Integrative Data Management for Reproducibility of Microscopy Experiments

Sheeba Samuel(✉)

Heinz-Nixdorf Chair for Distributed Information Systems,
Friedrich-Schiller University, Jena, Germany
sheeba.samuel@uni-jena.de

Abstract. Reproducibility is a fundamental factor in every domain of science since it allows scientists to trust data and results. The scientific community is interested in the results of experiments which are reproducible, reusable and understandable. In this paper, we present our work towards reproducibility of scientific experiments taking into account the use case of microscopy. We aim to analyze the components that are vital for reproducibility and to develop an integrative data management platform for scientific experiments. In this article, we show the use of Semantic Web technologies to conserve an experiment environment and its workflow. This allows scientists to ask queries related to an experiment and compare results. We present our approach for scientists to represent, search and share their experimental data and results to the scientific community for better data interoperability and reuse. Our overall goal is to extend data management and Semantic Web technologies to enable reproducibility.

Keywords: Reproducibility · Experiments · Ontology · Microscopy · Provenance

1 Introduction

Recent advancements in science and technology have brought a new range of challenges for scientists regarding the reproducibility of their research. Reproducibility has become a lot more difficult to achieve today because experiments and their setup have become much more complex. A sustainable, reliable and scalable data management platform is required for scientific experiments which generate a large volume of heterogeneous data. Apart from data management, it is required to ensure reproducibility of experimental data and results.

An experiment is said to be reproducible [18] when it can be repeated under different conditions to get the same results. This can occur when the experiment is carried out by another scientist in a different location using different devices and materials. To make an experiment reproducible, the provenance of the experiment and processing environment must be captured. Provenance is the source of information that is used to describe the entities and processes involved

E. Blomqvist et al. (Eds.): ESWC 2017, Part II, LNCS 10250, pp. 246–255, 2017.
DOI: 10.1007/978-3-319-58451-5_19

in generating a resource. The provenance of an experiment describes who performed the study and when, the materials used and how they were produced, the last time it was modified, the devices used and their settings, the experimental procedures used etc. [17]. Semantic Web-based representations of provenance help in better data interoperability and reuse.

The main focus of our research is to extend data management and Semantic Web technologies in order to better support reproducibility. In the initial phase, we work towards developing an integrative data management platform which can enable reproducibility of scientific experiments with the help of Semantic Web technologies. This research focuses on the use case of microscopy.

2 State of the Art

Scientists need to store information about an experiment and its workflow so that they can share this with other collaborators in an understandable way. This leads us to focus on three important aspects of our research which can aid work on the reproducibility of scientific experiments: (1) Scientific Data Management, (2) Semantic Web technologies for capturing provenance and (3) Scientific Workflows.

(1) *Scientific Data Management*
Advancements in data storage solutions allow successful preservation, processing and analysis of large volume and variety of data. Scientific data management brings challenges like scalability, heterogeneity, sharing, transformation and quality of data. Many platforms are developed to support scientific data management for general or specific requirements of projects such as the European Bioinformatics Institute[1] (EBI) for genomic data, BacDive[2] for bacterial data, BExIS[3] for biodiversity data, myExperiment[4] for sharing bioinformatics workflows. Demchenko et al. [6] show how cloud-based services provide support for scientific data infrastructures.
A number of platforms aim at providing data management and analysis of microscopic images. OMERO [1] and BisQue [9] are two examples for storage solutions for microscopy images. OMERO, developed by the OME consortium, is an open source client-server platform for visualization, management and analysis of images generated from a microscopy experiment. Since it provides a rich set of different image file formats and flexibility to extend features, we selected OMERO as a suitable data management platform to support microscopic data infrastructure.

