When is my information going to die?

Kresimir Duretec
Vienna University of Technology
Vienna, Austria
duretec@ifs.tuwien.ac.at

ABSTRACT

The switch from the paper to the digital form did not make our information accessible for ever but it has just introduced new threats. The question ”When is my information going to die?” still remains unsolved.The understanding of the two threat levels (physical and logical) to which every information stored in a digital form is exposed can enable a better understanding of the information life expectancy. While there is a good knowledge on how to address the problem on the physical level still there are significant gaps in work explaining how to do the same thing on the logical level. This work does not present a solution to the problem but only starts to define research questions. It begins with defining the top level question ”How can we calculate the life expectancy of information?” and then listing a series of sub question which need to be considered.

Keywords

Digital Preservation, Information Life Expectancy, Resources, Format Obsolescence

1. INTRODUCTION

Over the centuries librarians are determined to keep valuable information as much accessible as possible. The result of their work is an understanding of factors that shorten the information life span and measures how to prevent this. Because paper was used as a main information carrier, lots of the factors describe how does it react to different environmental conditions. From that knowledge, optimal conditions are devised to keep information available over longer periods. In the last decades, a major trend in storing information is the digitalization. While some paper documents and books are going through the process of digitalization, most of them are already born digital. Unfortunately, the switch from the paper to digital did not solve the main problem and made information accessible for ever. All it can be said is that it made some threats less important and, more importantly, introduced new ones.

Information stored in a digital form is exposed to two levels of threat to become lost: physical and logical. When a file is damaged on a physical level it means that it is not possible to retrieve the correct bits from the file. If it is damaged on a logical level it will mean that it is possible to get the correct file content but the information that it contains will not be understandable.

Physical threats rise from a hardware hosting the content because each hardware component despite its quality has a limited life span. That life span is expressed as the MTTF (Mean Time To a Failure). Taking MTTF in to an account, designers can calculate what is the expected time for the whole storage to lose some data. Still, they need to consider additional factors like fires and earthquakes which can also cause damaged on the physical level but are not modelled within the MTTF. In [2] a system is described that calculates the expected MTTF of a whole digital repository.

The logical level is more harder to understand. Even though all bits are unchanged in their places it still does not mean that a user will be able to access wanted information. This is because there could be a technological gap between the technology the user is using to access the information representation and the technology used for creating it.To prevent this content holders like national libraries and archives will take a number of precaution measures. One measure is the migration where a content created in one technology is transferred to another technology. Reasons for that are multiple and often ones are reducing the resources needed and changing the content to a more reliable technology. Because of the number of available possibilities preservation planning[1] is used to determine the optimal action. Furthermore in [3] a tool capable to estimate resource usage in future is explained.

While the physical level is covered with the MTTF as a help to estimate the time when certain information will be lost, there is no such measure on the logical level. Understanding the logical level is of the same importance and protecting information only on the physical level is not enough.

In this work, the problem of estimating life expectancy is addressed by listing research questions. The goal is not to list every possible question and give answers to them, but to guide possible future research to help in a better understanding of the information life expectancy and approaches to solve the problem of estimate it.

In the next chapter a list of possible research questions is given. It starts by defining the top level question and then a series of sub questions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IPRES 2012 November 2012, Toronto, Canada
Copyright 2012 ACM X-XXXXX-XX-X/XX/XX ...$10.00.

2. RESEARCH QUESTIONS

In this chapter the list of research questions is defined. They are expected to guide the possible future research in the area of the information life expectancy.
For the begin the top level question can be defined as:

How can we calculate the life expectancy of information?

To get a clear look at what is meant under the life expectancy a simple scenario is presented. Lets say that a person decided to use a tool X to create a document. Soon after the document was created this person finds out that the tool X is the only tool that can open and process the document. What is even worst there are no tools available that can migrate the document to another format. Soon the support for the tool X disappears and information stored in the document is doomed to be lost. It can be said that the life expectancy of information stored in that document is quite short.

The first step in better understanding the problem would be naming and understanding potential factors that affect the information accessibility. Therefore it is reasonable first to ask a question

What are the factors that affect information accessibility?

These factors will depend on the use cases. They can be divide to internal (to an organization) and external and here without trying to give a complete answer to the question a few of them are listed.

Limited resources

The first factor are limited resources. They affect the information life expectancy indirectly. Content holders could take some actions to optimize their resource usage but on the other side those actions could damage stored data. If they were not limited content holders could simple keep their content in every possible format and that would keep them on the save side of the information accessibility problem. There are two aspects.
The first aspect is the needed disk capacity. Several questions rise here.

What is the capacity needed to host a collection now and what will it be in the future?

There are two ways the needed capacity can change. A collection of documents can have a growth rate meaning new object are ingested over time. The question here is

How can ingest be predicted over time?

By solving this question one part of the collection growth can be solved. The remaining part is a result of preservation actions. It is known that at some points in future collections will be migrated to another formats. The question here is

How do preservation actions affect collection size?

Also different strategies like keeping all document copies or deleting everything except the originals and the current access copy affects the capacity so it is worth answering

How do different strategies affect the needed capacity over time?

The second are the computational resources. Every preservation action requires some time and because of the collection size the time needed to be executed on a whole collection can not be seen as a single point in the future. The limited computational resources also raise some questions. For example how to decide

What is better in terms of computational efficiency? To migrate objects upon ingest or to do it at certain periods?

