PARSE.Insight

Deliverable D4.3

Gap Analysis Final Report

Project Number	223758
Project Title	PARSE.Insight. INSIGHT into issues of Permanent Access to the Records of Science in Europe
Title of Deliverable	Gap Analysis Final Report
Deliverable Number	D4.3
Contributing Work package	WP4: Gap Analysis
Deliverable Dissemination Level	
Deliverable Nature	Report
Contractual Delivery Date	30st April 2010 (M26)
Actual Delivery Date	31st June 2010
Author(s)	Moritz Gomm (FUH), Holger Brocks (FUH), Eefke Smit (STM), Sabine Schrimpf (DNB), Jeffrey van der Hoeven (KB)

The PARSE.Insight project is partly funded by the European Commission under the 7th Framework Programme, Research Infrastructures.

Executive Summary

Deliverable 4.3 presents the final results of the gap analysis on digital preservation in Europe performed within the PARSE.Insight project. Gap analysis refers to 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.

Previous deliverables from this Work Package introduced the approach and tools for gap analysis, and applied them to one group of stakeholders, namely publishers of scientific journals. This deliverable reports on the gap analysis with respect to a different group of stakeholders, members of LIBER (the Association of European Research Libraries). It also updates the gap analysis for the publishers. By applying to two different groups, the approach is validated, and may thus be applied in different contexts.

The results confirm the insights on existing gaps within long-term preservation and foster a better understanding of their relevance and impact. The gap analysis tool offers a helpful method for discovering gaps, evaluating their relevance, and starting a discussion within and across communities on how to close them. The gap analysis thus supplements the roadmap and supports the project objectives by engendering awareness, developing knowledge, promoting implementation, and by accentuating the role of commitment towards a long-term preservation infrastructure.

The results are synthesized, by means of illustrative scenarios, into a set of conclusions that feed into the PARSE.Insight roadmap for the infrastructure of digital preservation in Europe.

Keywords:

Visual Gap Analysis, Tool Support, Publisher Gap Analysis, Libraries Gap Analysis, Vision on preservation

Contributors

Person	Role	Partner	Contribution
Moritz Gomm	Lead WP 4	FUH	Document owner and author
Björn Werkmann	Lead WP 4	FUH	Programming & GUI Design and Tool Documentation
Holger Brocks	Lead WP 4	FUH	Support and Management
Matthias Hemmje	Lead WP 4	FUH	Supervision
Eefke Smit		STM	Ideas and Documentation
Sabine Schrimpf		DNB	Analysis and Documentation of LIBER-Gap Analysis
Jeffrey van der Hoeven		KB	Workshop and Data

Document Approval

Person	Role	Partner
Matthias Hemmje	Lead WP 4	FUH
Holger Brocks		FUH

Distribution

Person	Role	Date	Partner
Moritz Gomm	Lead WP 4	15. Sept. 2009	FUH
PARSE-Mailing-List			

Revision History

Issue	Author	Date	Description
	Moritz Gomm	4.5.2010	Draft Version
	Holger Brocks	5.5.2010	Review
	Moritz Gomm	9.5.2010	Updates
	Simon Lambert	16.5.2010	Review
	Moritz Gomm	18.5.2010	Updates
	Eefke Smith	21.5.2010	Review
	Moritz Gomm	20.6.2010	Final Version

Table of Contents

1 Introduction: Purpose and Scope

The objective of work package 4 is to identify gaps in the European e-infrastructure for digital preservation based on the survey data from work package 3 ("insight report"). A gap in this context is the difference between the actual implementation of any relevant aspect of digital preservation and its objective requirement for a safe long-time preservation.

The following figure gives an overview of the phases in WP 4 and the gap analysis process. After the conceptualisation and implementation of the gap analysis tool it was first applied to the stakeholder-group of publishers (see the results in deliverable 4.2). These results were fed back and minor improvements of the IT tool were made. Finally a survey on preservation amongst libraries was used to validate the tool and the methodology in another stakeholder-group. These results are presented here along with a stepwise description of the gap analysis process.