(2) *Semantic Web Technologies for Capturing Provenance*
Semantic Web technologies provide possibilities to represent experiment data

[1] http://www.ebi.ac.uk/.
[2] http://bacdive.dsmz.de/.
[3] https://www.bexis.uni-jena.de.
[4] http://www.myexperiment.org.

in a more understandable and reusable way to scientists and in particular in a machine-understandable way. Various ontologies are developed and used in various domains to help scientists annotate resources and support data interoperability. PROV-O [10] is a general purpose ontology developed by the W3C working group to model the entities and the activities that produced them. It provides high flexibility for extension so that it can be used for specific applications.

PROV-O has been extended in various works for specific purposes. For example, Ciccarese et al. [3] evaluate why PROV-O was selected for capturing provenance of web resources. They present the PAV ontology which helps to capture provenance, authorship and versioning of the resources on the web. Compton et al. [4] combines the Semantic Sensor Networks Incubator Group's ontology (SSNO) and PROV-O to describe sensor data.

Several authors have built ontologies for microscopy. The Cellular Microscopy Phenotype Ontology (CMPO) [7] provides phenotypic observations related to cellular components. Another work [8] describes the development of an Ontology for an Integrated Image Analysis Platform to enable Global Sharing of Microscopy Imaging Data. They present an ontology to describe data from microscopic images by converting the Open Microscopy Environment (OME) data model to the Resource Description Framework (RDF) schema. These works focus on using ontologies for describing biological structures and annotating microscopic images.

Moreau in his paper [13] describes reproducibility semantics for the Open Provenance Model (OPM)[5]. It is a specification of a reproducibility service and defines reproducibility formally with a mathematical explanation of OPM graphs. This semantics which takes the form of a denotational semantics, is the basis of a theory of provenance-based reproducibility.

(3) *Scientific Workflows*

Scientific workflow management systems [5] model the flow of data through a series of computational steps performed in an experiment. Several workflow management systems have been developed over the past years which are either generic or specific to a domain. Systems like Kepler [11] and Vistrails [2] provide facilities to design, execute and rerun the scientific workflows and also provide a visual interface for composing workflows. Santana-Perez et al. [15] present a semantic approach to attain reproducibility of computational environments in scientific workflows by documenting the scientific workflow and conserving the execution environment using semantic vocabularies.

In spite of all the advantages and the richness in features, scientists are hesitant to abandon the tools they are used to and try new ones instead, thus, in many scientific communities, the uptake of scientific workflow management systems has been slow. This motivates the work for YesWorkflow [12] and noWorkflow [14]. YesWorkflow extracts comments from the scripts and provides a graphical rendering of the workflow. The noWorkflow tool

[5] http://openprovenance.org/.

transparently captures provenance from Python scripts and supports different kind of analysis on them. Recent work on combining YesWorkflow and noWorkflow captures provenance of results generated by scripts written by the scientists in an experiment [16]. This work takes benefits of both the systems by capturing provenance which is collected from the structure of scripts, events occurred during script execution, annotations in the comments of scripts and the files generated by the scripts.
We focus our research on end-to-end reproducibility of scientific experiments by integrating scientific data management and Semantic Web technologies.

3 Problem Statement

The main goal of our research is to enable reproducibility by extending data management and Semantic Web technologies. Inorder to start with the research, we focus on the use case of microscopy experiments. Recent advances in microscopy techniques make the study of biological systems more promising. The need for the research on this area arises from the Collaborative Research Center ReceptorLight[6]. Scientists from this joint project work together to understand the function of membrane receptors with the help of high-performance microscopy methods. Membrane receptors are protein molecules which receive chemical signals from outside a cell and distribute the signals to other parts of the cell. Through this project, scientists want to understand the minute interactions happening in the biological structures and processes. Using high-resolution imaging of these receptors, scientists can gain new insights on neurological autoimmune diseases and other areas.

Discussions with scientists from various domains working in this project brought light to the challenges they face related to the reproducibility aspects. One of the challenges faced by the scientific community is that most of the information related to the experiment are not integrated to the digital systems. They still use lab notebooks (analog) to record their data as they perform experiments. The information in these notebooks is of great value as they contain the description of experiment procedures, resources used, data and results. The difficulty arises when the data and results have to be shared between scientists from different locations. Shared understanding of data is essential so that the data can be reused for new experiments and analysis.

