PROPOSING A FRAMEWORK AND A VISUAL TOOL FOR ANALYSING GAPS IN DIGITAL PRESERVATION PRACTICE – A CASE STUDY AMONG SCIENTIFIC LIBRARIES IN EUROPE

Moritz Gomm
FernUniversität Hagen
Universitätsstrasse 1
58097 Hagen, Germany

Sabine Schrimpf
Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt/Main
Germany

Björn Werkmann
FernUniversität Hagen
Universitätsstrasse 1
58097 Hagen, Germany

Holger Brocks
FernUniversität Hagen
Universitätsstrasse 1
58097 Hagen, Germany

Matthias Hemmje
FernUniversität Hagen
Universitätsstrasse 1
58097 Hagen, Germany

ABSTRACT (150-200 words)

In this paper we present a case study and selected results from a research on digital preservation amongst digital libraries in Europe. We propose a framework for gap analysis in digital preservation encompassing the diffusion of preservation practices and the life-cycle of data. We also present a Gap Analysis Tool that we developed to support visual analysis of gaps in the implementation of digital preservation amongst communities. We discuss selected results from the application of the tool in the community of libraries in Europe.

The authors would like to thank Eefke Smit from STM, Jeffrey van der Hoeven and Tom Kuipers from KB, and the four unknown reviewers for their valuable input and feedback. The research presented here was co-funded by the EC (Project PARSE.Insight, FP7-2007-223758).

1. INTRODUCTION

The survey results from the PARSE.Insight Community insight study [6] reveal the status-quo in long-term preservation of digital data in a variety of countries and institutions. The Gap Analysis Framework uses the survey data and matches it against framework elements for supporting the identification and interpretation of gaps between the current situation and what is necessary to enable secure long-term preservation of digital assets, with respect to particular groups of stakeholders. The Gap Analysis Tool (GAT) enables users (domain and preservation experts) to interactively visualize the results of the analysis and allows them to carry out more specific investigations into the highlighted gaps.

1.1. Gap Analysis Framework

We developed a Gap Analysis Framework that encompasses the life-cycle of scientific data (creation and use of data, re-use, preservation, and publishing) and the diffusion of digital preservation within scientific communities (awareness, knowledge, implementation, and commitment). The two orthogonal dimensions form the Gap Analysis Framework and are visualised in Figure 1.

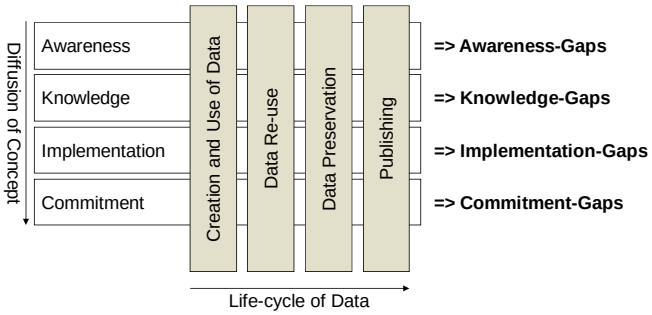

Figure 1: Gap Analysis Framework for digital preservation

While the four phases in the "life-cycle of data" are self explaining, the four aspects of the "diffusion of concept" need to be defined here:

- **Awareness** is the ability to perceive or to be conscious of the problems of long-term preservation in general.
- **Knowledge** is the sum of expertise and skills for the theoretical and practical understanding of long-term preservation issues. This includes knowledge about facts, information and means of long-term preservation.
- **Implementation** is the practical realization of means of long-term preservation including procedures, processes, systems and tools.
- **Commitment** is the willingness or pledge to preserve data.

1.2. Gap Analysis Tool

To support the application of the framework and to analyse the gaps in preservation practices we developed a tool within the EU-funded project "PARSE.Insight" [7]. The Gap Analysis Tool (GAT) analyses survey questionnaires used to gather information on preservation issues and calculate gaps in terms of the framework.

To allow for progressive refinement of search parameters and interactive data analysis [3], dynamic queries [10] and tight coupling [1] information visualization techniques have been employed. This way, immediate feedback is provided to enable interactive data analysis and visual scanning, to narrow down the choice of relevant information objects for a subsequent drill-down.

