The Story of an Experiment: A Provenance-based Semantic Approach towards Research Reproducibility

Sheeba Samuel[1], Kathrin Groeneveld[2], Frank Taubert[1], Daniel Walther[1], Tom Kache[2], Teresa Langenstück[2], Birgitta König-Ries[1], H. Martin Bücker[1], and Christoph Biskup[2]

[1] Institute for Computer Science, Friedrich Schiller University, Jena, Germany
{firstname.lastname}@uni-jena.de
[2] Biomolecular Photonics Group, Jena University Hospital, Friedrich Schiller University, Jena, Germany
{kathrin.groeneveld, teresa.langenstueck}@med.uni-jena.de,
{Tom.Kache, christoph.biskup}@uni-jena.de

Abstract. End-to-end reproducibility of scientific experiments is a key to the foundation of science. Reproducibility of an experiment does not necessarily guarantee the accuracy of its results, but it guarantees that the steps of an experiment can be repeated to a certain level of significance to generate similar results. Data provenance plays a key role in telling the story of an experiment which helps one step towards reproducibility. To convey the message of a story, it is essential to provide sufficient data and its flow along with its semantics. In this paper, we present a provenance-based semantic approach to explain the story of a scientific experiment with the primary goal of reproducibility. The REPRODUCE-ME ontology extended from PROV-O and P-Plan is used to represent the whole story of an experiment describing the path it took from its design to result. We visualize and evaluate the provenance lifecycle of a scientific experiment taking into account the use case of life science experiments.

Keywords: Provenance, Reproducibility, Experiment, Story, Ontology

1 Introduction

A story generally consists of the following components: plot, characters, background context, settings, events, conflicts, climax, and the final message. It is essential to know the characters, the context, and the flow of the story to understand its climax and message. Similarly, to make the story of a scientific experiment and its results understandable and reproducible, it is necessary to present its agents, execution, environmental attributes and workflow in a way that can be understood by the scientific community. According to [4], an experiment performed at time T with the environment setup E (e.g. settings) using data D (e.g. experiment materials, measurements) consisting of a sequence of steps S is said to be reproducible, if it can be executed at time $T' > T$ with environment setup E' (similar or identical to E) with a sequence of steps S' (modified from or equal to S) using data D' (similar or identical to D) with similar results. There are various challenges that hinder reproducibility of experiments which include integration of data generated from different devices, incomplete and uncertain provenance information, lack of documentation in digital media, lack of knowledge of the type of data and their formats and most importantly their semantics.

In life-science experiments, the preparation of experimental materials is very important. Scientists use the methods described in publications to prepare the specimens. But sometimes, the scientists fail to replicate the methods mentioned in the publications due to incorrect or incomplete data because of accidental omission or errors. Critical steps may not be included or fully described or the order of execution may be missing from the method description. Inconclusive results are often omitted. But such negative results are sometimes useful for experiments carried out by other scientists. Sources of reagents can also result in significantly different results [2]. So it is essential to know the attribution of these materials to replicate the method. Simple typographical errors or experiments done with different units like using 10 milligrams (mg) instead of 10 micrograms (μg) can make a difference. Thus, it is important to capture the provenance of an experiment with these fine details.

The aim of our work is to capture the provenance data from multiple resources of an experiment and provide the ability to visualize the data along with its semantics. Our contributions of this paper are as follows: *(i)* Identifying the components and competency questions needed to present and validate the story of a scientific experiment using microscopy experiments as an example. *(ii)* Presenting our provenance-based semantic approach using REPRODUCE-ME [13] ontology by extending PROV-O [9] and P-Plan [5] to represent the different paths taken from an input to an output of an experiment, the steps and the input and output variables of each step. *(iii)* Visualization of the provenance data of an experiment as a dashboard to the scientists in our prototype, CAESAR.

2 State of the Art

Missier [10] presents three main challenges for the practical usability of provenance data, one of which is the role of provenance in the reproducibility of the scientific processes. A well-developed approach to capture provenance is based on Scientific Workflow Management Systems. These are mature systems which ensure reproducibility in computational sciences by tracking and recording all the evolution of scientific workflows made by the user [11]. But the provenance captured by these systems is coupled to a workflow definition and does not include the detailed information of the execution environment (e.g., temperature) or the standard operating procedures followed in generating the materials of an experiment. They give more importance to provenance at the execution level of a workflow than on the entire lifecycle of an experiment. Many ontologies have also been introduced and developed to describe the workflow of computational experiments such as OBI [3] and CMPO [8]. However, these ontologies do not use the standard PROV model preventing the interoperability of the collected data.