Samples and resources used in the experiment cannot be preserved for long in many biological and medical studies, unlike the computational science domain. It is required to digitally conserve the execution environment of an experiment which consists of different devices, software and materials. Data collected by the experimenter from different data sources using USB sticks are stored in personal hard disks. There are greater chances of losing data and the different modifications of the data if they are not versioned.

Another challenge faced by the scientists is the big data. Each measurement of a microscopy experiment can produce terabytes of data and images. Scientists

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

have to make a choice to keep the data they are interested in and discard the rest of the data. So it is important to have a scalable and high-performance data management approach to handling the large volume of experimental data and results.

We have identified four main research questions based on the challenges faced by the scientists and ways to achieve the vision of reproducibility.

RQ-1 - How to capture the provenance of a scientific experiment through an integrative data management? Which are the vital components required to attain reproducibility?
RQ-2 - How to represent a scientific experiment and its execution environment with the help of Semantic Web technologies to enable reproducibility, data interoperability and reuse?
RQ-3 - How to provide a scalable and high-performance platform to handle the experimental data and results?
RQ-4 - How to enhance current Semantic Web languages with reproducibility qualifiers?

We focus our research based on the hypothesis that Semantic Web enabled scientific data management platform can be created that facilitates the reproducibility of scientific experiments.

4 Research Methodology and Approach

The research methodology and approach we present here are based on the research questions defined in Sect. 3. A high-level view of our research methodology in the first phase is illustrated in Fig. 1. The first step in our approach is to collect requirements and understand what type of information scientists need to reproduce the experiments. In this phase, it is required to identify the factors that can enable reproducibility. Continuous discussions with scientists help us to gain the domain knowledge and the challenges faced by them.

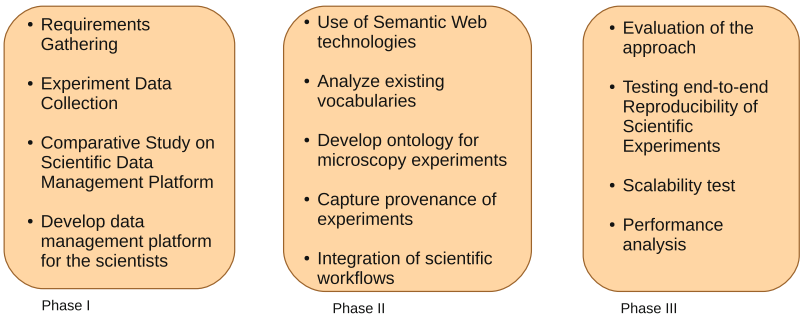

Fig. 1. Our research methodology and approach

For RQ-1, we explored the literature to find a suitable data management platform and selected a platform which can be extended to capture the provenance of microscopy experiments. We understood the deployment setup of the experiments and designed the system in such a way that scientists could conserve the experimental data along with the images in one place. In this phase, we consider different approaches to capture provenance and analyze whether those are sufficient to enable reproducibility. We are analyzing the various components required to enable reproducibility.

To answer RQ-2, we need to understand how data stored in computers in addition to their lab notebooks can benefit scientists. We realize the potential uses of semantic web technologies and how it can be used to describe the scientific workflow. Based on the literature review, we analyzed the different existing vocabularies in the scientific domain, particularly for microscopy. We comprehended the details of experiments we need to capture for our use-case and extended the existing ontology based on them. We will analyze whether all the questions related to reproducibility can be answered with this ontology or the ontologies need to be further refined. We will evaluate whether ontology-based representation of experimental data is enough to enable reproducibility.

For RQ-3, we will analyze the existing approaches for handling scalability. We will test the performance of the system and provide an optimal solution for scientists to handle their data.

For RQ-4, we will collect and analyze the questions that are asked by scientists concerning reproducibility. We will analyze how the questions and the approach be generalized for all the scientific domains. We will find out the type of qualifiers needed to extend Semantic Web languages. We will consider various methods to formalize the process needed to enable reproducibility.

