Ubiquitous Research Preservation: Transforming Knowledge Preservation in Computational Science

Sebastian S. Feger
CERN, Geneva and LMU Munich
sebastian.s.feger@cern.ch

Sünje Dallmeier-Tiessen
CERN, Geneva, Switzerland
sunje.dallmeier-tiessen@cern.ch

Pascal Knierim
HCUM - LMU Munich, Munich, Germany
pascal.knierim@um.ifi.lmu.de

Passant El.Agroudy
University of Stuttgart, Stuttgart, Germany
Passant.El.Agroudy@vis.uni-stuttgart.de

Paweł W. Woźniak
Utrecht University, Utrecht, the Netherlands
p.w.wozniak@uu.nl

Albrecht Schmidt
HCUM - LMU Munich, Munich, Germany
albrecht.schmidt@um.ifi.lmu.de

ABSTRACT

Research preservation is crucial for supporting researchers' sensemaking and knowledge sharing. However, human compliance to capturing strategies is a barrier for creating complete scientific repositories. In this paper, we introduce *Ubiquitous Research Preservation*, which we envision to automate preservation in computational science. We contribute a characterization of preservation processes, illustrate the spectrum of technology interventions and describe research challenges and opportunities for *Ubiquitous Research Preservation* in computation-based scientific domains.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; *Empirical studies in collaborative and social computing*; *Ubiquitous and mobile computing design and evaluation methods.*

KEYWORDS

Ubiquitous Research Preservation, Science Reproducibility, Sensemaking, Education, Reuse, Connected Devices, Computational Science.

1 INTRODUCTION

Preservation of scientific knowledge enables researchers to reflect on past choices and to share resources and findings with the scientific community. Yet, preserving and sharing research requires substantial efforts [1]. Studies have shown that documentation and preservation technology needs to ease scientists' efforts and make use of automated recording and processing mechanisms [3, 6, 8].

This paper focuses on research *preservation in computational and data-driven science*. Although barriers for capturing and sharing resources in computation-based science are rather low, availability and sharing of digital resources remains a major concern [2, 7]. In fact, shortcomings in personal repositories often require creative solutions[1].

Motivation and Background

Oleksik et al. [6] reported on their observational study on electronic lab notebooks (ELN) in a research organization. They found that the flexibility of digital media can lead to much less precision during experiment recording and that *"freezing"* parts of the record might be necessary. The authors stressed that *"ELN environments need to incorporate automatic or semi- automatic features that are supported by sophisticated technologies [...]."*

Studying the use of a hybrid laboratory notebook, Tabard et al. [8] found that *"users clearly do not want to focus on the process of capturing information."* Yet, they also noted that automated mechanisms can be intrusive and that users need to be in control of the recording and sharing. They illustrated the importance of reflection in the scientific process and highlighted how access to preserved, redundant information supports reflection, as *"scientists understand how their thoughts have evolved over time."*

In our ongoing research, we study practices around research preservation in High Energy Physics (HEP) [3, 4]. In an interview study with HEP data analysts [3], we found that lack of preservation and sharing highly impacts the ability to reuse and reproduce work in this data-intensive, computational environment. We also found that HEP data analysis work is based on common building blocks that foster implementation of automated recording strategies.

[1]The article "Local File History for Xcode - My solution" by Andrew Eades provides a good example of how local versioning of computational resources can be improved through existing technology (Retrieved February 24, 2020).

Box 1: Characterizing Researcher Interaction

Based on our research in experimental physics, we introduce and define two dimensions to characterize preservation practices from a researcher point of view: *Initiative* and *Resource Awareness*.

Initiative: Nominates the entity responsible for initiating a preservation process.

User-Initiated: The researcher is responsible for process initiation and control. User decides on suitable occasions.

Machine-Initiated: The machine initiates and controls processes. Decisions might be based on: workflow knowledge; pre-configured domain rules; and / or user-configured rules.

Resource Awareness: Describes how aware researchers are about the selection of resources in the preservation process.

Conscious: Only resources are preserved which are selected by the user.

Unaware: The user has no direct control over the resources that are preserved. However, he / she might have previously set rules for this process.

