Nanotate: Semantically annotating experimental protocols with nanopublications

Olga Giraldo[1][0000−0003−2978−8922], Miguel Ruano[2][0000−0002−7241−7089], Robin A. Richardson[3][0000−0002−9984−2720], Remzi Celebi[4][0000−0001−7769−4272], Michel Dumontier[4][0000−0003−4727−9435], and Tobias Kuhn[1][0000−0002−1267−0234]

[1] Department of Computer Science, VU Amsterdam, Amsterdam, Netherlands
[2] Universidad del Valle, Cali, Colombia
[3] Netherlands eScience Center, Amsterdam, Netherlands
[4] Maastricht University, Maastricht, Netherlands

Abstract. An experimental protocol describes a sequence of tasks executed to perform experimental research in biological and biomedical areas, e.g. genetics, immunology, neuroscience and virology. Such experimental protocols indicate, for each step, exactly how it should be executed, often including equipment, reagents, descriptions of critical steps, troubleshooting instructions, other kinds of tips, as well as any other information that researchers deem important for facilitating the reusability of the protocol. These protocols therefore have a clear systematic structure, but when published they are treated like any other scientific publication i.e. as a narrative text in HTML or PDF format. The formal structure is therefore not easily accessible and cannot be reused. This paper addresses this problem by extracting, representing and publishing steps from experimental protocols to make them Findable, Accessible, Interoperable, and Reusable (FAIR). Our work builds upon human annotations in combination with Named Entity Recognition delivering nanopublications. Our software toolkit, `Nanotate`, is based on a flexible web based annotation environment, namely Hypothes.is, the BioPortal NER web services and the nanopublications infrastructure. Our evaluation shows that our approach is viable and our tool user-friendly.

1 Introduction

Experimental protocols are documents providing detailed descriptions of the processes by means of which results, often data, are generated in biological sciences. For reproducibility purposes, both data and protocols describing the steps followed to obtain the data, should be available. The protocols often include equipment, reagents, critical steps, troubleshooting guidelines, tips and other information that facilitates reusability. Experimental protocols are described in natural language lacking of a formal structure; it is not surprising that important details are sometimes missing, e.g. time or temperature to centrifuge a sample, precise storage conditions of a suspension, or specific features of an equipment or reagent used.

To illustrate the importance and nature of experimental protocols, let us consider "*Identification of QTL Associated with Drought Tolerance in Common Bean*"; it involves the execution of a number of protocols, such as *sample preparation*, *DNA isolation* and *amplification of the DNA via PCR*. The study as a whole can be seen as a workflow that includes several steps, each of them a protocol on its own. Each protocol consists of a sequence of structured instructions. Each of the protocols and also each of their steps have inputs and outputs. In this sense, steps within each protocol could be seen as the smallest most granular part of the workflow. Following this analogy, understanding the study as a container of protocols, both steps within protocols as well as protocols on their own should be reproducible.

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) [21] compliant document [8]. In addition, when protocols are not properly documented or preserved, researchers can no longer interpret, communicate, or share information effectively. To address this problem, in this paper we focus on representing and publishing steps from experimental protocols as nanopublications [10], in order to make them FAIR and machine-consumable. Specifically, we provide a semantic annotation layer in order to improve search, organization, cataloging and maintenance of protocols.

Our approach makes use of domain expert annotations in protocols to extract information about steps, samples, instruments and reagents participating in each step. In this work, we investigate how we can facilitate the participation of the domain experts when extracting steps from experimental protocols while hiding the complexity of representing and publishing such artifacts. We developed `Nanotate`. `Nanotate` combines Named Entity Recognition and Human Based Annotation in a single annotation framework. The results suggest that our approach is practical, the semantic annotations rich, consistent between annotators and meaningful; and the interface is perceived to be usable and user-friendly.

2 Background

Open sharing of methods is an essential element to ensure reproducible research in life science, e.g. through repositories like BioProtocols[5], protocol exchange [13], protocols.io [19] and MethodsX[6]. Some existing approaches are focused on the semantic representation of protocols in order to provide all the information required for the replication of biomedical and biological protocols, e.g. the EXACT2 ontology [16] aimed at semantic extraction of knowledge of biomedical protocols. The SMART Protocols ontology [9] represents protocols in biological sciences as a workflow embedded within a document. Bioschemas also provides

[5] `https://bio-protocol.org/Default.aspx`
[6] `http://www.sciencedirect.com/science/journal/22150161`

a simple way to add structured data to web pages, such as the LabProtocol profile[7], which models the details of publications about experimental protocols.