The drill-down metaphor is based on a "tree visualization" that employs regular expand and collapse operations as well as degree-of-interest [4] based pruning of the tree, to ease navigation. This allows for access to the information domain starting on a category level (e.g. the "awareness dimension" of the framework), down to the level of actual data items (e.g. answers to the survey question "do you have a preservation policy in place").

The user interface of the tool shown in Figure 3 is divided into two areas: The "Filter Setting" to the left, and the "Analysis View" to the right showing the described tree visualization.

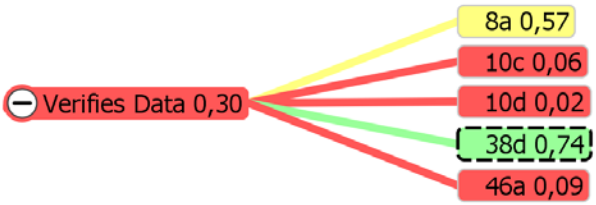

Figure 2: Gap Analysis Tree Detail: Leaf nodes to the right and the containing category to the left

Figure 2 shows an enlarged subsection of the tree with five leaf nodes and the containing category node. Leaves with dashed outline represent survey answers that indicate a problem (or gap), when chosen by a participant. Answers depicted with solid outline indicate that no gap exists. Each leaf node has a label like "8a 0.57", giving the name of the represented answer, followed by a computed gap value. The *leaf gap value* (*lgv*) is computed according to the following formula:

$$lgv = \begin{cases} leaf\ indicates\ gap, & \frac{(pc - ac)}{pc} \\ otherwise, & \frac{ac}{pc} \end{cases} \quad (1)$$

where *pc*, the *participant count*, is the total number of participants of the survey, and *ac*, the *answer count*, is the number of participants that gave this answer. Hence, a high gap value, close to 1, is a good sign, while a lower gap value, close to 0, is indicative of problems within the set of analyzed participants. The set of analyzed participants, and consequently the *answer count* may vary depending on the filter settings as described later.
The gap values computed for the leaf nodes are propagated towards the root node according to the following formula for *node gap values* (*ngv*):

$$ngv = \sum_{\substack{children \\ i\ of\ node}} i.gapValue * i.weight \quad (2)$$

The weights are chosen to sum to 1 to produce the average of the node values of the children. This was the case for all results presented here.
The nodes and edges of the tree are colored based on the gap values and the settings of a color slider. The color slider depicted in Figure 4 image for example, maps