5 Preliminary Results

Based on the literature review, we did a comparative study on two existing storage platform systems, OMERO [1] and BisQue [9]. We came to the conclusion to select OMERO for our requirements because of the richness in its features and higher flexibility to extend the platform. We collected data and requirements from the discussions we had with the scientists working in the CRC ReceptorLight project.

The initial goal of our approach is to document the description of an experiment and its execution environment conditions. The information includes experimental details, materials and devices used in the experiment and their settings, the time of each activity performed and the standard operating procedures used. The current prototype is developed [17] based on OMERO to achieve the initial goal.

We extended OMERO's server database to include the data schema provided by the scientists. OMERO's web client was extended to provide a facility to input the experimental data. It also helps the scientist to associate the experiment to the image results generated from the experiment. Users can view all the information of an experiment at one place.

Figure 2 presents the high-level system architecture developed for the data management of microscopy experiments. Scientists use different devices like a microscope, an electrophysiologic device for performing experiments. The workstations associated with them are installed with proprietary software of the device manufacturer. The description of the materials and the procedures used in the experiment and the execution environment parameters are noted down in the lab notebooks. The data and results obtained are stored in personal external devices. A desktop client was developed to deploy in the workstations associated with these devices as they are not connected to the internet due to security reasons. It allows the scientists to input the data as and when they perform experiments. Later they can upload this information to the server whenever the internet connection is available. Through the web client, they can share the experiment information to other scientists for their review. Based on the permissions and roles assigned to the scientists, they can view or edit the information related to the experiment. The web client also provides a facility to search experiment related information.

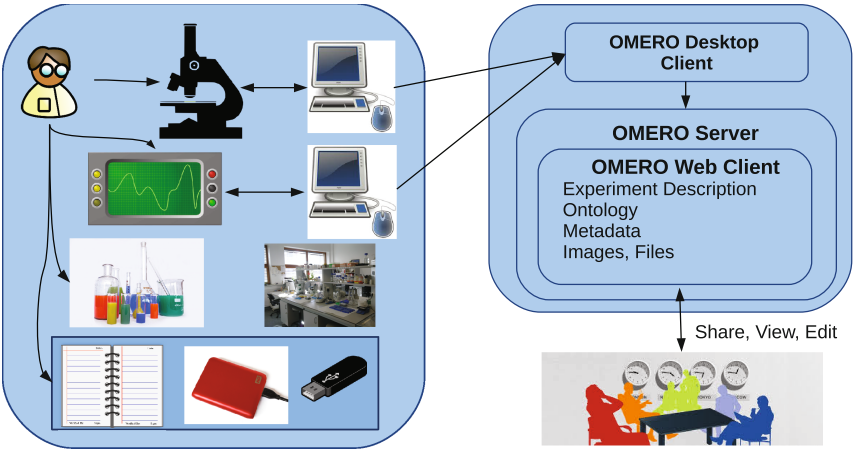

Fig. 2. The proposed approach

We have developed an ontology, REPRODUCE-ME (Reproduce Microscopy Experiments)[7] to describe an experiment and its execution environment. The ontology is built to capture the provenance of an experiment, the materials and devices used and their properties, standard operating procedures and the people who are responsible for an experiment. This is developed by extending the existing ontology, PROV-O. With the help of classes and properties of PROV-O and the new classes added in REPRODUCE-ME Ontology, it is possible to describe the entities, activities, agents and their role in a scientific experiment. The prefix "repr:" is used to indicate the namespace "http://fusion.cs.uni-jena.de/fusion/repr".

[7] http://fusion.cs.uni-jena.de/fusion/repr/.

Figure 3 shows the main concepts and properties of REPRODUCE-ME ontology. `prov:Entity`, `prov:Agents` and `prov:Activity` are the main concepts in PROV-O. The experiment materials and devices used in an experiment are extended from the concept `prov:Entity`. The people who are involved in producing materials and performing experiments are extended from `prov:Agents`. All the actions performed in an experiment are extended from the `prov:Activity` class. The standard operating procedures used in the experiment is extended from `prov:Plan`. Several objects and data properties are added to describe the experiment and its execution environment. Scientists can make semantic queries using SPARQL with the help of ontology.