Kery et al. [5] asked scientists to think about *"a magical perfect record"* in their study of literate programming tools. Participants created queries referring to *"many kinds of contextual details, including libraries used, output, plots, [...]."* Participants described their inability to find prior analyses and illustrated consequences. The authors found that in literate programming tools, *"version control is currently poor enough that records of prior iterations often do not exist."*

2 TECHNOLOGY INTERVENTIONS FOR RESEARCH PRESERVATION

To describe the spectrum of technology intervention in the preservation of machine-processed research, we characterize preservation efforts from a researcher point of view, taking into account our research in experimental physics [3, 4]. Researchers commonly document, preserve and possibly share information and resources in lab notebooks, cloud services or dedicated research preservation services (e.g. Figshare and Zenodo). Or, they decide to commit assets to repositories (e.g. GitHub). In either case, those actions are mostly USER-INITIATED. Scientists who — for any reason — decide to preserve or share their research make a CONSCIOUS selection of their study's data and materials. We assigned those characteristics to the dimensions INITIATIVE and RESOURCE AWARENESS, as described in Box 1.

Towards Ubiquitous Research Preservation

We characterize automated preservation strategies based on INITIATIVE and RESOURCE AWARENESS. In contrast to current user-initiated preservation efforts, machine-supported recording of workflows would be MACHINE-INITIATED. Here, researchers could be UNAWARE of continuous background preservation efforts. This envisioned transformation is based on the demonstrated need to support researchers through automated preservation processes.

Described dimensions and characteristics enable a wide spectrum of technology interventions, as depicted in Figure 1. For example, technology could implement a completely MACHINE-INITIATED/UNAWARE preservation of computational processes. Such an approach could guarantee (near-) continuous workflow recording, possibly taking inspiration from extreme forms of documentation like lifelogging.

Related work showed that *control* is an important factor in research preservation. Technology supporting USER-INITIATED/UNAWARE interactions might make an important contribution towards acceptance. For example, a researcher who considers a process to be relevant in the future, could start an application or execute a command that initiates recording of computational states and changes (see Figure 2). The researcher should be able to stop this process at any time.

MACHINE-INITIATED/CONSCIOUS interaction could also provide researchers with control. Here, the machine might actively propose users to preserve certain processes. This decision would need to be based on pre-defined triggers or in-depth workflow knowledge. A researcher might receive a notification detailing the proposed initiation of a preservation process or activity (see Figure 3).

We refer to the spectrum of technology interventions for machine-supported recording of computation-based research as **UBIQUITOUS RESEARCH PRESERVATION (URP)**. We define URP and URP technology in Box 2.

		Initiative	
		User-Initiated	**Machine-Initiated**
Resource Awareness	**Conscious**	*In today's focus:* General / Tailored services Journal / Conference repositories	Informs users about ongoing documentation and provides them with control Proposes to initiate documentation and highlights benefits / impact
	Unaware	The user initiates capturing of a process. She / He has control over the duration. Yet, the documentation is completely performed in the background.	Continuous process capturing over undefined period Possible inspiration from extreme forms of documentation (e.g. Lifelogging)

Figure 1: Characterizing preservation technology based on Initiative and Resource Awareness enables a wide spectrum of technology interventions.

```
user$ sci_recorder -all start
```

Figure 2: Speculative prototype of User-Initiated / Unaware interactions.

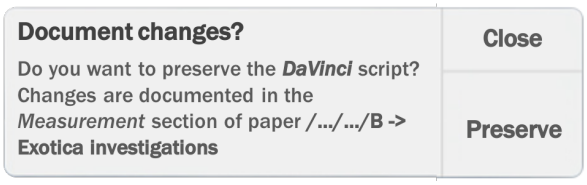

Figure 3: Machine-Initiated / Conscious interaction might provide needed control.

Box 2: Definitions

Ubiquitous Research Preservation (URP) refers to the machine-supported scientific knowledge recording and preservation process of computational workflows.

URP technology *initiates* and/or *controls* partial or complete preservation.

3 RESEARCH CHALLENGES AND FUTURE WORK

Our research and related studies illustrated various challenges resulting from automated recording strategies. Here, we expand on challenges and opportunities for research on URP technology:

Usefulness. To create complete *"magical records"*, preserved data need to be annotated, searchable and suitable for desired use cases. It will be important to manage the signal-to-noise ratio, as well as to find suitable ways for information discovery and presentation.

Generalizability. As URP technology profits from knowledge about research practices for recording and presenting information, development of assistive technology across heterogeneous environments needs to be further researched. Research questions include: How can technology assess researchers' practices, needs and integrate into their workflows? Can we create accessible templates based on learned and confirmed structures? How does technology adapt to scientific novelty and creativity?

Control. Acceptance of URP technology will depend on researchers' perceived control over the preservation process. Figure 4 shows our *<Recorder>* that continuously captures the screen and title of applications that the user selected for recording. Though we need to further evaluate the *<Recorder>*, it is clear that researchers want to control capturing and sharing. This conflict between exercising control over the preservation process and desired automated preservation requires further study.

