The ecology of longevity: the relevance of evolutionary theory for digital preservation

Doorn, Peter
peter.doorn@dans.knaw.nl
Data Archiving and Networked Services, Netherlands

Roorda, Dirk
dirk.roorda@dans.knaw.nl
Data Archiving and Networked Services, Netherlands

Software and data can be considered as digital organisms that function in a "digital ecosystem" of computers. The concept of ecology has been borrowed from biology by other disciplines for explaining or describing a variety of phenomena. In some cases ecology and other concepts from evolutionary theory are only used as metaphors; in other cases attempts have been made to apply an adapted version of the theory to the evolution of non-biological phenomena.

We think it makes sense to borrow the notions from evolutionary theory in thinking about digital longevity. In this paper we will explore the potential of Darwin's theory as an explanatory framework for digital survival.

The construction of a theoretical foundation serves to answer questions of why some criteria or characteristics will guarantee the "survival" of digital objects better than other ones. Taking an evolutionary view will also make clear that there is no such thing as digital permanence for eternity: some objects only have better *chances* to survive than other ones. In computing technology, there is a struggle of survival of the fittest going on. In this struggle, new technologies arise as modifications or adaptations of earlier technologies and the older ones die out when newer technologies are stronger or better suited for their tasks. The digital objects that are already in existence have to be adapted to the new technological surroundings, otherwise they become extinct.

Spencer originally coined the phrase "survival of the fittest" in 1864, drawing parallels between his ideas of economics and Darwin's theory of evolution, which is driven by "natural selection". With respect to digital objects, it is people who are actively or passively involved in making the selections, thus deciding which digital objects survive and which do not.

As electronic digital data can only be understood using computers and software to "translate" them into visible or audible form, the media and data formats that are specific for hard- and software change according to the same evolutionary principles. If we accept this view, we can start to ask ourselves: which characteristics (comparable to the "genetic properties" of living organisms) may influence the chances of digital survival of data? This is however not unproblematic, as it is typically with hindsight that we see which (traits in) a biological species have survived and how the evolutionary process took place. The explanatory power of evolution theory is a posteriori, not a priori. It may therefore be difficult to predict which traits are good for survival.

We will explore the possible use of the concept of evolvability, which is usually defined as the ability of a population of organisms to generate genetic diversity, hence giving a measure of an organism's ability to evolve. Maybe there is a parallel here with respect to digital objects. For instance, if we look at several formats for Microsoft Word (.doc, .rtf, .html, .xml (2003), .ooxml) then we see an increase in usability/interchangeability, and hence probably: evolvability.

We can break down the survival problem to questions concerning:

- The physical attributes of the media (tape, disk, etc.)
- The media format (density, size, etc.)
- The data content (integrity of the bits and bytes)
- The data format (the structure of the bits and bytes)
- The metadata content (the substantial description of the data)

- The metadata format (the format in which the metadata is described)
- The interlinking (the degree to which data is linked both internally and externally); a web of interlinked information is an ecosystem of its own.

We will demonstrate how digital preservation strategies such as technology preservation, software emulation, and data migration fit in an overarching evolutionary framework. The ecological approach also shows that it makes no sense to try to express the time horizon for the preservation of digital objects as a specific or indefinite period of time, but that we can better think in terms of "chances of survival".

The evolutionary framework can be used to argue why certain attributes and formats are more likely to survive than others. Also, analogous to natural selection, we will make clear that there is no single "best" strategy for survival of digital data. Some factors simply increase the chances of digital longevity, whereas other factors reduce these chances. Good factors for longevity may be bad for other desired characteristics. For example: stripping executable information from data improves its longevity, but hinders its functionality. It may also be so that some factors are intensely ambiguous for longevity. We may now think that "wrapping" text in WordPerfect in the 1990s was (with hindsight) not so good for survival, and that packaging it in Microsoft Word seems acceptable. This is probably related to the status (or market dominance) of the software packages. Similarly, packaging data in SGML in 1990 might have been not so good, while packaging it in XML in 2009 seems excellent. In the end, the environment determines what was good and what was bad for longevity.

So, when the whole "technological ecosystem" changes, what was well adapted before the change may appear to be ill suited in the next technological phase. Digital preservation strategies can use the principle of digital selection in order to maximize the adaptation of digital objects to their environment, thus increasing their chances of digital longevity.

Whether it makes sense to apply evolution theory to digital curation can be studied by looking at a number of parallels in other scientific domains. We will deal briefly with attempts to use Darwin's ideas in the social sciences and in technology. In the social sciences the idea of a "social ecology" was already applied and empirically tested in the 1920s by, among others, Robert Park and Ernest Burgess of the "Chicago School" of urban ecology. With respect to man-created systems it is probably better to use the ideas on evolution by Lamarck. Lamarckism is the idea that an organism can pass on characteristics that it acquired during its lifetime to its offspring (also known as heritability of acquired characteristics or soft inheritance).

Several researchers have proposed that Lamarckian evolution may be accurately applied to cultural evolution. Human culture can be looked upon as an ecological niche-like phenomenon, where the effects of cultural niche construction are transmissible from one generation to the next. Ecological notions on the evolution of software, in which ideas and characteristics of programming languages compete with each other, have been formulated in information science. Inheritance is an important concept with an evolutionary basis.

The development of open source software has also been described as evolving in a Lamarckian fashion. Ensuring free access and enabling modification at each stage in the process means that the evolution of software occurs in the fast Lamarckian mode: each favourable acquired characteristic of others' work can be directly inherited.

Kauffman and Dennett point out the parallels between biological evolution and technological evolution. They distinguish two stages: (i) explosion of the number of greatly different designs when there are still many unoccupied niches; (ii) microevolution where the existing designs are optimised for competition in existing niches.

It is also useful to compare the selection and survival of digital information with that of analogue information. In both cases there is "information selection" and evolution. What makes the digital world so different from the analogue world?

Next we will treat a few examples of the evolution of computing technologies, software, file formats and data sets, which will illustrate how well evolutionary theory is suited for

explaining what has happened empirically. It makes sense to look backwards and use evidence from the – still very short – historical evolution of computing technology since the 1940s and '50s. Can we still read the first image, the first word processor file, the first database, the first web page, email or pdf-file? If yes, how come this is still possible? If not, why? And how did the formats of those data types (probably equivalent to the taxonomical rank of the *genus* in biology) evolve, and which abandoned file formats ("*species*") can still be read today?

Maybe there is a manifestation of genotype/ phenotype here, where the application can be considered as the phenotype. Applications struggle with each other in the "econosphere" of consumers. The surviving application dictates the data format. If there are two strong surviving applications in a domain, you get peer-to-peer data convertors. If there are many, weaker survivors, you get interchange formats, which are better for preservation. A familiar lesson is not to rely on monopolists.

Finally, there is a marked tendency among data curators to set criteria for "trusted digital repositories", the nature reserves of the digital world. On the basis of the evolutionary ideas expressed in this paper, it makes sense not to make such criteria too narrow. They should be chosen in such a way that they make use of the "natural" evolutionary processes of technology and digital objects, making sure that what is threatened by extinction can be rescued in an effective way.