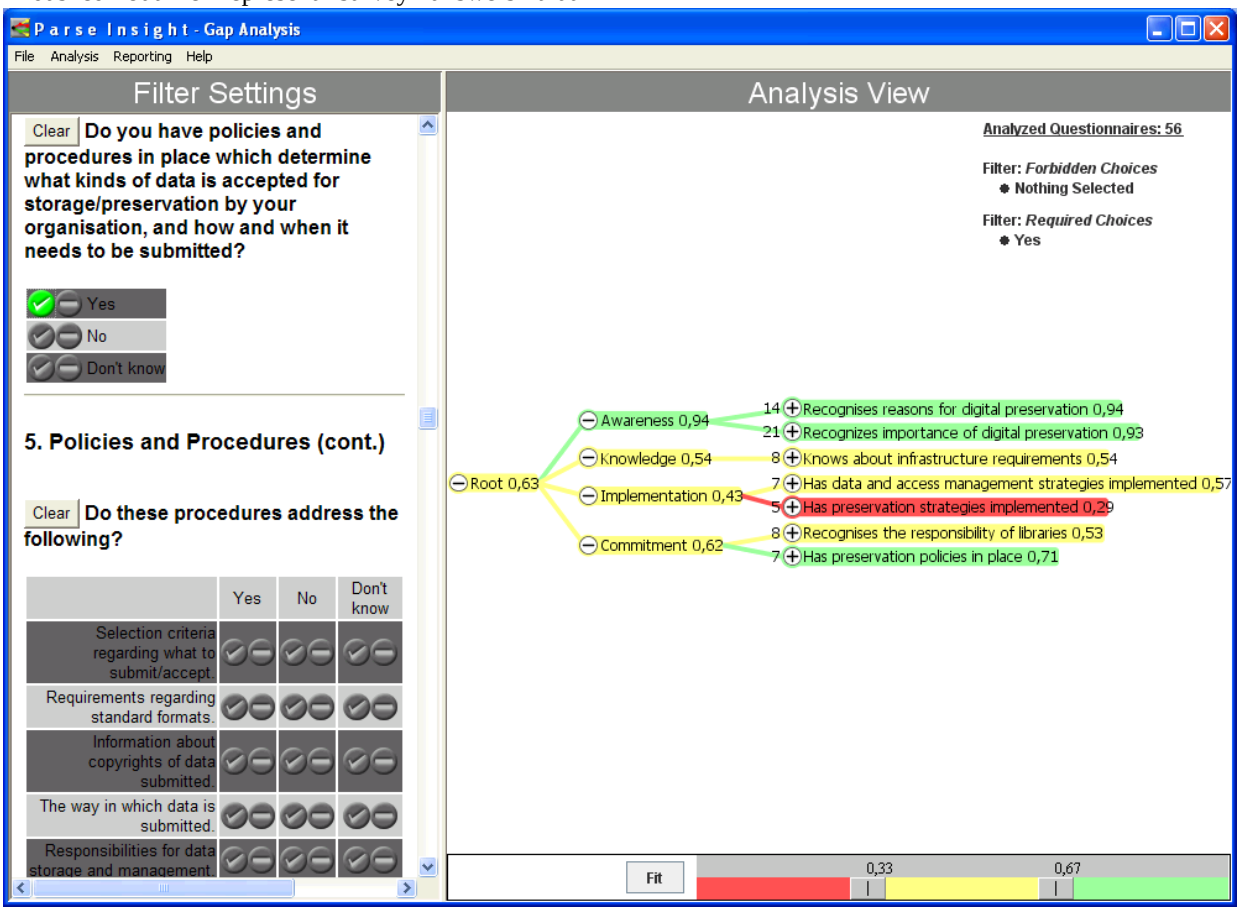

Figure 3: User Interface of the Gap Analysis Tool

values from 0 to 0.23 onto the color red, while values between 0.65 and 1 will be shown in green. Intermediate values will be shown in yellow. This reflects the meaning of the gap value by showing gaps in red, as is also visible from Figure 2

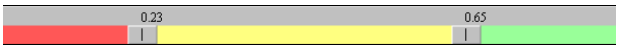

Figure 4: slider

As mentioned above, the set of analysed participants is affected by the *filter settings* view, which shows a list of all survey questions and corresponding options for answers. Modifying these settings, it is possible to change the *answer count*, i.e., the subset of participants that are taken into account for the analysis view. In the beginning all participants are included. By clicking on the check mark button that is associated with every possible answer, the analysis view is configured to take only participants into account that chose the associated answer. This way, certain groups among the participant, e.g. those storing only certain types of data, can be analyzed separately.

2. CASE STUDY

2.1. Introduction

Our tool was applied to a variety of communities. Here we want to present the case study on a survey among Libraries using data from LIBER (Ligue des Bibliothèques Européennes de Recherche). LIBER is the main research libraries network in Europe and encompasses more than 400 national, university and other libraries in 45 countries [5]. The survey was conducted as part of the PARSE.Insight Community insight study [8].

The tree structure for the survey data was modelled by the German National Library (Deutsche Nationalbibliothek, DNB) – a scientific library – and reviewed by individual LIBER members.
In the following we first present assumptions that were drawn from the LIBER-survey using classical empirical analysis methods before applying the Gap Analysis Tool. The results of our analysis were reviewed by individual LIBER members. Review was conducted on a voluntary basis, preceded by a call for review from the LIBER secretariat to the LIBER Working Group on Preservation and Digital Curation.