Provenance is an important aspect in experiments, and Reproduce Microscopy Experiments (REPRODUCE-ME) [15] is an ontology for the semantic documentation of provenance in microscopy experiments. To make scientific workflows open and FAIR, a semantic model to publish scientific workflows as FAIR data has been proposed [6]. Several tools for the manual annotation of biomedical documents have been developed in order to facilitate search and retrieval of semantic content in this kind of documents. Some examples are Bionotate [5], Semantator [17] and Brat [18], and they are based on a strategy to search terms in thesauruses or ontologies by finding occurrences of a concept chain in a text fragment using coincidence of terms.

3 Nanotate Approach and Implementation

To address the problem of extracting, representing and publishing steps from experimental protocols we developed `Nanotate`, a web based tool that facilitates the publication of annotated steps as nanopublications. The goal of our approach is to allow end-users to directly publish nanopublications about annotated protocol steps. Our approach is highly scalable; `Nanotate` extends an existing annotation framework, `Hypothes.is`, which is widely used, and has an extensive community of users, thus supporting developers and end users. More importantly, `Hypothes.is` is highly adaptable; further adaptations of the interface are possible while reusing the backend and authentication mechanisms.

`Nanotate` prioritizes Human based annotation in specific parts of the scientific discourse; in this case the identification of steps in experimental protocols and the material entities participating in each one of them (samples, equipment, reagents). `Nanotate` incorporates capability for the automatic recognition of samples, equipment and reagents with classes from 8 ontologies (OBI [2], SP [9], BAO [1], EFO [12], CHEBI [11], UBERON, NCBI TAXON [7] and ERO [20]), by way of the BioPortal API.

3.1 Nanotate architecture

The Nanotate architecture is shown in Figure 1. The workflow architecture starts on the protocols website, called here "annotated site". We replaced the hypothes.is annotator and sidebar with our own annotation user interface. We use the BioPortal API to consult the ontologies with terminology related to the samples, equipment and reagents to be annotated. To produce nanopublications about individual annotated steps we use the Python libraries nanopub[8] and fairworkflows[9], which allow for searching and publishing nanopublications and support the construction of FAIR scientific workflows using nanopublications[14]. Below, we further describe the `Nanotate` architecture components. In order to

[7] `https://bioschemas.org/profiles/LabProtocol/0.6-DRAFT-2020_12_08/`
[8] `https://github.com/fair-workflows/nanopub`
[9] `https://github.com/fair-workflows/fairworkflows`

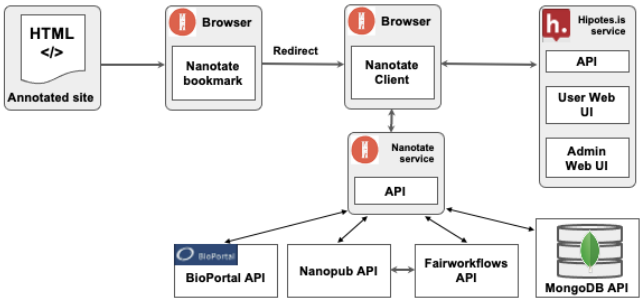

Fig. 1. Nanotate architecture

explain each stage of the process, a running example is available online[10].

