4811-7-12>.

- [77] J. H. Lee, E. R. Daugharthy, J. Scheiman, R. Kalhor, T. C. Ferrante, R. Terry, B. M. Turczyk, J. L. Yang, H. S. Lee, J. Aach, K. Zhang, and G. M. Church, "Fluorescent in situ sequencing (fisseq) of rna for gene expression profiling in intact cells and tissues", *Nat. Protocols*, vol. 10, no. 3, pp. 442–458, Mar. 2015. [Online]. Available: <http://dx.doi.org/10.1038/nprot.2014.191>.
- [78] W. Zhang, S. E. Nilson, and S. M. Assmann, "Isolation and whole-cell patch clamping of arabidopsis guard cell protoplasts", *Cold Spring Harbor Protocols*, vol. 2008, no. 6, pdb.prot5014, 2008. DOI: 10.1101/pdb.prot5014. eprint: <http://cshprotocols.cshlp.org/content/2008/6/pdb.prot5014.full.pdf+html>. [Online]. Available: <http://cshprotocols.cshlp.org/content/2008/6/pdb.prot5014.abstract>.
- [79] J. Cao, X. Zhu, and X. Yan, "Fluorescence microscopy for cilia in cultured cells and zebrafish embryos.", *Bio-protocol*, vol. 4, no. 14, 2014. [Online]. Available: <http://www.bio-protocol.org/e1188>.
- [80] B. Chazotte, "Labeling golgi with fluorescent ceramides", *Cold Spring Harbor Protocols*, vol. 2012, no. 8, pdb.prot070599, 2012. DOI: 10.1101/pdb.prot070599. eprint: <http://cshprotocols.cshlp.org/content/2012/8/pdb.prot070599.full.pdf+html>. [Online]. Available: <http://cshprotocols.cshlp.org/content/2012/8/pdb.prot070599.abstract>.
- [81] M. Javelle, C. F. Marco, and M. Timmermans, "In situ hybridization for the precise localization of transcripts in plants", *Journal of Visualized Experiments : JoVE*, no. 57, p. 3328, 2011. DOI: 10.3791/3328. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3308598/>.
- [82] Pharm.Eur., *What are the regulatory definitions for "ambient", "room temperature" and "cold chain"?*, Retrieved on 24/01/2018 from <https://www.gmp-compliance.org/gmp-news/what-are-the-regulatory-definitions-for-ambient-room-temperature-and-cold-chain>, 2017. [Online]. Available: <https://www.gmp-compliance.org/gmp-news/what-are-the-regulatory-definitions-for-ambient-room-temperature-and-cold-chain>.
- [83] WHO, *Guidelines for the storage of essential medicines and other health commodities*, Retrieved on 24/01/2018 from <http://apps.who.int/medicinedocs/en/d/Js4885e/>, 2003. [Online]. Available: <http://apps.who.int/medicinedocs/en/d/Js4885e/>.
- [84] USP, *Packaging and storage requirements*, Retrieved on 24/01/2018 from http://www.drugfuture.com/Pharmacopoeia/USP35/data/v35300/usp35nf30s0_c659.html, 2018. [Online]. Available: http://www.drugfuture.com/Pharmacopoeia/USP35/data/v35300/usp35nf30s0_c659.html.
- [85] N. Rohland and M. Hofreiter, "Ancient dna extraction from bones and teeth", *Nat. Protocols*, vol. 2, no. 7, pp. 1756–1762, Jul. 2007. [Online]. Available: <http://dx.doi.org/10.1038/nprot.2007.247>.
- [86] E. Varkonyi-Gasic, R. Wu, M. Wood, E. F. Walton, and R. P. Hellens, "Protocol: A highly sensitive rt-pcr method for detection and quantification of micrornas", *Plant Methods*, vol. 3, no. 1, p. 12, 2007, ISSN: 1746-4811. DOI: 10.1186/1746-4811-3-12. [Online]. Available: <http://dx.doi.org/10.1186/1746-4811-3-12>.
- [87] BFO, *The Basic Formal Ontology (BFO)*, Retrieved on 24/01/2018 from <http://ifomis.uni-saarland.de/bfo/>, 2018. [Online]. Available: <http://ifomis.uni-saarland.de/bfo/>.

- [88] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse, "Relations in biomedical ontologies", *Genome Biology*, vol. 6, no. 5, R46, 2005, ISSN: 1474-760X. DOI: 10.1186/gb-2005-6-5-r46. [Online]. Available: <https://doi.org/10.1186/gb-2005-6-5-r46>.
- [89] JoVE, *Journal of Visualized Experiments*, 2017. [Online]. Available: <https://www.jove.com/>.
- [90] protocols.io, *Open Access Repository Of Science Methods*, Retrieved on 24/01/2018 from <https://www.protocols.io/>, 2018. [Online]. Available: <https://www.protocols.io/>.
- [91] MIBBI, *Minimum Information for Biological and Biomedical Investigations*. Retrieved on 07/07/2017 from <https://biosharing.org/collection/MIBBI>, 2017. [Online]. Available: <https://biosharing.org/collection/MIBBI>.
- [92] O. Flórez-Vargas, M. Bramhall, H. Noyes, S. Cruickshank, R. Stevens, and A. Brass, "The quality of methods reporting in parasitology experiments", *PLoS ONE*, vol. 9, no. 7, Jul. 2014, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0101131.
- [93] C. Kilkenny, N. Parsons, E. Kadyszewski, M. F. W. Festing, I. C. Cuthill, D. Fry, J. Hutton, and D. G. Altman, "Survey of the quality of experimental design, statistical analysis and reporting of research using animals", *PLoS ONE*, vol. 4, no. 11, e7824, 2009. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2779358/>.
- [94] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "PubChem Substance and Compound databases", *Nucleic Acids Research*, vol. 44, no. D1, pp. D1202–D1213, 2016, ISSN: 0305-1048.
- [95] Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He, and J. Zhang, "Pubchem bioassay: 2017 update", *Nucleic Acids Research*, vol. 45, no. D1, p. D955, 2017. DOI: 10.1093/nar/gkw1118. eprint: http://oup/backfile/content_public/journal/nar/45/d1/10.1093_nar_gkw1118/3/gkw1118.pdf. [Online]. Available: [+http://dx.doi.org/10.1093/nar/gkw1118](http://dx.doi.org/10.1093/nar/gkw1118).
- [96] UniProt, *taxonomy database*, Retrieved on 07/12/2017 from <http://www.uniprot.org/help/taxonomy>, 2017. [Online]. Available: <http://www.uniprot.org/help/taxonomy>.
- [97] C. Goble, *DON'T PUBLISH. RELEASE!*. Retrieved on 07/07/2017 from <https://www.force11.org/presentation/dont-publish-release>, In Visions of the Future, FORCE11, AMSTERDAM, NL, 2013., 2017. [Online]. Available: <https://www.force11.org/presentation/dont-publish-release>.
- [98] N. Yachie, R. B. Consortium, and T. Natsume, "Robotic crowd biology with maholo labdroids", *Nat Biotech*, vol. 35, no. 4, pp. 310–312, Apr. 2017. [Online]. Available: <http://dx.doi.org/10.1038/nbt.3758>.
- [99] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher,

M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The fair guiding principles for scientific data management and stewardship", *Scientific Data*, vol. 3, p. 160 018, 2016. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/>.

Chapter 3

Using Semantics for Representing Experimental Protocols

Background: An experimental protocol is a sequence of tasks and operations executed to perform experimental research in biological and biomedical areas, e.g. biology, genetics, immunology, neurosciences, virology. Protocols often include references to equipment, reagents, descriptions of critical steps, troubleshooting and tips, as well as any other information that researchers deem important for facilitating the reusability of the protocol. Although experimental protocols are central to reproducibility, the descriptions are often cursory. There is the need for a unified framework with respect to the syntactic structure and the semantics for representing experimental protocols.

Results: In this paper we present “*SMART Protocols ontology*”, an ontology for representing experimental protocols. Our ontology represents the protocol as a workflow with domain specific knowledge embedded within a document. We also present the *Sample Instrument Reagent Objective* (SIRO) model, which represents the minimal common information shared across experimental protocols. SIRO was conceived in the same realm as the Patient Intervention Comparison Outcome (PICO) model that supports search, retrieval and classification purposes in evidence based medicine. We evaluate our approach against a set of competency questions modeled as SPARQL queries and processed against a set of published and unpublished protocols modeled with the SP Ontology and the SIRO model. Our approach makes it possible to answer queries such as *Which protocols use tumor tissue as a sample*.

Conclusion: Improving reporting structures for experimental protocols requires collective efforts from authors, peer reviewers, editors and funding bodies. The SP Ontology is a contribution towards this goal. We build upon previous experiences and bringing together the view of researchers managing protocols in their laboratory work. Website: smartprotocols.github.io.¹

¹<https://smartprotocols.github.io>

3.1 Background

Experimental protocols are fundamental information structures that support the description of the processes by means of which results are generated in experimental research [1]. Experimental protocols describe how the data were 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 [2]. Nowadays, such protocols are generally written in natural language and presented in a “*recipe*” style, so as to make it possible for researchers to reproduce the experiments.

The quality of experimental protocols reported in articles is a cause of concern. Reproducibility, central to research, depends on well-structured and accurately described protocols. [3] found that 4 percent of the 271 journal articles assessed did not report the number of animals used anywhere in the methods or the results sections. Assessing statistical significance requires to know the number of animals participating in an experiment; it is also necessary if the experimental methods are to be reproducible, reused and adapted to similar settings. High-quality description of experimental methods is also critical when comparing results and integrating data. In an effort to address the problem of inadequate methodological reporting, journals such as Nature Protocols [4], Plant Methods (Methodology) [5] and Cold Spring Harbor Protocols [6], have guidelines for authors that include recommendations about the information that should be documented in the protocols. The ISA-TAB also illustrates work in this area; it delivers metadata standards to facilitate data collection, management and reuse from “*omic-based*” experiments [7]. The BRIDG initiative [8] aims to formalize a shared view of the dynamic and static semantics of protocol-driven research. The BioSharing initiative [9], is a catalog of standards promoting the representation of information in the life, environmental and biomedical sciences [9]. STAR [10] is an effort that proposes to “*Empowering Methods*” offering an overview of resources used in a study. Ontologies such as EXACT [11], [12] aim to formalize the description of protocols focusing on experimental actions; the BioAssay Ontology (BAO) [13] describes biological screening assays and their results; the eagle-i resource ontology (ERO) [14] represents some aspects related to protocols.

Here we present SMART Protocols ontology (henceforth SP), our ontology for representing experimental protocols; we aim to “*facilitate the semantic representation of experimental protocols*”. Our representation makes it possible to answer queries such as “*Which protocols use “tumor tissue” as a sample?*”, “*Retrieve the reagents and the corresponding information from the manufacturers for a specific protocol*”, “*retrieve the diseases caused by the reagents used in a specific protocol*”. These and other queries can be processed at our SPARQL endpoint². The SP Ontology provides the structure and semantics for data elements common across experimental protocols. For representing reagents, samples, instruments and experimental actions we reuse ontologies such as the Chemical Entities of Biological Interest (ChEBI) [15], NCBI taxonomy [16]–[18], the Ontology for Biomedical Investigations (OBI) [19], the BioAssay Ontology (BAO), The Experimental Factor Ontology (EFO) [20], eagle-i resource ontology (ERO), Cell Line Ontology (CLO) [21], [22], and EXACT. We also reuse and extend classes from the Information Artifact Ontology (IAO) [23]. In this paper we

²<http://smartprotocols.linkeddata.es/sparql>

TABLE 3.1: Repositories and number of protocols analyzed

Repository	Bio Tech	CSH	CP	GMR	JoVE	NPE	PM	PO	SP	CIAT
No. of protocols	6	9	25	5	21	13	12	5	4	75
Total	175									

The protocols are available at: <https://smartprotocols.github.io/>

also present the SIRO model; this is a minimal information model for the representation of *Samples Instruments Reagents Objective* (hence SIRO). This model has been conceived in a way similar to that of the **Patient Intervention Comparison Outcome** (PICO) model; it helps to frame questions and provides an anchor for the records [24]. SIRO facilitates classification and retrieval without exposing the content of the document. In this way, publishers and laboratories may keep the content private, exposing only the information that describes the sample, instruments, reagent and objective of the protocol. As an illustration, in this paper we use the protocol “*Extraction of total RNA from fresh/frozen tissue (FT)*” [25] as a running example. We represent this protocol with the SP ontology and SIRO.

3.2 Methods

Our SMART Protocols ontology [26] is based on an exhaustive analysis of 175 published and unpublished experimental protocols (see Table 3.1 in Domain Analysis and Knowledge Acquisition, DAKA); we also analyzed on-line repositories and guidelines for authors. For the development of the SP Ontology [1] we have followed the practices recommended by the NeOn methodology [27], as well as those reported by García [28]. For example, we used conceptual maps to better understand the correspondences, relations and possible hierarchies in the knowledge we were representing. The stages and activities we implemented throughout our ontology development process are illustrated in Figure 6.1 and explained below. For the ontology development process we also considered the guidelines from the OBO foundry [29].

3.2.1 The Kick-off, Scenarios and Competency Questions

In the first stage, we gathered motivating scenarios, competency questions, and requirements. We focused on the functional aspects that we wanted the ontology to represent. Domain experts were asked to provide us with a list of competency questions, these are presented in our website³. Some of the competency questions we gathered include, “retrieve the protocols using a given sample” and, “which protocols can I use to process this sample given that I only have X and Z reagents”. Competency questions were initially used to scope the domain for which we were developing the ontology; these questions were also used during the evaluation.

3.2.2 Conceptualization and Formalization

In this stage we identified reusable terminology from other ontologies; for supporting activities throughout this stage we used BioPortal [30] and Ontobee [31]. We also looked into minimal information standards [32], guidelines and vocabularies representing research activities [33]–[35]. Issues about axioms required to represent this

³<https://smartprotocols.github.io/queries/>

FIGURE 3.1: Developing the SMART Protocols ontology, methodology

domain were discussed and tested in Protégé v. 4.3 and 5.0 [36]; during the iterative ontology building, classes and properties were constantly changing. We identified, and explain below, three main activities throughout this stage, namely: Domain Analysis and Knowledge Acquisition (DAKA), Linguistic and Semantic Analysis (LISA), Iterative ontology building and validation (IO).

Domain Analysis and Knowledge Acquisition, DAKA

We manually reviewed 175 published and unpublished protocols from topic areas such as molecular biology, cell and developmental biology, biochemistry, biotechnology, microbiology and virology, as well as guidelines for authors from journals. The unpublished protocols (75 in total) were collected from four laboratories located at The International Center for Tropical Agriculture (CIAT)[37]. The published protocols (open access) were gathered from 9 repositories; table 3.1 presents the list of journals and the number of protocols that we analyzed. We used these sources to prepare a checklist with data elements that were required in guidelines for authors and also present in published protocols –see Annex 1⁴. This was the seed for our discussions with domain experts.

Our domain analysis focused on gathering terminology and data elements, idem higher abstractions that could be used to group terminology. Domain experts were bringing their protocols and discussing specific issues, e.g. what was missing for applying a particular protocol. As the discussions were progressing, published and unpublished protocols were added to the mix. Due to time constraints domain experts were not required to work before or after the workshops. Olga Giraldo was the facilitator for the DAKA activities. This made the processes with domain experts

⁴<https://smartprotocols.github.io/annex/>

more efficient because she has extensive experience in laboratory practices. Ten domain experts participated in DAKA; they all had hands-on experience in areas such as molecular biology, virology, plant breeding, biochemistry, clinical microbiology and pathology. From DAKA we confirmed most of the data elements in our initial checklist and identified clusters of terminology, e.g. samples and instruments. The output of this activity was an improved checklist and relations to the information in the protocols. This output was used as input for the linguistic analysis.

Linguistic and Semantic Analysis, LISA

From our corpus of protocols we selected 100 documents; these represented the topic areas for which we had domain experts. We tried to have some complex and lengthy protocols involving several procedures and technologies; for instance, protocols describing the development of an SNP genotyping resource [38] and protocols describing the construction of an RNA-seq library [39]. We also worked with simpler protocols such as sample preparation or DNA extraction protocols. The terminology gathered in DAKA was discussed with domain experts and analyzed against existing ontologies; BioPortal and Ontobee were used to browse the ontologies in order to determine how terms were related to biomedical ontologies and which were the ontologies that could be relevant for this work.

Throughout this activity we also addressed the representation of workflows in the protocols. This was particularly problematic because domain experts did not agree on how granular the descriptions of the workflows and the relation between steps needed to be, how to indicate order in the sequence of operations and, what information was obligatory in the description of the steps. In this activity we used an on-line survey that helped us to determine and validate what data elements were necessary and sufficient for the description of the protocols –see Annex 2³. We used the outputs from DAKA in the survey and asked participants to indicate whether a particular data element was relevant or not; an invitation to participate was circulated over mailing lists, participants did not have to disclose their identity. Twenty participants filled up the survey; this survey helped us to informally validate the outputs from DAKA and also gave us another perspective about relevant data elements in the description of protocols. Results from the survey are available in Annex 3³.

From this activity, we identified linguistic structures that authors were using to represent actions. We were interested in understanding how verbs were representing actions and what additional information was indicating the attributes for actions. For instance, “*Fresh-leaf tissue (0.2 g) was **ground** in a 1.5-mL Eppendorf tube with a **micropestle** and preheated freshly prepared 800 uL **extraction buffer** was immediately added to the tube*” [40] is a commonly used cell disruption step in nucleic acids and protein extraction protocols. In our corpus of documents, these steps were usually described using verbs like “break, chop, grind, homogenize”. There are also common methods for specific operations; for instance, for breaking the cells the methods were “blending, grinding or sonicating” the sample. The sequence of instructions had an implicit order that was not always clearly specified as authors sometimes hide it in the narrative. There is, however, an input-output structure. Actions in the workflow of instructions are usually indicated by verbs; accurate information for implementing the action implicit in the verb was not always available. For instance, structures such as “Mix thoroughly at room temperature”, “Briefly spin the racked tubes” are common in our dataset. The instructions always have actions and participants, which

may be samples, reagents, instruments and/or measures. This was particularly useful in the definition of our workflow; the pattern that emerged is discussed in the "Results" section. In this activity we also identified document-related data elements; for instance, roles for authors, e.g. validator, statistical reviewer. We also identified the ontologies that could represent the concepts we were working with. A draft ontology with the seminal terminology and initial classification was the output from LISA; this output was further refined during the iterative ontology building stage.

Iterative ontology building and validation, IO

The draft ontology from LISA was incrementally growing in complexity, number of concepts and relations. The knowledge engineer conducted continuous evaluations of the draft ontologies against competency questions. The ontology models were shared with domain experts, they reviewed the drafts, gave feedback and the ontology was updated.

As we were building ontology models, we identified the modularity needed to represent experimental protocols. From our models, we conceptualized the protocols as workflows embedded within documents. Thus, the document module of SP ontology (henceforth SP-Document) was designed to provide a structured vocabulary that could represent information for reporting an experimental protocol. The workflow module of SP ontology (henceforth SP-workflow) delivers a structured vocabulary to represent the sequence of actions in the execution of experimental protocols. The main outcome from this activity was an ontology with the SP-Document and SP-workflow modules and their corresponding classes and object properties. Our ontologies were developed using OWL-DL. We used the Protégé editor versions 4.X and 5; the Protégé plug-in OWLViz [41] was used to visualize the model.

3.2.3 Ontology Evaluation

during the evaluation process, we addressed issues related to the syntax, the conceptualization and formalization. We also verified whether the competency questions could be resolved by representing experimental protocols using the ontology and having the resulting RDF in a SPARQL endpoint.

We evaluated the syntax of the ontology using The Ontology Pitfall Scanner (OOPS) [42]; it was useful to detect and correct anomalies or pitfalls in our ontologies [43]. For instance, the identification of incomplete inverse object properties, lack of domain and range, missing annotations and issues in naming conventions. The resulting ontology from the "Conceptualization and Formalization" phase was evaluated by 10 domain experts. They were asked to determine if the proposed classes in the ontology could represent the information from a set of 13 protocols that we selected for this purpose. A list of the protocols as well as results from this evaluation are presented in Annex 4³.

We also tested the capability of the SMART Protocols ontology to answer the competency questions specified by domain experts; does the ontology represent enough information to answer these types of questions? do the answers require a particular level of detail or representation of a particular area? This part of the evaluation entailed the transformation of 10 experimental protocols to RDF⁵. These were uploaded in our SPARQL endpoint and the queries were formalized in SPARQL; a complete list of SPARQL queries has been made available².

⁵<https://smartprotocols.github.io/protocolsrdf/>

3.3 Results

3.3.1 The SMART Protocols ontology

Our ontology reuses BFO; we are also reusing the ontology of relations (RO) [44] to characterize concepts. In addition, each term in the SP ontology is represented with annotation properties imported from OBI Minimal metadata[45]. The classes, properties and individuals are represented by their respective labels to facilitate readability. The prefix indicates the provenance for each term; for instance, the prefix `sp` is used to identify classes and object properties from SP ontology. For the object properties we are using italics, words or phrases representing instances are in between quotation marks, e.g. “RNA extraction”, instance of the class `sp:lab procedure` 3. In this section we use the protocol “*Extraction of total RNA from fresh/frozen tissue (FT)*” [25] as a running example to represents the document and workflow aspects of a protocol. Our ontology is available in BioPortal ⁶, github ⁷ and also is registered at vocab.linkeddata.es ⁸. vocab.linkeddata.es is a list of vocabularies developed by the Ontology Engineering Group (OEG). A graphical illustration of the ontology can be found at Annex 5 ³.

The Document Module

The document module of the SP ontology [46] aims to provide a structured vocabulary of terms to represent information for 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 represents metadata elements as classes, some of them are: `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`. We have used the SP-Document module to represent our running example, the results are presented in Table 3.2 and Figure 6.2; metadata elements are organized in SP-Document as information content entities. In order to facilitate the use of identifiers for the material entities like reagents and equipments, we created the object property *sp:has catalog number* and the class `sp:catalog number`. In this way a relation is established between the reagent or equipment and the corresponding manufacturer.