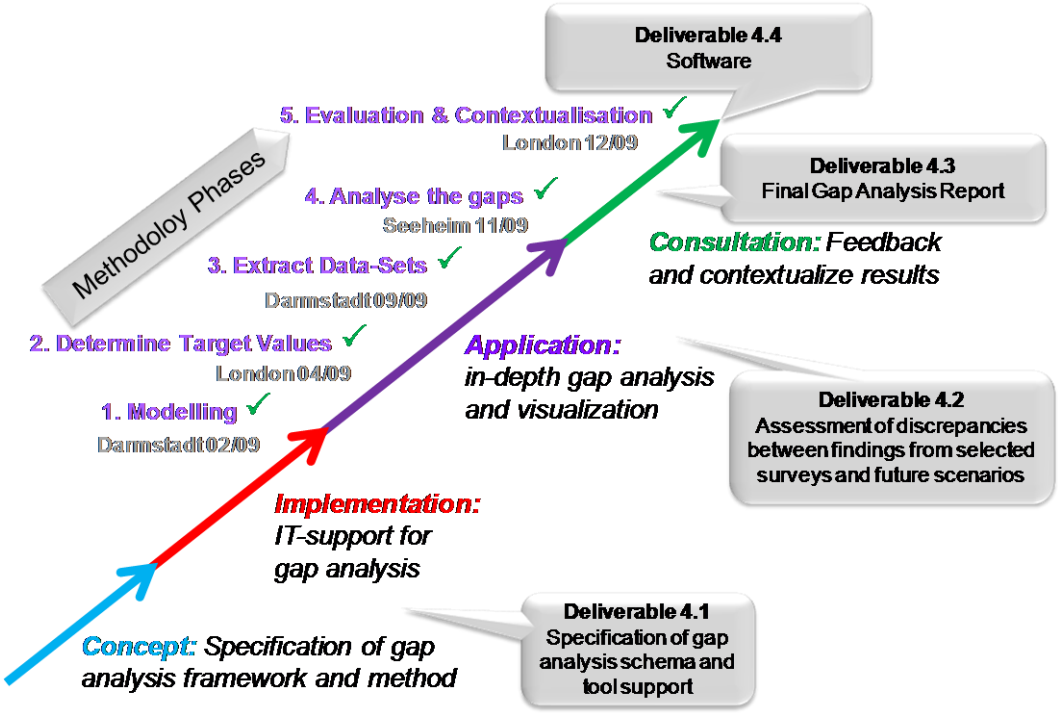


Figure 1: Timeline and project phases

The objective of deliverable 4.3 is to document the usage of the gap analysis framework and the implemented software. In addition to the stakeholder-group of publishers (deliverable 4.2) the libraries are analysed in this deliverable. Furthermore the process of the gap analysis is demonstrated and documented in a stepwise procedure. This is a generic procedure that will be usable in other contexts and for other groups of stakeholders for future research in digital preservation and beyond.

2 Evolution of the Gap Analysis Framework and Tool

The Gap Analysis Framework encompasses a stepwise, systematic procedure for assessing gaps in the status-quo of permanent access and gives insights into the needs for the European e-infrastructures landscape of the future. As an addition to the Draft Gap Report (D4.2) the dimensions are defined more explicitly here.

2.1 Evolution of Gap Analysis Framework

The Gap Analysis Framework encompasses the life-cycle of scientific data (creation, preservation and publishing, re-use and use of data) and the diffusion of long-term preservation within scientific communities (awareness, knowledge, implementation and commitment). The two orthogonal dimension of the gap analysis framework are visualised in Figure 2 (for more details on the framework see deliverable 4.1).