Generation of annotations (from HTML to JSON). In this process participate the following components: i) Annotated site, ii) `Nanotate` bookmark, iii) `Nanotate` client and iv) BioPortal API. i) Annotated site: We start on the web page of the protocol to be annotated. `Nanotate` takes as input full experimental protocol documents in HTML. ii) `Nanotate` bookmark: `Nanotate` does not use the hypothes.is client ('annotator' and 'sidebar'); instead, it uses the bookmarklet approach in a similar way to The HelloWorldAnnotated demo[11]. The "`Nanotate bookmark`" is used to redirect from an "`Annotated Site`" to the "`Nanotate client`" through a URL that contains data about the selected text to be annotated. iii) `Nanotate` client: Our UI works as a template guiding users through the annotation process and posting the annotations to hypothes.is. The UI includes a tag box to add one or multiple tags from the available options ("*sample*", "*equipment*", "*reagent*", "*input*", "*output*", "*step*"). After labeling the selected text, a JSON file in which hypothes.is stores the annotation is generated. iv) BioPortal API: The next step is adding context to the tagged text by using ontology terms. `Nanotate` facilitates the connection with the BioPortal API for consulting the 8 ontologies mentioned above. When it is possible to link a tagged text to an ontology term, the resulting JSON includes the URI of the ontology term. All annotations about steps and the material entities participating on each one of them are posted to Hypothes.is.

Generation of nanopublications (from JSON to RDF). Here is generated the publication of each annotated step and their components as nanopublications. In this process participate the component Nanopub library. To generate the nanopublications, the users just press the button "`Nanopub`" located in each annotated step. The nanopub library was configured to group annotations according to their text position. These are grouped, if an annotation that does not

[10] `https://git.io/JDFOV`
[11] `https://github.com/judell/HelloWorldAnnotated`

have the tag "*step*" is contained in an annotation that does have a tag "*step*". All annotations are sent to `Nanotate` via API where validation is done and subsequent publication. In addition, the published nanopublications are stored locally in a MongoDB data storage.

Generation of RDF workflows. Once having all the nanopublications for the individual steps, the next step is creating the corresponding workflow. Here, `Nanotate` client, the users consult the nanopublications and press the button "`new workflow`". Then, users register the fields: i) label: to add a name to the workflow to be created and ii) description: short description about the new workflow. Finally, the users select the nanopublications that are part of the workflow. The resulting nanopublications are then published as above. The published nanopublications about workflows are also stored locally in a MongoDB data storage.

`Nanotate` is a free and open source tool. The code behind the tool is available on Github `https://github.com/nanotate-tool`. End-users can install it by creating a bookmark. The documentation about how to install and use the `Nanotate` is available at `http://doi.org/10.5281/zenodo.5101941` A running instance of the tool can be found at `https://nanotate.bitsfetch.com/`

4 Evaluation Design

To assess our approach we carried out a controlled annotation study with domain experts. We evaluated the nature and consistency of the annotations and, the subjective usability of the tool. The methodology is presented below.

Materials. We worked with i) six open access protocols in molecular biology from Bio-protocols and Nature Protocol Exchange; See Table 1. ii) human annotators: Three annotators with experience in laboratory techniques. iii) the annotation tool, `Nanotate`: a web-based tool used in this study. During the training sessions the participants learnt how to use the tool. iv) training documentation: We provided the participants with a detailed training document, how to install and use the tool. It gives examples of samples, equipment, and reagents and how this information should be annotated. And v) a questionnaire to evaluate a subjective usability of the `Nanotate` tool: A usability questionnaire with ten standard questions from the System Usability Scale (SUS) [4]. Experts in life sciences rated the tool following the standard ten questionnaire items on a five-point scale ranging from "strongly agree" to "strongly disagree". The questionnaire can be found online[12].

Methods. Our controlled annotation has a series of activities organized in the following stages: i) training session, ii) assignment of protocols to annotators,

[12] `https://forms.gle/o6JYQ7xY7wVFsqqG8`

Table 1. Set of annotated protocols

#	Protocol ID	Source	Steps
1	DOI:10.21769/BioProtoc.323	Bio-protocols	7
2	DOI:10.21203/rs.2.1347/v2	Nature Protocol Exchange	12
3	DOI:10.1038/protex.2013.007	Nature Protocol Exchange	11
4	DOI:10.21203/rs.2.1645/v2	Nature Protocol Exchange	8
5	DOI:10.21769/BioProtoc.1751	Bio-protocols	16
6	DOI:10.21769/BioProtoc.1077	Bio-protocols	11

iii) review of annotations and iv) generating data for the analysis. In the first stage, a virtual session was organised with each annotator in order to train them in the use of the `Nanotate` tool and give them some tips about good annotation practices. The meetings were carried out by using Google Hangouts. In the second stage, the six protocols presented in table 1 were annotated by three human annotators. Then, in the third stage, virtual meetings were scheduled with annotators in order to solve doubts and inconsistencies. Specifically, when we had to deal with nanopublications validation problems –A subset of protocol steps were annotated but the corresponding nanopublication was generated incorrectly, probably due to problems in the HTML of those protocols. Finally, the data in the form of the semantic annotations structured as nanopublication as well as the answers to the usability questionnaire were analyzed. We focused on tag distribution, ontology coverage, completeness analysis, and inter-annotator agreement.