The Workflow Module

The SP ontology also considers the protocol as an executable element to be carried out and maintained by humans. The workflow module [47] is a descriptive model for workflows; it is not a workflow programming language. The workflow module represents the procedures, subprocedures, actions (or verbs), experimental inputs (samples/specimens) and other participants such as reagents and instruments. Experimental protocols often include a set of laboratory procedures; these transform inputs into outputs. Our running example (see Figure 6.3 and Table 3.3), includes 3 laboratory procedures: `sp:lab procedure 1` (“Protocol overview”, indicating how to process the sample), `sp:lab procedure 2` (“Prior to RNA extraction: cleaning

⁶<http://bioportal.bioontology.org/ontologies/SP>

⁷<https://smartprotocols.github.io/>

⁸<http://vocab.linkeddata.es/SMARTProtocols/>

TABLE 3.2: Metadata represented in SP-Document

Metadata	Caption
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 2–21) is taken from the method supplied with TRIzol reagent Invitrogen, Paisley, UK)."
Metadata for materials	
sp:specimen name	"tumor tissue"
sp:reagent name	"TRIzol", "Chloroform", "Ethyl alcohol", "Isopropyl alcohol"
sp:manufacturer name	"Invitrogen", "Sigma-Aldrich"
sp:equipment or supplies name	"Tissue storage container", "Homogenizer blades", "Forceps", "Scalpel", "Scalpel holder"

process of equipment") and sp:lab procedure 3 ("RNA extraction"). The first column in Table 3.3 includes the procedures from our running example. The second column includes subprocedures or instructions for each procedure.

The class sp:lab procedure 1 ("Protocol overview") has a tumor tissue (nci:tumor tissue) as an input (*sp:has experimental input*); in a similar way, the lab procedure 1 has a homogenized tissue (sp:homogenized tissue) as an output (*sp:has output*). The laboratory procedure 1 includes 3 subprocedures (or steps/instructions) indicating how to manipulate and prepare the sample, namely: sp:lab subprocedure 1.1, sp:lab subprocedure 1.2 and sp:lab subprocedure 1.3. The order in which these subprocedures should be executed is represented by the BFO property *is preceded by* and *precedes*. The class sp:lab procedure 2 ("Prior to RNA extraction: cleaning process of equipment") is a recipe describing how to clean the equipment to be used during the RNA extraction protocol. This recipe includes 3 steps, sp:lab subprocedure 2.1, sp:lab subprocedure 2.2 and sp:lab subprocedure 2.3.

The class sp:lab procedure 3 ("RNA extraction") has the homogenized tissue (output from the lab procedure 1) as an input and, the class chebi:RNA as an output. It includes 20 subprocedures, these are not represented in the Figure 6.3 due to lack of space. We propose the classes sp:laboratory procedure and sp:laboratory subprocedure for the representation of procedures and subprocedures. The object property, *sp:has procedure*, is used to characterize the laboratory procedures that are part of the execution of an experimental protocol (sp:experimental protocol execution); the object property *sp:has subprocedure*, is used to characterize the

FIGURE 3.2: SP-Document module. This diagram illustrates the metadata elements described in Table 2. The classes, properties and individuals are represented by their respective labels.

subprocedures that are part of a given procedure. Procedures have inputs and outputs, subprocedures have participants. For cases where authors only have an extensive list of steps, the SP ontology considers these as subprocedures under a procedure container. In this way we are representing protocols with only a long list of steps as well as those with groups of steps. This also allows us to represent more complex protocols that usually result from merging several protocols.

We are representing antibodies, cell lines and plasmids as material entities. We are using *ro:derives from* to indicate that it derives from an organism; similarly, we are using the *obi:has_role* to indicate the role that it plays, as understood by the author of the protocol.

3.3.2 Evaluation

Syntax

OOPS allowed us to identify the lack of domain and range in the object properties *ro:part_of* and *ro:has_part*; these were imported from the Relations Ontology (RO). We verified in the original ontology and these two properties do not have domain and range [48]. OOPS was useful for verifying the syntax of the ontology.

Conceptualization and Formalization

The resulting ontology was evaluated by 10 domain experts, they were asked to determine whether the resulting ontology was representing the information items from experimental protocols. This evaluation was satisfactory because the information from the protocols was represented in the ontology. Interestingly, resulting from

TABLE 3.3: Procedures and subprocedures from “Extraction of total RNA from fresh/frozen tissue (FT)”

Procedure	Subprocedure
Protocol overview (sp:lab procedure 1)	Recover tumor tissue at the time of surgery, trim into 1-cm ³ fragments, and immerse immediately in TRIzol reagent prior to freezing at -80° .
	Thaw and weigh tissue prior to RNA extraction, working quickly.
	Use a tissue power homogenizer (or a mortar and pestle) to homogenize tissue by hand.
Prior to RNA extraction: cleaning process of equipment (sp:lab procedure 2)	Autoclave or wash equipment (i.e., tissue storage container, homogenizer blades, forceps, scalpel holder) in Neutracon solution for 2–4 h.
	Rinse equipment well in 1% SDS (prepared using DEPC-treated or other nuclease-free water).
	Rinse in 100% ethanol and leave to air-dry.
RNA extraction (sp:lab procedure 3)	Homogenize sample using tissue homogenizer.
	Add 0.2 mL chloroform per 1 mL TRIzol and cap tube tightly.
	Add 0.5 mL isopropyl alcohol per 1 mL TRIzol.
	Add 1 mL 75% ethanol per 1 mL TRIzol and vortex for 10 s.

this evaluation we could identify some issues related to the way published and unpublished protocols were described. For instance, published protocols don’t have any information that facilitates the identification of roles; for instance, who is the chief scientist, who did the statistical validation, who was the lab scientist, etc. Identifying these roles was considered as important because it is an indication of quality control in the development of the protocol; this data element was identifiable in unpublished protocols and it is part of our ontology. Unpublished protocols usually have version information, as well as a short description of the roles played by those who are using, developing, standardizing or modifying the protocols.

From this evaluation it was also evident that published protocols were not consistent in the data elements that they use to represent the experimental protocol. For instance, some of the protocols had an explicit description of “advantages” and “application of the protocol”, while some others did not provide this information. A similar situation was found with respect to information about limitations. The bibliographic metadata that was identified includes, title, author, subject area and protocols identifiers (IDs). These were not always available; in the case of unpublished protocols the ID was sometimes an internal code. Although the class author identifier (sp:author identifier) could not be instantiated, we decided to leave it in the ontology because it was deemed important. Published and unpublished protocols have authors as literal values without any relation to IDs.

FIGURE 3.3: SP-Workflow module. This diagram illustrates the meta-data elements described in Table 3. The classes, properties and individuals are represented by their respective labels.

Published and unpublished protocols often report the name of the materials but not the manufacturer and the corresponding identifier, this is usually the catalog number. This information is frequently available and it is always necessary when trying to reuse a protocol, the SP Ontology models these data elements. Alert messages, hints, pause points, cautions or troubleshooting were represented in SMART Protocols ontology and validated by the domain experts. Although the description of the work steps, procedures, subprocedures and recipes varied across the protocols, the data elements describing the workflow could be easily represented in our ontology.

We also asked domain experts to instantiate the classes with text from the protocols. They were selecting excerpts of text and assigning classes to these narratives, e.g. "This is a simple protocol for isolating genomic DNA from fresh plant tissues" was classified as an objective, "DNA from this experiment can be used for all kinds of genetics studies, including genotyping and mapping" was classified as an application. They were also selecting some specific words and classifying them; for instance, "Isopropanol" was classified as a reagent, "mortar and pestle" was classified as an equipment. Information related to the overall objective of the protocol, applications, advantages, limitations and provenance was represented in our ontology; these data elements were validated by domain experts as they were mapping them to the ontology. Information about the sample (strain, line, genotype, developmental stage, organism part, growth conditions, treatment type and quantity used) was identified in published and unpublished protocols and could easily be mapped to the ontology.

Materials were also identified and mapped; interestingly, domain experts recognized different types of materials, for instance, instruments (including laboratory consumables), reagents, kits and software. In the resulting ontology we included

“reagent” and “kit” under material entities; this made it easier for domain experts to identify terminology related to these classes. Published and unpublished protocols don’t differentiate across reagents, recipes, and kits; these are all usually listed under “Reagents”. However, domain experts reusing the protocols understand these under different categories. Reagents are understood as “ready to use”, often purchased; they also included mixtures prepared in the lab under reagents. Reagents are substances used in a chemical reaction to detect, measure, examine, or produce other substances [49]. Kits were considered as “gear consisting of a set of articles or tools for a specified purpose”. For instance, the Qiagen RNeasy Spin mini is a kit for purification of RNA from cells and tissues. However, a kit could also be an instrument; for instance, a digital recording transcribing kit, an instrument used to digitally record speech for transcription.

Recipes were identified as the most appropriate part of the protocol for including the details indicating how to prepare a particular solution, media, buffer, etc. The recipes could also describe how to make something; for example, “recipes describing how to clean laboratory equipment before starting the execution of a procedure”, see lab procedure 2 in our running example (Figure 6.3 and Table 3.3); a recipe also is a way to include details regarding, e.g., the setup of HPLC separation methods. We classified the term “recipe” as a textual entity. The execution of a recipe was also considered, we included the term “recipe execution” as a planned process.

Competency questions

The RDF generated from instantiating the ontology was loaded in our SPARQL endpoint; the competency questions were then executed against this dataset. In general the expected information was retrieved; however, as domain experts were looking at the results, they started to reformulate the questions by asking for more information. For instance, domain experts asked for reagents to be linked to catalogs from the manufacturers or to resources like PubChem [50]. They were also interested in linking the samples/organisms to DBPEDIA [51] and NCBI taxonomy database [17], [18]; similarly, safety information was deemed as another case for establishing links between entities in the protocol and other information resources in the web. Some queries making use of linked data resources via federated queries illustrate this requirement; as additional information was necessary, we were looking into linked data resources that could complement the retrieved information. Queries like “Retrieve all the reagents and the information about where to buy them” illustrate how we were making use of other information resources; federated queries, see ², are retrieving complementary information from linked data resources such as DBpedia, Uniprot [52], PubChem, SNOMED over BioPortal and ChEBI. Some of the federated queries are presented in Table 3.4.

3.4 Applying the SMART Protocols Ontology to the Definition of a Minimal Information Model

Initially we developed the SP ontology and then the SIRO model. As we were representing the protocols as RDF we were also analyzing the competency questions; by doing so we saw a common pattern. From our competency questions, 17.4 percent were related to Samples, 8.7 percent were related to Instruments, 34.8 percent were related to Reagents Figure 6.4. Furthermore, although the description of the workflow varies across our evaluation corpus, these data elements were always present.

We focused on the manual identification of commonalities, the very minimal information shared across our corpus of documents. We then classified these data elements by mapping them to the SP ontology. This allowed us to determine higher abstractions to which the terminology could be mapped, e.g., “sample”, “reagent” and “instrument”. Domain experts discussed the granularity of the workflow description, whether the limitations of the protocol should or should not be reported, how to report the application of the protocol, etc. However, there was no disagreement about the need to report the objective of the protocol, e.g. “*method for the production of 3D cell lysates that does not compromise cell adhesion before cell lysis*”. Unlike samples, instruments and reagents, the objective is not always easily identifiable; it may be scattered throughout the document. It is, however, an important element; the description of the objective makes it easier for the readers to decide on the suitability of the protocol for their experimental problem. The SIRO model is illustrated in Fig 6.5.

FIGURE 3.4: Distribution of SIRO elements

3.4.1 The Sample Instrument Reagent Objective (SIRO) Model

SIRO represents the minimal common information shared across experimental protocols. It serves two purposes. First, it extends available metadata for experimental protocols, e.g. author, title, date, journal, abstract, and other properties that are available for published experimental protocols. SIRO extends this layer of metadata by aggregating information about Sample, Instrument, Reagent and Objective –hence the name. Categories and instances of the data elements for SIRO are presented in table 3.5. Second, SIRO makes it possible to frame and answer queries based on the minimal common data elements in experimental protocols. This facilitates finding specific protocols; if the owner of the protocol chooses not to expose the full content, as it is the case of publishers and/or laboratories, SIRO may be exposed without compromising the full content of the document. For instance, queries such as “*retrieve protocols that use samples from the rodent order*” or “*retrieve protocols that use Nucleic acid purification kits*” are executed using information that is also part of the SIRO model. Retrieving information related to steps, procedures, and recipes is

FIGURE 3.5: The SIRO model

only possible if the protocol is public, e.g. open access. In our case, CIAT facilitated some protocols for which only SIRO elements could be exposed; steps, alert messages and troubleshooting were considered as sensible information that should not be publically available.

3.4.2 Evaluating the SIRO Model

For evaluating SIRO we extracted and populated the SIRO model with the RDF dataset that we used for the evaluation of the SP ontology. As the SIRO model does not expose the whole content of the protocol we also added five unpublished, private, protocols to the dataset. In total, for this evaluation we have 15 protocols in the SPARQL endpoint ⁴. For those queries involving instances of SIRO, we could satisfactorily retrieve the information required by the competency questions. Moreover, as SIRO complements bibliographic metadata information, the wealth of queries can be expanded. For instance:

- Retrieve the protocols and the list of reagents for documents authored by Yoshimi Umemura.
- Retrieve the protocols authored by Yoshimi Umemura and Beata Dedicova using rice leaves as sample.
- Retrieve the common reagents across the protocols "[Bio101] Subcutaneous Injection of Tumor Cells" and "Scratch Wound Healing Assay".

3.5 Discussion

3.5.1 SMART Protocols Ontology

We propose the SP ontology to represent experimental protocols. It reuses the meta-data structure, as well as some classes and properties, from OBI. It also builds upon experiences such as the BioAssay Ontology (BAO), The Experimental Factor Ontology (EFO), eagle-i resource ontology (ERO) and also the EXACT ontology. The SP Ontology also considers reporting structures such as ARRIVE, BRIDG as well as those from BioSharing. For representing “instruments”, “reagents/chemical compounds”, “organisms” and “sample/specimen” we reuse, amongst others, NCBI taxonomy, Cell Line Ontology (CLO) and Chemical Entities of Biological Interest (ChEBI). Our results indicate that the SP ontology makes it possible to represent all the data elements in the experimental protocols that we have analyzed.

3.5.2 Modularization of the SP ontology

Modularization, as it has been implemented in SP, facilitates specializing the ontology with more precise formalisms. For instance, reagents, instruments and experimental procedures (actions), may be instantiated based on the activities carried out by a particular laboratory. We have two main modules in our ontology, the SP-Document and the SP-Workflow modules. The document module address issues related to archiving and representing the narrative. The workflow module aims to deliver a reusable executable object. In this way we make it possible for protocols to “*be born semantics*”. To “*be born semantics*” delivers a self-describing workflow embedded within a document from the onset. As a document, it is easily managed and understood by humans. As a self-describing workflow embedded within a document it is easily processed by machines. Our representation has some limitations with respect to machine processability; for instance, it is not suitable for robots to interpret it.

The document module facilitates archiving; publishers and laboratories can extend it depending on their use cases. The workflow module delivers an extensible representation describing the sequence of activities in an experimental protocol. Actions, as presented by [11], 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) are also necessary. Our workflow representation makes it possible to link procedures and subprocedures to reagents, instruments, samples, recipes, hints, alert messages, etc. This is particularly useful because procedures and subprocedures can easily be reused and adapted; also, it allows researchers to retrieve very specific information and aggregate other data elements as it is needed. Formalizing workflows has an extensive history in Computer Science; not only in planning but also in execution –as in Process Life-cycle Management and Computer Assisted Design/Computer Assisted Manufacturing. The SP-workflow module helps to formalize the workflow implicit in protocols; our workflow specification has some limitations. For instance, loops, conditionals and other workflow constructs are currently being formalized as new use cases are identified. Our workflow constructs are easily extensible; we are also evaluating formal workflow languages for processes and adapting these to the biomedical scenario. Overcoming the limitations in the description of the workflow will make it possible to have an accurate representation of the protocol as an executable object for machines to fully process -including robots. The workflow nature

implicit in experimental protocols should also be intelligible and manageable by humans; we are currently exposing the protocols in a format, RDF, that machines can understand for web purposes, e.g. discovery, interoperability.

3.5.3 Limitations

Describing samples was particularly difficult because attributes like strain, line or genotype, developmental stage, organism part, growth conditions, age, gender, pre-treatment of the sample and volume/mass of sample, etc, are important depending on the experiment and the type of sample. Reagents and instruments were easier to describe as they only require the commercial name, manufacturer and identification number. However, linking reagents and instruments to other information resources is not as simple. Manufacturers don't always offer Application Programming Interfaces (APIs) that make it possible to resolve these entities against their websites. For our experiment we had to scrape these websites in order to build the links. Furthermore, they don't always use controlled vocabularies, common identifiers or describe chemicals in the same way; this made it difficult to search across their catalogs. Sigma-Aldrich and PubChem link to each other and PubChem has links to several manufacturers and vendors, this was deemed useful by domain experts. Linking was not initially considered by domain experts in their early competency questions; however, when they saw the answers for their queries, their expectation for linking data grew. In order to meet this demand, we re-formulated the queries by adding some external resources. This was received with satisfaction by domain experts; however, the expectation for more data continued growing. The use of external data sources was problem dependent, so were the external data sources to use.

3.5.4 The SIRO model, application of the ontology

The SIRO model for minimal information breaks down the protocol in key elements that we have found to be common across our corpus of experimental protocols: *i*) Sample/ Specimen (**S**), *ii*) Instruments (**I**), *iii*) Reagents (**R**) and *iv*) Objective (**O**). Exposing SIRO makes it possible for laboratories and publishers to present key elements that frame questions often asked by researchers when searching for experimental protocols. SIRO was tested and results were satisfactory. External sources of information, e.g. vendor information from PubChem, can also be used to enrich SIRO elements. By extending the bibliographic metadata, SIRO is also extending the wealth of queries being supported; it provides specific information that is relevant to the description of the protocol.

3.6 Conclusions

Experimental protocols are central to reproducibility and they are widely used in experimental laboratories. Our ontology and minimal information model have been validated with domain experts; our evaluations indicate that the SP ontology can represent experimental workflows and also that retrieving specific information from protocols represented with the SP ontology is possible. Both, the ontology and the SIRO model are easily adaptable. Experimental protocols describe step by step "*how to do or how to execute*" an experimental procedure. In our conceptualization experimental protocols have a document and a workflow component; as workflows embedded within documents, the experimental protocols should have complete information that allows anybody to recreate an experiment.

Our approach facilitates the generation of a self-describing document. It makes it possible to present meaningful information of experimental protocols without compromising the content. More importantly, it makes it possible to anchor information retrieval within a context that is meaningful for experimental researchers, e.g. reagents, samples and instruments participating in subprocedures. Queries such as *“What DNA extraction protocol is used on rice samples?”*, *“what amount of leaf tissue to use?”* are common for experimental researchers; answering these is possible with the SP ontology. In laboratory settings experimental protocols are usually managed just like any other document. However, these are plans for the execution of experiments; resources are allocated based on specifics described in the workflows of experimental protocols. The SMART Protocols approach generates a computable document that may interoperate with, for instance, inventories or Laboratory Information Management Systems (LIMS). Thus making it easier for researchers to plan according to available resources.

Harmonizing efforts such as EXACT, OBI, STAR [10], BRIDG and SMART Protocols ontology is important because without a clear semantics, reporting structure and a minimal information model for experimental protocols these will remain highly idiosyncratic. Moreover, without such consensus the experimental record will remain highly fragmented and therefore not easily processable by machines or reproducible by humans. Efforts such as the Resource Identification Initiative (RRId) [32], [53] and identifiers.org [54], [55] are central in the preservation of the experimental record; it is important that these efforts start to address reagents and instruments more broadly as these resources don’t always have identifiers. Being able to review the data makes it possible to evaluate whether the analysis and conclusions drawn are accurate. However, it does little to validate the quality and accuracy of the data itself. The data must be available, so does the experimental protocol detailing the methodology followed to derive the data. Journals and funders are now asking for datasets to be publicly available; there have been several efforts addressing the problem of data repositories; if data must be public and available, shouldn’t researchers be held to the same principle when it comes to methodologies? Openness and reproducibility are not only related to data availability; when replicating research, being able to follow the steps leading to the production of data is equally important.

The SP ontology is a digital object that follows the FAIR Principles [56]. Our ontology is **findable**; it is registered at Bioportal ⁵, it is also available in github ⁶ and the vocab.linkeddata.es ⁷. The ontology is documented to facilitate the **reusability**; classes and object properties are documented with annotation properties imported from the OBI Minimal metadata. Reusing the ontology is easy as it has “preferred terms”, “definitions”, “definition sources”, “example of use”, “alternative terms”, etc; this makes it easier for others to know the context of the terminology as well as the suitability for addressing other use cases. The SP ontology was developed in OWL-DL and it is licensed under a Creative Commons Attribution 4.0 International License; in this sense SP ontology is **interoperable** and **accessible**.