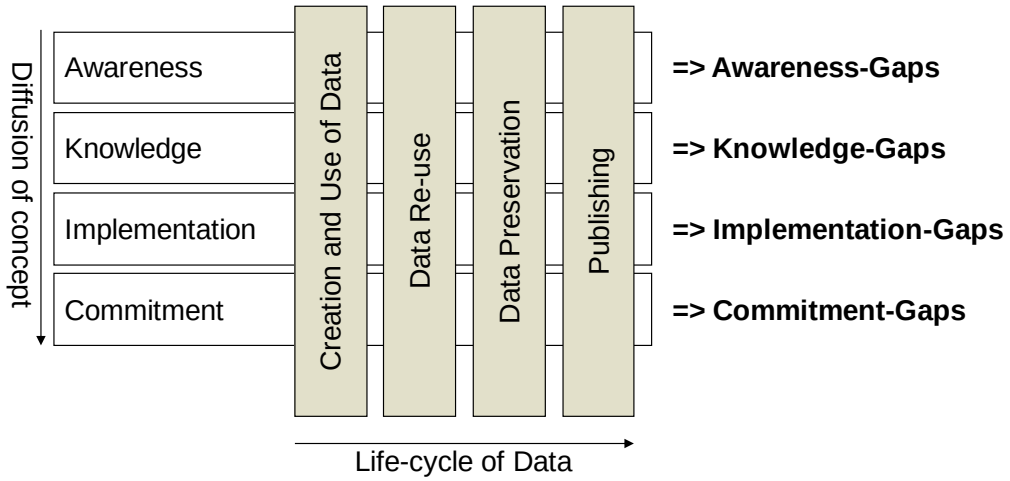


Figure 2: Gap Analysis Framework

One important Feedback from the validation-process was the need for a coherent definition and delimitation of the four dimensions. These are now explicitly defined as follows:

- **Awareness** is the ability to perceive or to be conscious of the problems of long-term digital preservation in general.
- **Knowledge** is the sum of expertise and skills for the theoretical and practical understanding of long-term digital 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.

In deliverable D4.2 the stepwise procedure of the gap analysis to transform survey data into the gap analysis framework was documented. In chapter 3 of this deliverable the gap analysis methodology is demonstrated using a scenario from the project.

2.2 Evolution of Software Components and Architecture

Building on the gap analysis methodology and the results obtained in its validation here the technical updates made on the Gap Analysis Tool (GAT) are summarized. A few minor technical improvements were made for better reuse of the tool for future surveys.

2.2.1 Single command-line interface

- Single command-line interface integrating:
  - Conversion of SurveyMonkey files for Gap Analysis Tool project file
  - Excel file creation to collect Expert feedback
  - Excel file analysis to incorporate feedback
  - Tree structure definition to capture categories
  - Configure custom names and labels

2.2.2 Ensure more robust export processing for online survey database

- Consolidate inconsistent character encodings/ file formats
  - UTF8, UTF16 Little Endian with Byte Order Mark (BOM), without BOM
  - Various HTML entity encodings
- Detect and convert for later use in Gap Analysis Tool

3 Step-by-Step Description of the Gap Analysis

The Gap Analysis can be based on any survey that asks questions about long-term preservation, such as the surveys from work package 3 ("insight report"). In the following the preparation and usage of the tool is described step-by-step using the data of a survey by the Association of European Research Libraries (LIBER).

3.1 Tool Preparation

At first the items from the survey on which the gap analysis is based on, had to be **mapped to the gap analysis framework** by domain experts (in this case Members of LIBER) to feed the survey data into the tool. This means assigning each question of the survey – if possible - to the corresponding gap categories, as shown in the following example:

Example:

The answer to Question 14 in the LIBER-Questionnaire:

"Do you have security protocols that protect stored data from unauthorized modification, damage or deletion?"

is an indicator for the **Implementation-dimension** and is mapped correspondingly:

- a "yes" to this item *decreases* the gap value in the implementation dimension while
- a "no" *increases* the corresponding gap value.

The resulting structure (called the "domain tree") represents the *status-quo-information* in 1:n-relationships and is shown in Figure 3Figure 3.