Access to primary research data is important for scientists. Many public repositories have been created to host and share computational experiments with the aim of preservation of provenance components. The environment myExperiment [6] is an online web-portal which provides researchers the facility to share their scientific workflows. Image Data Resource (IDR) [17] is another public repository for sharing imaging data and links data from public chemical databases with controlled ontologies. Most of these platforms either focus on publishing datasets or the workflow part of an experiment.

Capturing provenance of non-computational parts of an experiment when not using a scientific workflow management system is challenging. We present our prototype which is a scientific data management platform where a user can document the experimental data like in a laboratory notebook. The semantics of the experimental data is then represented using an ontology to represent the whole story of an experiment including the path taken, the steps etc. Our prototype is different because we represent the whole picture of an experiment using provenance standards like PROV-O and P-Plan as well as give equal importance to the computational and non-computational provenance part of the experiment, unlike other systems, which target mostly the computational part. We also give equal importance to the agents involved directly or indirectly in an experiment because that may also affect the reproducibility of experiments.

3 Provenance-based Semantic Approach

The motivation of our work arises from the Collaborative Research Center (CRC) ReceptorLight[3] where scientists from two universities[4] , two university hospitals[5] and a non-university research institute[6] work together to understand the function of membrane receptors and develop high-performance microscopy techniques. Interviews with the scientists in the CRC as well as a workshop conducted to foster reproducible science[7] helped us to understand the different scientific practices followed in their experiments and their requirements of reproducibility and data management. We collected a list of competency questions from these oral interviews from the scientists from various projects performing different kinds of experiments. The relevance of these questions is further supported by their large overlap with competence questions obtained in other contexts, e.g. the provenance challenge [12]. From the collected questions, we selected the ones which were commonly told by scientists and generalized them. Here we present the

[3] `http://www.receptorlight.uni-jena.de/`

[4] `https://www.uni-jena.de/`, `https://www.uni-wuerzburg.de/`

[5] `http://www.uniklinikum-jena.de/`,`http://www.ukw.de/`

[6] `http://www.ipht-jena.de/`

[7] `http://fusion.cs.uni-jena.de/bexis2userdevconf2017/workshop/`

most common competency questions of which answers are required to describe and validate the story of an experiment.

1. What are the input and output variables of an experiment?
2. Which are the methods and standard operating procedures used?
3. Which are the files and materials that were used in a particular step?
4. Which are the steps involved in an experiment which used a particular material?
5. What is the complete path taken by a scientist for an experiment?
6. Which are the instruments that are associated with an experiment and their settings when the output was generated?
7. Which are the agents directly or indirectly responsible for an experiment?
8. Who created this experiment and when? Who modified it and when?
9. Which are the publications or external resources that were referenced in each step of an experiment?
10. List all the experiments which use growth protocol (EFO_0003789) and studies on "Homo sapiens" and resulted in phenotype "shorter prophase" which passed the quality control.

Question 10 is an example query specific to life science experiments. To answer these kinds of competency questions, we developed an ontology to represent the conceptual model of an experiment.

3.1 Development of REPRODUCE-ME ontology

The REPRODUCE-ME ontology[8] is extended from PROV-O to represent all entities, agents, activities and their relationships [13]. It also extends P-Plan to represent the steps of the activities or events involved in an experiment in detail. Figure 1 shows an excerpt of the classes and properties of the REPRODUCE-ME ontology depicting the lifecycle of a scientific experiment. We explain how we added classes and properties to the ontology to answer each of the competency questions.
To answer the competency questions from 1 to 5, we describe an *Experiment* as *p-plan:Plan* which in turn also is a *prov:Entity* by inference rules. The object property *p-plan:isSubPlanOfPlan* is used to associate an experiment with its subplans. Each subplan of an experiment consists of several smaller steps *p-plan:Step* which uses input and output variables *p-plan:Variable* which are represented using *p-plan:hasInputVar* and *p-plan:hasOutputVar* object properties. For example, the *HighContentScreening* is a step of *Experiment* and *ImageAcquisition* step has *Image* as an output variable. The complete path taken by an experiment is described by ordering these steps using the object property *p-plan:isPrecededBy*. For example, the execution order of cells of a Jupyter Notebook[9], a subplan of an experiment, is described using the *p-plan:isPrecededBy* property [14].