TABLE 3.4: Queries making use of external resources. Queries are available at <https://smartprotocols.github.io/queries/>

Competency question	Was the question answered?	Other Information Resources	SPARQL	Comment
Retrieve all the protocols that use mouse as a sample	Yes. Could there be a short description about the organism and also, mouse is too specific, I may also be interested in rats and other rodents.	The DBPEDIA property <code>dbo:order</code> includes individuals that belong to the order rodents, e.g. rats, hamsters, squirrels, etc. DBPEDIA also has <code>dbo:abstract</code> , this property allows us to retrieve information about rodents.	Query#1. Retrieve all the protocols with samples that belong to the Rodent order and also retrieve information for these samples	Additional information was useful but basic
Retrieve all the reagents used in the protocols	Yes. It is also useful to know where to buy these products.	PubChem has a list of vendors for some reagents. For instance, for sodium chloride it has more than ten vendors. Also, we are resolving the entities against the websites of the manufacturers.	Query #4. Retrieve all the reagents along with the different web sites to buy them and all the different manufacturers registered for every reagent	Additional information was useful
Retrieve the protocols in which Bromophenol blue is used	Yes. Could the applications for the reagent be included in the answer?	ChEBI is an external resource that has the applications for some reagents.	Query #23 Retrieve the protocols in which Bromophenol blue is used and tell me about the application of Bromophenol blue	Additional information was useful
Retrieve the steps that have CAUTIONS as alert messages from the protocol "X"	Yes. I would also like to have the diseases caused by this reagent	In this case we are making use of Biportal and SNOMED (<code>causative_agent_of</code>).	Query #14. Retrieve all the diseases caused by the reagents in the protocol "Extraction of total RNA from fresh/frozen tissue (FT)"	Additional information was useful

TABLE 3.5: SIRO Elements

Sample	Whole organism	Scientific name: <i>Arabidopsis thaliana</i> , <i>Oriza sativa</i> , <i>mangifera indica</i> , <i>Mus musculus</i> . Common names: Mouseear Cress, rice, mango, mouse.
	Anatomical part	Leaf, stem, cells, tissues, membranes, organs, skeletal system, muscular system, nervous system, reproductive system, cardiovascular system, etc.
	Biomolecules	Nucleic acids: Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Proteins: enzymes, structural or support proteins (keratin, elastin, collagen), antibodies, hormones, etc.
	Body fluids	Blood serum, saliva, semen, amniotic fluid, cerebrospinal fluid, gastric acid, etc.
Instrument	High-throughput equipment	Liquid Handling Platforms, Real-Time PCR Detection System, Microplate Reader, etc.
	Instruments	Goggles, Bunsen burner, spot plate, pipet, forceps, test tube rack, mortar and pestle, etc.
	Laboratory glassware	Beaker, Erlenmeyer flask, graduated cylinder, volumetric flask, etc.
	Standard equipment	Balances, shakers, centrifuges, refrigerators, incubators, thermocyclers, fume hood, etc.
	Consumables	Weighing dishes, pipette tips, gloves, syringes, petri dishes, test tubes, micro centrifuge tubes, glass slides, filter paper, etc.
Reagents	Chemical compound/Substance	Glucose, ethanol, glycerol, chloroform, acetic acid, isopropyl alcohol, etc.
	Solutions / buffers	70% ethanol, 10X PCR buffer, phenol:chloroform:isoamyl alcohol, etc.
	Cell culture media	Nutrient media, minimal media, selective media, differential media, etc.
Objective	Part of discourse	Here we present a detailed protocol for Smart-seq2 that allows the generation of full-length cDNA and sequencing libraries by using standard reagents

Bibliography

- [1] O. Giraldo, A. Garcia, and C. Oscar, "SMART Protocols: SeMAntic Representation for Experimental Protocols", 4th Workshop on Linked Science 2014-Making Sense Out of Data (LISC2014), Riva del Garda, Trentino, Italy, 2014.
- [2] L. G. Acevedo, A. L. Iniguez, H. L. Holster, X. Zhang, R. Green, and P. J. Farnham, "Genome-scale ChIP-chip analysis using 10,000 human cells", *Biotechniques*, vol. 43, no. 6, pp. 791–797, 2007.
- [3] C. Kilkenny, W. J. Browne, I. C. Cuthill, M. Emerson, and D. G. Altman, "Improving bioscience research reporting: The arrive guidelines for reporting animal research", *PLoS Biol*, vol. 8, no. 6, e1000412, 2010.
- [4] (). Nature protocols, guide to authors. <http://www.nature.com/nprot/info/gta.html>. Accessed 7 May 2016.
- [5] (). Plant methods – BioMed Central, submission guidelines. <http://plantmethods.biomedcentral.com/submission-guidelines/preparing-your-manuscript/methodology>. Accessed 7 May 2016.
- [6] CSH-Protocols, *Cold Spring Harbor Protocols, Instructions for Authors*, Retrieved on 07/07/2013, 2013. [Online]. Available: <http://cshlpress.com/cshprotocols/>.
- [7] P. Rocca-Serra, S.-A. Sansone, and M. Brand, *Release candidate 1, ISA-TAB v1.0 specification document, version 24th*. 2008, p. 36. [Online]. Available: http://isatab.sourceforge.net/docs/ISA-TAB_release-candidate-1_v1.0_24nov08.pdf.
- [8] *Biomedical Research Integrated Domain Group*. [Online]. Available: <https://bridgmodel.nci.nih.gov/> (visited on 02/10/2017).
- [9] C. F. Taylor, D. Field, S.-A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C. A. Ball, P.-A. Binz, M. Bogue, T. Booth, A. Brazma, R. R. Brinkman, A. Michael Clark, E. W. Deutsch, O. Fiehn, J. Fostel, P. Ghazal, F. Gibson, T. Gray, G. Grimes, J. M. Hancock, N. W. Hardy, H. Hermjakob, R. K. Julian, M. Kane, C. Kettner, C. Kinsinger, E. Kolker, M. Kuiper, N. L. Novere, J. Leebens-Mack, S. E. Lewis, P. Lord, A.-M. Mallon, N. Marthandan, H. Masuya, R. McNally, A. Mehrle, N. Morrison, S. Orchard, J. Quackenbush, J. M. Reecy, D. G. Robertson, P. Rocca-Serra, H. Rodriguez, H. Rosenfelder, J. Santoyo-Lopez, R. H. Scheuermann, D. Schober, B. Smith, J. Snape, C. J. Stoeckert, K. Tipton, P. Sterk, A. Untergasser, J. Vandesompele, and S. Wiemann, "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project", *Nature Biotechnology*, vol. 26, no. 8, pp. 889–96, 2008.
- [10] E. Marcus, "A STAR Is Born", *Cell*, vol. 166, no. 5, pp. 1059–1060, 2016, ISSN: 00928674.

- [11] L. N. Soldatova, W. Aubrey, R. D. King, and A. Clare, "The EXACT description of biomedical protocols", *Bioinformatics*, vol. 24, no. 13, pp. i295–i303, 2008, ISSN: 1367-4803. [Online]. Available: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn156>.
- [12] L. N. Soldatova, D. Nadis, R. D. King, P. S. Basu, E. Haddi, V. Baumlé, N. J. Saunders, W. Marwan, and B. B. Rudkin, "EXACT2: the semantics of biomedical protocols", *BMC Bioinformatics*, vol. 15, no. Suppl 14, S5, 2014, ISSN: 1471-2105. [Online]. Available: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-S14-S5>.
- [13] S. Abeyruwan, U. D. Vempati, H. Küçük-McGinty, U. Visser, A. Koleti, A. Mir, K. Sakurai, C. Chung, J. A. Bittker, P. A. Clemons, S. Brudz, A. Siripala, A. J. Morales, M. Romacker, D. Twomey, S. Bureeva, V. Lemmon, and S. C. Schürer, "Evolving BioAssay ontology (BAO): Modularization, integration and applications.", *Journal of Biomedical Semantics*, vol. 5, no. Suppl 1 Proceedings of the Bio-Ontologies Spec Interest G, S5, 2014.
- [14] C. Torniai, M. Brush, N. Vasilevsky, E. Segerdell, M. Wilson, T. Johnson, K. Corday, C. Shaffer, and M. Haendel, "Developing an application ontology for biomedical resource annotation and retrieval: Challenges and lessons learned", in *Proceedings of the Second International Conference on Biomedical Ontology: July 26-30, 2011; Buffalo, NY. 2011*, http://icbo.buffalo.edu/ICBO-2011_Proceedings.pdf, vol. 833, 2011, pp. 101–108.
- [15] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, "The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013", *Nucleic Acids Res*, vol. 41, pp. D456–63, 2013.
- [16] S. Federhen, "Type material in the NCBI Taxonomy Database", *Nucleic Acids Res*, vol. 43, pp. D1086–98, 2015.
- [17] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank", *Nucleic Acids Research*, vol. 37, no. Database, pp. D26–D31, 2009, ISSN: 0305-1048.
- [18] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmsberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye, "Database resources of the National Center for Biotechnology Information", *Nucleic Acids Research*, vol. 37, no. Database, pp. D5–D15, 2009, ISSN: 0305-1048.
- [19] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, L. Fan, J. Fostel, G. Fragoso, F. Gibson, A. Gonzalez-Beltran, M. A. Haendel, Y. He, M. Heiskanen, T. Hernandez-Boussard, M. Jensen, Y. Lin, A. L. Lister, P. Lord, J. Malone, E. Manduchi, M. McGee, N. Morrison, J. A. Overton, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, D. Schober, B. Smith, L. N. Soldatova, C. J. Stoeckert, C. F. Taylor, C. Torniai, J. A. Turner, R. Vita, P. L. Whetzel, and J. Zheng, "The Ontology for Biomedical Investigations", *PLOS ONE*, vol. 11, no. 4, Y. Xue, Ed., e0154556, 2016, ISSN: 1932-6203.

- [20] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, "Modeling sample variables with an Experimental Factor Ontology", *Bioinformatics*, vol. 26, no. 8, pp. 1112–1118, 2010.
- [21] S. Sarntivijai, Y. Lin, Z. Xiang, T. F. Meehan, A. D. Diehl, U. D. Vempati, S. C. Schürer, C. Pang, J. Malone, H. Parkinson, Y. Liu, T. Takatsuki, K. Saijo, H. Masuya, Y. Nakamura, M. H. Brush, M. A. Haendel, J. Zheng, C. J. Stoeckert, B. Peters, C. J. Mungall, T. E. Carey, D. J. States, B. D. Athey, and Y. He, "Clo: The cell line ontology", *Journal of Biomedical Semantics*, vol. 5, pp. 37–37, 2014. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4387853/>.
- [22] S. Sarntivijai, Z. Xiang, T. Meehan, A. Diehl, U. Vempati, S. Schurer, C. Pang, J. Malone, H. Parkinson, B. Athey, and Y. He, "Cell line ontology: Redesigning the cell line knowledgebase to aid integrative translational informatics", *Neoplasia*, vol. 833, pp. 25–32, 2011, ISSN: 1522-8002.
- [23] (). Information Artifact Ontology (IAO). <https://github.com/information-artifact-ontology/IAO/>. Accessed 7 May 2016.
- [24] J. M. Coggan, "Evidence-based practice for information professionals: A handbook", *Journal of the Medical Library Association*, vol. 92, no. 4, pp. 503–503, Oct. 2004. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC521524/>.
- [25] K. M. Linton, Y. Hey, S. Dibben, C. J. Miller, A. J. Freemont, J. A. Radford, and S. D. Pepper, "Extraction of total RNA from fresh/frozen tissue (FT)", *The International Journal of Life Science Methods*, p. 53, 2010.
- [26] (). SMART Protocols project in Github. <https://github.com/oxgiraldo/SMART-Protocols>. Accessed 7 May 2016.
- [27] M. C. Suarez-Figueroa, A. Gomez-Perez, and M. Fernandez-Lopez, "The neon methodology for ontology engineering", in *Ontology Engineering in a Networked World*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–34, oeg, ISBN: 978-3-642-24794-1. [Online]. Available: <http://oa.upm.es/21469/>.
- [28] A. G. Castro, P. Rocca-Serra, R. Stevens, C. Taylor, K. Nashar, M. A. Ragan, and S.-A. Sansone, "The use of concept maps during knowledge elicitation in ontology development processes – the nutrigenomics use case", *BMC Bioinformatics*, vol. 7, no. 1, p. 267, 2006, ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-267. [Online]. Available: <https://doi.org/10.1186/1471-2105-7-267>.
- [29] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The obo foundry: Coordinated evolution of ontologies to support biomedical data integration", *Nat Biotech*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007. [Online]. Available: <http://dx.doi.org/10.1038/nbt1346>.
- [30] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.", *Nucleic acids research*, vol. 39, no. Web Server issue, W541–5, 2011, ISSN: 1362-4962.

- [31] Z. Xiang, C. Mungall, A. Ruttenberg, and Y He, "Ontobee: A Linked Data Server and Browser for Ontology Terms", *Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO)*, July 28-30, 2011, Buffalo, NY, USA, pp. 279–281, URL: <http://ceur-ws.org/Vol-833/paper48.pdf>. Accessed 26 Oct 2016.
- [32] *Publishing in the 21st century: Minimal (really) data standards | FORCE11*. [Online]. Available: <https://www.force11.org/node/4145> (visited on 02/09/2017).
- [33] P. Zimmermann, B. Schildknecht, D. Craigon, M. Garcia-Hernandez, W. Gruissem, S. May, G. Mukherjee, H. Parkinson, S. Rhee, U. Wagner, and L. Hennig, "MIAME/Plant – adding value to plant microarray experiments", *Plant Methods*, vol. 2, no. 1, p. 1, 2006, ISSN: 17464811.
- [34] S. A. Bustin, V. Benes, J. A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M. W. Pfaffl, G. L. Shipley, J. Vandesompele, and C. T. Wittwer, "The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments", *Clinical Chemistry*, vol. 55, no. 4, pp. 611–622, 2009, ISSN: 0009-9147.
- [35] F. Gibson, L. Anderson, G. Babnigg, M. Baker, M. Berth, P.-A. Binz, A. Borthwick, P. Cash, B. W. Day, D. B. Friedman, D. Garland, H. B. Gutstein, C. Hoogland, N. A. Jones, A. Khan, J. Klose, A. I. Lamond, P. F. Lemkin, K. S. Lilley, J. Minden, N. J. Morris, N. W. Paton, M. R. Pisano, J. E. Prime, T. Rabilloud, D. A. Stead, C. F. Taylor, H. Voshol, A. Wipat, and A. R. Jones, "Guidelines for reporting the use of gel electrophoresis in proteomics", *Nature Biotechnology*, vol. 26, no. 8, pp. 863–864, 2008, ISSN: 1087-0156.
- [36] *Protégé*. [Online]. Available: <http://protege.stanford.edu/> (visited on 04/11/2017).
- [37] CIAT, *International Center for Tropical Agriculture (CIAT)*, 2017. [Online]. Available: <https://ciat.cgiar.org/>.
- [38] E. Bachlava, C. A. Taylor, S. Tang, J. E. Bowers, J. R. Mandel, J. M. Burke, and S. J. Knapp, "SNP Discovery and Development of a High-Density Genotyping Array for Sunflower", *PLoS ONE*, vol. 7, no. 1, P. K. Ingvarsson, Ed., e29814, 2012, ISSN: 1932-6203. [Online]. Available: <http://dx.plos.org/10.1371/journal.pone.0029814>.
- [39] L. Wang, Y. Si, L. K. Dedow, Y. Shao, P. Liu, and T. P. Brutnell, "A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq", *PLoS ONE*, vol. 6, no. 10, M. E. Hudson, Ed., e26426, 2011, ISSN: 1932-6203. [Online]. Available: <http://dx.plos.org/10.1371/journal.pone.0026426>.
- [40] S. Hasan, J. Prakash, A. Vashishtha, A. Sharma, K. Srivastava, F. Sagar, N. Khan, K. Dwivedi, P. Jain, S. Shukla, *et al.*, "Optimization of dna extraction from seeds and leaf tissues of chrysanthemum (*chrysanthemum indicum*) for polymerase chain reaction", *Bioinformation*, vol. 8, no. 5, p. 225, 2012.
- [41] (). OWLViz. <http://protegewiki.stanford.edu/wiki/OWLViz>. Accessed 30 May 2016.
- [42] (). OOPS! (OntOlogy Pitfall Scanner!) <http://oops.linkeddata.es/>. Accessed 30 May 2016.

- [43] M. Poveda-Villalon, M. Suarez-Figueroa, and A. Gomez-Perez, "Validating Ontologies with OOPS!", Knowledge Engineering and Knowledge Management, A. ten Teije, et al., Editors, Springer, Berlin, Heidelberg, 2012, pp. 267–281.
- [44] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, and C. Rosse, "Relations in biomedical ontologies", *Genome Biology*, vol. 6, no. 5, R46, 2005.
- [45] *OBI Minimal metadata - OBI Ontology*. [Online]. Available: <http://obi.sourceforge.net/ontologyInformation/MinimalMetadata.html> (visited on 02/09/2017).
- [46] (). Documentation of SMART Protocols Ontology: Document Module. http://vocab.linkeddata.es/SMARTProtocols/myDocumentation_SPdoc_18Abril2017/index_SPdoc_V4.0.html. Accessed 26 Oct 2017.
- [47] (). Documentation of SMART Protocols Ontology: Workflow Module. http://vocab.linkeddata.es/SMARTProtocols/myDocumentation_SPwf_19Abril2017/index_SPwf_V4.0.html. Accessed 26 Oct 2017.
- [48] *Relations Ontology*. [Online]. Available: <http://obofoundry.org/ontology/ro.html> (visited on 04/13/2017).
- [49] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013", *Nucleic Acids Research*, vol. 41, no. D1, pp. D456–D463, 2013, ISSN: 0305-1048.
- [50] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "PubChem Substance and Compound databases.", *Nucleic acids research*, vol. 44, no. D1, pp. D1202–13, 2016, ISSN: 1362-4962.
- [51] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data", *The semantic web*, pp. 722–735, 2007.
- [52] T. U. Consortium, "Uniprot: The universal protein knowledgebase", *Nucleic Acids Research*, vol. 45, no. D1, p. D158, 2017.
- [53] A. Bandrowski, M. Brush, J. S. Grethe, M. A. Haendel, D. N. Kennedy, S. Hill, P. R. Hof, M. E. Martone, M. Pols, S. Tan, N. Washington, E. Zudilova-Seinstra, N. Vasilevsky, and Resource Identification Initiative Members are listed here: <https://www.force11.org/node/4463/members>, "The Resource Identification Initiative: A cultural shift in publishing", *F1000Research*, vol. 4, p. 134, 2015, ISSN: 2046-1402.
- [54] N. Juty, N. Le Novère, and C. Laibe, "Identifiers.org and MIRIAM Registry: Community resources to provide persistent identification", *Nucleic Acids Research*, vol. 40, no. D1, pp. D580–D586, 2012, ISSN: 03051048.
- [55] *Identifiers.org*. [Online]. Available: <http://identifiers.org/> (visited on 02/10/2017).

- [56] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The fair guiding principles for scientific data management and stewardship", *Scientific Data*, vol. 3, p. 160018, 2016. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/>.

Chapter 4

Laboratory Protocols in Bioschemas

Search Engine Optimization (SEO) techniques are widely used by webmasters in order to increase online visibility over unpaid search results often referred to as “natural”, “organic”, or “earned” results. SEO techniques make extensive use of structured data; the easier it is for search engines to “understand” the aboutness of a website, the easier it is for specific products or web pages to be found. Schema.org is a collaborative project providing schemas for semantically structuring data in web pages; this effort was initiated by Yahoo!, Google, and Microsoft. Schema.org provides a hierarchical set of vocabularies to embed metadata in HTML pages for an enhanced search and browsing experience. Bioschemas is a biomedical community effort that aims to bring the main driving idea behind schema.org to data providers in life science. By publishing a controlled vocabulary for embedding metadata in web pages, very often from database records being rendered on the fly, it is expected that search engines will be able to determine whether a web page refers to a single protein, a gene, or a protein-protein interaction network. It is also expected that the availability of such markup will make it easier for search agents to summarize information in a way similar to that offered by infoboxes in Wikipedia.

4.1 Introduction

Bioschemas is a community initiative aiming to extend schema.org in order to improve data discoverability and interoperability in Life Sciences [1]. Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond [2]. Schema.org was initiated by Yahoo!, Google, and Microsoft; these are amongst the major search engines on the WWW [3], [4]. It provides a hierarchical set of controlled vocabularies, semantic markup, to embed metadata in HTML pages for an enhanced search and browsing experience. Such semantic markup makes it easier for search agents to “understand” the content of web pages; it also establishes a clear separation between content and layout [5]. RDFa- Lite, Microdata and JSON-LD as lower semantic techniques have gained more attention by Web users to markup Web pages and even emails based on Schema.org

Bioschemas is expected to be a major contributor for life-sciences-controlled vocabularies to schema.org. As a shared vocabulary, schema.org heavily relies on the input from the community. A shared vocabulary makes it easier for content masters to decide on a common schema; it is in this spirit that Bioschemas and schema.org are expected to work together. Schema.org defines common generic types like “events” and “datasets” which can be used not just in life sciences but in many other disciplines. Bioschemas is working on specifications to improve the description of generic types in life sciences [6]. Successive editions and releases are expected to be included in schema.org throughout 2018. The main outcome of Bioschemas is a collection of specifications that provide guidelines to facilitate a more consistent adoption of schema.org markup within the life sciences [7]; specifications and guidelines are available at [8].

4.2 Why semantic structuring?

Publishing structured content has the advantage of making it simpler for search engines to better understand the content without layout-related considerations. Mature Web applications such as Web search are increasingly seeking to use the structured content, if any, to power richer and more interactive experiences [5]. Anchoring the structure with semantics, as that provided by a controlled vocabulary, facilitates a number of important features, namely: .

- Search and retrieval
- Reusability and interoperability
- Personalization
- Quick summarization

4.3 Bioschemas at a glance

Bioschemas inherits from schema.org types and properties. Types are generic descriptors that can be further specialized. The broadest item type is Thing, which has four properties: name, description, url, and image. Properties are attributes that make the type a concrete “thing” that can exist, thus discoverable by search agents, in a web page. More specific types share properties with broader types. For example,

a Place is a more specific type of Thing, and a LocalBusiness is a more specific type of Place. More specific items inherit the properties of their parent. (Actually, a LocalBusiness is a more specific type of Place and a more specific type of Organization, so it inherits properties from both parent types.) Taken verbatim from [9].

The types in schema.org are organized into a hierarchy. Each class may have one or more supertypes. Relations are polymorphic; they have one or more domains and one or more ranges. The type hierarchy is meant more as an organizational tool to help browse the vocabulary than as a formal ontological representation.

Bioschemas is built upon the principles laid down by schema.org, in order to be fully aligned with schema.org it reuses generic types whenever possible. For instance, it reuses DataCatalog and Dataset; it adds new properties to others such as CreativeWork and proposes new types such as BioChemEntity, DataRecord, and Lab-Protocol [1]. An overview of the main types involved in Bioschemas is presented in Figure 4.1.