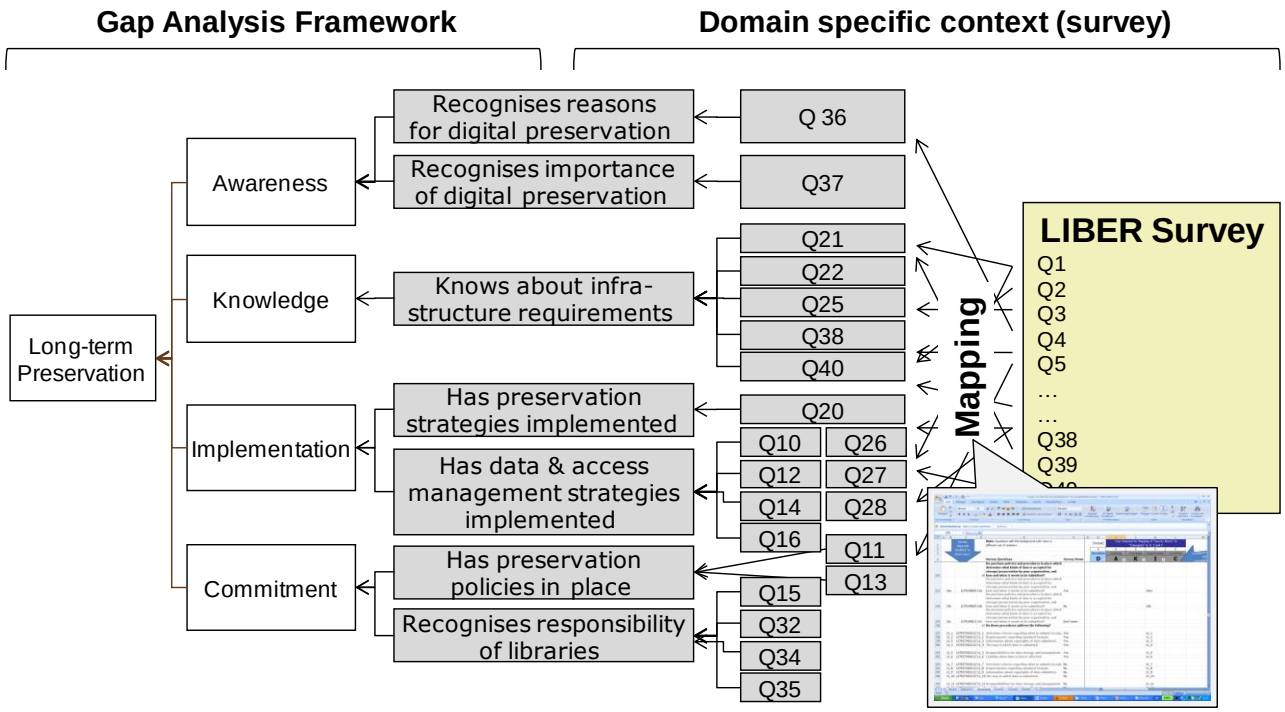


Figure 3: Mapping the survey to the framework to obtain the "domain tree"

To support the mapping process a Microsoft Excel tool was developed, since most of the users and researchers are used to the Microsoft Excel interface. The user indicates for each question item if the target-values it is a positive (+1) or negative (-) indicator for the corresponding dimension.

After the mapping process a structured text file has to be prepared, which is readable by the Gap analysis Tool.[1] For this the hierarchies in the four dimensions and the assigned question items are coded and related to the appropriate sub-category as shown in the example below:

Example:

....

name: "Awareness",

children: [{ name: "Recognises reasons for digital preservation", children: "36" },

{ name: "Recognises importance of digital preservation ", children: "37" }],

name: "Knowledge", ...

....

After all relevant items from the questionnaire are mapped to the corresponding sub-categories in the four gap dimensions the "gap analysis tree" is ready. Next the data-sets from the survey are loaded into the tool using the file-dialog shown in Figure 4.

[1] JSON (JavaScript Object Notation) - a lightweight data-interchange format - is used here.

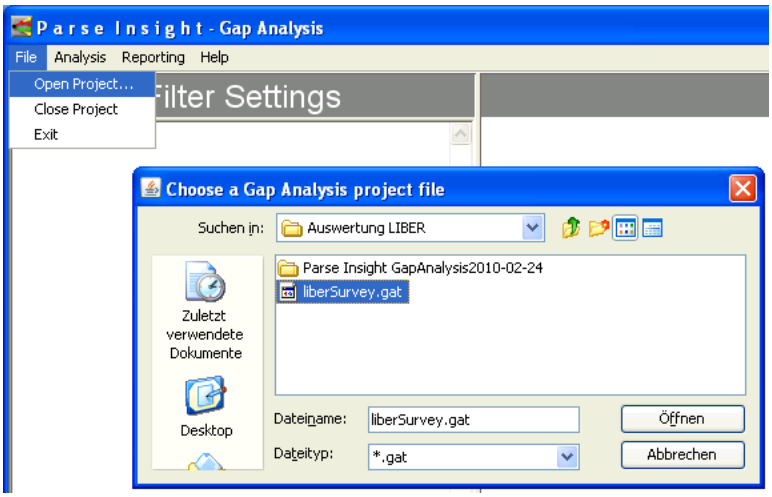


Figure 4: Selecting the Gap Analysis project file

All question items of the survey are converted into interactive controls displayed as "**Filter Settings**" on the left side of the tool.