5 Results

5.1 Tag distribution

A total of 232 part-of-speech entities were tagged with one of the six available categories (sample, equipment, reagent, input, output, step). The results of this stage are summarized in Figure 2 (the full data is available online[13]). In this study, 6.4% (17) of the samples were also tagged as "input"; and 5.7% (15) of the samples were also tagged as " output" (see Figure 2). 22 parts-of-speech were just tagged as "sample". These results indicate that the categories "input" and " output" were the least used to tag parts-of-speech. The results indicate that the 65 steps could be identified and annotated (see Figure 2). Finally, 19.7% (52) parts-of-speech were tagged as "equipment" and 21.6% (57) parts-of-speech were tagged as "reagent" (see left-hand side of Figure 2).

5.2 Annotations mapped to ontology terms

Parts-of-speech tagged as "*sample*", "*equipment*" and "*reagent*" were mapped to ontology terms that come from the 8 aforementioned ontologies available in

[13] `http://doi.org/10.5281/zenodo.5089323`

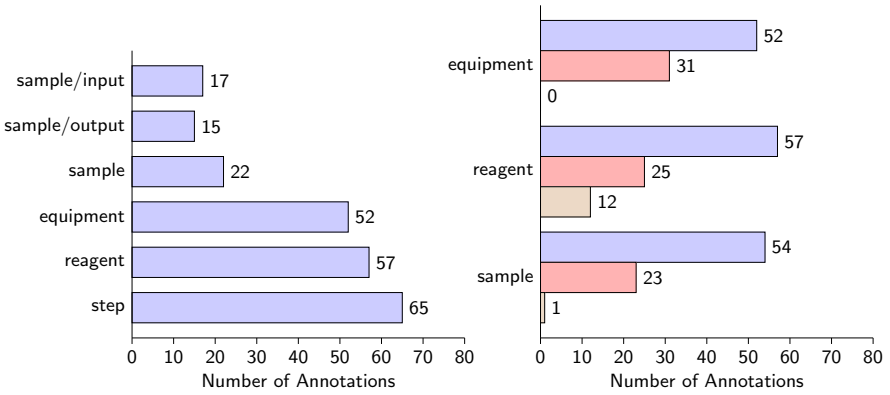

Fig. 2. Tag distribution (left). Comparison to ontology matching (right), total number of annotations (blue), those linked to ontology terms (red) and those with an ontology class available (yellow)

BioPortal. The results of this stage are summarized in Figure 2 (right; the full data is available online[14]). 59.6 % (31) of the text tagged as "*equipment*" were mapped to ontology terms from OBI (8), BAO (20), SP (2) and ERO (1). 43.9% (25) of the text tagged as "*reagent*" could be mapped to ontology terms from CHEBI (17), OBI (3), SP (4) and BAO (1). 39.7% (23) of the text tagged as "*sample*" could be mapped to ontology terms from NCBI TAXON (3), UBERON (1), CHEBI (11), BAO (2), SP (5) and EFO (1). We manually analyzed the cases that we not mapped to an ontology term to find out about the reasons for that. We found the following four main reasons: i) terminology not represented in the available ontologies; for instance reagent manufacturer names (e.g.: TRIsol) and acronyms or short names (e,g,: PBMCs, PBS). ii) name of specific type of reagents (e.g.: extraction buffer, this is a buffer), and samples (e.g.: human venous blood, this is a blood). In this second case, just superclasses are available in the ontologies used (See yellow bar in Figure 2). These ontologies do not reach a high level of granularity. iii) Plural words (like mosquitoes, bacteria), are not represented in our subset of ontologies, and iv) Typos, for instance Micropistle *instead of* Micropestle.

