A Biological Perspective on Digital Preservation

Michael J. Pocklington
Department of Genetics
University of Leicester
Leicester LE1 7RH, UK
m.pock@me.com

Anna Grit Eggers
Göttingen State and University Library
Georg August Universität
37070 Goettingen, Germany
eggers@sub.uni-goettingen.de

Fabio Corubolo
IPHS, University of Liverpool
Waterhouse Building
Brownlow Street
Liverpool L69 3GL, UK

Jens Ludwig
Göttingen State and University Library
Georg August Universität
37070 Goettingen, Germany
ludwig@sub.uni-goettingen.de

Mark Hedges
King's College London
Strand
London WC2R 2LS, UK
mark.hedges@kcl.ac.uk

Sándor Darányi
Swedish School of Library and Information Science
University of Borås
Allégatan 1, 50190 Borås, Sweden
sandor.daranyi@hb.se

ABSTRACT

Successful preservation of Digital Objects (DOs) ultimately demands a solid theoretical framework. Such a framework with a high degree of generality emerges by treating DOs as containers of functional genetic information, exactly as in the genomes of organisms. We observe that functionality links survival in organisms and utility in DOs. In both cases, functional information is identifiable in principle by the consequence of its ablation. In molecular biology, genetic ablations (mutations) and environmental ablations (experimental manipulations) are used to construct interaction maps fully representing organismic activity. The equivalent of such interaction maps are dependency networks for the use of DOs within their Digital Environment (DE). In the poster we will present early work on the application of the theoretical background. It includes first results from a case-study examining a software-based art preservation scenario (SBA) developed as part of the PERICLES FP7 project [1].

General Terms

Theory of digital preservation, preservation strategies and workflows.

Keywords

Digital ecosystems, digital preservation, niche, interaction map, significant environment information, sheer curation.

1. INTRODUCTION

Many active research programs exploit equivalences between biological objects and digital objects, up to and including, in the position taken by strong artificial life, the assumption of indistinguishability. The latter follows from the recognition that life is not dependent on any particular underlying medium, but is instead a property of evolving information-processing structures [2]. DP can not avoid such a viewpoint by internalisation and a retreat to technical issues, since it is embedded within policies and technologies that are themselves subject to the most rapid type of evolutionary change.

Scholars of culture have long debated the existence of autonomous informational processes in human society, and it seems likely that these become entangled with DOs, which inevitably evolve as technology advances. This brings issues for DP that may be best considered from a biological perspective. This is not merely a conceptual position: informational viruses and instant stock-trading algorithms can not be ignored, and seem to possess an autonomous evolutionary status. Despite repeated attempts at a generalised biological or Darwinian perspective of human organisational entities such as DOs, no consensus has been reached, even as to the best way to proceed.

DP is uniquely in a position of having to deal with DOs across the entire realm of human activity; they replicate, behave, consume resources, mutate, get selected, and evolve, demanding a meta-view of biology-based informational concepts. A key element for such a meta-view can be provided by systems biology. Systems biologists have found a way of visualising functions such as biochemical pathways and behaviours by interfering with genetic and environmental information, revealing the underlying structure of that information, in the form of genetic interaction maps. Similar methodology could be applied to DOs, to the benefit of their long term use, and reuse.

2. THEORETICAL FRAMEWORK

Underlying the existence of all biology is the specific context enabling organisms to survive, which is their niche. To call this an "environment" would be glib, as the niche is more than a regional container, but a specified provision of resources contingent upon the appropriate behaviour. We can operationally define information allowing survival by removing it one piece at a time. Traditionally we would call the removal of information "mutation" if a change was made to the genome, and "experimental manipulation" if it was made to the niche. Generally we may call such perturbations *ablations*. Equivalently, a philosopher might talk of *counterfactuals*, i.e., what would happen if such-and-such an element of a system were missing. This is what is done in the high-throughput molecular biology laboratory. Large numbers of ablations are produced independently and in pairwise combination, allowing the definition of genetic interaction maps, defining the underlying information-processing structure of the organism. If we make enough independent recordable ablations, we can operationally define all the informational elements comprising the organism. Notice how ablations define the information that matters, that which confers real meaning -life or death - to the organism. Just as importantly, the procedure defines the ablations that do not matter.

If we can do enough experiments (i.e, with enough independent ablations) we can achieve full definition of the organism as an informationally closed entity. In other words, if we could continue obtaining ablations, we would reach a point where we would get no new ones. Any suggestion to the contrary would be to posit the inability to obtain an ablation, and this objection would be self-defeating, since if an ablation could not occur, it could not have an affect on the entity. Similarly, any objection as to the ability to define the niche takes the objector beyond the agreed definition of the niche in question.

We suggest the same process could allow the visualisation in principle of the dependency networks for the use of DOs have the appearance of genomic information; indeed, we could ground our position in the following example, in which we obtain a definite genetic interaction map for a known DO.