The calculated gaps of the currently selected data-sets are visualized in the "**Analysis View**" on the right side of the tool. After start the tool visualizes the entire data-sets with no filter set as indicated in the upper right corner (see Figure 5). The calculation of the gaps is as follows: For any survey item (e.g. leaf in the tree) all respondents are counted, that have selected the corresponding item. The sum is divided by the total number of surveys. For example in the picture below (Figure 5) only 33% of the participating libraries see libraries responsible for preservation, thus the awareness among this group seems to be relatively low: the leaf is thus coloured red. Since the second aspect of "awareness" results to a much better value of 0,46 (yellow) the overall gap value for "awareness" is 0,40 (the average of its two sub-categories), which is considered a moderate gap (coloured yellow).

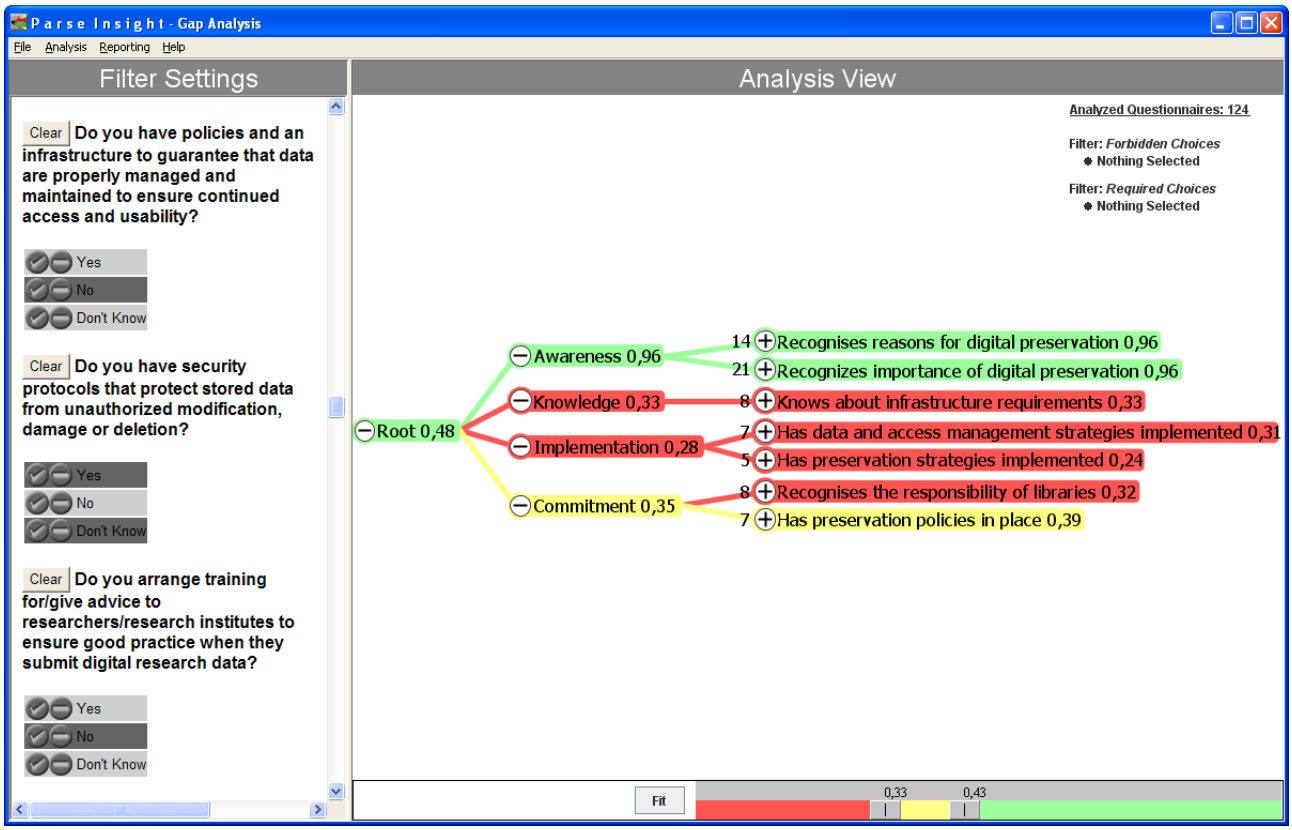


Figure 5: Gap Analysis Tool after Start-up (from the LIBER-Study)

The expert can now start to **analyze and compare gaps** by changing the selection criteria to drill in and out of the entire data set. Thus more general gaps can be separated from gaps in special areas or sub-groups. The gap thresholds (indicated by the colours red, yellow, and green) can be varied using the sliders at the bottom of the tool.