[8] https://w3id.org/reproduceme

[9] https://jupyter.org/

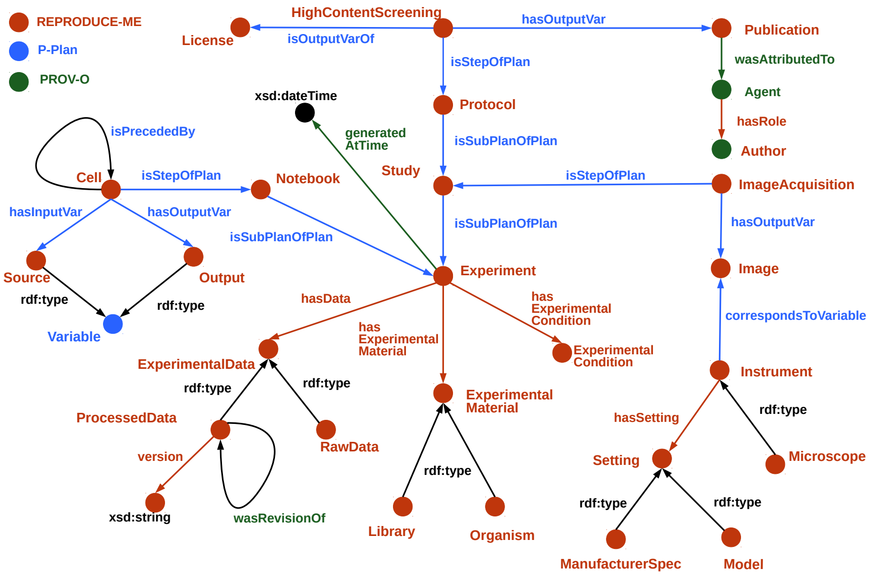


Fig. 1. The story of a scientific experiment depicted using the REPRODUCE-ME ontology

Images are an integral part of life-science experiments which involve microscopes for their acquisition. The acquisition, analysis, and annotation properties of a biological microscopic image are added as part of the ontology using the OME Data Model [7]. The class *Image* represents all the features of an image. For question 6, various instruments used in these experiments like microscopes are added as a subclass of *Instrument*, while the settings of these instruments are added as *p-plan:Variable*. Each of the instruments has *ManufacturerSpec* which consists of *Manufacturer*, *Model*, *SerialNumber* and *LotNumber*. Apart from this information, these instruments have certain attributes of their own. For example, *Laser* has data properties like *Wavelength*. In addition to the static properties of an instrument, some properties or settings are changed during an experiment to capture the image in a particular way. These properties are represented through *Settings*.

To answer the competency questions from 7 to 9, the following classes were added. A story needs characters to proceed. The agents are represented through the *prov:Agent*. Each person has a *prov:Role* in the experiment like *Experimenter*, *Distributor*, or *Manufacturer*. The resources and devices used in an experiment are either distributed by *Distributor* or manufactured by *Manufacturer*. Sometimes these resources are produced in the laboratory using a method represented in a *Publication*.

The time is another important factor in the story of an experiment. The data properties like *prov:generatedAtTime*, *receivedAtTime*, and *modifiedAtTime* are used to describe the time of each event. The *ResearchGroup* and *ResearchProject* represents the plot which describes the group and the project/institute for which the experiment was performed. The results of an experiment are the main outcome of an experiment which is represented by *Output*. The final message of the story of an experiment is represented using *Rating*, which is the rating given to the experiment and its description represents if any problems occur during its execution.

The *StandardOperatingProcedure* is a *p-plan:Plan* which describes the procedure of a method. The *File* is a variable which is also an experimental data. One variable is associated with another variable using the object property *reference*. Some variables are also added as *prov:Entity* so that properties associated with agent, entity and time can also be used. For example, *File* is a *p-plan:Variable* as well as a *prov:Entity*. Together, all the constructs described so far enable answering questions like Question 10.

3.2 Development of the semantic-based scientific data management platform

For the experimental data management, we present our prototype, CAESAR (**C**oll**A**borative **E**nvironment for **S**cientific **A**nalysis with **R**eproducibility), which is extended from OMERO [1]. The OMERO software, developed by the Open Microscopy Environment Consortium, is an open-source imaging database platform for experimental biology. The framework has a plugin architecture with a rich set of features including analyzing and modifying images and supporting over 140 image file formats using BIO-Formats [1]. BIO-Formats is an image translation library which reads and converts the proprietary microscopy data to an open standard model which then can be used by other tools. With the help of BIO-Formats, OMERO automatically extracts the image acquisition data including the data about devices and their settings.