2.2. Assumptions from survey results

The following results from the survey attracted attention:

a) The great majority of the LIBER libraries recognize the reasons for and the importance of digital preservation. **Awareness** seems to be high.
b) The majority of the participating libraries believes that an international infrastructure would help to guard against the threats of digital preservation (66 %). Furthermore, the majority of libraries is convinced that more is needed for digital preservation, above all more resources, more knowledge, more digital repositories, and more training opportunities. **Knowledge** about digital preservation requirements seems to be high, too.
c) The majority of libraries claim that they do already have policies and an infrastructure in place (59 %). However, only 27 % believe that the tools and the infrastructure available to them is sufficient for their digital preservation objectives, as opposed to 56 % who believe not so. There seems to be an **implementation** gap.
d) The majority of the libraries consider National libraries and research libraries responsible for digital preservation. Additionally, for about 75% of the participating libraries, funding for digital preservation is and will also in the future be an issue. This shows that there is a lot of **commitment**.

The gap analysis tool then was used by the DNB staff to check these assumptions and to render some findings more precisely.

2.3. Preparation of the Gap Analysis Tool

A total of 70 items were identified and grouped according to the framework (see Table 1). The selection of question items and grouping into categories was subject of the review by LIBER members.

Dimension	Sub-categories for survey items
Awareness	Recognises reasons for digital preservation Recognises importance of digital preservation
Knowledge	Knows about infrastructure requirements
Implementation	Has data/access management strategies Has preservation strategies implemented
Commitment	Recognises the responsibility of libraries Has preservation policies in place

Table 1: Sub-categories from the LIBER-survey

2.4. Gap Analysis of the LIBER data

The visualization of the base data gives a slightly different picture from the assumptions above (see Figure 5): Only awareness is – as assumed – marked with a positive gap value (green colour), while commitment is on a modest level (yellow) and knowledge is even low (red). The implementation gap that was assumed can be confirmed.

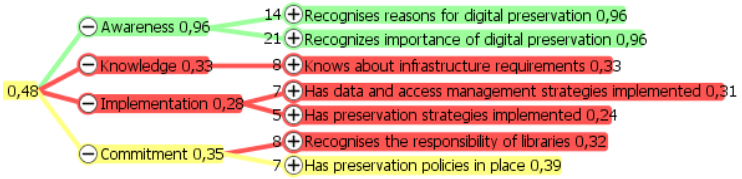

Figure 5: Visualization of the libraries survey data

The experts now analysed the data further by drilling in and out and selecting different subsets of the data which caught their interest. For example if they want to find out

the differences in the community between those who have appropriate "policies and procedures" in place compared to those, that haven´t, they only need to change the filter setting for the corresponding question and compare the visual results (see Figure 6). Other filters that were set include for example:

- The kind of data stored in organisations (data-sets, e-books, or e-journals)
- The volume of data stored in organisations
- Preservation strategy in place
- The kind of preservation strategy in place (migration, emulation, or outsourced to third party)
- Confidence that the organisation's infrastructure will scale with future requirements
- Opinion what is needed to guarantee reliable preservation measures (we distinguished between training, more resources, more repositories/archives)

The time and effort required for the entire analysis was about one personal month plus a few hours of technical support. The external reviews of the results took approximately half a day per reviewer. It should be noted, that the effort was relatively high because feedback from the experts was also used to further improve the Gap Analysis Tool.

2.5. Findings

2.5.1. Policies and Infrastructure

A clear relation between selection policies and the level of implementation and commitment could be shown. Libraries that have thought of what kind of content they add to their collections and documented that in writing in their selection policies are more committed to the task of digital preservation and are better prepared in terms of implementation – although there remains a gap in terms of implemented preservation strategies.

What seems to be important is the fact that there are selection policies in place. The kind of material, however, that libraries collect does not seem to have a heavy impact on libraries' preparedness for digital preservation. No matter if they are focussing on more traditional publication types like e-books and journals or on for libraries unfamiliar data sets – the gaps remains almost the same.

2.5.2. Amount of Data

In contrast, the amount of data that a library currently stores seems to have an impact on the gaps in preservation. The larger the amount of data that a library has to deal with, the smaller the gaps in the area of implementation and commitment are. There is a direct relation between the fact that a library stores data and feels responsible. Another relation could be shown between the amount of data and the implementation of data and access management strategies.

2.5.3. Preservation Strategies

Since a gap is indicated in the area of implementation of preservation strategies in all analyses it is instructive to look in more detail at those institutions that have already implemented preservation strategies in comparison with those that have not implemented the respective strategies.