The simple collection growth will also lead to an increased need for the computing power.

What is the expected computational load over time?

Answering those and similar questions content holders can get a good understanding of the needed resources. They could also consider option to outsource some of their operations to some cloud providers.

Format and component relationships

There is a significant difference between information stored in a format that has only one tool which is able to process it and no possibilities to migrate it to another format and information stored in a format which has a number of tools that can process it and has a number of possibilities for a migration. Further more in the first case the life expectancy of information is quite simple to calculate while in the former case it is not clear how to calculate it because of complex relationships. As a start a question can be posed.

How can we model relationships between formats and components which are capable to process them?

Once modelled these relationships will not be static. The old tools will disappear and the new ones will appear. Same is valid for formats. This will also affect the life expectancy so the question

Once modelled how can we calculate relationships in the future?

needs to be answered as well.
Before answering that question format/component obsolescence needs to be addressed.

How can we calculate expected format/component obsolescence periods?

To calculate obsolescence periods a better understanding of the obsolescence itself is required.

What is obsolescence formally and what are its causes?

What does it mean one format is obsolete? Is it when nobody is using it or when the number of users is under certain threshold. What are the causes for a format to become obsolete? Is it a simple appearance of a new format or users need some properties that the old version does not have and the new one has?

Preservation actions quality

Last factor that will be presented here is the quality of preservation actions. It is a well known fact that each migration results in a loss of information. In some cases that can be neglected and in some cases not.

What is the confidence level that a migration tool produces the object with same properties as the original?.

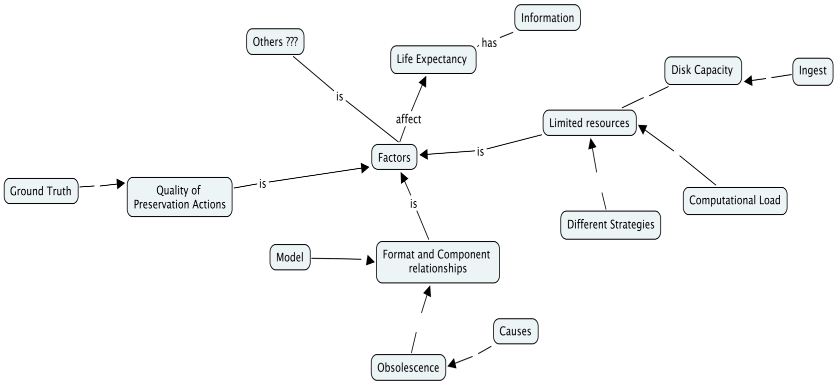

Figure 1: Research question concepts

For example we plan to migrate a text document from a file format A to a file format B using a tool X. If we know that the text document contains only text and that the tool X 100% successfully migrates file format A to the file format B when it contains only text we can conclude that the text document in the file format B will contain the same information as the text document in the file format A. But what if our document contains images and we know the successfulness of the tool when a document contains images is 95%. What is the probability the new document contains the same information as the original one? Now a question can be asked.

How can we automatically generate the knowledge to tell us with which confidence certain migration tool successfully migrates an object with known properties?

This knowledge is currently based on experience and experiments conducted by different sides. Often it is not automatically generated and there exists only limited testing possibilities.
The migration process can not be seen as a simple invocation of one tool. It is a workflow where different tools like characterization and quality assurance tools are called in different steps. The problem which rises here is the verification of these tools. Do they return the right properties or not? It is not possible to use a specific collection for the verification because it is not known which properties have the documents in that collection. It is not possible to use the characterization tools to find out because they are not verified. So the need for collections with known ground truth appears.

How can we create collections of files with known ground truth?

Once characterization and quality assurance tools are verified they can be used to verify migration tools.
Another problem that rises here is how to create meaningful knowledge about tools. Knowing the fact the characterization component failed to correctly characterize certain file is not very helpful information. Also not a single property will cause problems in a characterization but the combination of different properties. Therefore there is a need for more meaningful knowledge about the possible defects in tools.
The question is

How can we make correlations between different combinations of properties and outcomes of preservation tools?

This knowledge would enable calculating the outcome of a preservation action on a certain object with a given set of properties. When this will be possible it will be also possible to understand how a certain document will be affected by the migration process.

3. CONCLUSION

In this work the information life expectancy is presented as a research challenge. The switch from the paper to the digital form which is happening over the last decades did not make the information accessible for ever but just introduced new threat factors and made some less important as before. From the two levels of threat (physical and logical) that every content is exposed to the focus is on the logical level. The reason are major gaps between the work already done. To have a complete understanding of the information life expectancy additional questions need to be answered. This work tries to address them as much as possible. The research challenge concept discussed in this work can be seen in the Figure 1. It is important to note that this is by no means the complete list of the questions and by starting to answer them new ones are expected to appear. Also there is a possibility some of them are already partially or completely answered.
Once answered new insights about the information stored will be gained and content holders will have better opportunities to decide what is the best thing to do for their content.

4. REFERENCES

[1] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *IJDL*, 10(4):133–157, 2009.

[2] A. Crespo and H. Garcia-Molina. Modeling archival repositories for digital libraries. In *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '00, pages 190–205, London, UK, UK, 2000. Springer-Verlag.

[3] C. Weihs and A. Rauber. Simulating the effect of preservation actions on repository evolution. In *Proc. of iPRES 2011*, pages 62–69, Singapore, 2011.