FIGURE 4.1: General overview of Bioschemas and the LabProtocol profile

The Bioschemas' profiles define a community-agreed layer over the Schema.org model providing additional constraints. These constraints capture (i) the information properties agreed on by the community which are minimum (M), recommended (R), or optional (O), (ii) the cardinality of the property, i.e. whether it is expected to occur once or many times, and (iii) associated controlled vocabulary terms drawn from existing ontologies.

4.3.1 Experimental Protocols and Bioschemas

The LabProtocol profile in Bioschemas is a generic model that can be used to exchange information about a laboratory protocol, e.g. purpose, instruments, sample, software application, etc. Experimental protocols are a special type of scientific publication that focuses on the process; the current version of the LabProtocol profile does not fully address the representation of an experimental protocol as a publication. Experimental protocols are information structures that provide descriptions of the processes by means of which results, often data, are generated in experimental research. The protocols often include equipment, reagents, critical steps, duration, troubleshooting, tips, and all the information that facilitates reusability [6].

The focus of the LabProtocol profile, *idem.* mandatory data elements, is on the description of the SIRO model namely, sample, instruments, reagents and purpose (objective). The description should make it possible for search agents to summarize the content of an experimental protocol, thus, making it easier for the end user to, for instance, decide which protocol to reuse or how protocols are similar in the suggested reagents. The LabProtocol profile extends and adapts the SIRO [6] model to the specifics of Bioschemas and schema.org. The proposed markup makes it easier for end users to compare protocols based on the descriptors. The profile has been designed to be aligned with existing ontologies while preserving the design principles behind schema.org. It has also been designed to be easily extensible and adaptable; for instance, if only protocols meant to be followed by robots are to be considered.

4.4 Developing the LabProtocol profile

The development of this vocabulary followed some of the methodological principles suggested for developing ontologies -see [6], [10], [11]. For instance, competency questions and scenarios of use were discussed from the beginning. These were the basis for the formalization of the use cases. Also, ontology classes and properties were reused whenever possible; however, these were stripped of their logical formalisms and only the concepts as terms were used. The domain analysis and knowledge acquisition phases as described by Garcia et al [10] were adapted for this scenario. For instance, the domain analysis focused on one specific question, "How is the type of interest being described by major data providers related to the type of interest?". The knowledge acquisition focused on schema.org by asking, "is this a new type?", "are there closely related types?", "are there reusable properties?". There was a strong community component in the development of this profile; it represents the agreement of the working group led by the author of this thesis. When developing specifications for Bioschemas, the community is not limited to that within Bioschemas; in this specific case, there was also interaction with the larger schema.org community. Below the steps that we followed throughout our development process.

1. The development process started by proposing to create a new working group (WG). The aim of this WG is to bring forward a Bioschemas specification with a specific type, LabProtocol. New specifications usually start by raising the issue to be addressed in GitHub¹ and by doing so generating a WG with a specific aim.

¹<https://github.com/BioSchemas/bioschemas/issues>

2. How is the “type of interest” being described by major data providers? This involves visiting sites that host or list these items; particular attention should be paid to the properties and data fields being used. The spreadsheet “Training Material Descriptions Review”² is suggested as a starting point for this task.
3. Then, use cases should be defined; those for the LabProtocol profile are available here³; they are also described below as of 27/12/2018. These use cases focus on three main features, findability, summarization and linking to external resources.

A **Findability:** As a user, I would like to search for protocols according to:

- availability of reagents and/or equipment in my lab, also
- the sample to be tested, and/or
- the overall objective of the protocol

Searching for protocols using these four elements helps me (as a user) to decide on the suitability of the protocol(s) that should be executed during an experiment or assay.

B **Summarization:** As a user, whenever I look for a lab protocol (i.e., experimental protocol), I would like to see a quick summary with information related to: the sample to be tested (e.g., whole organisms, anatomical parts, biomolecules, body fluids), the list of equipment (including standard and high-throughput equipment, consumables), the software used, the list of reagents (e.g., bought ready-to-use, solutions or mixtures prepared in the lab, media, buffers, kits), the overall objective of the protocol (to know about the suitability), the protocol identifier (which helps me to find a protocol), and the license.

C **Link to external resources** As a user of a protocol, I would like to know where I can buy equipment and/or reagents used in a particular protocol. Rationale: For this reason, it is important to include information about the manufacturers, catalog number, and homepage of equipment and reagents. As a user, I would like to know previous uses/applications of a protocol; who has used it, who has built derivations of it, and are those derivations related to the sample, instrument, reagent or objective? Finally, as a user of the protocol, I would like to know where the data is derived from the application of the protocol –if publicly available, then where?

4. The use cases help to define the properties that describe the proposed type. Look for common properties used across websites and determine the most important properties shared across these sites. Consider whether there are other properties missing, and consult with the community to see if other properties would be useful.
5. Search through schema.org⁴ and determine if there is a type that fits your use cases. Browse the list of types⁵ at schema.org for feasible types to be reused.

²https://docs.google.com/spreadsheets/d/1cQ6mDbsG_cMX2EDAN8xH6-9yMRba8-rEr1PeP8HTs8A/edit#gid=0

³<http://bioschemas.org/useCases/LabProtocols/>

⁴<http://schema.org/>

⁵<https://schema.org/docs/full.html>

6. Start mapping. Investigate whether it is possible to match existing properties from schema.org to the properties you need; it is always possible to ask the schema.org community to adopt new properties if they are not available. New suggestions should be kept as generic as possible; they should be useful across domains. For this mapping process, it is suggested to use a spreadsheet and add specialized descriptions marginalities, cardinalities and Controlled Vocabularies (CVs). The spreadsheet with the mappings for the LabProtocol type is available here⁶.
7. Look at existing specifications and follow the structure; the LabProtocol⁷ specification is available and open for comments. This increases the outreach to the community, that of Bioschemas as well as that of schema.org.
8. Start using the specification in the real world by describing possible scenarios of use with a specific end user in mind.
9. Contact schema.org and explain the result of the specification; highlight new properties, types, and type extensions.

FIGURE 4.2: A general overview of the development process

⁶<https://docs.google.com/spreadsheets/d/1RWYIphvcBMH18SLJ15-xRMZIOYaYrtJtClBS0OPL4xQ/edit#gid=1261485211>

⁷<http://bioschemas.org/types/LabProtocol/>

4.5 Results, The Labprotocol Profile

4.5.1 Mandatory properties

The four elements from the SIRO model [6] (sample, instrument, reagent, and objective) were proposed as mandatory properties in the LabProtocol profile. The property “instrument” is reused from schema.org, and the “purpose” is reused from the health-lifesci.schema.org extension. See Table 4.1.

TABLE 4.1: Mandatory properties proposed to represent the LabProtocol type

Mandatory properties from [LabProtocol]					
Property	Expected type	Description	CN	MG	CV
instrument	Thing, Text or URL	The object that helped the agent perform the action. e.g. John wrote a book with a pen. Bioschema usage. For LabProtocols it would be a laboratory equipment use by a person to follow one or more steps described in this LabProtocol.	many	M	OBI, ERO, EFO
purpose	Text	Defined in the health-lifesci-schema.org extension. A goal towards an action is taken. Can be concrete or abstract	one	M	SMART Protocols
reagent	PhysicalEntity, Text or URL	Reagents used in the protocol, ChEBI and PubChem entities can be used whenever available. Commercial names are also acceptable (URL if possible)	many	M	ChEBI, PubChem
sample	PhysicalEntity, Text or URL	Sample used in the protocol. It could be a record in a Dataset describing the sample or a physical object corresponding to the sample or a URL pointing to the type of sampled used.	many	M	NCBI taxonomy, UBERON, PO

The terminology related to these four elements is represented ontologies or controlled vocabularies like SMART Protocols [6], OBI [12], ERO [13], EFO [14], CHEBI [15], and NCBI taxonomy [16], [17].

4.5.2 Recommended properties

Eight properties were proposed as “recommended properties”. Two of them were reused from the set of Thing properties described in schema.org.

TABLE 4.2: Thing properties from schema.org proposed as recommended properties

Properties from [Thing]					
Property	Expected type	Description	CN	MG	CV
description	Text	A description of the item. Bioschemas usage. Use in LabProtocol to include the step by step process followed in this protocol.	one	R	
identifier	Property-Value, Text, URL	The identifier property represents any kind of identifier for any kind of Thing, such as ISBNs, GTIN codes, UUIDs etc. Schema.org provides dedicated properties for representing many of these, either as textual strings or as URL (URI) links. See background notes for more details.	one	R	

Four properties come from the set of CreativeWork properties described in schema.org.

TABLE 4.3: CreativeWork properties from schema.org proposed as recommended properties

Properties from [CreativeWork]					
Property	Expected type	Description	CN	MG	CV
citation	CreativeWork or URL	A citation or reference to a creative work, such as a publication, web page, scholarly article, etc.	many	R	
license	Creative-Work or URL	A license document that applies to this content, typically indicated by URL.	one	R	
isPartOf	Creative-Work	Indicates a CreativeWork that this Creative-Work is (in some sense) part of.	many	R	
hasPart	Creative-Work	<p>Indicates a CreativeWork that is (in some sense) a part of this CreativeWork.</p> <p>A particular case in Bioschemas is LabProtocol where document parts or sections are used to described advantages (situations the Protocol has been successfully employed), limitations (situations the Protocol would be unreliable or otherwise unsuccessful), applications (Applications of the protocol list the full diversity of the applications of the method and support if is possible to extend the range of applications of the protocol. e.g. northern blot assays, sequencing, etc.), and outcomes (outcome or expected result by a protocol execution).</p> <p>Bioschemas usage. For LabProtocol, in the applicationType, please use⁸ for advantages, ⁹ for limitations, ¹⁰ for applicability, and ¹¹ for outcomes.</p>	many	R	SMART Protocols ¹²

The last two properties proposed as recommended are the types Duration and SoftwareApplication from schema.org.

⁸<http://purl.org/net/SMARTprotocol#AdvantageOfTheProtocol>

⁹<http://purl.org/net/SMARTprotocol#LimitationOfTheProtocol>

¹⁰<http://purl.org/net/SMARTprotocol#ApplicationOfTheProtocol>

¹¹<http://purl.org/net/SMARTprotocol#OutcomeOfTheProtocol>

¹²<http://bioportal.bioontology.org/ontologies/SP>

TABLE 4.4: Types from schema.org proposed as recommended properties

Properties from [Thing]					
Property	Expected type	Description	CN	MG	CV
duration ¹³	Duration ¹⁴	The time it takes to actually carry on the protocol, in ISO 8601 duration format.	one	R	
software	SoftwareApplication	An application that can complete the request.	many	R	

4.6 Discussion

In this chapter, the LabProtocol profile has been presented; it is work in progress that is part of Bioschemas and, therefore, it has been designed to be compliant with schema.org. The need for simplicity drives the development of schema.org, as developers stated on the schema.org website, “In creating schema.org, one of our goals was to create a single place where webmasters could go to figure out how to mark up their content, with reasonable syntax and style consistency across types. This way, webmasters only need to learn one thing rather than having to understand different, often overlapping vocabularies...” [18]. Schema.org is meant to be used by non-domain experts. It is a general purpose mid-level ontology, i.e., an ontology that neither tries to be too abstract or all-encompassing nor describes a knowledge field in depth [19]. Following Ronallo, “it is unrealistic for them (Schema.org) to try to support every vocabulary in use. Schema.org is an attempt to define a broad, Webscale, shared vocabulary focusing on popular concepts. [...] A central goal of having such a broad schema all in one place is to simplify things for mass adoption and cover the most common use cases” [20].

Understanding Schema.org as explained above, the LabProtocol profile is a lightweight vocabulary meant for non experts to use in order to mark up content. It does not have the formality of the SMART Protocols Ontology [6], nor does it provide an extensive list of information items such as those available in the reporting guideline published by Giraldo et al [21]. The LabProtocol profile reuses from these experiences; it presents a different layer of semantics, that of a compact, generic, and schema.org-compliant vocabulary. The LabProtocol profile is not an attempt to create a logically consistent model of the experimental protocols, but instead to provide a shared conceptual structure that facilitates data interoperability between Web content and applications that consume this data, the main purpose being discoverability and rapid summarization.

Laboratory protocols are very often private documents; this should not prevent these documents from being discovered by search engines. The SIRO model [6] was proposed as a minimal information model that describes four elements that investigated protocols had in common. Following up on that work, the LabProtocol profile delivers a markup model that search engines can use to summarize and discover laboratory protocols. SIRO is minimal, whereas the LabProtocol profile is more comprehensive and thought for schema.org purposes. The LabProtocol profile may be used by publishers; it can easily be understood by webmasters so that during the

¹³<http://schema.org/duration>

¹⁴<http://schema.org/Duration>

publication stage in the publication workflow, such mark up can be added to the content. Having such mark up does not mean that the whole content is open and free; only the mark up can be exposed without compromising the full content of the publication.

4.7 Conclusions and Future Work

Schema.org and Bioschemas are purpose-bound simplifications of a domain of interest; in this case, the boundaries of the models are defined by the main purpose behind schema.org and Bioschemas, that of discoverability and rapid summarization of content. Evaluating Bioschemas will take time because a valid metric has to be defined with regards to a purpose. Criticizing the models as if they were ontologies is not valid because they are not; by the same token, deficiencies in the models, e.g. coverage, granularity, etc, should only be accounted for if they are related to the area of application for which the model is intended. The community nature of efforts like Bioschemas make it more difficult to have metrics of success, other than those related to adoption. Adoption should not be confused with popularity; ultimately, schema.org is a tool with a specific purpose, that of search engine optimization. A type may be popular, but if the popularity is not translated in a sustained feature for the end user that warrants increasing traffic or retaining and attracting new users, then the popularity will fade away. The validity of the models proposed by Bioschemas will ultimately be evaluated by the larger community of data providers and consumers of information; the evolution of the vocabularies will reflect this evaluation. As webmasters see an increase in traffic that is related to the use of this markup, users will demand more accuracy in the search results and infoboxes associated with these results.

Biotea [22], [23] as well as the Nature Publishing Group Linked Data platform [24], [25] are delivering conceptual models for biomedical publications in RDF. These models are taking into account metadata, references, content, and biomedical annotations. Bioschemas provides a simple way to add structured data to web pages; in particular, the LabProtocol profile models specifics in a type of publication, that related only to experimental protocols. Aligning these models to Bioschemas represents an opportunity to easily add a semantic layer to publications, making them FAIRer [26] and more machine-consumable. Having scholarly data in Bioschemas opens up possibilities for smarter recommendation systems and literature-based knowledge graphs and discovery.

Bibliography

- [1] L. Garcia, O. Giraldo, A. Garcia, and M Dumontier, "Bioschemas: Schema.org for the life sciences", *Proceedings of SWAT4LS*, 2017.
- [2] *Home – schema.org*. [Online]. Available: <https://schema.org>.
- [3] A. Chris, *Top 10 search engines in the world*. [Online]. Available: <https://www.reliablessoft.net/top-10-search-engines-in-the-world/>.
- [4] C. Forsey, *The top 7 search engines, ranked by popularity | page*. [Online]. Available: <https://www.reliablessoft.net/top-10-search-engines-in-the-world/>.
- [5] R. V. Guha, D. Brickley, and S. Macbeth, "Schema.org: Evolution of structured data on the web", *Commun. ACM*, vol. 59, no. 2, pp. 44–51, Jan. 2016, ISSN: 0001-0782. DOI: 10.1145/2844544. [Online]. Available: <http://doi.acm.org/10.1145/2844544>.
- [6] O. Giraldo, A. García, F. López, and O. Corcho, "Using semantics for representing experimental protocols", *Journal of Biomedical Semantics*, vol. 8, no. 1, p. 52, 2017, ISSN: 2041-1480. DOI: 10.1186/s13326-017-0160-y. [Online]. Available: <https://doi.org/10.1186/s13326-017-0160-y>.
- [7] *Bioschemas*, 2018. [Online]. Available: <http://bioschemas.org/>.
- [8] *Bioschemas - bioschemas specifications*, 2018. [Online]. Available: <http://bioschemas.org/specifications/>.
- [9] *Getting started - schema.org*. [Online]. Available: https://schema.org/docs/gs.html#schemaorg_types.
- [10] A. G. Castro, P. Rocca-Serra, R. Stevens, C. Taylor, K. Nashar, M. A. Ragan, and S.-A. Sansone, "The use of concept maps during knowledge elicitation in ontology development processes – the nutrigenomics use case", *BMC Bioinformatics*, vol. 7, no. 1, p. 267, 2006, ISSN: 1471-2105. DOI: 10.1186/1471-2105-7-267. [Online]. Available: <https://doi.org/10.1186/1471-2105-7-267>.
- [11] M. C. Suarez-Figueroa, A. Gomez-Perez, and M. Fernandez-Lopez, "The neon methodology for ontology engineering", in *Ontology Engineering in a Networked World*, M. C. Suarez-Figueroa, A. Gomez-Perez, E. Motta, and A. Gangemi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–34, ISBN: 978-3-642-24794-1. DOI: 10.1007/978-3-642-24794-1_2. [Online]. Available: https://doi.org/10.1007/978-3-642-24794-1_2.
- [12] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, L. Fan, J. Fostel, G. Fragoso, F. Gibson, A. Gonzalez-Beltran, M. A. Haendel, Y. He, M. Heiskanen, T. Hernandez-Boussard, M. Jensen, Y. Lin, A. L. Lister, P. Lord, J. Malone, E. Manduchi, M. McGee, N. Morrison, J. A. Overton, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, D. Schober, B. Smith, L. N. Soldatova, C. J. Stoeckert Jr., C. F. Taylor, C. Torniai, J. A. Turner, R. Vita, P. L. Whetzel, and J. Zheng, "The ontology for biomedical investigations",

- PLOS ONE*, vol. 11, no. 4, pp. 1–19, Apr. 2016. DOI: 10.1371/journal.pone.0154556. [Online]. Available: <https://doi.org/10.1371/journal.pone.0154556>.
- [13] C. Torniai, M. Brush, N. Vasilevsky, E. Segerdell, M. Wilson, T. Johnson, K. Corday, C. Shaffer, and M. Haendel, “Developing an application ontology for biomedical resource annotation and retrieval: Challenges and lessons learned”, English (US), in *CEUR Workshop Proceedings*, vol. 833, 2011, pp. 101–108.
- [14] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, “Modeling sample variables with an experimental factor ontology”, *Bioinformatics*, vol. 26, no. 8, pp. 1112–1118, 2010. DOI: 10.1093/bioinformatics/btq099. eprint: /oup/backfile/content_public/journal/bioinformatics/26/8/10.1093/bioinformatics/btq099/2/btq099.pdf. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btq099>.
- [15] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, “The chebi reference database and ontology for biologically relevant chemistry: Enhancements for 2013”, *Nucleic Acids Research*, vol. 41, no. D1, pp. D456–D463, 2013. DOI: 10.1093/nar/gks1146. eprint: /oup/backfile/content_public/journal/nar/41/d1/10.1093/nar/gks1146/2/gks1146.pdf. [Online]. Available: <http://dx.doi.org/10.1093/nar/gks1146>.
- [16] S. Federhen, “Type material in the ncbi taxonomy database”, *Nucleic Acids Research*, vol. 43, no. D1, pp. D1086–D1098, 2015. DOI: 10.1093/nar/gku1127. eprint: /oup/backfile/content_public/journal/nar/43/d1/10.1093/nar/gku1127/2/gku1127.pdf. [Online]. Available: <http://dx.doi.org/10.1093/nar/gku1127>.
- [17] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmsberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye, “Database resources of the national center for biotechnology information”, *Nucleic Acids Research*, vol. 37, no. suppl₁, pp. D5–D15, 2009. DOI: 10.1093/nar/gkn741. eprint: /oup/backfile/content_public/journal/nar/37/suppl_1/10.1093/nar/gkn741/2/gkn741.pdf. [Online]. Available: <http://dx.doi.org/10.1093/nar/gkn741>.
- [18] *Faq - schema.org*. [Online]. Available: <https://schema.org/docs/faq.html>.
- [19] M. Antoniazzi, “Mapping the torch ontology to schema.org master thesis”, Master’s thesis, Oslo and Akershus University College of Applied Sciences, 2015.
- [20] J. Ronallo, “Html5 microdata and schema.org”, *Code4Lib Journal*, vol. 16, 2012.
- [21] O. Giraldo, A. Garcia, and O. Corcho, “A guideline for reporting experimental protocols in life sciences”, *PeerJ*, vol. 6, e4795, May 2018, ISSN: 2167-8359. DOI: 10.7717/peerj.4795. [Online]. Available: <https://doi.org/10.7717/peerj.4795>.

- [22] A. Garcia, F. Lopez, L. Garcia, O. Giraldo, V. Bucheli, and M. Dumontier, "Biotea: Semantics for pubmed central", *PeerJ*, vol. 6, e4201, Jan. 2018, ISSN: 2167-8359. DOI: 10.7717/peerj.4201. [Online]. Available: <https://doi.org/10.7717/peerj.4201>.
- [23] L. J. Garcia Castro, C. McLaughlin, and A. Garcia, "Biotea: Rdfizing pubmed central in support for the paper as an interface to the web of data", *Journal of Biomedical Semantics*, vol. 4, no. 1, S5, 2013, ISSN: 2041-1480. DOI: 10.1186/2041-1480-4-S1-S5. [Online]. Available: <https://doi.org/10.1186/2041-1480-4-S1-S5>.
- [24] *Press release archive: About npg*, 2012. [Online]. Available: https://www.nature.com/press_releases/linkedata.html.
- [25] scigraph, *Scigraph*, Retrieved on 02/02/2019 from <https://scigraph.springernature.com>, 2019. [Online]. Available: <https://scigraph.springernature.com/explorer/>.
- [26] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The fair guiding principles for scientific data management and stewardship", *Scientific Data*, vol. 3, p. 160018, 2016. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/>.