Figure 6 gives an overview of the entire gap analysis process which is described in detail in deliverable 4.1.

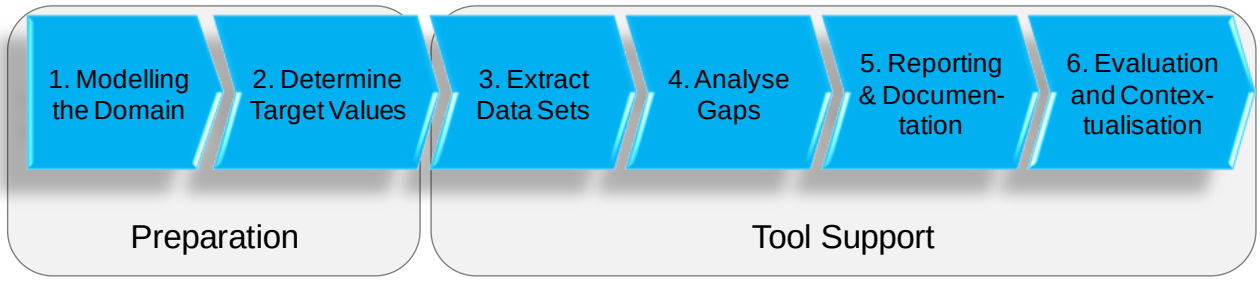


Figure 6: Overview of the Gap Analysis Process

3.2 Tool Usage

The usage of the analysis tool shall now be demonstrated with the LIBER-study.

Before using the tool it was assumed by the experts, that awareness, knowledge and commitment amongst the scientific libraries are high, while they are still facing an implementation-gap. The visualization of the base data gives a different picture (see Figure 5 above): 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.

In order to get a more differentiated and sophisticated picture, the experts then for example compare those libraries that stated explicitly that they do or do not have preservation policies and an infrastructure for digital preservation in place. For this the correspondin "Filter Settings" are set (see left side of Figure 7). Immediately the new view is shown on the left also indicating how many respondents belong to the subset (see right side of Figure 7).