CAESAR extends the OMERO framework so that scientists can document their experimental data along with their images [16]. The platform provides a form-based provenance capture system where a user can record experimental data like in their laboratory notebook. Each experiment has multiple steps where each step follows a Standard Operating Procedure (SOP). These procedures can either be files or Jupyter Notebooks. The user can upload the files and link the experiment materials used in each step of an experiment at a specific step.

The form also provides auto-completion of data. For example, if a user enters a Chemical Abstracts Service (CAS) number of a chemical compound, the web-client calls the CAS registry web service and auto-fills the chemical formula, molecular weight, and structural formula of the compound. The form also provides a virtual keyboard so that the user can insert special symbols, sub and superscript text since they are widely used in the documentation of the experimental data which include chemical formula and other structures. The prototype employs the user and group management provided by the OMERO platform. Groups in OMERO enable sharing of data between users. Users have role and permissions to restrict the modification of data. A user may belong to one or many groups. The data are shared between the users in the same group in the same OMERO server. The data can be made available to members of other groups based on the permission level of the group. Based on the permission level, if a user does not have the right to modify other member's data, then that user can propose changes to the experiment. This request is done using the proposal feature provided by CAESAR. The members of the current group or other groups can provide suggestions and propose the experimental data. The owner of the experiment receives those suggestions as proposals. The user has two options: First, to accept the proposal and add it to the current experimental data. Second, to reject the proposal and delete the proposal. CAESAR also provides a facility for the user to view the version history of an experiment to see all the changes made in its description.

In order to capture the provenance of the computational part of a scientific experiment, JupyterHub[10] is installed and connected to CAESAR so that users can create new notebooks, run and share them [14]. We capture the provenance of the execution of interactive notebooks used in a scientific experiment over the course of time with the help of ProvBook [15], an extension of Jupyter Notebook. The stored provenance data, also available as RDF, include the start and end time of the execution, the total time it took to run the cell and the input and output of the cell. The provenance difference feature provided by ProvBook provides the users to compare their results with the original results of the author of the notebook and detect which factors affected their results. In this way, we capture and link the provenance of both the computational and non-computational part of a scientific experiment to represent its story.

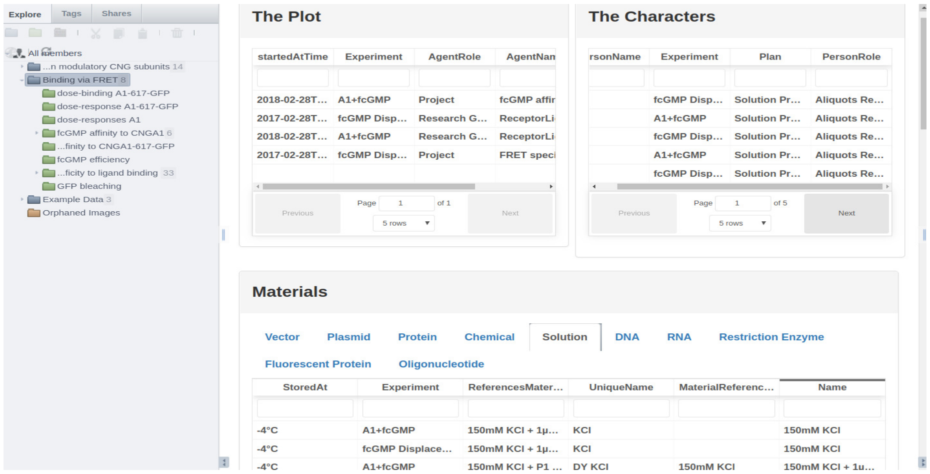
The data in the OMERO server which include the image metadata and experimental data are stored in the PostgreSQL relational database. To semantically represent this data and avoid replicating the data, we used ontology-based data access techniques to convert the relational database data to RDF data [13]. We created a mapping between the conceptual layer and the relational database layer. So now we have a mapping between the OMERO database and the experimental database. Every field in the database is mapped to the REPRODUCE-ME ontology. The ontology-based mapping is done using Ontop[11]. The mapping with Ontop created new terms in the ontology based on the mapping. Manual intervention was needed to remove the unnecessary terms and add the correct terms from the REPRODUCE-ME ontology.