Chapter 5

BioH, The Smart Protocols Annotation Tool

BioH specializes the hypothes.is annotation tool to the biomedical domain. There are a significant number of projects delivering automatic annotation infrastructures; thus, making it possible for domain experts to annotate their documents. We are interested in having human annotation co-existing with automatic annotation pipelines over a single interface and using an open standard that also facilitates sharing, collaborating and discovering. BioH is built upon the idea that annotations are central for researchers to collaborate, create, discover, share and re-use Knowledge. BioH is built over an open source community project; it facilitates the integration of several annotation tools and exports annotations in a standardized format also making them available as linked data over a SPARQL endpoint. BioH builds upon previous work, hypothes.is and extensions to hypothes.is that we have built to support human validation of results from automatic annotation pipelines.

5.1 Introduction

The benefits of “ harnessing the collective intelligence” have been proven across the web e.g. Waze¹, Google Maps². Community based annotation is exemplary of the Web 2.0/3.0 phenomena [1]–[3]; using the Web to “harness” the collective intelligence is central to the business models for the evolution of the Web 3.0 [4], [5]. Within this realm, Google maps and Waze are among the most successful public-sourcing annotation systems; users can easily create, share, edit and discover annotations (geographical annotations) that are interesting and relevant to their particular situations. Community Based sharing and discovering grows rapidly and spontaneously; it responds to the need for structuring and classifying information. By doing so, it facilitates information retrieval, generation of task specific networks and serendipitous relations.

The biomedical domain has not been alien to this tendency; there have been various efforts in harnessing the collective intelligence in the biomedical domain. Researchers are constantly annotating literature; marginal notes, post-its, comments and highlights are frequently added to papers. However, there is little or no interoperability across these efforts. For instance, the DAS WriteBack [6], WikiProteins [7], The GENIA corpus [8], Biotea [9], and many others have built their own annotation infrastructures for sequences, literature, pathways, networks, etc . Although Biomedical information is highly interrelated, the lack of an interoperable infrastructure make it difficult to use discovery tools across the diaspora of annotations. Furthermore, the lack of such shareable infrastructure increases the costs by forcing everyone to rebuild similar infrastructures over and over again.

Web.hypothes.is [10] delivers a generic open infrastructure for annotating text over a web interface. It makes it possible for annotation projects to adapt the infrastructure while preserving the interoperability across annotations. Automatic annotation pipelines make it possible to identify entities or relations; these workflows are usually used to process batches of papers. Domain expert annotation and automatic annotation pipelines are not always available over the same user experience; they should coexist and benefit from each other. By combining domain expert and automatic annotation over one single end-user tool, the SP annotation tooling is facilitating the generation of better annotated corpus and simplifying the process for validating the results from automatic annotation pipelines. Moreover, by having such tool communities can organize annotations (hackathon principles applied to annotations) with very specific targets, e.g. annotating orphan diseases, annotating the Population Intervention Comparison Outcome (PICO) model in evidence based medicine [11].

Throughout this chapter, the emphasis is on the annotation of experimental protocols, focusing on the identification of Samples, Instruments, Reagents and Objectives (SIRO); for samples, instruments and reagents we were interested in enriching an existing ontology representing experimental protocols. The annotation of objectives had a different purpose, we wanted to gather a dataset of objectives so that we could understand the discursive structure and train machine learning algorithms in the automatic identification of objectives. Some of the additions we developed for this case include, interoperability with BioPortal, adequation of the user interface in order to constrain the annotations to specific facets, a reporting module so that we could analyse the annotations, interoperability with pubannnotator [12], a simple

¹<https://www.waze.com>

²<https://www.google.com/maps>

transformation module so that we could expose the annotations over a SPARQL endpoint using the open annotation framework [13], e.g. vocabulary and data model. As this annotation task feeds an ontology, we are calculating the inter annotator agreement as a parameter of quality before moving the terms to the ontology.

The annotation of biological information is a common task; it can be either manual or automatic. Manual annotation refers to the actions of domain experts annotating; this is usually time consuming and expensive, however, it produces high quality resources, *idem*. Gold standards. Although automatic annotation does not always produce high quality annotations, it allows large scale processing. A combination of the two types of annotation is required in order to balance the needs for both high quality annotation and large scale processing. Manual annotation thus becomes a quality control mechanism for the information obtained by automatic methods. Such hybrid approach is the one supporting the tool presented in this chapter. The approach presented here is easily usable in other domains and it also covers PDFs. It is expected that by enabling the coexistence between authoritative-human annotations and NLP or NER based annotations we will facilitate higher levels of rigor and reproducibility in biomedicine, particularly in the area of biocuration, and hope to support authoritative annotation in other science and research areas.

In this chapter the annotation tool is presented. First, a brief description of the requirements is provided. Then, the tool is presented, functionality is explained and a specific annotation case is described. By reusing the web.hypothes.is infrastructure our tool also contributes to the overall pool of annotations available over the web.hypothes.is API. More importantly, by reusing an existing proven annotation infrastructure our tool and approach reaches out to a larger group of users who are also interested in the annotation of biomedical documents. Also, reusing and tapping on an existing community make it faster and simpler the software development process.

5.2 The SIRO Curation Model

The Sample Instrument Reagent Objective model [14] has been proposed to annotate experimental protocols. the SIRO model summarizes the content of an experimental protocols and also makes it possible to establish relations to entities in the web that are easily resolved by an existing information resource. For instance, Sample:rat liver could be resolved as follows.

FIGURE 5.1: From general to specific, navigating an ontology

Similar to annotations based on the PICO model, researchers benefit from having a comprehensive curatorial annotation tool that integrates automatic annotation and allows them to validate whatever comes from services such as the Ontology Look up Service [15] or the NCBO annotator [16]. Annotations are stored and exposed as linked open data. Bio-H will also make it possible for researchers to annotate with

Research identifiers from resources such as identifiers.org. Our approach reuses results from efforts such as that of Research Resource Identifier (RRID)₁₀₁₁ [17]. See appendix A and B for more information about the annotation process.

FIGURE 5.2: What and how to annotate using BioH

5.3 The Tool

Web.hypothes.is has been modified to support the SIRO curation model. The tool supports a simple user experience that relies on extensive experience from the Web.hypothes.is project. As previously stated, the user selects the text snippet, defines the facet to which it refers to (e.g. sample, instrument, etc.) and then if necessary expands on the annotation. Fig. 5.2 presents the user interaction with the tool. The architecture is presented in figure 5.3.

5.3.1 Architecture

We have modified the data model and user interface of hypothes.is. We are not capturing free text comments on specific areas of the text; instead, we are asking users to classify text snippets according to a predefined set of facets, e.g. sample, instruments and reagents. We have Elastic Search as our backend database; a python service transforms the data into RDF triplets and exposes these over our SPARQL endpoint. We are representing the annotations using the Web Annotation Data Model [18]. The components and data flow is presented in Figure 5.3.

FIGURE 5.3: Architecture and components of the BioH annotation tool

5.4 Discussion and Concluding Remarks

In this chapter the tool BioH has been presented. This tool was designed to support SIRO annotation in order to gather specific terminology and also to build the gold standard as presented in chapter 6. The tool also supports the validation of automatic annotation. For instance, the document may be pre annotated and these annotations could be validated by domain experts simply by choosing valid/no valid for each annotation. If necessary, domain experts could also expand where necessary.

The process of document annotation very often requires significant effort and time from multiple domain experts to work iteratively and collaboratively to identify and resolve discrepancies. The process is expensive; this BioH can also be used to support such scenarios.

Bibliography

- [1] A. Darwish and K. Lakhtaria, "The impact of the new web 2.0 technologies in communication, development, and revolutions of societies", *Journal of Advances in Information Technology*, vol. 2, Nov. 2011. DOI: 10.4304/jait.2.4.204-216.
- [2] T. Gruber, "Collective knowledge systems: Where the social web meets the semantic web", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, pp. 4-13, Jan. 2007. DOI: 10.2139/ssrn.3199378.
- [3] E Feigenbaum *et al.*, "Computer-assisted semantic annotation of scientific life works", 2007.
- [4] *What is web 2.0 - o'reilly media*. [Online]. Available: <https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=2>.
- [5] S. S. Gunasekaran, S. A. Mostafa, and M. S. Ahmad, "Using the internet as a collective intelligence platform in harnessing issues on climate change", in *Information Technology and Multimedia (ICIMU), 2014 International Conference on*, IEEE, 2014, pp. 130-135.
- [6] G. A. Salazar, R. C. Jimenez, A. Garcia, H. Hermjakob, N. Mulder, and E. Blake, "Das writeback: A collaborative annotation system", *BMC bioinformatics*, vol. 12, no. 1, p. 143, 2011.
- [7] B. Mons, M. Ashburner, C. Chichester, E. van Mulligen, M. Weeber, J. den Dunnen, G.-J. van Ommen, M. Musen, M. Cockerill, H. Hermjakob, *et al.*, "Calling on a million minds for community annotation in wikiproteins", *Genome biology*, vol. 9, no. 5, R89, 2008.
- [8] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus—a semantically annotated corpus for bio-textmining", *Bioinformatics*, vol. 19, no. suppl_1, pp. i180-i182, 2003.
- [9] A. Garcia, F. Lopez, L. Garcia, O. Giraldo, V. Bucheli, and M. Dumontier, "Biotea: Semantics for pubmed central", *PeerJ*, vol. 6, e4201, 2018.
- [10] *Hypothesis – the internet, peer reviewed*, 2018. [Online]. Available: <https://web.hypothes.is>.
- [11] X. Huang, J. Lin, and D. Demner-Fushman, "Evaluation of pico as a knowledge representation for clinical questions", in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2006, 2006, p. 359.
- [12] *Pubannotation*. [Online]. Available: <http://pubannotation.org/>.
- [13] *Welcome*. [Online]. Available: <http://www.openannotation.org/>.
- [14] O. Giraldo, A. García, F. López, and O. Corcho, "Using semantics for representing experimental protocols", *Journal of Biomedical Semantics*, vol. 8, no. 1, p. 52, 2017, ISSN: 2041-1480. DOI: 10.1186/s13326-017-0160-y. [Online]. Available: <https://doi.org/10.1186/s13326-017-0160-y>.

- [15] J. S. et al., "A new ontology lookup service at embl-ebi", *Proceedings of SWAT4LS*, 2015.
- [16] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.", *Nucleic acids research*, vol. 39, no. Web Server issue, W541–5, 2011, ISSN: 1362-4962.
- [17] A. Bandrowski, M. Brush, J. S. Grethe, M. A. Haendel, D. N. Kennedy, S. Hill, P. R. Hof, M. E. Martone, M. Pols, S. C. Tan, *et al.*, "The resource identification initiative: A cultural shift in publishing", *Journal of Comparative Neurology*, vol. 524, no. 1, pp. 8–22, 2016.
- [18] B. Y. Robert Sanderson. Paolo Ciccarese., *Web annotation data model*, Retrieved on 03/03/2019 from <https://www.w3.org/TR/annotation-model/>, 2017. [Online]. Available: <https://www.w3.org/TR/annotation-model/>.

Chapter 6

Generating a Gold Standard Corpus for Experimental Protocols

Corpus of documents are necessary for training and evaluating information retrieval algorithms; these standardized collections are called Gold Standard Corpus (GSC). This chapter presents the development of a GSC for experimental protocols; the collection has 58 fully annotated protocols. The focus of this GSC is the annotation of samples, instruments and reagents. The general annotation tool *hypothes.is* was adapted for this specific scenario as described in Chapter 5. Domain experts were presented with the tool and, a clear explanation of the task. A high inter-annotator agreement was achieved; the vocabularies were thus enriched. Website: <https://smartprotocols.github.io/gold-standard/>

6.1 Introduction

Information in the biomedical domain is usually encoded in natural language; publications are conceived as a channel for humans to disseminate knowledge. Nowadays sharing and reusing knowledge are, however, activities that require software facilitate information retrieval. For instance, researchers should be able to automatically extract facts, find and generate hypothesis from existing literature, etc. Simple tasks such as finding sample information or the workflow for preparing a recipe in “Materials and Methods” demand a significant effort; very often domain expertise is required. Natural language processing (NLP) and text mining are commonly used to facilitate extracting meaningful information from large corpus of specialized text. NLP applications require training sets or information that is more structured than what we find in existing literature. Collections of annotated documents, gold standard corpus (GSC) are necessary to train algorithms [1]; these annotations usually require the participation of domain experts. Errors in training corpus significantly affect the outcome of these algorithms.

In life sciences there are many highly specialised gold standard corpus; these range from gene-disease to protein-protein interactions. For instance, the PennBioIEOncology [2] GSC comprises approximately 327,000 specialized terms, the NCBI disease corpus [3] contains 6892 mentions of diseases, the SCAI-Test corpus [4] has 1206 annotations of chemicals, the Genia [5] has 2000 MEDLINE abstracts with almost 100,000 annotations. The CRAFT corpus [6] consists of 67 biomedical journal articles of various biological domains (plus unpublished articles) with more than 100,000 annotations. Recognizing genes, chemicals and diseases are amongst the most common applications of biomedical named entity recognition (NER). Developing highly effective tools to automatically detect biological concepts depends on the availability of high quality annotated corpus. In this chapter the development of a novel and highly specialized corpus is presented; first a description of materials and methods, as well as the workflow that was followed is provided. Then, results and discussion follow.

6.2 Materials and Methods

In this section the corpus of documents is presented. We are including, Digital Object Identifiers (DOIs), domains, how was the corpus put together, criteria for the corpus, annotators (how were they selected, were they all in one place, areas of expertise), aspects evaluated in the phase of annotations (correctness in the annotation of the terms and their synonyms, failures in the recognition of terms in the texts, and identification of terms incorrectly annotated, namely a word with different meaning), tool supporting the annotation, how was the analysis of the data thus gathered done. We then present the workflow we followed.

6.2.1 Materials

Corpus of documents

From the 530 protocols initially analyzed in this research (see chapters 2 and 3), 58 of them were selected to be fully annotated. This subset of protocols coming from BioTechniques [7], MethodsX [8], Bio-Protocol [9], the Journal of Biological Methods [10], Cold Spring Harbor Protocols [11], Current Protocols [12], Genetic and Molecular Research [13], Journal of Visualized Experiments [14] and Nature Protocols [15]

(Table 6.1). The corpus includes from simple (few steps with few decision points) to complex protocols in topic areas such as molecular biology, cell and developmental biology, biochemistry, biotechnology, microbiology and virology.

Journal	No. of protocols
Bio-Protocols	19
Biotechniques	4
Cold spring harbor protocols	7
Current protocols	3
Genetic and Molecular Research	3
Journal of Biological Methods	4
Journal of Visualized Experiments	11
MethodsX	6
Nature Protocols	1

TABLE 6.1: Corpus of annotated protocols

Annotators

Thirty four annotators participated in the annotation of protocols. The skills of the annotators was diverse, from students to university professors; their areas of expertise were related to the content of the protocols. The annotation sections were carried out in several institutions (Table 6.2).

Institution	No. of annotators
Centro de Bioinformática y Biología Computacional de Colombia	8
Universidad del Valle, Colombia	11
Database Center for Life Science (DBCLS), Robotic Biology Institute (RBI), Spiber, Yachie-Lab, Universidad de Tokyo, Japan	14
Universidad Santiago de Cali, Colombia	1

TABLE 6.2: Number of annotators by institution

Annotation guidelines

The annotation guidelines were defined for this part of the research. The guidelines were influenced by the terminology to be annotated, peculiarities observed in protocols and practical issues e.g. browser issues, to ensure a consistent annotation. In addition, good practices about how to annotate were indicated, they are: i) Read the document, ii) mark each occurrence of an entity, iii) reduce the noise in the annotation, iv) add a comment indicating an annotation or decision that was hard to make, v) time to solve doubts, vi) Notify the finalization of the annotation task. The annotation guidelines are available in the appendix B and zenodo [16]. At the beginning of each annotation session these guidelines were revised; doubts were clarified. The annotation of each protocol was done by three annotators. The annotation tasks were focused terminology related to Sample / Organism, Instruments, Reagents and statements describing the overall Objective of the protocol.

Annotation tool

The annotation was supported by adapting the generic annotation tool *hypothes.is* [17] (also known as Hypothesis). The tool is presented in Chapter 5.

6.3 Methods

The annotation workflow is presented in 6.1, as well as more detailed information about the process.

FIGURE 6.1: An overview of the annotation process

The annotation task on the corpus of experimental protocol was performed in two stages. At the beginning of each annotation task annotators were given a training in the use of the tool and the annotation workflow. In the first stage annotators were working individually, the annotation tasks could be saved and re addressed later. Annotators were reading and annotating the documents, three per document. The annotators could use the comment feature in the tool indicating the issues they were finding through the process. This feature was also used to direct comments to a specific user, e.g. the supervisor. This facilitated the communication.

The second stage involved the participation of a supervisor; during these sessions issues with the annotations were discussed and addressed. Annotators could never see the annotations of another annotator. The reviewer and annotators scheduled meetings to discuss the annotations, the process and issues with the tooling; both parts could be agree on the annotation, and remove or modify some annotations. The issues with the annotations were not influenced by the supervisor; the task of the supervisor was more that of a facilitator. Each annotator examined and edited his or her own annotations after the a brief discussion; examples of use always proved to be useful.

Once the annotation task was done, the inter-annotator agreement was calculated; for more information about the inter-annotation agreement see the Results section. Annotators received a gift, as an incentive, for their participation in the process. The parts mentioning materials were categorized as follows. These categories pre-existed the annotation task. They are the product of the extensive review of experimental protocols through chapters 2 and 3.

1. samples or specimens: (i.e., Whole organism (mouse, rice,...), anatomical parts (tissues, membranes, organs,...), Biomolecules (nucleic acids, proteins, ...));
2. instruments (i.e., high-throughput equipment (Liquid Handling Platforms, Real-Time PCR Detection System,...), laboratory glassware (Beaker, Erlenmeyer flask,...), Standard equipment (Balances, shakers, centrifuges,...), software (WinGene 2.31, BioLign 4.0.6), and consumables (pipette tips, gloves, syringes, microcentrifuge tubes,...));
3. Reagents (i.e., Chemical compound/Substance (Glucose, ethanol, glycerol, chloroform,...), Solutions / buffers (70% ethanol, 10X PCR buffer,...)).

Each snippet of text related to materials are linked to a standardized vocabulary or ontology term. Then we compared the automatic annotation against those received from the manual annotation process. The overall objective of the protocols

was also annotated. The objective of a protocol is a formal statement describing the goal (i.e., “Here we present a detailed protocol for Smart-seq2 that allows the generation of full-length cDNA and sequencing libraries by using standard reagents” [18]). We decided to annotate the overall objective so that during the comparison we could have some added context to the captured data object word-surrounding of the word.

FIGURE 6.2: Workflow summarizing annotation sections

6.4 Results

The corpus has 58 fully annotated experimental protocols; the annotation effort focused on the identification of samples (including specimens), instruments, reagents and paragraphs describing instructions -idem experimental actions. The overall objective of the protocol could not be annotated in 12 protocols (6 from bio-protocols, 2 from biotechniques, 2 from cold spring harbor protocols and 2 from Journal of visualized experiments) see Table 6.3 because it was not clearly specified. An explicit statement about this metadata "not found" in these documents was added by the annotators. Annotators agreed that the objective could be inferred by a human (and not by machines), after reading the document. Annotators suggested that the title of the protocol could be annotated as the objective, very often the title was descriptive of the scope of the protocol.

	Journal	ID
1	Bio-protocols	10.21769/BioProtoc.128 (ref 19)
2	Bio-protocols	10.21769/BioProtoc.322 (ref 20)
3	Bio-protocols	10.21769/BioProtoc.181 (ref 21)
4	Bio-protocols	10.21769/BioProtoc.213 (ref 22)
5	Bio-protocols	10.21769/BioProtoc.1076 (ref 23)
6	Bio-protocols	10.21769/BioProtoc.1371 (ref 24)
7	Biotechniques	https://doi.org/10.2144/000113294 (ref 25)
8	Biotechniques	10.2144/000113260 (ref 26)
9	Cold Spring Harbor Protocols	10.1101/pdb.prot5014 (ref 27)
10	Cold Spring Harbor Protocols	10.1101/pdb.prot5393 (ref 28)
11	Journal of visualized experiments	10.3791/683 (ref 29)
12	Journal of visualized experiments	10.3791/54231 (ref 30)

TABLE 6.3: Protocols where the objective could not be annotated

As it was described in materials and methods section, each text was manually annotated by three domain experts. Thirty four annotators participated in the process, their areas of expertise were related to the scope of the protocols. Differing annotations were discussed so that a consensus could be reached; this consensus seeking process defined the second phase in the annotation process. The aim of the consensus was to focus on the main issue at hand; very often annotators discussed materials that could also be samples. The supervisor only helped the annotator in focusing on the role so that the annotation could move forward. In this setting, a high inter-annotator agreement was observed [19]. The public release of the corpus contains 1769 concepts related to Samples, Instruments and Reagents. These entities were mapped to concepts represented in ontologies like NCBI Taxon (organisms) [20]–[22], UBERON (anatomical parts) [23], ERO [24], ChEBI [25], PubChem (reagents or chemical compounds) [26], OBI [27], BAO [28], EFO [29] (instruments) and SMART Protocols ontology. A total of 6 gazetteers related to samples/organisms, instruments, reagents/chemical compounds and experimental actions (verbs), were thus gathered.

FIGURE 6.3: Architecture for generating the gazetteers

The gazetteers are used by GATE. 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 parts such as instruments, instructions, reagents). Some of our gazetteers had over half a million terms; the default ANNIE Gazetteer were used to build the gazetteers with less than 1 million terms per ontology and subdomain (Figure 6.3). These gazetteers make it possible to search and categorize meaningful segments in the text. The terms in the gazetteers are structured with metadata such as definition, URIs, provenance, synonyms, etc. By combining NLP and semantics it is possible to, for instance, disambiguate, “centrifuge” as an action and “centrifuge” as an instrument.

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 (Figure 6.3); this allows the extraction of the root for each token to be analyzed by a Morphological Analyzer. A large KB Gazetteer was used to store sets of over 1 million terms related to organisms (Figure 6.3). To facilitate data storage, a non-relational database were used and connected it to GATE.