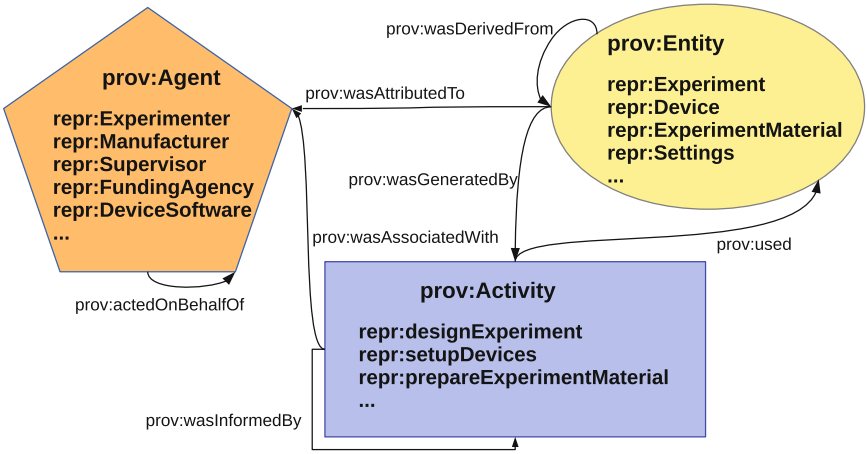

Fig. 3. A part of REPRODUCE-ME ontology

6 Evaluation Plan

To evaluate the research questions mentioned in Sect. 3, we will validate the approach using microscopy experiments. Later it will be validated by scientists from other domains. The validation of the approach will be achieved when scientists from different locations can reproduce the experiments based on the shared description provided by the REPRODUCE-ME ontology. The ontology developed for the microscopy experiments will be continuously validated, revised and corrected by the team of scientists. The detailed list of queries that a scientist would like to pose based on an experiment will be collected from the scientific community. The competency questions that the ontology can support will be clearly listed. We will evaluate whether the provenance captured through the prototype is sufficient enough to attain end-to-end reproducibility of scientific experiments.

Scalability, performance and data quality will also be considered during the evaluation. The current prototype [17] will be manually tested by the domain scientists. Test cases will be formulated to validate the system and the results of

queries. We are interested in scientists testing our system with a large amount of data and different experiment setups.

7 Conclusions

Reproducible research brings great value and benefits for the scientific community. The overall objective of our research is to extend data management and Semantic Web technologies to enable reproducibility. In this research, we aim to examine the suitable components required for reproducibility. As a first step towards reproducibility, we developed an integrative platform for the scientists to capture the provenance of the experiment. We built an ontology for the microscopy experiments with a focus on capturing the description of experiment and execution environment conditions. This allows better interpretation of data from different scientists in a collaborative project. The system allows scientists to query the experimental data through SPARQL queries and get results without worrying about the underlying technologies. We will consider ways to enhance Semantic Web languages with reproducibility qualifiers. Scalability, performance and quality tests will be conducted to handle the sheer volume of data generated from each experiment.

Acknowledgements. This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in Project Z2 of the CRC/TRR 166 High-end light microscopy elucidates membrane receptor function - ReceptorLight. I thank Birgitta König-Ries and H. Martin Bücker for their guidance and feedback for the research plan. I thank Christoph Biskup and Kathrin Groeneveld from the Biomolecular Photonics Group at University Hospital Jena, Germany, for providing the requirements to develop the proposed approach and validating the system.