3.3 Visualization of Provenance Data

Visualization of provenance data is another important feature of CAESAR provided to the user using a dashboard. Figure 2 shows a part of the project dashboard. Data are visualized at the project level, where multiple experiments are evaluated together. The competency questions described in Section 3 were converted to SPARQL queries and the answers to these questions are represented as tables in the dashboard. The dashboard provides a panel for each component of a story. The data are represented as data tables so that users can search and

[10] http://jupyter.org/hub

[11] http://ontop.inf.unibz.it/



The Plot

startedAtTime	Experiment	AgentRole	AgentNan
2018-02-28T...	A1+fcGMP	Project	fcGMP affir
2017-02-28T...	fcGMP Disp...	Research G...	ReceptorLi
2018-02-28T...	A1+fcGMP	Research G...	ReceptorLi
2017-02-28T...	fcGMP Disp...	Project	FRET speci

The Characters

rsonName	Experiment	Plan	PersonRole
	fcGMP Disp...	Solution Pr...	Aliquots Re...
	A1+fcGMP	Solution Pr...	Aliquots Re...
	fcGMP Disp...	Solution Pr...	Aliquots Re...
	A1+fcGMP	Solution Pr...	Aliquots Re...
	fcGMP Disp...	Solution Pr...	Aliquots Re...

Materials

StoredAt	Experiment	ReferencesMater...	UniqueName	MaterialReferenc...	Name
-4°C	A1+fcGMP	150mM KCl + 1µ...	KCl		150mM KCl
-4°C	fcGMP Displace...	150mM KCl + 1µ...	KCl		150mM KCl
-4°C	A1+fcGMP	150mM KCl + P1 ...	DY KCl	150mM KCl	150mM KCl + 1u...

Fig. 2. The Project Dashboard in CAESAR

filter the data. The plot panel provides the dates, research group, and research project associated with an experiment. The characters panel provides data about the agents responsible for an experiment. The materials panel displays the materials used in an experiment. The external resources and files panel display the publications and files associated with the materials and each step of an experiment. The devices panel shows all the devices associated with an experiment and the settings panel displays all the settings including settings of the devices. In addition to the dashboard, we also provide a SPARQL query editor with SPARQL templates so that answers to the questions like 10 can be obtained. In addition to the existing parameterized SPARQL templates, the user can also write their own SPARQL queries related to experimental data and get results.

3.4 Evaluation

We conducted both a user and a data-based evaluation on two different datasets to check whether the data along with the ontology can explain the story of an experiment. The REPRODUCE-ME ontology, the supplementary materials used for the evaluation and the results are publicly available[12]. All the evaluations are based on the list of competency questions described in Section 3. The metrics used for the evaluation were the usability of the system and whether the competency questions were answered. For the user-based evaluation, a group of four scientists working with high-end light microscopy techniques from the project B1 of the CRC ReceptorLight evaluated the platform. Around ten different experiments of type Fluorescence Resonance Energy Transfer (FRET) and confocal Patch Clamp Fluorometry (cPCF) were used for the evaluation. Eleven different types of experiment materials like chemicals, proteins, solutions etc. were used as input of the experiments and around 70 microscopic images generated from

[12] `https://w3id.org/reproduceme/research`

instruments with different settings were used for the evaluation. The results of SPARQL queries in the dashboard were manually compared and their correctness was evaluated by the domain experts. The evaluation results show that the dashboard provided them with a complete overview of the experiments. Since none of them (like most scientists) possesses SPARQL knowledge, such a complete overview could not have been gained without the dashboard. The ability to filter the results in each table of the dashboard also helped them to search their queries.
The other evaluation of the ontology was done with the data from the Image Data Repository (IDR) which currently consists of around 35 imaging experiments [17]. The metadata of each imaging study from the IDR datasets was extracted and described in RDF using the REPRODUCE-ME ontology with scripts[13]. The SPARQL queries generated from the competency questions were executed also on this data to check whether the ontology can be used to describe the story of other types of experiments as well. To illustrate the story of such an experiment, we take an example from IDR. The study "Focused mitotic chromosome condensation screen using HeLa cells" (idr0002[13]) is an *Experiment* which consists of *ImagingStudy* as a step. There are around 1160 *Image*s which are the output variables of *ImagingStudy.* The *Publication* and the *ProcessedData* file are also the output variables of this experiment. Each *Image* in the experiment is annotated with the *GeneIdentifier* and *Phenotype.* The *Experiment* is attributed to several agents who take the role of *Submitter* of the experiment, *Manufacturer* of the *Library* used and *Author* of *Publication.* The *Experiment* has several *Protocol*s, which describe the various instructions followed in the experiment. The results from data-based evaluations show that the provenance-based semantic system helps in providing the whole story of an experiment along with all its dependencies.