From the gazetteers, linguistic patterns were identified so that The Iterative Rule Writing step (see Figure 6.4) could begin. Gazetteers based on ontologies have context; rules making use of these gazetteers find meaningful parts of the text. In this sense, GATE allows more than just named entity recognition (NER).

FIGURE 6.4: Example illustrating 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.

The rules for the NLP layer are encoded in JAPE (Java Annotation Patterns Engine). In this stage, were designed rules to automate the identification of meaningful elements in the narrative e.g. instructions, are characterized, then rules are written, tested and improved. Ontologies and domain terminology were mapped to the corresponding vocabularies (Figure 6.5).

Precision, Recall and F1 score were calculated to compare the automatic annotation with that from domain experts. “Instruments” was the type of entity that was best annotated with more than 60% of the annotations presenting a high F1 score (>0.70). Reagents presented a 59% and samples 45% [30]. The gazetteers failed in the following cases: i) words with typos (e.g. centrifuge vs centifuge), ii) words with different meaning (e.g., 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

FIGURE 6.5: Example illustrating a rule designed to find and annotate statements related to cell disruption

the short word for “catalog”) ; iii) or those cases where multiple samples, reagents or instruments in the same statement (e.g. MiSeq, HiSeq or NextSeq next-generation sequencing platform *instead of* MiSeq sequencing platform, HiSeq sequencing platform, next-generation sequencing platform) were mentioned.

6.5 Discussion

The NLP layer makes use of the semantics that have been defined. The gazetteers are currently reusing terminology from EFO, ERO, OBI, NCBI Taxonomy and ChEBI. The infrastructure presented throughout this chapter, brings together semantics and NLP. Thus, making it possible to retrieve specific information from the content of the protocols. The NLP layer is able to extract annotations about samples, instruments, reagents and instructions automatically.

In this thesis were encountered issues with the free narrative often used for describing the objectives. Identifying and classifying actions based on verbs and attributes, e.g. units of measure, instruments, reagents, etc has also been difficult. Other efforts addressing the representation of experimental protocols, e.g. EXACT, have also reported similar problems; as this research is using the semantics in support of NLP tasks, some of these issues have been solved by using the rule engine in GATE; as new experimental actions are identified, rules are improved.

6.6 Conclusions

Here is presented a set of gazetteers to extract automatically terms related to samples, instruments, and reagents. In addition, a set of rules were developed to extract instructions from protocols in life sciences. Annotation techniques were used in this thesis because is powerful mechanism for storing, reusing and analyzing information. Despite a small collection of protocols (58 fully annotated documents), the results shown a high inter-annotator agreement; good practices related to the development of an annotated corpus of documents were put into action. They are, i) a clear annotation tasks (what and how annotate), ii) low ambiguity of the data, iii)

the participation of annotators experts in life sciences, iv) three-annotators per document (randomly selected) and v) two annotation phases. The construction of this annotated corpus was a laborious, very costly and time-consuming process because the agreement of the annotators depended on the ambiguity of the data, skill of the annotators and the task at hand. In addition, processing and interpreting text automatically in natural language are challenging tasks, because the meaning of terms is context dependent [1].

Summarizing, our approach combines NLP and semantics; it facilitates the generation of a self-describing document as it helps to automatically identify fragments that are often hidden in the narrative. It makes it possible to present meaningful information of experimental protocols without compromising the content of the whole document. Publishers may choose to publish only experimental actions as nanopublications without releasing the rest of the text. More importantly, it makes it possible to anchor information on a context that is meaningful for experimental researchers, such as samples, instruments and reagents.

Bibliography

- [1] L. Wissler, M. Almashraee, D. Monett, and A. Paschke, "The gold standard in corpus annotation", Jun. 2014. DOI: 10.13140/2.1.4316.3523.
- [2] P. W. Mark Liberman Mark Mandel, "Pennbioie oncology 1.0", *Linguistic Data Consortium*, 2008.
- [3] R. I. Doğan, R. Leaman, and Z. Lu, "Ncbi disease corpus: A resource for disease name recognition and concept normalization", *Journal of biomedical informatics*, vol. 47, pp. 1–10, Feb. 2014. DOI: 10.1016/j.jbi.2013.12.006. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/24393765>.
- [4] *Corpora for chemical entity recognition - fraunhofer scai*. [Online]. Available: <http://www.scai.fraunhofer.de/chem-corpora.html>.
- [5] J. D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpus—a semantically annotated corpus for bio-textmining", *Bioinformatics*, vol. 19, no. suppl_1, pp. i180–i182, Jul. 2003. [Online]. Available: <http://dx.doi.org/10.1093/bioinformatics/btg1023>.
- [6] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake, and L. E. Hunter, "Concept annotation in the craft corpus", *BMC Bioinformatics*, vol. 13, no. 1, p. 161, 2012. DOI: 10.1186/1471-2105-13-161. [Online]. Available: <https://doi.org/10.1186/1471-2105-13-161>.
- [7] *Biotechniques.com*. [Online]. Available: <https://www.biotechniques.com/>.
- [8] *MethodsX, MethodsX, Instructions for Authors*, 2014. [Online]. Available: <https://www.elsevier.com/journals/methodsx/2215-0161/guide-for-authors>.
- [9] *Bio-protocol, Bio-protocol*, 2017. [Online]. Available: <http://www.bio-protocol.org/Default.aspx>.
- [10] *Journal of Biological Methods*. [Online]. Available: <http://www.jbmethods.org/jbm> (visited on 02/09/2017).
- [11] CSH-Protocols, *Cold Spring Harbor Protocols*, Retrieved on 06/12/2017, 2017. [Online]. Available: <http://cshprotocols.cshlp.org/>.
- [12] *Current Protocols - Wiley Online Library*. [Online]. Available: <https://currentprotocols.onlinelibrary.wiley.com/>.
- [13] GMR, *Genetics and Molecular Research*, 2017. [Online]. Available: <http://www.geneticsmr.com/>.
- [14] JoVE, *Journal of Visualized Experiments*, 2017. [Online]. Available: <https://www.jove.com/>.
- [15] *Nature-Protocol, Nature Protocol*, 2017. [Online]. Available: <https://www.nature.com/nprot/>.
- [16] O. Giraldo, *Guidelines to annotate experimental protocols*, Dec. 2018. DOI: 10.5281/zenodo.2171281. [Online]. Available: <https://doi.org/10.5281/zenodo.2171281>.

- [17] *Hypothesis – the internet, peer reviewed*, 2018. [Online]. Available: <https://web.hypothes.is>.
- [18] S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg, “Full-length rna-seq from single cells using smart-seq2”, *Nature Protocols*, vol. 9, 171 EP –, Jan. 2014. [Online]. Available: <https://doi.org/10.1038/nprot.2014.006>.
- [19] O. Giraldo, *Fleiss kappa of protocols*, Nov. 2018. DOI: 10.5281/zenodo.1489112. [Online]. Available: <https://doi.org/10.5281/zenodo.1489112>.
- [20] S. Federhen, “Type material in the NCBI Taxonomy Database”, *Nucleic Acids Res*, vol. 43, pp. D1086–98, 2015.
- [21] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank”, *Nucleic Acids Research*, vol. 37, no. Database, pp. D26–D31, 2009, ISSN: 0305-1048.
- [22] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye, “Database resources of the National Center for Biotechnology Information”, *Nucleic Acids Research*, vol. 37, no. Database, pp. D5–D15, 2009, ISSN: 0305-1048.
- [23] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel, “Uberon, an integrative multi-species anatomy ontology”, *Genome Biology*, vol. 13, no. 1, R5, 2012, ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-1-r5. [Online]. Available: <http://dx.doi.org/10.1186/gb-2012-13-1-r5>.
- [24] C. Torniai, M. Brush, N. Vasilevsky, E. Segerdell, M. Wilson, T. Johnson, K. Corday, C. Shaffer, and M. Haendel, “Developing an application ontology for biomedical resource annotation and retrieval: Challenges and lessons learned”, in *Proceedings of the Second International Conference on Biomedical Ontology: July 26-30, 2011; Buffalo, NY. 2011*, http://icbo.buffalo.edu/ICBO-2011_Proceedings.pdf, vol. 833, 2011, pp. 101–108.
- [25] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, and C. Steinbeck, “The ChEBI reference database and ontology for biologically relevant chemistry: Enhancements for 2013”, *Nucleic Acids Res*, vol. 41, pp. D456–63, 2013.
- [26] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, “PubChem Substance and Compound databases.”, *Nucleic acids research*, vol. 44, no. D1, pp. D1202–13, 2016, ISSN: 1362-4962.
- [27] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, L. Fan, J. Fostel, G. Fragoso, F. Gibson, A. Gonzalez-Beltran, M. A. Haendel, Y. He, M. Heiskanen, T. Hernandez-Boussard, M. Jensen, Y. Lin, A. L. Lister, P. Lord, J. Malone, E. Manduchi, M. McGee, N. Morrison, J. A. Overton, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, D. Schober, B. Smith, L. N. Soldatova, C. J. Stoeckert, C. F. Taylor, C. Torniai, J. A. Turner, R. Vita, P. L. Whetzel, and J. Zheng, “The Ontology for Biomedical Investigations”, *PLOS ONE*, vol. 11, no. 4, Y. Xue, Ed., e0154556, 2016, ISSN: 1932-6203.

- [28] S. Abeyruwan, U. D. Vempati, H. Küçük-McGinty, U. Visser, A. Koleti, A. Mir, K. Sakurai, C. Chung, J. A. Bittker, P. A. Clemons, S. Brudz, A. Siripala, A. J. Morales, M. Romacker, D. Twomey, S. Bureeva, V. Lemmon, and S. C. Schürer, "Evolving BioAssay ontology (BAO): Modularization, integration and applications.", *Journal of Biomedical Semantics*, vol. 5, no. Suppl 1 Proceedings of the Bio-Ontologies Spec Interest G, S5, 2014.
- [29] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson, "Modeling sample variables with an Experimental Factor Ontology", *Bioinformatics*, vol. 26, no. 8, pp. 1112–1118, 2010.
- [30] O. Giraldo, *Precision, recall and f1 score*, Nov. 2018. DOI: 10 . 5281 / zenodo . 1753520. [Online]. Available: <https://doi.org/10.5281/zenodo.1753520>.

Chapter 7

Semantics at Birth, the SMART Protocols Publication Platform

Semantic publishing has various meanings in scholarly communication; in general, the implementations of semantic publishing are not about publishing semantic representations for the scholarly paper. Instead, semantic publishing has been more related to semantically post-process published narratives; the semantic processing involves, annotations with ontologies in order to support discovery, via conceptual queries, and interlinking, via linked data. The SP approach to semantic publishing does not process the final output from the publication workflow; our approach considers the generation of a semantic document through the publication workflow in order to have semantics at birth. In this way semantics is not a post mortem process, after it has been published, but rather it is an integral part of the generation of the document. Scientific publications aggregate data by encompassing it within a persuasive narrative. The SP approach addresses the problem of supporting such aggregation over a document that is to be born semantic, interoperable and conceived as an aggregator within a web-of-data publishing workflow. The SMART Protocols (SP) approach delivers the tooling necessary for authors to generate this type of documents. Existing ontologies, data structures, standards and, Application Programming Interfaces are brought together in order to facilitate the assemblage, identification and characterization of these arrangements in the document -in this case the experimental Protocol. Website; <https://smart-protocols.firebaseio.com/login>

7.1 Introduction

Research papers are ill-suited to continue being considered as the most valuable scholarly channel of communication; although publication is a key part of the scientific process, it has not fundamentally changed over the years. Nowadays, communicating scientific outcomes requires having the content available for humans in a way so that it can be navigated and managed and also for machines in a way so that it can be easily processed -e.g. by providing methods to automatically organize reported scientific findings [1]. Scientific publications should now be conceived for the web of data instead of simply being printed articles. “Perhaps the most important shortcoming of the current publication system is that scientific papers do not come with formal semantics that could be processed, aggregated, and interpreted in an automated fashion” [1].

Bringing the direct advantages of adding semantics from the authoring process is particularly difficult because of the burden it imposes on the author -extra work, lack of appropriate tooling, no clear outputs from the publication workflow, disagreements on the terminology, etc. Hiding the complexity of semantic publishing while making it part of the publication workflow is therefore key for effectively engaging authors in making scholarly work born semantic. The SMART Protocols Publication Workflow (SPPW) address the problem of supporting the generation of scientific documents that are born interoperable and with a well-defined semantics. It delivers an aggregative, interoperable and, semantically defined structure upon which meaningful objects are brought for experimental protocols. The SPPW uses the SP ontology [2] as a template for representing experimental protocols while underneath generating RDF and annotating specific components with ontologies such as the NCBI taxon [3], ChEBI [4], PubChem [5] and other relevant resources. As the experimental protocol is intrinsically related to data, the SPPW also facilitates the generation of Research Objects (ROs) [6] and Distributed Scholarly Compound Objects (DISCOs) [7]. The final object is an aggregation of assertions expressed in RDF and exposed as linked data as well as HTML and PDF; the aggregated object is automatically deposited in Zenodo (RDF). In addition, the SPPW publishes a layer of metadata that is enriched by bio-schemas; thus making it easier for search engines to process these documents.

Other approaches, most notably Biotea [8], [9] have addressed this problem as a mortem issue; publications are semantically enriched after they have been authored and published. There are a number of problems with this approach; for instance, the dependence on the accuracy of Named Entity Recognition (NER) algorithms.

Publishers are acknowledging the importance of making content more machine friendly by adding semantics for scientific publications; they are actively improving programmatic access to their products and using ontologies to annotate their content and exposing it as linked data. For instance, Nature Publishing Group (NPG) recently released 20 million Resource Description Framework (RDF) statements, including primary metadata for more than 450,000 articles published by NPG since 1869. The dataset is limited to basic citation information (title, author, publication date, etc.), identifiers, and Medical Subject Headings (MeSH) terms. Their data model makes use of vocabularies such as the Bibliographic Ontology (BIBO) [10], Dublin Core Metadata Initiative (DCMI) [11], Friend of a Friend (FOAF) [12], and the Publishing Requirements for Industry Standard Metadata (PRISM) [13] as well as ontologies that are specific to NPG [14]. Similarly, Elsevier provides an Application Programming Interface (API) that makes it possible for developers to build specialized applications. The Cochrane Collaboration [15] illustrates a very important

example where scientific documents are semantically annotated by domain experts and full text content is released as linked data [16], [17]. The Linked Data portal for the Semantic Web Journal [18] exposes the publication and review data from the journal into the linked data cloud.

Semantic annotations and linked data technology make it possible to use ontology concepts to formulate queries built from concepts, e.g. retrieve papers about “calcitonin and kidney injury together with Uniprot proteins that have calcitonin binding as molecular function as well as the calcitonin resource description from from DBPedia [19]”. The queries can easily be expanded by adding concepts and data sources. The biomedical linked data infrastructure facilitates to expand the query by indicating data sources capable of resolving specific parts of it; this is supported by the SPARQL specification [20]. Semantic annotations also make it possible to compare sections from different papers, e.g., “what chemical entities do papers X and Y have in common in the Methods section”. The semantics in SP facilitates the definition of granular queries focusing on entities in specific sections; for instance, it allows us to retrieve papers based on very detailed information beyond just matching entities; for instance, retrieve protocols for “*real time PCR and chromatography methods*” will result in documents in which real time PCR methods and chromatography methods are presented and not just mentioned.

This chapter is organized as follows, first the concept of semantic publishing for experimental protocols is explained. Then, The ontologies used and the architecture are presented. The implementation as well as the publication workflow follows then. Discussion and conclusions are the final part of this chapter.

7.2 Semantic Publishing for Experimental Protocols

A semantic document is one where human-readable knowledge is augmented to enable its interpretation by machine. Semantic publishing is often referred to as “anything that enhances the meaning of a published journal article, facilitates its automated discovery, enables its linking to semantically related articles, provides access to data within the article in actionable form, or facilitates integration of data between papers” [21]. This definition is limited because it assumes that the semantics is an add on to the paper; the SPPW delivers the semantics as an integral part of the digital object, in this case an experimental protocol. It is also too inclusive because it assumes the semantics as a post mortem process -idem. after the paper has been published; not a natural part of the pre, publication and post publication process. Moreover, this definition falls short because it is assuming semantics as simply an annotation; it does not convey the semantics as the natural aggregator of assertions for both, humans and machines.

The operational definition supporting the SP semantic publishing platform for experimental protocols is as follows. Semantic publishing is the workflow that aggregates detailed and well characterized semantic interoperable assertions in a way so that these are intelligible for humans and procesable by machines. Unlike other approaches, the SP semantic publishing workflow manages the semantics as the assertions are being created by domain experts. In addition, it assumes that the aggregation of assertions does not end at the time of publication; it assumes that the assertions in the published object will continue to evolve. The formalization of knowledge in the SP semantic publishing platform does not start from the statements written in natural language; instead, it frames the natural language statement

within a predefined semantic structure. Annotations are therefore assertions aggregated to the object being published but they are not the only semantics in the object. The experimental protocol becomes a container for semantics instead of a closed box with semantics as secondary usage and an unintended consequence.

Experimental protocols are a type of publication with a strong workflow component. Also, as the main purpose of an experimental protocol is to facilitate reproducing an experiment, these objects are rich in reagents, samples, instruments and actions as well as parameters making actions possible -e.g. heat at 80C for 24 hours. The structure implicit in experimental protocols makes them ideal candidates for a semantic publication workflow that delivers and manages the semantics at birth. By the same token, experimental protocols are central for reproducibility, they are frequently related to datasets, materials and methods sections in publications, and analysis software. All these being objects that are not necessarily part of the protocol; very often these are related objects living somewhere else in the web. Thus, making the data bundle necessary in order to produce a comprehensive usable research object.

7.2.1 Preserving the Resource Map for a Protocol

Our approach builds upon the deliverables of the RMap project [7]. RMaps capture and preserves maps of scholarly works; RMaps are based on a simplified version of the Object Reuse and Exchange (ORE) [22] specification, that of the Distributed Scholarly Compound Objects (DiSCOs) [23]. The SP approach represents relations across digital assets using DISCOs. These are graphs representing aggregations of related scholarly resources. For example, a single DiSCO might represent, an article, its related datasets, and software – as well as any useful context information describing those resources. When created, DiSCOs are assigned a unique identifier that can be used to retrieve them later. They also have a status, and Event (provenance) information. Research Objects are a specification that is similar to that of DISCOs; both represent bundles of digital objects. For practical purposes both representations are considered equivalent throughout this research. Hence, we represent the bundles of digital objects using ROs and DISCOs. Our main purpose is to express and preserve the map of items related to an experimental protocol. These may be datasets, presentations, videos, marginal notes, etc. The generic representation of a DiSCO for protocol is depicted in Fig. 7.1.

FIGURE 7.1: General view for an RMap represented as a Disco. In this figure, assets related to a protocol are presented. Small icons were taken from www.flaticon.com

The RMap presented in Fig. 7.1 follows the best practices as indicated by the RMap project [24]. In our case we are using the SMART Protocols ontology and extending the DiSCO specification with relations such as `fabio:hasManifestation` and `fabio:isManifestationOf` from the SPAR Ontologies [25]. In this case the `fabio:hasManifestation` and `fabio:isManifestationOf` is used to represent the relation between the protocol and the various formats in which it is available.

7.3 Results

The central component of the data acquisition tool is the OWL ontology. It provides a *semantically enriched* data entry form that can be used for aggregation and querying. Some specific parts of the text thus captured are automatically annotated; this enriches the semantics of the resulting document. For instance, reagents are also annotated against the web services offered by PubChem [5]. In addition to this kind of resources, that is publicly available, the annotation is also supported by a combination of gazetteers and rules as described in chapter 6.

For the application scenario, protocols from the GlycoScience Protocol Online Database [26] were used. As an on-line protocol, GlycoPOD has a unique architecture that made it easier for a parser to extract the information and populate the ontology with minimal human intervention. The GlycoPOD DB has a well standardized structure with the "Protocols (main text)", which consist of "introduction", "protocol" and "references". In "protocol", all the experimental procedures are documented in flow charts with comments. If you click "Parts", all the experimental procedures listed in "protocol" are broken down into individual experimental elements (parts).

The semantic supporting the SP approach delivers a fine granular characterization of steps; these can also be grouped as objects that may be independent, albeit related by the provenance, from the rest of the protocol. In this way only the workflow component of the protocol together with the execution parameters may be exported. It is also possible to establish a relation between the steps or, set of steps or, the whole workflow and a file, e.g. the spreadsheet with the results for a specific step. The data is bundled as a DiSCO as well as an RO. The SP platform generates a unique identifier for the protocol as well as for components grouped and exported as ROs or DiSCOs. The SP platform allows users to record runs of protocols; thus, the users generate a snapshot that makes it possible for anyone to verify experimental reproducibility.

7.3.1 Architecture and Data Workflow

The application is connected to ORCID ID services; in this way authors are identified with their ORCID Ids. Then, the structure of the SMART Protocols ontology, presented in chapter 3, guides the data capture. The reagent name resolver from PubChem and the organism name resolver from Uniprot (base on the NCBI Taxon [27] [28]) are used throughout the data capture process for resolving reagents and samples. The gazetteers developed in chapter 6 are also being used to annotate samples, for specific sample names with no reference in the NCBI Taxon, reagents, for reagents with no reference in the PubChem DB and, instruments. A general overview of the architecture is presented in 7.2

The protocol is data that is both structured and semantically enriched. It is possible to query the content using SPARQL or, classify it into criteria representing powerful OWL expressions. For example, the code snippet 1 presents a simple SPARQL

FIGURE 7.2: General Architecture for SMART Protocols

query that returns all instances of protocols where rodent was used as a sample; more complex queries can be found at ¹.

```
PREFIX sp: <http://purl.org/net/SMARTprotocol#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX ro: <http://www.obofoundry.org/ro/ro.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
```

```
SELECT ?title ?specimenName
WHERE {
  ?protocol sp:hasTitle ?title_uri .
  ?title_uri rdf:value ?title .
  ?protocol sp:hasExperimentalInput ?specimens .
  ?specimens a sp:SpecimenList .
  ?specimens ro:has_part ?specimenNameUri .
  ?specimenNameUri rdf:value ?specimenName .
  ?specimen sp:hasName ?specimenNameUri .
  SERVICE <https://dbpedia.org/sparql>
  {
    ?externalUri dbo:order <http://dbpedia.org/resource/Rodent> .
    ?externalUri rdfs:comment ?dbpediaDesc .
```

¹<https://smartprotocols.github.io/>

```

    FILTER(lang(?dbpediaDesc) = 'en')
  }
  ?specimen owl:sameAs ?externalUri .
}
```

LISTING 7.1: SPARQL query "Retrieve all the protocols with samples that belongs to the Rodent order"

The application interoperates with several web applications. For instance, ZENODO is used for publishing the research object bundle; the DISCO formalism is also being used. The application builds the research object using various formats for representing the protocol, e.g. RDF, HTML, PDF, JSON LD. Protocols are also available over a SPARQL endpoint. The figures below illustrate the publication process 7.3, 7.4 .