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. Nat. Methods **9**(3), 245–253 (2012)
2. Callahan, S.P., Freire, J., Santos, E., Scheidegger, C.E., Silva, C.T., Vo, H.T.: VisTrails: visualization meets data management. In: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, SIGMOD 2006, pp. 745–747. ACM, New York, NY, USA (2006). http://doi.acm.org/10.1145/1142473.1142574
3. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J., Goble, C., Clark, T.: PAV ontology: provenance, authoring and versioning. J. Biomed. Seman. **4**(1), 37 (2013). http://dx.doi.org/10.1186/2041-1480-4-37
4. Compton, M., Corsar, D., Taylor, K.: Sensor data provenance: SSNO and PROV-O together at last. Terra Cognita and Semantic Sensor Networks, pp. 67–82 (2014)
5. Curcin, V., Ghanem, M.: Scientific workflow systems - can one size fit all? In: 2008 Cairo International Biomedical Engineering Conference, pp. 1–9 (2008)

6. Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A., de Laat, C.: Addressing big data challenges for scientific data infrastructure. In: Proceedings of the 4th IEEE International Conference on Cloud Computing Technology and Science, pp. 614–617 (2012)
7. Jupp, S., Malone, J., Burdett, T., Heriche, J.K., Williams, E., Ellenberg, J., Parkinson, H., Rustici, G.: The cellular microscopy phenotype ontology. J. Biomed. Seman. **7**(1), 28 (2016). http://dx.doi.org/10.1186/s13326-016-0074-0
8. Kume, S., Masuya, H., Kataoka, Y., Kobayashi, N.: Development of an ontology for an integrated image analysis platform to enable global sharing of microscopy imaging data. In: Proceedings of the ISWC 2016 Posters & Demonstrations Track co-located with 15th International Semantic Web Conference (ISWC 2016), Kobe, Japan, October 19, 2016 (2016). http://ceur-ws.org/Vol-1690/paper93.pdf
9. Kvilekval, K., Fedorov, D., Obara, B., Singh, A., Manjunath, B.: Bisque: a platform for bioimage analysis and management. Bioinformatics **26**(4), 544–552 (2010)
10. 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)
11. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y.: Scientific workflow management and the kepler system. Concurrency Comput. Pract. Experience **18**(10), 1039–1065 (2006). http://dx.doi.org/10.1002/cpe.994
12. McPhillips, T., Song, T., Kolisnik, T., Aulenbach, S., Belhajjame, K., Bocinsky, K., Cao, Y., Chirigati, F., Dey, S., Freire, J., et al.: YesWorkflow: a user-oriented, language-independent tool for recovering workflow information from scripts (2015). arXiv preprint arXiv:1502.02403
13. Moreau, L.: Provenance-based reproducibility in the semantic web. Web Seman. Sci. Serv. Agents World Wide Web **9**(2), 202–221 (2011)
14. Murta, L., Braganholo, V., Chirigati, F., Koop, D., Freire, J.: noWorkflow: capturing and analyzing provenance of scripts. In: Ludäscher, B., Plale, B. (eds.) IPAW 2014. LNCS, vol. 8628, pp. 71–83. Springer, Cham (2015). doi:10.1007/978-3-319-16462-5_6
15. Pérez, I.S., Pérez-Hernández, M.S.: Towards reproducibility in scientific workflows: An infrastructure-based approach. Sci. Program., 243180:1–243180:11 (2015). http://dx.doi.org/10.1155/2015/243180
16. Pimentel, J.F., Dey, S., McPhillips, T., Belhajjame, K., Koop, D., Murta, L., Braganholo, V., Ludäscher, B.: Yin & yang: demonstrating complementary provenance from noWorkflow & YesWorkflow. In: Mattoso, M., Glavic, B. (eds.) IPAW 2016. LNCS, vol. 9672, pp. 161–165. Springer, Cham (2016). doi:10.1007/978-3-319-40593-3_13
17. Samuel, S., Taubert, F., Walther, D., König-Ries, B., Bücker, H.M.: Towards reproducibility of microscopy experiments. D-Lib Mag. **23**(1/2) (2017)
18. Taylor, B.N., Kuyatt, C.E.: Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. Tech. rep., NIST Technical Note 1297 (1994)