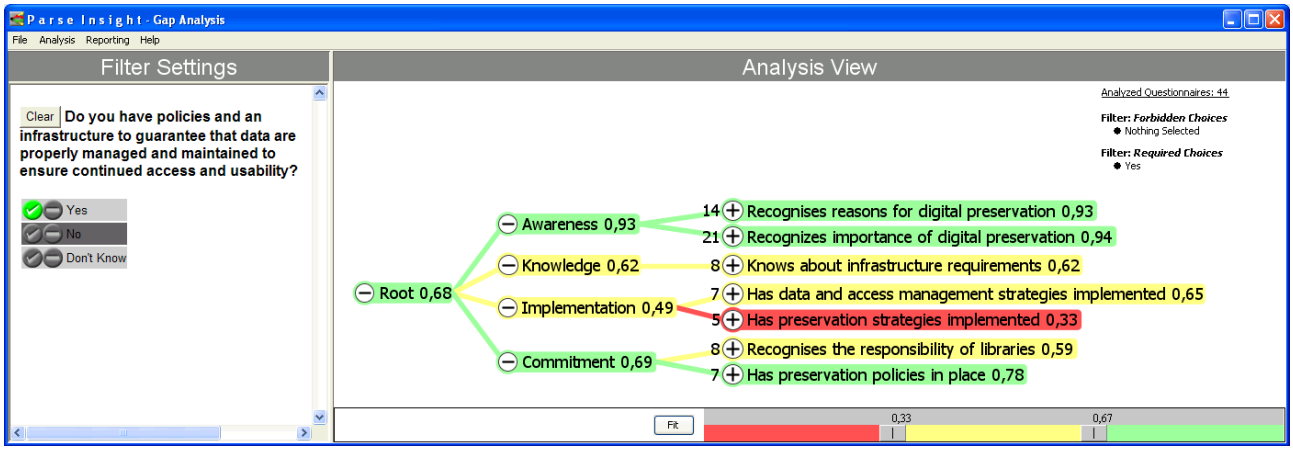


Figure 7: Example of a drill down into the LIBER data set

The experts can analyse the data by drilling in and out and selecting subsets of the data which catch their interest. For example if the researcher 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. For this the filter on the item "yes" to question *"Do you have policies and procedures in place which determine what kinds of data is accepted for storage/preservation by your organisation, and how and when it needs to be submitted?"* is selected and the result is compared to the filter set to "no".
The two results are shown in the following table. Obviously the organisations, that don´t have these policies and procedures in place have a far bigger gap in preservation, due to weaknesses in implementation and commitment.

2	"Do you have policies and procedures in place which determine what kind of data is accepted for storage/preservation by your organisation, and how and when it needs to be submitted?"
Yes	
No	

For **reporting and documentation** the user can produce screenshots and paste them into the report. It turned out to be advantageous for communicating and understanding the results

to compare settings in tables as shown in the example above (for more see the analysis in section 4 below).

Finally the **evaluation and contextualisation** of the results should be done within the community. This can either be in form of a workshop or – as done in LIBER – by communicating the results among the experts and gathering their feedback. This phase can also be used to estimate the impact of gaps and derive ideas on how to close them.

4 Application of the Gap Analysis to the User Group "Scientific Libraries"

4.1 Introduction

LIBER (Ligue des Bibliothèques Européennes de Recherche) is the main research libraries network in Europe and encompasses more than 400 national, university and other libraries in 45 countries. LIBER aims at representing the interests of European research libraries, their universities and their researchers. Amongst others it promotes efficient information services, access to research information, in any form whatsoever, and preservation of cultural heritage.

As a comparison group to the lead-user-group "Publishers", the community of scientific libraries (as a subset of the community "data managers") was selected. Reasons were that the group of the scientific libraries is congruent with the LIBER community and is as such well defined and easily controllable. The existing knowledge of the work package members in this domain is high and the LIBER office was willing to contribute to the analysis. Only the statistical basis from WP 3 was not as good as that of the publishers. The number of skipped questions was significantly higher because more questions were "optional", which means they did not require an answer.

The domain tree was modelled by the Deutsche Nationalbibliothek (DNB) staff –a scientific library – and reviewed by LIBER members. Again the results from WP3 were used first to get the picture of the status-quo of preservation in this specific domain. The following assumptions were drawn from the survey results.

4.2 Assumptions before application of the Gap Analysis Tool

The following results from the LIBER survey (N=124) attracted attention:

- The great majority of the LIBER libraries recognize the reasons for and the importance of digital preservation. Awareness seems to be high.

- The majority of the participating libraries believe 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.

- 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.

- 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 was used and helped to check these assumptions and to render some findings more precisely.

4.3 Gap Analysis

A total of 70 items were identified and grouped under the following dimensions and sub-categories. The selection of question items and grouping into categories was subject of the review by LIBER members.

Dimension	Sub-categories and question items
Awareness	SC: Recognises reasons for digital preservation [-36_15 to -36_28] SC: Recognises importance of digital preservation [-37_15 to -37_35]
Knowledge	SC: Knows about infrastructure requirements [21a] [25a] [22a] [38a] [40a-d]
Implementation	SC: Has data and access management strategies implemented [10a] [12a] [14a] [16a] [26a] [27a] [28a] SC: Has preservation strategies implemented [20a-d] [-20e]
Commitment	SC: Recognises the responsibility of libraries [15a] [34h] [34i] [35h] [35i] [32_1] [32_2] [32_3] SC: Has preservation policies in place [11_1 to 11_6] [13a]