5.3 Completeness checks

As the above analyses focused only on the annotated entities, we should also have a look at completeness. We are looking for indications of the extent to which entities were missed. We found that first steps should always specify an input, while last steps specify an output. In the set of analyzed protocols, we found that 100% of first steps included an annotated input; 83.33%, five of the six last steps, included an annotated output. For equipment annotations,

[14] http://doi.org/10.5281/zenodo.5089726

steps in such protocols almost always involved the use of some equipment. We checked how often these were correctly covered. We found that in 55,38% (36) of the steps, at least one of the equipment was not explicitly mentioned in the text. The domain experts are able to infer the use of some equipment, due to the narrative; for instance, use of experimental actions like "*centrifuge*", and additional information such as temperature of centrifugation (e.g.: 4°C) help to deduce the use of a "*refrigerated centrifuge*" in a particular step (more examples are available in a table online[15]). This table include nanopublication links where inferred equipment are missing in the assertion. We can therefore conclude that the completeness is far from perfect, but the annotations cover a substantial part of the mentioned entities.

5.4 Inter-annotator agreement

Annotators categorized the annotations into the different workflow component classes (i.e., step, reagent, equipment, input, output) with perfect agreement. However, in some cases, the spanning text of these annotations overlapped but did not match perfectly. We observed that the annotators chose the same spanning text to annotate "*step*" and "*sample/input*" (samples that were also tagged as input). Partial matches were identified where one or more annotators highlighted slightly different text. That was the case of the 9.6% of the text tagged as "*equipment*", 15,4% of the text tagged as "*sample*", 17,5% of the text tagged as "*reagent*" and 20% of the text tagged as "*sample/output*" (samples that were also tagged as output). As was presented in subsection 5.2, we found a low incidence in annotations linked to ontology terms. However, the annotators have a high agreement on linking annotations to the ontology terms (Fleiss's Kappa: 0.70). The full data, including the set of annotations used to calculate inter-annotator agreement is available online[16].

5.5 Subjective usability by questionnaire (SUS)

The table https://doi.org/10.5281/zenodo.5528946 summarizes the results of the SUS. The participants were three annotators of the annotation study described above and one additional expert who used the tool but did not participate in the annotation study. Overall, the tool got a SUS score of 93.12%; between "excellent" and "best imaginable" on the adjective scale [3]. From these results it is clear that the tool was well received and that users hardly experienced problems using it.

6 Discussion and conclusions

This work involves manual annotation of protocols in order to extract information about samples, instruments and, reagents participating in each step. The

[15] https://doi.org/10.5281/zenodo.5528934
[16] http://doi.org/10.5281/zenodo.5095720

methodological aspects involved the participation of domain experts, reusing existing resources while focusing on how end users could make that valuable information from experimental protocols readily available in a semantic manner. We developed `Nanotate`, a tool to publish nanopublications from annotated protocol steps. It extends the Hypothesis platform by making it compliant with the nanopublication workflow. It also adds NER capabilities from BioPortal in a single annotation framework.

From this experiment we concluded that the `Nanotate` tool, the guidelines about `what` and `how` to annotate, and the knowledge that comes from experts in laboratory techniques were key in achieving a high consistency in the annotations and the subsequent nanopublications. This is a consequence of a standardized annotation process to publish individual protocol steps and their participants (samples, equipment, reagents). We also found some missing elements in protocols. For example, no inputs or outputs. Some equipment could be inferred by experts but it could not be annotated because there was no explicit mention to it. Such inaccuracies limit the reproducibility and reusability of protocols. It is necessary to improve reporting structures for experimental protocols, this requires collective efforts from authors, peer reviewers, editors and funding bodies [9]. The main limitation in this study was the lack of a large number of annotators and protocols. Our study started during the first year of the covid pandemic and it was difficult to plan the virtual sessions and find annotators. `Nanotate` facilitates the annotation of web protocols in HTML format. As a future work, we want to facilitate the annotation of protocols available in other formats.

Acknowledgments

This work was supported by the Dutch Research Council (NWO)(No. 628.011.011).