4 Conclusions and Future Work

Data provenance is a key factor towards reproducibility of scientific experiments. In this paper, we present a provenance-based semantic approach to explain the story of a biological experiment from its plot to its output. The REPRODUCE-ME ontology extended from the existing ontologies PROV-O and P-Plan, is used to represent a whole picture of an experiment including the plot, characters, settings, plans, steps, input and output variables. The ontology and the prototype are validated through answering the competency questions. In future work, we will focus on the scalability and performance of the system.

Acknowledgements

This work is supported by the Deutsche Forschungsgemeinschaft (German Research Foundation, CRC/TRR 166 "High-end light microscopy elucidates membrane receptor function - ReceptorLight", projects Z2 and B1).

[13] The subsets of idr datasets converted to RDF are available here `https://github.com/Sheeba-Samuel/REPRODUCE-ME/`

References

1. Allan, C., Burel, J.M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., MacDonald, D., Moore, W.J., Neves, C., Patterson, A., et al.: OMERO: flexible, model-driven data management for experimental biology. Nature methods 9(3), 245–253 (2012)
2. Baker, M.: Reproducibility crisis: Blame it on the antibodies. Nature 521(7552), 274–276 (2015)
3. Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M.H., Bug, B., Chibucos, M.C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., et al.: The ontology for biomedical investigations. PloS one 11(4), e0154556 (2016)
4. Chirigati, F., Freire, J.: Provenance and Reproducibility, pp. 1–5. Springer New York, New York, NY (2017)
5. Garijo, D., Gil, Y.: Augmenting PROV with plans in P-Plan: scientific processes as linked data. CEUR Workshop Proceedings (2012)
6. Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., et al.: myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic acids research 38(suppl_2), W677–W682 (2010)
7. Goldberg, I.G., Allan, C., Burel, J.M., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P.K., Swedlow, J.R.: The open microscopy environment (OME) data model and XML file: open tools for informatics and quantitative analysis in biological imaging. Genome biology 6(5), R47 (2005)
8. Jupp, S., Malone, J., Burdett, T., Heriche, J.K., Williams, E., Ellenberg, J., Parkinson, H., Rustici, G.: The cellular microscopy phenotype ontology. Journal of Biomedical Semantics 7(1), 28 (2016)
9. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., Zhao, J.: PROV-O: The PROV Ontology. W3C Recommendation 30 (2013)
10. Missier, P.: The lifecycle of provenance metadata and its associated challenges and opportunities. In: Building Trust in Information, pp. 127–137. Springer (2016)
11. Missier, P., Woodman, S., Hiden, H., Watson, P.: Provenance and data differencing for workflow reproducibility analysis. Concurrency and Computation: Practice and Experience 28(4), 995–1015 (2016)
12. Moreau, L., Ludäscher, B., et al.: Special issue: The first provenance challenge. Concurrency and computation: practice and experience 20(5), 409–418 (2008)
13. Samuel, S., König-Ries, B.: REPRODUCE-ME: ontology-based data access for reproducibility of microscopy experiments. In: The Semantic Web: ESWC 2017 Satellite Events, Portorož, Slovenia. pp. 17–20 (2017)
14. Samuel, S., König-Ries, B.: Combining P-Plan and the REPRODUCE-ME ontology to achieve semantic enrichment of scientific experiments using interactive notebooks. In: The Semantic Web: ESWC 2018 Satellite Events. pp. 126–130 (2018)
15. Samuel, S., König-Ries, B.: ProvBook: Provenance-based semantic enrichment of interactive notebooks for reproducibility. In: Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with ISWC 2018, Monterey, USA (2018), `http://ceur-ws.org/Vol-2180/paper-57.pdf`
16. Samuel, S., Taubert, F., Walther, D., König-Ries, B., Bücker, H.M.: Towards reproducibility of microscopy experiments. D-Lib Magazine 23(1/2) (2017)
17. Williams, E., Moore, J., Li, S.W., Rustici, G., Tarkowska, A., Chessel, A., Leo, S., Antal, B., Ferguson, R.K., Sarkans, U., et al.: Image data resource: a bioimage data integration and publication platform. Nature methods 14(8), 775 (2017)