Let us consider a DO which is an actual recorded DNA sequence, such as the yeast genome obtained by DNA sequencing methodology, currently found within a digital library such as GenBank [3]. We could in principle synthesise DNA from this digital library information [4], insert it into a yeast cell lacking its own DNA, and use this cell to inoculate a culture within a growth chamber (its niche). If this culture performs its usual behaviour within its niche, it verifies the authenticity of the digital information. The test is straightforward: we do not need to know what behaviour to look for, we just compete the yeast (culture) containing the synthesised DNA with the wild or natural yeast (culture).

We could obtain a genetic interaction map of this DO, by performing ablations on the digital sequence of the object as well as on its digital environment, to figure out the boundaries of its Significant Properties (SP). Unimportant environment information disappears in this process, but environment information that matters - Significant Environment Information (SEI) - crystallise out in the interaction map, especially that information that influences the SP of the DO.

We could then perform the same test for functionality as above, and get the same map as we obtained using ablations in the natural DNA. This conceptual procedure would tell us that the DO maintains its significant properties; but its utility here is not in merely confirming the functional content of the information. Instead, it forces the recognition that the information in the DO, if it is to have utility at all, is exactly that prescriptive information that is contained in the natural DNA; the DO and the information in the genome are one and the same thing, by operational definition.

Thus, at least for yeast, the DO corresponding to the genome can be rendered into a useful map by ablation. In this case the recognition is made clear by employing the known identity between a DO and an organism's DNA sequence. Benefit may be gained by application of the heuristic. We *know* that every DO has a definite niche in its usage dependencies, just like an organism; we *know* that every DO comprises prescriptive information, SP, SEI and other bits in the DO and the environment, just like the genome of an organism; we *know* that just like in the genome, some of that information is historical nonsense, the "other bits", while other information (SP and the SEI) is crucially important. Crucially, this depends, on where we decide to throw the boundaries of the niche, which this perspective forces us to be clear about.

We analysed the kinds of information extractable from a DO and its environment to improve its chances of being useful in the long term. By this perspective we came to the conclusion that it is possible to extend the SP framework beyond the DO to its environment. This is the SEI [5] for a DO, defined as all the information needed, based on a particular purpose being addressed, to make use of it. Thereby SEI is a broad super-set of the existing SPs from where we adopt the concept of intended purpose, but extended to the whole DO environment and not just for the DO's intrinsic properties. See Fig.1.

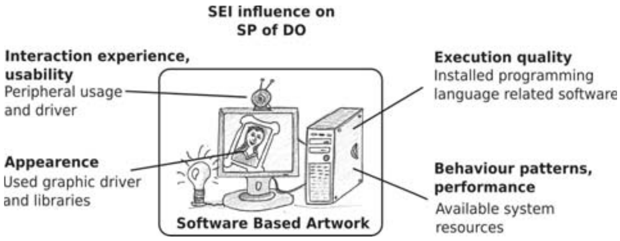
SEI influence on
SP of DO
Interaction experience, usability
Peripheral usage and driver
Appearence
Used graphic driver and libraries
Software Based Artwork
Execution quality
Installed programming language related software
Behaviour patterns, performance
Available system resources

Figure 1. SEI influences SP

Further we exemplified the above finding on a real-life software-based art preservation scenario using PERICLES Extraction Tool [5], a tool to extract SEI from the DE of a DO in a sheer curation [6] scenario, to improve the DOs reuse and the preservation of its SP that are influenced by the SEI. Sheer curation is a parallel to DE where organisms cannot be observed reliably outside of their niches, this resulting in an unavoidable loss of important information. To map their connectedness, a software agent observed and collected information about interactions between the DO and its immediate surroundings. By observing such interactions one can obtain a series of observations for further analysis and recognise functional dependencies. Such information cannot be reliably reconstructed after the DO is archived. It has to be extracted from the "live" system when the user is present, and preserved together with the DO.

Our theoretical model is visualised with the aid of this example on our corresponding poster.

3. ACKNOWLEDGMENTS

This work was supported by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES.

4. REFERENCES

1. http://pericles-project.eu/
2. Fernando, C., Kampis, G., and Szathmáry, E. 2011. Evolvability of natural and artificial systems. In *Proceedings of the European Future Technologies Conference and Exhibition.*
3. GenBank ®: http://www.ncbi.nlm.nih.gov/genbank/
4. http://www.ncbi.nlm.nih.gov/genome/15
5. Corubolo, F., Eggers, A.G., Hasan, A., Hedges, M., Waddington, S., and Ludwig, J. 2014. A pragmatic approach to signifcant environment information collection to support object reuse, in IPRES 2014 proceedings.
6. http://alimanfoo.wordpress.com/2007/06/27/zoological-case-studies-in-digital-curation-dcc-scarp-imagestore/