FIGURE 7.3: A view of the publication process

FIGURE 7.4: Publishing a narrative as data

7.4 Discussion

In this paper, the SMART Protocols publication system for RDF-based form generation and data acquisition has been presented. The generated triplets as well as research objects and/or DISCOs are available over a SPARQL endpoint. In addition, data is also available from ZENODO. The document is related with the generated data for each step; it is also annotated with external services such as PubChem for reagents.

The aggregative nature of the scientific documents, that is as an aggregator of assertions, is lost in the current publication workflow. The content is flat and remains flat throughout the publication workflow. Semantic enrichment happens after the paper has been published and with little or no input from the author. Moreover, the relation between the paper and data elsewhere in the web is lost -most often than not for ever. The work presented in this chapter addresses these issues and delivers a genuine semantic document as defined here and as well as defined by Kuhn and Dumontier [1].

The semantic representation of SMART Protocols defines data items as granular as possible. This makes it possible to re-purpose the triples as needed. Two scenarios where the triplets were re-purposed are described below. These scenarios were not implemented in full; they were analyzed as manually generated data. Given the current technology implementing them is feasible.

7.4.1 Granular preservation over Hyperledger

This first scenario responds to the need for preserving specific parts of the experimental record that are related to the execution of a protocol. For instance, researchers may be required to provide an immutable record for outcomes derived from certain steps, e.g. agarose gel images. In order to address this use case we decided to use blockchain technology; this is widely used for preserving financial assets, e.g. cryptocurrencies. Blockchains can be used to record promises, trades, transactions or simply items we never want to disappear. Mirrored exactly across all nodes in a given network, it allows everyone in an ecosystem to keep a copy of the common system of record. Nothing can ever be erased or edited. Blockchain and cryptocurrencies are often discussed in similar contexts, but they are not the same thing. Distributed ledgers, e.g. blockchains, don't require a cryptocurrency to work. For our early testing we decided to use Hyperledger, this is a free open source distributed ledger framework. As we are testing we are working with Hyperledger Composer; this allows us to define the business network that we want to implement. In our case, it is simply a model that records a transaction, that of storing a DiSCO.

The data object "Protocol-run-Step+data" is passed from the SP platform to the blockchain implemented in Hyperledger Composer [29]; in this way an immutable record is generated. In this way we are creating a trusted source of evidence where the data bundled "Protocol-run-Step+Data" is stored. The degree of granularity for the DiSCO is adjustable. In this way administrators may choose to preserve in Hyperledger [30] only the data bundle "Protocol+Data" with no information indicating the step in which the data is generated. Others may choose to be as granular as the SMART Protocols representation allows. The use of blockchain technology is not mandatory; the data bundles can be stored using other technologies, e.g. Ethereum [31].

7.4.2 Nanopublications from SMART Protocols

A nanopublication is the smallest unit of publishable information: an assertion about anything that can be uniquely identified and attributed to its author [32]. In Fig. 7.5 we represent an experimental procedure as a nanopublication. As SMART Protocols represents statements, these are meaningful assertions within the discourse of an experimental protocol then these assertions can be re purposed as needed. In this case, we are simply embedding an experimental action within the structure of a nanopublication. Experimental actions indicate what to do (verb), what to use (equipment and sample), how to do it (parameters for equipment, or experimental conditions). They are fully contained units; having these as nanopublications may help to better identify them and also to reuse them. We are not formally aligning the ontologies in this example; however, conceptually it works and developing the machinery to extract nanopublications from SMART Protocols RDF should not pose a significant challenge.

Notes:

- (i) Each part should correspond to an ontological term. For instance “incubate” is an **occurrent** in the Basic Formal Ontology while “tubes” is a **continuant** in the same ontology
- (ii) Materials can also be included, i.e., what is in the tubes

FIGURE 7.5: Nanopublications from a procedure

7.5 Conclusions and Final Remarks

The SMART Protocols publication tool for generating Web forms based on RDF ontologies, and acquire instance data that relates to ontology terms with potentially complex class expressions associated with them has been presented in this chapter. The SMART Protocols ontology, presented in chapter 3, is used to guide the data capturing process by specifying the structure of Web-based forms and the structure of data to be acquired from the forms, and to semantically enrich the data elements in the forms. The system is designed to deliver a true semantic publication, as defined earlier in this chapter. The system is query friendly as it also publishes over a SPARQL endpoint. The publication process presented here is a novel publication paradigm that delivers semantics at birth. In this way the annotation of entities is not a post mortem task but one where the author takes an active role with minimal overhead. As a consequence, the data acquired is fully integrated into the ontologies; thus enabling inferences and queries on the data that are otherwise not easy. This

system also reuses the work presented in chapter 4, the LabProtocol Bioschemas profile; the HTML is marked with this extension.

The early experiments presented in sections 7.4.1 and 7.4.2 indicate feasible research paths that are worth looking into. They are not meant to be conclusive; both approaches are related to the concept of a semantic publication and are related to the SMART Protocols approach to semantic publications. The work presented in this chapter focuses on one single very specific type of scientific document. The approach can be easily applied to other domains, as Goncalves et al have already demonstrated [33].

Bibliography

- [1] M. D. Tobiasa Kuhn, "Genuine semantic publishing", *Data Science*, vol. 1, pp. 139–154, 2017.
- [2] O. Giraldo, A. García, F. López, and O. Corcho, "Using semantics for representing experimental protocols", *Journal of Biomedical Semantics*, vol. 8, no. 1, p. 52, 2017, ISSN: 2041-1480. DOI: 10.1186/s13326-017-0160-y. [Online]. Available: <https://doi.org/10.1186/s13326-017-0160-y>.
- [3] S. Federhen, "The ncbi taxonomy database", *Nucleic Acids Research*, vol. 40, no. D1, pp. D136–D143, 2012. DOI: 10.1093/nar/gkr1178. eprint: /oup/backfile/content_public/journal/nar/40/d1/10.1093_nar_gkr1178/2/gkr1178.pdf. [Online]. Available: <http://dx.doi.org/10.1093/nar/gkr1178>.
- [4] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner, "Chebi: A database and ontology for chemical entities of biological interest", *Nucleic Acids Research*, vol. 36, no. suppl₁, pp. D344–D350, 2008. DOI: 10.1093/nar/gkm791. eprint: /oup/backfile/content_public/journal/nar/36/suppl_1/10.1093_nar_gkm791/2/gkm791.pdf. [Online]. Available: <http://dx.doi.org/10.1093/nar/gkm791>.
- [5] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "PubChem Substance and Compound databases.", *Nucleic acids research*, vol. 44, no. D1, pp. D1202–13, 2016, ISSN: 1362-4962.
- [6] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. García-Cuesta, J. M. Gomez-Perez, G. Klyne, K. Page, M. Roos, J. E. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C. Goble, "Workflow-centric research objects: First class citizens in scholarly discourse", vol. 903, May 2012.
- [7] *The rmap project: Linking the products of research and scholarly communication*, 2015. [Online]. Available: https://www.stm-assoc.org/2015_04_22_Annual_Conference_DiLauro_Linking_data_and_publications.pdf.
- [8] A. Garcia, F. Lopez, L. Garcia, O. Giraldo, V. Bucheli, and M. Dumontier, "Biotea: Semantics for pubmed central", *PeerJ*, vol. 6, e4201, 2018.
- [9] L. J. Garcia Castro, C. McLaughlin, and A. Garcia, "Biotea: Rdfizing pubmed central in support for the paper as an interface to the web of data", *Journal of Biomedical Semantics*, vol. 4, no. 1, S5, 2013, ISSN: 2041-1480. DOI: 10.1186/2041-1480-4-S1-S5. [Online]. Available: <https://doi.org/10.1186/2041-1480-4-S1-S5>.
- [10] *Bibliographic ontology specification | the bibliographic ontology*, 2009. [Online]. Available: <http://bibliontology.com/specification.html>.
- [11] *Dcmi: Dcmi metadata terms*, 2012. [Online]. Available: <http://dublincore.org/documents/dcmi-terms/>.

- [12] *Hypothesis – the internet, peer reviewed*, 2018. [Online]. Available: <http://www.foaf-project.org/>.
- [13] *Prism – idealliance*. [Online]. Available: <https://www.idealliance.org/specification/prism>.
- [14] *Press release archive: About npg*, 2012. [Online]. Available: https://www.nature.com/press_releases/linkedata.html.
- [15] *Cochrane | trusted evidence. informed decisions. better health*. [Online]. Available: <https://www.cochrane.org/>.
- [16] *Cochrane*. [Online]. Available: <http://linkedata.cochrane.org>.
- [17] C. Mavergames, S. Oliver, and L. Becker, “Systematic reviews as an interface to the web of (trial) data: Using pico as an ontology for knowledge synthesis in evidence-based healthcare research.”, in *SePublica*, 2013, pp. 22–26.
- [18] *Linked data portal for the semantic web journal*, 2012. [Online]. Available: <http://semantic-web-journal.com/SWJPortal/>.
- [19] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, *et al.*, “Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia”, *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.
- [20] *Sparql working group*, 2013. [Online]. Available: https://www.w3.org/2009/sparql/wiki/Main_Page.
- [21] D. SHOTTON, “Semantic publishing: The coming revolution in scientific journal publishing”, *Learned Publishing*, vol. 22, no. 2, pp. 85–94, 2009. DOI: 10.1087/2009202. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1087/2009202>. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1087/2009202>.
- [22] O. A. Initiative, *Open archives initiative object reuse and exchange*, Retrieved on 07/07/2017 from <http://www.openarchives.org/ore/1.0/toc>, 2014. [Online]. Available: <http://www.openarchives.org/ore/1.0/toc>.
- [23] RMaps, *Rmaps glossary*, Retrieved on 07/07/2017 from <http://www.openarchives.org/ore/1.0/toc>, 2017. [Online]. Available: <http://www.openarchives.org/ore/1.0/toc>.
- [24] *Rmap disco design best practices*, 2019. [Online]. Available: <http://tiny.cc/e4uk5y>.
- [25] *Spar ontologies*, 2019. [Online]. Available: <http://www.sparontologies.net/>.
- [26] *[about database]:glycoscience protocol online database*. [Online]. Available: <https://jcggdb.jp/GlycoPOD/aboutDatabase>.
- [27] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank”, *Nucleic Acids Research*, vol. 37, no. Database, pp. D26–D31, 2009, ISSN: 0305-1048.
- [28] E. W. Sayers, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko, and J. Ye, “Database resources of the National Center for Biotechnology Information”, *Nucleic Acids Research*, vol. 37, no. Database, pp. D5–D15, 2009, ISSN: 0305-1048.

- [29] hyperledger.org, *Hyperledger composer*, Retrieved on 07/07/2018, 2017. [Online]. Available: <https://www.hyperledger.org/projects/composer>.
- [30] —, *Hyperledger*, Retrieved on 07/07/2018, 2017. [Online]. Available: <https://www.hyperledger.org>.
- [31] ethereum.org, *Ethereum, blockchain app platform*, Retrieved on 07/07/2018, 2017. [Online]. Available: <https://www.ethereum.org/>.
- [32] <http://nanopub.org>, *What is a nanopublication*, Retrieved on 07/07/2018 from [:/nanopub.org/wordpress/?page_id=65](http://nanopub.org/wordpress/?page_id=65), 2017. [Online]. Available: http://nanopub.org/wordpress/?page_id=65.
- [33] R. S. Gonçalves, S. W. Tu, C. I. Nyulas, M. J. Tierney, and M. A. Musen, "An ontology-driven tool for structured data acquisition using web forms", *Journal of biomedical semantics*, vol. 8, no. 1, p. 26, 2017.

Chapter 8

Discussion and Conclusions

8.1 Summary

This thesis has primarily dealt with the problem of semantics for necessary and sufficient documentation for biological experiments; the focus of this work has been on one specific document, the experimental protocol. This research presents a detailed description of three different but complementary layers of semantics for experimental protocols, idem a checklist, an ontology and a specification for bioschemas. Throughout this investigation the complementary nature across these three entities has been demonstrated. Moreover, this research has also addressed the problem of using these layers of semantics in order to build a publication workflow that delivers semantics at birth. This novel publication paradigm addresses the problem of semantics making use of resources that the bio community has been building, idem checklists, ontologies and markup specifications.

The introductory chapters have investigated methodological aspects about building semantic artifacts for representing experimental protocols. The methodological aspects always involved the participation of domain experts, reusing existing resources and more importantly focusing on making it clear for the end user what the the win-win situation. This research is based upon real cases in which researchers were involved; this allowed the author to gain a deeper insight into the context in which solutions were to play a role. It also forced the author to focus on deliverables that made sense to the domain expert.

Later in this thesis the problem shifts from building semantic layers into using these layers in a coherent manner and thus proposing a solution for the problem of having data without accurate descriptions that indicate how these data were produced. From the conception of this research the author focused on the premise that data availability is not a proxy for data reproducibility. The bundle “protocol-data” using specifications such as that of Research Objects within a fully semantic framework have the advantage of producing an object that is for the web of data, idem for machines to process, while preserving the human in the process. This approach made it possible for the author to investigate and propose a novel publication workflow for experimental protocols -documents to be born semantics.

The discussion and conclusions are organized as follows; initially a summary of the thesis is presented. In the subsection *Reusable data* the author discusses the implications of the SMART Protocols approach to the production of reusable data. Then, in *Using the Semantic Layers*, the author discusses how the different layers of semantics presented in chapters 2,3 and 4 could be used. As the experimental protocol is just a part of the experimental record a part of the discussion addresses issues related to LIMS.

8.2 Reusable Data

Open science requires open data; equally true, open is not necessarily related to making data freely available. High quality data is expensive; an experimental record that facilitates reproducibility is not easy to maintain. We (scientists and the public at large) need access to the data supporting each scientific finding. We also need to understand how the data was collected and what each data element represents. Reuse of data in the current scientific ecosystem, is fraught with complications in accessing and understanding data, resulting in limited reuse, and the potential for false findings when data is reused. Scientists create data using a wide variety of processes. Scientists separate the recording of measurements (often called “the data”) from the specifications of the conditions under which the recordings were made (often called “Materials and Methods” in a scientific paper). The figure below depicts the current state and some of the issues arising throughout the ecosystem.

FIGURE 8.1: Reusable data

Specifications of data collection are typically not machine readable, other than as text, and are often incomplete as a result of the norms of the scientific publication process. Measurements are shared as data with little specification. Scientists share data to a large collection of archives, each with its own disclosure, indexing, formatting, and curation requirements, and with potential reuse limitations. Archives then provide data to potential reuses of data in a variety of formats, and under a variety of use restrictions. Specifications of the data are provided as “documentation,” or “data dictionaries.” Preparing data for reuse involves mastering the various formats of the measurements, discovering the definitions of each data element, aligning multiple archived data sets with respect to common data elements, and eventually producing a combined data set that can be reused.

This research has addressed the problem of representing the experimental protocol; implicit in the bottom up approach taken by the author of this investigation lays the convenient incorporation of data specifications, that is, metadata, often called semantics, into the scientific record. The conditions under which measurements are taken are best known as the measurements are taken. Thus, the author advocates for data to be “born semantic” at the point of primary data collection, for measurements to be taken in the context of their semantics, and to never be separated from their

semantics. Moreover, the author advocates aggregating statements in a publication workflow that makes the process transparent to the end user.

The vast amount of scientific data that has already been collected, and that much scientific data is collected in settings where data specifications may need to be added to data already collected is acknowledged by the author. Such post mortem “semantic enrichment,” is possible using the same bottom-up approach. The metadata will be not be 100% complete at any point through the digital continuum. The scientific process is one which requires continuous reexamination of findings, often with the application of new concepts. Previously collected data will often be further enriched. The aggregative nature of semantics in the form of statements that are intelligible for humans and machines alike facilitate this evolution. Having the data-protocol semantics at birth makes this progression much simpler and easier to manage.

Many semantic models, with various degrees of formality, exist for the definition of scientific data elements, context of scientific discovery, and provenance of scientific data. Incentives for producing complete, machine readable semantic content are the subject of intense discussion as issues of reuse and reproducibility. A missing piece, addressed through this research work, is a tool for the creation of “born semantic” data. Capturing data is inherently related to specific points in the execution of the experimental process and thus of the digital continuum of the experimental record. The SMART Protocols approach makes it natural to declare the workflow, specify those points at which data occurs, and for those points, semantically enrich the measurement provided by the researcher. Data entry, as described, generates “born semantic” data-protocol as a natural part of the scientific process.

Identity metadata of investigators, the protocol, the sponsors, and objects of study greatly improves the “findability” of the data. Access is improved by using the tool to publish its data-protocol to archives, e.g. ZENODO. Interoperability is improved by using standard metadata definitions and representations. Reuse is improved by never separating the measurements from the descriptions of the measurements. These four principles Findable, Accessible, Interoperable, and Reusable are known as the FAIR data principles.

8.2.1 Using the Semantic Layers

In this investigation it has been implicitly argued that the integration of information in molecular bioscience (and, by extension, in other technical fields) is a deeper issue than access to a particular type of data and relations for such data type across other databases. Integrating information in the bio domain has to support research endeavours. The problem of experimental reproducibility is related to a rupture in the digital continuum of the experimental record. Integrating information across the experimental record involves addressing existing practices that make sense for laboratory researchers but may not be good enough for the current demands on experimental reproducibility. Data should be available but, how the data was produced should also be available.

By using the semantic layers built throughout this investigation the digital continuum, data-experimental protocol, is preserved. Furthermore, by producing documents that are intelligible for both humans and machines, the author of this research work is understanding the web as a platform. Privacy issues, e.g. the need to keep the actual workflow private, has also been addressed through this research work. Laboratory Information Management Systems (LIMS) are a special kind of biological information systems as they in principle organise the information produced by laboratories. Once this information has been organised the analysis process takes

place, discovering relations, being able to publish data and, preserving the digital continuum becomes more and more important. The publication workflow proposed by the author makes it possible for LIMS producers to reuse the protocols because they are structured as an aggregation of machine processable statements.

Equally important, the publication workflow proposed by the author makes it possible for publishers, as well as for any laboratory, to deliver a document that is semantics at birth. By doing so, the document becomes processable at various levels. The more formal, that of the ontology, makes it possible to facilitate the execution of complex queries that are common when working in the laboratory. Also, it facilitates to establish relations between steps and specific research outcomes, e.g., an image. The Bioschemas specification for laboratory protocols makes it possible for publishers to present their content to the web for search purposes. The more casual level of semantics, that of the checklist, helps workbench researchers to keep track of their processes in a simple and intuitive manner; one that only formalises a common lab practice.

LIMSs should use domain-specific terminology in order to automatically annotate the experimental record. These vocabularies should be shared across the community so exchanging information might be a simpler task. LIMS should aggregate statements, a la nanopublications. In order for information to be shared the semantic layers should be independent from the LIMS; different LIMS should be able to share a standard vocabulary. This ensures the independence between both the conceptual and the functional model – researchers may use different LIMS but still name things with a consistent vocabulary. In the same vain this may allow to share experiments in the form of customizable “templates”. This is at the core of the novel publication workflow presented in this thesis. The data capturing templates are highly customisable and largely dependent on publicly available domain-specific vocabularies. In this sense the results presented in this thesis should be further evaluated within the context of a LIMS. Accurate annotation of the experimental record, thus preserving the digital continuum, also includes intermediary results and research outcomes of all sorts. The semantics upon which annotation relies should make it possible to move in the direction of better documentation facilitating experimental reproducibility. This sort of large-scale data integration can be achieved by the use of a data integration engine based on graph theory. The result will be a better understanding of the meaning of the results of a wide variety of experiments and the increased ability to develop further hypothesis and experiment validators in silico.

8.2.2 Concluding remarks

This thesis has made it possible to have a better understanding about reusable data; not just available data but more importantly data in the context within which it was produced. There are several issues that require more work. The automatic annotation when producing the semantic document depends on the knowledge encoded in the ontologies. There is the need for better and more exhaustive ontologies describing samples, instruments and reagents. Reagents and instruments should come directly from suppliers; in order for this information to be easily reusable manufacturers and vendors should agree on how to describe these products and how to publish this information so that it is easy for web agents to make extensive reuse of it. Lists of samples, e.g., the NCBI taxon, are in any case incomplete in terms of synonyms, acronyms and common English names. Perhaps it is more important to agree on how to describe the sample, e.g., common name, acronym, also known as,

etc., and make it possible for communities to maintain these resources, a la wiki. Thus empowering the community.

Representing experimental actions require more accurate ontologies. Currently, available terminological resources make it difficult to reduce the ambiguity of the language. In order to support the automatic analysis of lab data and the processes by means of which these data was produced, software agents need more comprehensive ontologies. The Ontology for Biomedical Investigations (OBI) is moving in the direction of representing biomedical experiments; it is a step in the right direction. However, the lack of direct relation between biomedical ontologies and LIMS generate a disconnection that is necessary to bridge. Biomedical ontologies need to focus on end user needs (in this case the end user being the bench biologist); by the same token, manufacturers and suppliers of reagents, LIMS and equipment should have a clearer role in steering the development of biomedical ontologies so that these can easily be used in their products. This is in the best interest of open science; laboratory equipment and information systems should make use of open standards.