There is a "commitment gap" in the category "recognises the responsibility of libraries" which cannot be explained

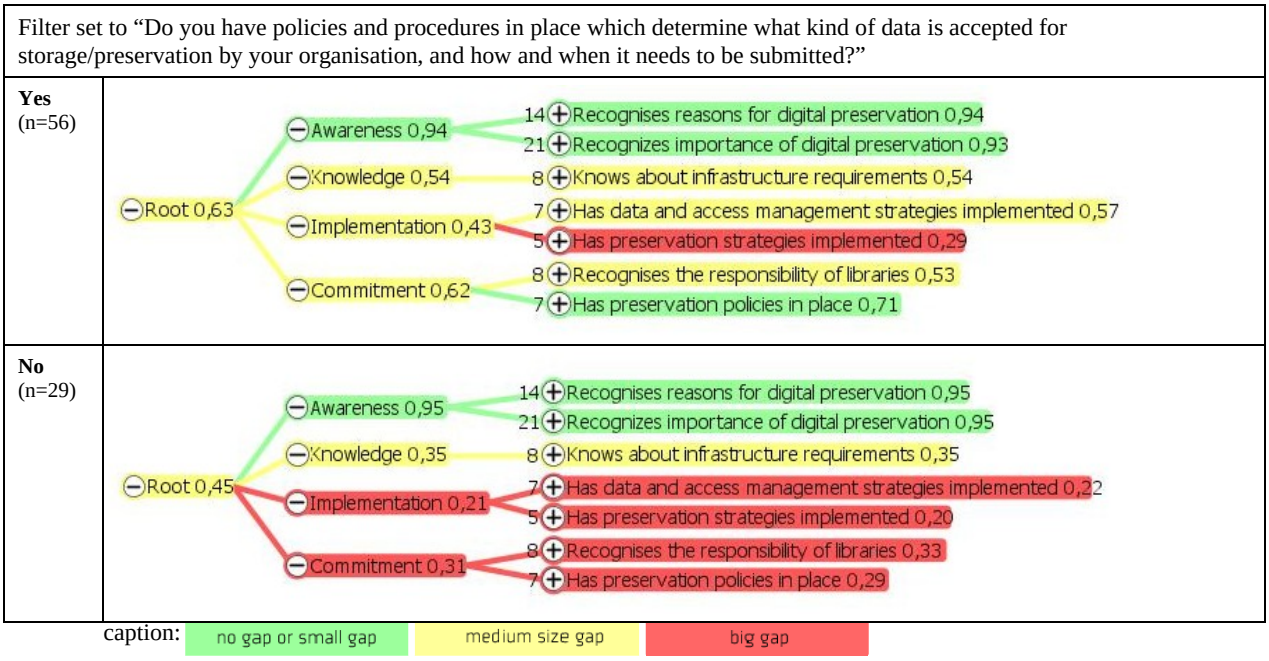

Figure 6: Example of visual analysis: comparing two groups of respondents

easily. The gaps get even bigger when we look at those institutions that state explicitly that they do not have preservation strategies in place. Here awareness remains high, but the values in knowledge, implementation and commitment are significantly lower in comparison to the institutions that have strategies implemented. The relation is obvious: Implementation of preservation strategies requires knowledge, results in implementation and facilitates commitment.

What should concern the library community is the fact that the institutions without implemented strategies are far behind in most categories. In order to catch up, they need to start with building up knowledge.

2.5.4. Scalability

When we compared those libraries that are confident that their infrastructure will scale with future requirement with those that are not so confident, we find a distinction mainly in the areas of knowledge and commitment. Again, the area of preservation strategies catches the eye. While there is only a small gap at those institutions that feel prepared, it is the largest gap at those institutions that feel not prepared for future requirements.

2.6. Summary

The framework and the gap analysis tool allowed deeper insight into the gaps within the scientific libraries community and showed some relation between gaps that were not obvious before.

The first visualization of results indicated larger gaps than could be expected from a simple review of the survey results. It must be acknowledged, though, that many survey participants had skipped many answers that were not mandatory, while skipped answers were counted as negative answers. For future study designs that make use of the Gap Analysis Tool we will take this finding into account and exclude optional questions as far as possible from the surveys.

