Digital Preservation as a Science

Artur Kulmukhametov
Vienna Univertity of Technology
Vienna, Austria
artur.kulmukhametov@tuwien.ac.at

ABSTRACT

Digital Preservation is a domain within ICT, that is mainly based on empirical knowledge. One way to introduce theoretical fundamentals is to study accumulated experience and best practices, and expose a common framework from them. In this paper, an idea of having metrics as common objective assessment of capabilities of tools, environments and processes is described. The goal is to have DP well-described and structured field, where main terms, decisions and processes would be theoretically stated and proven, just like the science does. In order to identify challenges coming with the idea, research questions are stated.

Keywords

digital preservation, tools, business processes, capabilities, metrics

1. INTRODUCTION

Digital Preservation (DP) is an emerging field within ICT, that has it's own unique set of features, goals, tasks and challenges. DP is dissolved in Computer Science, Software Engineering, Business Informatics, Management and Economics. There are different people involved from different domains with own goals, requirements, tasks and problems. All this affects to a broad poorly structured domain.

The idea is to focus the attention of the DP community to the problem of absence of the domain's fundamentals. To date, there is no common theory of preservation, tools, ground truth, benchmarks, not even a common vocabulary.

As a question of creating theoretical grounds is raised, it is useful to refer to works that describe different aspects of theories, their types, behaviors, relations to other theories and limitations. Shirley Gregor [Gre06] introduces a taxonomy of information systems theories. It is suggested, that a selection of a type of theories while development of one influences understanding of phenomena and approaches to be chosen and applied.

There is a European project, called SCAPE, that produce tools, services and work-flows for the DP community. These are necessary elements and enablers that make DP possible and operational. But also the mentioned problem, the lack of theoretical knowledge and systematic scientific approaches in DP, is noticed and brought to light in SCAPE [BPS$^+$12]. More precisely, experimentation, simulation, prediction, hypothesis testing and model building are addressed as key aspects that are poorly referred in the domain. The most attention is paid to experimentation part, benchmarking: key elements, barriers and research questions are defined.

Also there is another project called BenchmarkDP that is developing a systematic approach to assess and compare processes, systems, organisational capabilities within DP[PVA$^+$13]. This project is not alone, as there are people, such as in [Moo08], addressing gaps in the domain. They are making attempts to characterize DP, explain basic principles and define relations within the domain.

2. RESEARCH CHALLENGES

Having a lot of approaches that are solving different problems is a great thing. DP is getting rich with tools, methods, systems and processes at hand. However an organisation has to decide on their own what is an appropriate solution for them that fits their requirements. Unless there is a well-described, structured framework that could help defining useful properties and measure capabilities of these solutions, provide guides how to evaluate them on which datasets etc. So that the organisation could choose the best solution according to it's goals.

What is needed in order to do DP successfully?
How to measure it objectively?

Obviously, DP is a complex domain, where addressing only technical issues is not enough. It is necessary to pay attention to business processes of institutions, their requirements, goals and policies[BR11]. There should be ways to measure metrics, on one side, of software and tools and, on another side, organisational capabilities and processes. These are two different aspects of DP, but definitely they have relations. Although, usually these relations are not easy to define explicitly and they may differ from institutional requirements. An observation doesn't have to follow an expectation.

Having an ability to expose this correlation, features and dependencies within will help describe solutions and processes objectively and independently on others. Answering

IPRES '13 Lisbon, Portugal

these questions will allow organisations get an expertise of how they are doing DP. But is it not enough? Also it would be great to know *how well* they are doing it. Assume there are two preservation institutions and they were both studied according to our objective examination. There are two surveys, and a goal is to select an institution with a higher probability that data will be safe for 50 years. Then we are coming to the following questions:

Do we know how to compare results of an objective examination?
Which tool is better for given goals?
What about comparison of metrics for evaluation of business processes?
How well an institution is doing DP in achieving it's goals?

Within the current research the problem of discovering and specifying metrics, that would reflect previously mentioned properties, is addressed. This would allow the community to have a systematic framework. It will help denoting key metrics of processes, work-flows, software solutions and their possible values. By gaining additional knowledge the framework will allow giving an expertise of how to assess and improve mentioned processes.

For example, speaking about DP tools, we are working on verifying of a correctness of digital image migrations. Having this quality assurance tool will make it possible to get an assessment (a normalized value from [0,1]) for a migration tool under review. Furthermore, having several results, a ranking of tools is feasible with respect to specified requirements, such as:

- the best tool that covers a migration from 10 most popular image formats, or
- the best tool that migrates from JPG, and so on..

However this is the easiest metric to discover.

What about other metrics?

Regarding processes, having results of such measurements it is becoming feasible to choose a best way among existing ones. In general, it would lead to a proven and verified strategy for capabilities improvement within an organization, ability to choose from existing approaches, that were already verified and proved their effectiveness.

Also answers to the questions and accumulated knowledge would allow building theories that would explain why certain events or issues occur, under which circumstances and what are solutions for them. That is extremely important for Preservation Planning, where making right decisions is difficult, but still is a crucial task.

3. CONCLUSIONS

Currently, research in DP is driven mainly by empirical knowledge, obtained from experience, case studies and lessons learned. This approach is not bad, as it collects all best practices. However, key factors of DP as a science, such as repeatability and reproducibility, are not confirmed yet. There are no solid experimentations conducted and recognized in the DP community. Organizations usually try to solve their DP tasks on their own rather than using already existing well-defined solutions. There is a risk, that DP will be just a collection of use cases without much understanding of core mechanisms and ability to explain them. Unless, there will be changes introduced.

4. ACKNOWLEDGEMENTS

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

5. REFERENCES

[BPS+12] Christoph Becker, Norman Paton, Rainer Schmidt, Natasa Milic-Frayling, Andreas Rauber, and Brian Matthews. D3.1 open research challenges and research roadmap for scape. http://www.scape-project.eu/deliverable/d3-1-open-research-challenges-and-research-roadmap-for-scape, December 2012.

[BR11] Christoph Becker and Andreas Rauber. Preservation decisions: terms and conditions apply. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 67–76. ACM, 2011.

[Gre06] Shirley Gregor. The nature of theory in information systems. *Mis Quarterly*, 30(3):611–642, 2006.

[Moo08] Reagan Moore. Towards a theory of digital preservation. *International Journal of Digital Curation*, 3(1):63–75, 2008.

[PVA+13] Diogo Proença, Ricardo Vieira, Gonçalo Antunes, Miguel Mira da Silva, José Borbinha, Christoph Becker, and Hannes Kulovits. Evaluating a process for developing a capability maturity model. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 1474–1475. ACM, 2013.