The publication paradigm presented in this thesis could also be applicable to other types of documents. Understanding and supporting the aggregation of assertions should be more important than having extensive narratives telling incomplete stories primarily for humans. This approach forces post mortem modifications over the content, e.g., semantic annotations. This is due, in part, to the legacy of the Gutenberg printer paradigm; the content is not meant to be processed by machines. This should change; the need for reusable data, as well as the need for better experimental reproducibility will move us in the direction of machine processable data. Furthermore, up to now publication workflows have been generic. Implementations have been built to support the publication of archaeology data just as they support the publication of nanosciences. New publication workflows are slowly focusing on specific types of content. For instance, the Journal of Open Source Software (JOSS) uses git technology in order to support the publication of scientific software only. This makes it possible to better tailor the technology for specifics within that content. The emergence of Jupiter notebooks also illustrates this trend.

Chapter 9

Future Work

The research presented in this thesis has raised some important questions. There are several lines of research arising from this work which should be pursued.

Firstly, the automatic identification of experimental actions, as well as the automatic reconstruction of experimental workflows from existing publications remains a significant challenge. This thesis took a first step in organizing the vocabularies and using a rule based system in order to recognize experimental actions, instruments, samples and reagent but the need for improved NLP methods in combinations with better ontologies remains valid.

A second line of research which follows from chapters 2, 3 and 4 is to investigate more agile methods for end user engagement in the process of maintaining terminological resources. It is also important to understand that in many instances we don't need overly axiomatized ontologies but comprehensive terminological resources.

The SIRO model presented in this thesis has captures most of the information in a structured manner. However, structuring the objective remains a challenge; an important one because the objective is a very specific data item that should be processable by machines. The identification and semantic characterization of the objective of a protocol is a third line of research identified by the author. This thesis addressed the semantics for experimental protocols; the understanding of experimental protocols through this research is basic and limited to the classical protocols used by workbench biologists. There is the need for more research on data coming from sensors; the workflow nature of the protocol remains the same but, the nature of sensor data and their inner protocols may have an impact on the SP representation. By the same token, there is the need to have one single representation for workflows executed by robots and those meant for humans. This thesis made some progress in this area, it specifies the protocol as a machine readable object; but, having one single unified representation for humans and workflows is needed. Robots and sensors are becoming common across laboratories, is one single representation for the protocols reasonable? How to manage these workflows? This is a fourth line of research arising from this thesis.

The representation of the workflow through this thesis needs to be better thought for more complex workflows -fifth line of research. There is also the need to address the problem of delivering workflow constructs to the end user so that these can be easily configured by domain experts. Workflows bring together reagents, samples and instruments; these are all items commonly managed by inventories in laboratories. How could these, in principle independent information systems interoperate so that the end user deals with one single experience? Future work in blockchain technology for managing research assets is something we have already started to address, with significant achievements. We are currently working with blockchain4openscience.com in generating the kind of infrastructure that should allow researchers to have assets in a distributed ledger.

Appendix A

User guide for the SMART Protocols Annotation Tool

User guide of The SMART Protocols Annotation Tool

This video is an introduction to the basic functionality of our annotation tool.

1

Go to:

<http://labs.linkingdata.io:9000/dist/dev/#/signup>

Use: **Safari or Google Chrome**

2

Type the **email** and **password** as **they were** assigned; then **log in**.

Welcome to the Smart
Protocols Annotation
Tool

Email

Password

Log in Register

3

Select a document...

Clicking on the title

Choose files

Welcome

Defining transcribed regions using RNA-seq

Brian T Wilhelm, Samuel Marguerat, Ian Goodhead, Jürg Bähler, 2010, 'Defining transcribed regions using RNA-seq', *Nature Protocols*, vol. 5, no. 2, pp. 255-266

Copy bibtex to clipboard!

Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing

Andreas Mayer, L Stirling Churchman, 2016, 'Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing', *Nature Protocols*, vol. 11, no. 4, pp. 813-833

Copy bibtex to clipboard!

Bioportal tag cloud

RNA | 155

DNA | 78

PROTOCOL | 66

GENOME | 55

AGE | 51

DATA | 48

BASE | 39

CDNA | 39

GENE | 35

ADD | 33

Bioportal tag cloud

RNA | 347

CEL | 154

CELL | 153

DNA | 152

MIX | 112

HUMAN | 82

SOLUTION | 78

MIC | 71

FRACTION | 67

ADD | 66

4

Read the document from start to end, making no annotations to get an understanding of the processes involved in the protocol.

5

Read the document a second time and annotate!

The process is simple...

- a. Detect the word or phrase of interest,
- b. Use the highlighting tool and highlight.

a. Select the text by using the mouse, then...

b. Click on the annotate button to highlight the text.

6

Once the text is highlighted in yellow, a tab is displayed on the right side of the screen.

Please, Sign in again...

a. Click on the **sign in** button

Page: 1 of 12 Automatic Zoom Public Sign up / Sign in

To annotate this document create a free account or sign in

Username or email address:

Password:

Forgot your password?

Sign In

You must be signed in to create annotations.

Next-generation sequencing technologies are revolutionizing genomics research. It is not possible to obtain a complete DNA sequence within a week without time-consuming cloning or massive infrastructure applied to the development of 'RNA-seq' techniques for sequencing cDNA from various or entire transcriptomes. These methods provide unprecedented resolution and depth of data of gene expression, discovery of novel transcripts and exons, and measurement of splicing **protocol for nonstrand-specific transcriptome sequencing via RNA-seq, describing the library bioinformatic analysis procedure.** While sample preparation and sequencing take a fairly short time, the downstream analysis is by far the most challenging and time-consuming aspect and can take the experimental objectives.

INTRODUCTION
Surveying the first resolution with Recent data from portions of genomic

b. Introduce the **email** and **password** assigned; then **sign in**.

To add a tag hit space in the tab “Hit space to add a tag”

Then select from a drop-down list one of the following 4 options:

- Sample/Organism
- Instrument
- Reagent
- Objective

Page: 1 of 17
Automatic Zoom
Public

Published online 18 June 2015; doi:10.1038/nprot.2015.064

Structure-seq is a high-throughput and quantitative method that provides genome-wide single-nucleotide resolution. Structure-seq can be performed both *in vivo* and *in vitro* to study relationships, RNA regulation of gene expression and RNA processing. Structure-seq can be carried out by a molecular biologist with a basic understanding of bioinformatics. Structure-seq begins with a single-hit kinetics conditions. Certain chemical modifications, e.g., methylation of the Watson and cytosine residues by dimethyl sulfate, result in a stop in reverse transcription. Modified RNA is then reverse transcribed using random hexamer primers, which minimizes 3' end bias; reverse transcriptase is then chemically modified residue. Resultant cDNAs are amplified by adapter-based PCR and subjected to high-throughput sequencing, subsequently allowing retrieval of the structural information on a genome-wide scale. In contrast to information only on individual transcripts, a single structure-seq experiment provides information on RNA structures in ~1 month. Although the procedure described here is for *Arabidopsis thaliana*, the procedure is widely applicable, thereby opening new avenues to explore RNA structure–function relationships.

oxgiraldo
Only me

Arabidopsis thaliana

B I " " " "

Hit space to add a tag

Save Cancel

a. Hit space to add a tag; then...

America, Inc. All rights reserved.

Page: 1 of 17
Automatic Zoom
Public

Published online 18 June 2015; doi:10.1038/nprot.2015.064

Structure-seq is a high-throughput and quantitative method that provides genome-wide single-nucleotide resolution. Structure-seq can be performed both *in vivo* and *in vitro* to study relationships, RNA regulation of gene expression and RNA processing. Structure-seq can be carried out by a molecular biologist with a basic understanding of bioinformatics. Structure-seq begins with a single-hit kinetics conditions. Certain chemical modifications, e.g., methylation of the Watson and cytosine residues by dimethyl sulfate, result in a stop in reverse transcription. Modified RNA is then reverse transcribed using random hexamer primers, which minimizes 3' end bias; reverse transcriptase is then chemically modified residue. Resultant cDNAs are amplified by adapter-based PCR and subjected to high-throughput sequencing, subsequently allowing retrieval of the structural information on a genome-wide scale. In contrast to information only on individual transcripts, a single structure-seq experiment provides information on RNA structures in ~1 month. Although the procedure described here is for *Arabidopsis thaliana*, the procedure is widely applicable, thereby opening new avenues to explore RNA structure–function relationships.

oxgiraldo
Only me

Arabidopsis thaliana

B I " " " "

Sample/Organism
Instrument
Reagent
Objective

b. Choose one of the 4 options

INTRODUCTION

RNA structure has a vital role in diverse biological processes, including transcription, translation, polyadenylation and splicing. This approach for use with other methods such as 1-cyclohexyl-(2-n

8

Save the tag ...

9

Add a **comment** to say more about an annotation or decision that was hard to make

a. Select the text by using the mouse, then...

Poly(A) selection ● **TIMING** ~2–3 h

▲ **CRITICAL** Enrich mRNA with a kit that reliably provides high quality mRNA. We use the poly(A) Purist kit.

19| Use 2 ml of extracted total RNA (1.2 mg total) from the poly(A) Purist kit, according to the instructions in the manual.

20| Add 1/10 the volume of 2 M sodium chloride, 1 µl of GlycoBlue and 2.5 volumes of 95% ethanol to the mixture for at least 1 h on powdered dry ice or overnight at –80 °C.

■ **PAUSE POINT** The poly(A) RNA can be stored in a –80 °C freezer overnight.

b. Click on the **annotate** button to highlight the text.

total RNA (1.2 mg total) from the 600 µg/ml RNA stock from Step 18 according to the instructions in the manual.

2 M sodium chloride, 1 µl of GlycoBlue and 2.5 volumes of 95% ethanol to the mixture for at least 1 h on powdered dry ice or overnight at –80 °C to precipitate the RNA. The poly(A) RNA can be stored in a –80 °C freezer overnight.

RNA for 20 min at 16,000g at 4 °C. Discard the supernatant and rinse the pellet. Air-dry or briefly vacuum-concentrate the pellet and dissolve it in 1 µl of water by a NanoDrop spectrophotometer. The yield of the enriched mRNA is ~1.2 mg.

c. add a tag; as was indicated in step 7, then...

oxgiraldo
Only me

2 M sodium chloride

B I Preview

i'm not sure if this reagent is correctly annotated

Reagent ✕ Hit space to add a tag

Save ▼ Cancel

d. add a comment, and...

e. save

Appendix B

Guidelines to annotate experimental protocols using the SIRO model

Guidelines to annotate experimental protocols

SCOPE

We are manually annotating experimental protocols in life sciences. We want to identify words or phrases that can be related to: i) the **Sample(s)** tested in a protocol, ii) **Instruments** used, iii) **Reagents** employed, and the overall iv) **Objective** of a protocol –**SIRO** elements. Before reading this document please look at the slides illustrating how to use the tool.

These **four** elements are common across protocols in life sciences. The manual identification of these elements will help us to: i) enrich our controlled vocabularies and, ii) facilitate information retrieval.

WHAT SHOULD BE ANNOTATED?

Words or phrases related to:

- Sample(s), specimen(s) or organism(s) to be tested.
- Instruments (including software) and consumables.
- Reagents, chemical compounds, solutions or mixtures used.
- Objective or purpose of the protocol.

SOME EXAMPLES...

The **sample** tested in a protocol may be an organism or a part of it. Some examples include:

SAMPLE	Whole organism	Scientific name: <i>Arabidopsis thaliana</i> , <i>Oriza sativa</i> , <i>mangifera indica</i> , <i>Mus musculus</i> . Common name: Mouseear Cress, rice, mango, mouse.
	Anatomical part	leaf, stem, cells, tissues, membranes, organs, skeletal system, muscular system, nervous system, reproductive system, cardiovascular system, etc.
	Biomolecules	<u>Nucleic acids</u> : Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). <u>Proteins</u> : enzymes, structural or support proteins (keratin, elastin, collagen), antibodies, hormones, etc.
	Body fluids	Blood serum, saliva, semen, amniotic fluid, cerebrospinal fluid, gastric acid, etc.

The **instruments** used in protocols include high-throughput equipment, software and consumables. Some examples:

INSTRUMENTS	High-throughput equipment	Liquid Handling Platforms, Real-Time PCR Detection System, Microplate Reader, etc.
	Instruments	Goggles, Bunsen burner, spot plate, pipet, forceps, test tube rack, mortar and pestle, etc.
	Laboratory glassware	Beaker, Erlenmeyer flask, graduated cylinder,

		volumetric flask, etc.
	Standard equipment	Balances, shakers, centrifuges, refrigerators, incubators, thermocyclers, fume hood, etc.
	Consumables	Weighing dishes, pipette tips, gloves, syringes, petri dishes, test tubes, micro centrifuge tubes, glass slides, filter paper, etc.

The **reagents** used in protocols include buffers, solutions, culture media and kits. Some examples:

REAGENTS	Chemical compound/ Substance	Glucose, ethanol, glycerol, chloroform, acetic acid, isopropyl alcohol, etc.
	Solutions / buffers	70% ethanol, 10X PCR buffer, phenol:chloroform:isoamyl alcohol, etc.
	Kits	Nucleic acid purification kits, virus purification kits, PCR screening kits, etc.
	Cell culture media	Nutrient media, minimal media, selective media, differential media, etc.

The **objective** of a protocol is a formal statement describing the goal; what we do want to achieve by executing the protocol? An example:

Part of the speech	Source
"Here we present a detailed protocol for Smart-seq2 that allows the generation of full-length cDNA and sequencing libraries by using standard reagents"	doi:10.1038/nprot.2014.006

ANNOTATION PROCESS

The annotators should carry out the following steps in this specific order:

1. **Read the whole document.** Read the document from start to end, making no annotations to get an understanding of the processes described in the protocol.
2. **Mark the entities.** Read the document a second time and annotate it. The process is simple:
 - a. detect the word or phrase of interest,
 - b. highlight by selecting the text of interest with the mouse, then
 - c. use one of the labels (tags) that you will see on the right hand side of the screen. The tags are sample, reagent, instrument, objective

This annotation task is just about attaching a label (sample, instrument, reagent or objective) to a word or phrase. Examples:

- “Yeast” → sample
- “70% Ethanol” → reagent
- “PCR apparatus” → instrument
- “We present a protocol for isolating DNA from ancient bones” → objective

d. **Mark each occurrence of an entity.** For example if the word “ethanol” appears in different parts of the document, make sure to mark each occurrence. Please, bear in mind that sometimes samples, equipment and reagents are not listed in the materials sections.

e. **Reduce the noise in the annotation.** Avoid excess. Don’t mark an entire paragraph that includes more than one entity associated to the same label (see row 3 in the table below). Also, avoid making incomplete annotations (see the first row of the table).

✓ Good annotations	✗ Bad annotations
9 Add glycerol to 15% or DMSO to 7% (vol/vol),	9 Add glycerol to 15% or DMSO to 7% (vol/vol)
This protocol was developed for single cells but works equally well on purified total RNA of a few picograms or more. A flow-	This protocol was developed for single cells but works equally well on purified total RNA of a few picograms or more. A flow-

2.1 Annotating

2.1.1 Sample/Specimen/Organism

In the Nature Protocol (doi:10.1038/nprot.2007.427), the authors use several names for the same sample. Some of them are:

1. “collection of gene-deletion mutants in *Saccharomyces cerevisiae*”
2. “yeast deletion mutants”

Note: the annotators should annotate the different names given to the sample tested.

Let’s focus on “collection of gene-deletion mutants in *Saccharomyces cerevisiae*”, this could be annotated as follows and both are correct:

- “collection of gene-deletion mutants in *Saccharomyces cerevisiae*” OR
- “collection of gene-deletion mutants in *Saccharomyces cerevisiae*”

Please, annotate other potential samples that could be used. For instance, in doi:10.1038/nprot.2007.427:

“...the general method of studying pooled samples with barcode arrays can also be adapted for use with other types of samples, such as mutant collections in other organisms, short interfering RNA vectors and molecular inversion probes.”

In this text there are 3 samples that could be tested:

1. *mutant collections in other organisms,*
2. *short interfering RNA vectors,*
3. *molecular inversion probes.*

Therefore, you should add three annotations, each one tagged as Sample/Organism, as follows:

“...the general method of studying pooled samples with barcode arrays can also be adapted for use with other types of samples, such as mutant collections in other organisms, short interfering RNA vectors and molecular inversion probes.”

A **bad annotation practice**, is highlighting a part of speech that includes the 3 samples and add a unique tag, for example:

*“...the general method of studying pooled samples with barcode arrays can also be adapted for use with other types of samples, such as mutant collections in other organisms, short interfering RNA vectors and molecular inversion probes.” **

* Note: in our database, each annotation is an entry. If the text is wrongly annotated, then the database of annotations will consider that the protocol “x” has only one sample/Organism entry.

2.1.2 Instruments

In protocols (such as doi:10.1038/nprot.2007.427), it is sometimes mentioned the capacity of the containers used. Some examples are:

- 96-well microliter plates
- 250 ml flasks
- 0.5 ml microfuge tubes

When this information is available, the annotation should include the storage capacity of the containers used and described in the protocol.

Please, avoid highlighting the entire list of instruments/equipment for adding a unique tag; the database of annotations will consider that the protocol “x” has only one Instrument entry.

Examples of good a bad practices annotating instruments:

Good annotation of instruments	Bad annotation of instruments
<p>(❖)</p> <ul style="list-style-type: none"> 48-well plates (Greiner, part no. 677102) 250 ml culture flasks 0.5 ml microfuge tubes suitable for boiling (Eppendorf 0.5-ml Safe-Lock microcentrifuge tubes; Sigma, cat. no. T8911) 	<ul style="list-style-type: none"> 48-well plates (Greiner, part no. 677102) 250 ml culture flasks 0.5 ml microfuge tubes suitable for boiling (Eppendorf 0.5-ml Safe-Lock microcentrifuge tubes; Sigma, cat. no. T8911)
<p>Equipment (❖)</p> <ul style="list-style-type: none"> Aerodisc 0.2 mm filters (Pall Life Sciences, cat. no. 4192) Syringes (Becton Dickinson, cat. no. 309653) Nunc Omni trays (VWR, cat. no. 62409-600) 96-well pin tool (V&P Scientific, cat. no. VP407A) 	<p>Equipment (⌘)</p> <ul style="list-style-type: none"> Aerodisc 0.2 mm filters (Pall Life Sciences, cat. no. 4192) Syringes (Becton Dickinson, cat. no. 309653) Nunc Omni trays (VWR, cat. no. 62409-600) 96-well pin tool (V&P Scientific, cat. no. VP407A)

❖ each line or bullet point is annotated as an instrument.

⌘ the full list was highlighted and annotated as an instrument

2.1.3 Reagents

The protocols should include a list of the chemical and biochemical compounds and solutions (e.g. buffers, Cell culture media) used in a protocol. Sometimes, some chemical compounds are used as a solute in a solution. Please annotate the different concentration grades of the reagents (e.g. G418, 1,000X G418 stock (200 mg ml⁻¹), 1,000X G418 stock –these are highlighted in blue in the example below). Illustrating how to annotate see the following example from doi:10.1038/nprot.2007.427. Other reagents, also listed in the example, are highlighted in green

“REAGENTS

.

G418 (Agri-Bio, cat. no. 3000)

.

REAGENT SETUP

1,000X G418 stock (200 mg ml⁻¹) Dissolve 5 g of G418 in 25 ml of dH₂O. Filter-sterilize using a 0.2 mm filter and a syringe. Shield from light by wrapping bottle in foil. Store at 4 °C.

YPD + 200 µg ml⁻¹ G418 rectangular plates Mix 10 g of yeast extract, 20 g of peptone, 20 g of dextrose, 20 g of agar and 1 liter of dH₂O to a 2-liter flask with a stir bar. Autoclave. Allow media to cool to approximately 50 °C with gentle stirring. Add 1 ml of 1,000X G418 stock. Stir gently for an additional 1 min to ensure that the drug is evenly mixed. Pour into Nunc Omni trays, 50 ml per tray; sufficient for approximately 20 plates. Store at 4 °C.

YPD liquid + 200 lg ml⁻¹ G418 Mix 10 g of yeast extract, 20 g of peptone, 20 g of dextrose and 1 liter of dH₂O to a 1-liter bottle. Autoclave. Allow media to cool to approximately 50 °C. Add 1 ml of 1,000X G418 stock. Store at 4 °C.

Up primer mix Dissolve uptag and Buptagkanmx4 each in dH₂O at 100 pmol µl⁻¹ and mix in a 1:1 ratio. Note that the uptag oligo is also used in mixed oligonucleotides, so take care to leave enough for use in both mixes. Store at 20 °C.”

If you want to include the quantity used for reagent, it is also fine. For instance:

- 20 g of peptone **instead of** 20 g of peptone
- 1 liter of dH₂O **instead of** 1 liter of dH₂O

2.1.4 Objective

The objective or goal of the protocol in most cases is described in the abstract section. Please, highlight only the part of the text that describes the objective of the protocol. For instance,

“... from a single pooled culture. Here, we present protocols for the study of pooled cultures of tagged yeast deletion mutants with a tag microarray. This process involves...”

3. **Add a comment indicating an annotation or decision that was hard to make.** Comment in the annotation tool; the tool has a comment field above the labels. Please use it; this will be useful for us in order to understand and respond to those difficulties.

4. **Time to solve doubts.** Planning personalized sessions (via skype, slack, email, etc.) to solve doubts is simply a matter of contacting me (oxgiraldo@gmail.com). These sessions help us to know, for example:

- any questions or uncertainty that you may have about the annotations and guidelines,
- anything unclear or ambiguous in the guideline,
- things that you consider important and that were not covered by the labels for the annotations in this task.
- Bugs in the annotation tool and/or issues.

5. **Notify the finalization of the annotation task.** Each annotator should notify via email the finalization of his/her annotation task.