However, the gap analysis with the tool proved the assumption right that there is mainly an implementation gap, which can be explained with a gap in the implementation of preservation strategies. The gap analysis furthermore indicated a relation between missing preservation strategies and little knowledge and commitment within the respective libraries.

The results also indicated that there is a difference between large and small archiving facilities: The more data a library has to store, the lesser its gaps in the areas of knowledge, implementation and commitment are, hence the better it is prepared for digital preservation. The results indicate in a similar way that libraries with preservation and selection policies in place have smaller preservation gaps than those who have not. The largest difference is between those libraries that have or have not implemented preservation strategies.

Overall, the results indicate a gap between well prepared and less prepared libraries. The less prepared libraries must be attentive that they do not fall behind. Means to close these gaps are discussed in the PARSE.Insight Roadmap [9].

In general, the Gap Analysis Framework and Tool can be used for assessing current preservation practices and benchmarking progress within or compare results between given communities of practice. Of course the basic prerequisite is the availability of sufficient survey data on which the gap analysis can be based.

The analysis indicated that the four aspects of the framework dimension regarding the “diffusion of concept of digital preservation” are not fully independent of each other. From our research results they seem to be interrelated as follows:

a) Implementation requires basic knowledge
b) Knowledge hardly exists without awareness.
c) Commitment requires awareness and can be strengthened by knowledge
d) Commitment can exist without implementation which then is considered a “lip service”
e) Systems can be in implemented without being used, if the commitment of using them is missing.
f) Commitment can be found on a corporate level (e.g. policies) and on a personal level (willingness to use the implemented systems)

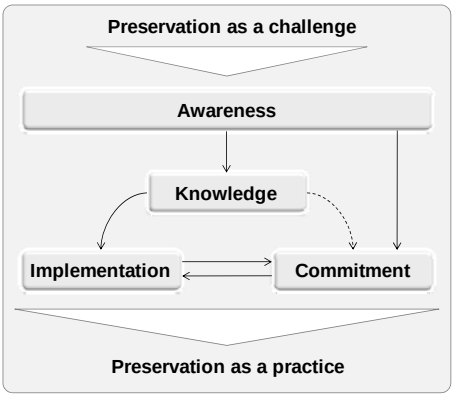

Figure 7: Relations and dependencies of the aspects of the diffusion of the concept of digital preservation

In further research projects we will refine the Tool and investigate how it can be integrated with other tools such as the AIDA-Toolkit (Assessing Institutional Digital Assets) [2] for analysis of institutional levels.

3. REFERENCES

[1] Ahlberg, Chr. and Shneiderman, B. (1994): *Visual Information Seeking: Tight coupling of dynamic query filters with starfield displays*, Proc. of ACM CHI94 Conference (April 1994), 313-317 + color plates.

[2] AIDA Self-Assessment Toolkit, URL: http://aida.jiscinvolve.org/wp/

[3] Card, S., Mackinlay, J., and Shneiderman, B. (1999): *Readings in Information Visualization: Using Vision to Think.* Morgan-Kaufmann.

[4] G.W. Furnas, *Generalized fisheye views*, *SIGCHI Bull.*, vol. 17, 1986, pp. 16-23.

[5] Ligue des Bibliothèques Européennes de Recherche (LIBER), Website: http://www.libereurope.eu.

[6] PARSE.Insight: *Insight Report. Insight into digital preservation of research output in Europe, 2010. URL: http://www.parse-insight.eu/downloads/ PARSE-Insight_D3-6_InsightReport.pdf*

[7] PARSE.Insight: *Insight into issues of Permanent Access to the Records of Science in Europe.* Project No. 223758. (EU-funded Project)

[8] PARSE.Insight: *survey results*, URL: https://www.swivel.com/people/ 1015959-PARSE-insight/group_assets/public

[9] PARSE.Insight: *Science Data Infrastructure Roadmap, 2010. URL: http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf*

[10] Shneiderman, B., Williamson, Chr., and Ahlberg, Chr. (1992): Dynamic Queries: *DataBase Searching by Direct Manipulation*. In Proc. of Human Factors in Computing Systems, CHI '92, ACM Press, 1992, pp. 669-670.