Integration. The landscape of connected devices that measure, generate or process scientific data is large and diverse. Devices range from desktop computers to microscopes and sensors. Integrating all those data sources into the preservation process poses further challenges regarding user control and system architectures. As depicted in Figure 5, some devices will implement URP strategies. And even though our examples and developments are mostly limited to computer applications, a wide variety of connected devices can offer URP by directly communicating with repository servers. Other devices can be connected to URP technology, which acts as a proxy in the preservation process.

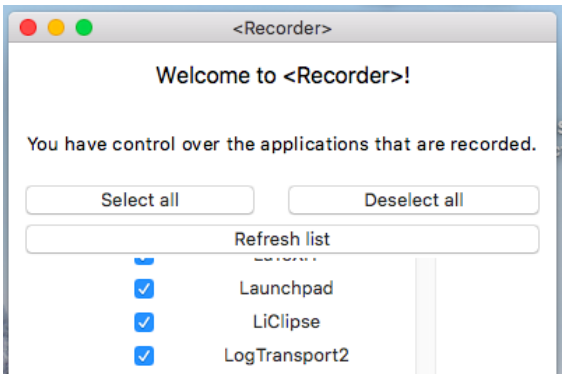

Figure 4: The <Recorder> continuously captures screens and titles of applications that a user selected for preservation.

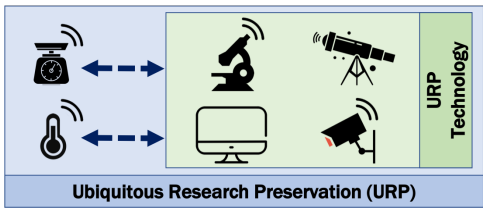

Figure 5: URP technology needs to *integrate* data sources from various types of devices or even *be part of* those devices.

4 DISCUSSION AND CONCLUSION

We described our past and current efforts aiming to spark discussions and further research on machine-automated preservation in computation-based science. We illustrated a broad spectrum of technology interventions that we refer to as *Ubiquitous Research Preservation (URP)*. We expect URP to make a positive impact on researchers' ability to reflect on past processes, to provide training material and to improve the reproducibility of their work. Yet, we do not intent to oversimplify complex use cases. Preservation is a first step towards supporting those, but it is not the only requirement. In particular, the decision to share resources does not only depend on the effort to preserve data, but on various other factors, including competition, fear of judgement and privacy policies.

We described four major research challenges, crucial for the design and acceptance of URP technology. *Usefulness* and *control* will be crucial for the acceptance and use of URP systems. *Generalizability* needs to be considered, to provide fast and wide access to URP tools and to include even branches of science and organizations that find it challenging to spend considerable resources on the development and adaptation of URP systems. Finally, the diverse landscape of connected, data-producing or data-processing devices needs to be *integrated* into URP systems. Developments and URP architectures must not be limited to computer applications.

Our research focuses on computational science, as automated, machine-supported knowledge preservation promises to best map experimental processes and resources. Yet, as all science became to varying degrees connected to computation, we expect URP to profit scientific domains beyond computational science. Similarly, URP is likely to impact technology users well beyond science.

ACKNOWLEDGMENTS

This work has been sponsored by the Wolfgang Gentner Programme of the German Federal Ministry of Education and Research (grant no. 05E15CHA).

REFERENCES

[1] Christine L Borgman. 2007. *Scholarship in the digital age: information, infrastructure, and the internet.* MIT Press, Cambridge.

[2] Florian Echtler and Maximilian Häußler. 2018. Open Source, Open Science, and the Replication Crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, alt02.

[3] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Paweł W. Woźniak. 2019. Designing for Reproducibility: A Qualitative Study of Challenges and Opportunities in High Energy Physics. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI'19* (2019). https://doi.org/10.1145/3290605.3300685

[4] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Paweł W. Woźniak, and Albrecht Schmidt. 2019. Gamification in Science: A Study of Requirements in the Context of Reproducible Research. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI'19* (2019). https://doi.org/10.1145/3290605.3300690

[5] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E. John, and Brad A. Myers. 2018. The Story in the Notebook: Exploratory Data Science using a Literate Programming Tool. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI'18* (2018), 1–11. https://doi.org/10.1145/3173574.3173748

[6] Gerard Oleksik, Natasa Milic-Frayling, and Rachel Jones. 2014. Study of electronic lab notebook design and practices that emerged in a collaborative scientific environment. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14* (2014), 120–133. https://doi.org/10.1145/2531602.2531709

[7] Victoria Stodden and Sheila Miguez. 2013. Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. (2013).

[8] Aurélien Tabard, Wendy E Mackay, and Evelyn Eastmond. 2008. From individual to collaborative: the evolution of prism, a hybrid laboratory notebook. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*.