References

1. Abeyruwan, S., Vempati, U.D., Küçük-McGinty, H., Visser, U., Koleti, A., et al.: Evolving BioAssay ontology (BAO): modularization, integration and applications. Journal of Biomedical Semantics **5**(Suppl 1 Proceedings of the Bio-Ontologies Spec Interest G), S5 (2014)
2. Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M.H., Bug, B., et al.: The Ontology for Biomedical Investigations. PLOS ONE **11**(4), e0154556 (apr 2016)
3. Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Intl. Journal of Human–Computer Interaction **24**(6) (Jul 2008). https://doi.org/https://doi.org/10.1080/10447310802205776
4. Brooke, J.: Sus: A quick and dirty usability scale. Usability Eval. Ind. **189** (11 1995)
5. C, C., T, M., A, B., P, W.D., L, P.: Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. J Biomed Inform **42**(5), 967–977 (2009). https://doi.org/https://doi.org/10.1016/j.jbi.2009.02.001
6. Celebi, R., Rebelo Moreira, J., Hassan, A.A., Ayyar, S., Ridder, L., et al.: Towards fair protocols and workflows: the openpredict use case. PeerJ. Computer science **6**, e281–e281 (2020). https://doi.org/https://doi.org/10.7717/peerj-cs.281

7. Federhen, S.: Type material in the NCBI Taxonomy Database. Nucleic Acids Res **43**, D1086–98 (2015)
8. Giraldo, O., Garcia, A., Corcho, O.: A guideline for reporting experimental protocols in life sciences. PeerJ **6**(e4795) (May 2018). https://doi.org/https://doi.org/10.7717/peerj.4795
9. Giraldo, O., García, A., López, F., Corcho, O.: Using semantics for representing experimental protocols. Journal of Biomedical Semantics **8**(1), 52 (Nov 2017). https://doi.org/https://doi.org/10.1186/s13326-017-0160-y
10. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication **30**, 51–56 (2010). https://doi.org/https://doi.org/10.3233/ISU-2010-0613
11. Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., et al.: The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic Acids Res **41**, D456–63 (2013)
12. Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., et al.: Modeling sample variables with an Experimental Factor Ontology. Bioinformatics **26**(8), 1112–1118 (2010)
13. Protocols, N.: Introducing the new protocol exchange site. Nat Protoc **14**(1945) (Jun 2019). https://doi.org/https://doi.org/10.1038/s41596-019-0199-6
14. Richardson, R.A., Celebi, R., van der Burg, S., Smits, D., Ridder, L., Dumontier, M., Kuhn, T.: User-friendly composition of fair workflows in a notebook environment. In: Proceedings of the 11th on Knowledge Capture Conference. p. 1–8. K-CAP '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3460210.3493546, https://doi.org/10.1145/3460210.3493546
15. Samuel, S., König-Ries, B.: Reproduce-me: Ontology-based data access for reproducibility of microscopy experiments. In: Blomqvist, E., Hose, K., Paulheim, H., Ławrynowicz, A., Ciravegna, F., Hartig, O. (eds.) The Semantic Web: ESWC 2017 Satellite Events. pp. 17–20. Springer International Publishing, Cham (2017)
16. Soldatova, L.N., Nadis, D., King, R.D., Basu, P.S., Haddi, E., et al.: Exact2: the semantics of biomedical protocols. BMC Bioinformatics **15**(14), S5 (2014). https://doi.org/https://doi.org/10.1186/1471-2105-15-S14-S5
17. Song, D., Chute, C.G., Tao, C.: Semantator: annotating clinical narratives with semantic web ontologies. (article). AMIA Jt Summits Transl Sci Proc **2012**, 20–29 (2012)
18. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107. Association for Computational Linguistics, Avignon, France (Apr 2012), https://aclanthology.org/E12-2021
19. Teytelman, L., Stoliartchouk, A., Kindler, L., Hurwitz, B.L.: Protocols.io: Virtual communities for protocol development and discussion. PLoS biology **22**(14(8)) (Aug 2016). https://doi.org/10.1371/journal.pbio.1002538, https://doi.org/10.1371/journal.pbio.1002538
20. Torniai, C., Brush, M., Vasilevsky, N., Segerdell, E., Wilson, M., et al.: Developing an application ontology for biomedical resource annotation and retrieval: Challenges and lessons learned, vol. 833, pp. 101–108 (2011)
21. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3**(160018) (Mar 2016). https://doi.org/https://doi.org/10.1038/sdata.2016.18