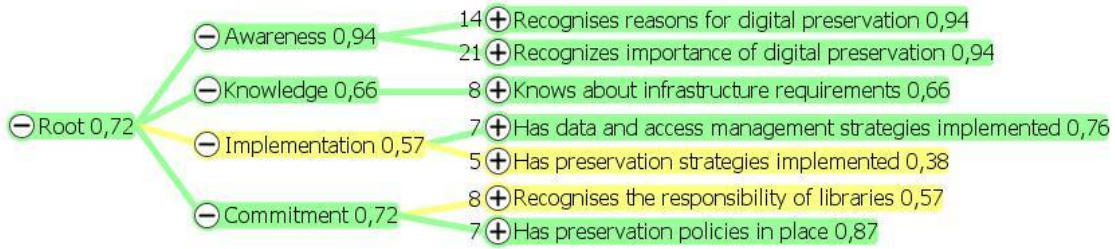
4.3.1 Analysis of overall gaps

Feeding the tool with the data of the 124 datasets from the survey gives the following results:

The visualization gives a different picture than the assumptions that were stated above: Only awareness is visualized as positive (green colour) as was assumed. Commitment is on a modest level (yellow), due to some preservation policies that libraries have implemented. The tool indicates a larger gap in the area of knowledge than pure looking at the survey results had indicated. The implementation gap that was assumed can be proven.

4.3.2 Policies and Infrastructure

The high knowledge and implementation gaps can be explained with the fact that many participants did not answer the respective questions and skipped answers were counted as negative answers. The picture gets different (and more sophisticated) when we compare those libraries that stated explicitly that they do or do not have preservation policies and an infrastructure for digital preservation in place:

1	"Do you have policies and an infrastructure to guarantee that data are properly managed and maintained to ensure continued access and usability?"
Yes	
No	

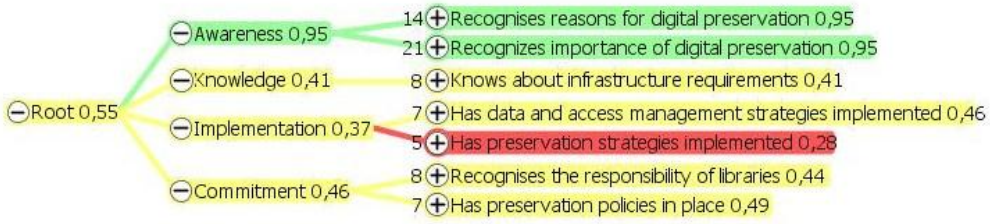
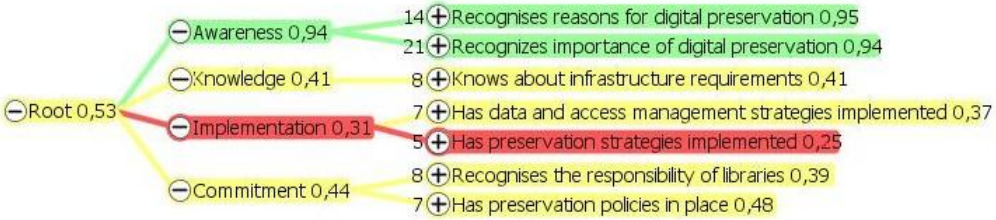
Or those that stated that they do or do not have selection policies and procedures in place:

2	"Do you have policies and procedures in place which determine what kind of data is accepted for storage/preservation by your organisation, and how and when it needs to be submitted?"
Yes	
No	

A clear relation between selection policies and the level of implementation and commitment can 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'

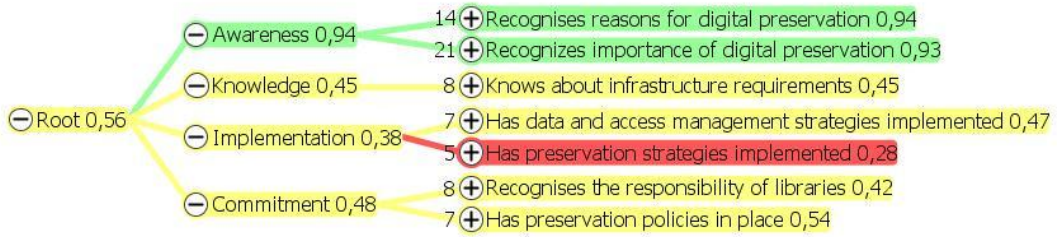
preparedness for digital preservation. No matter if they are focussing on more traditional publication types like e-books and journals or on for libraries unfamiliar data sets – the picture remains almost the same:

3	"Which kind of digital material is stored at your organisation?" (Multiple answers were possible
Data Sets	
E-Books	
Journals	

4.3.3 Amount of Data

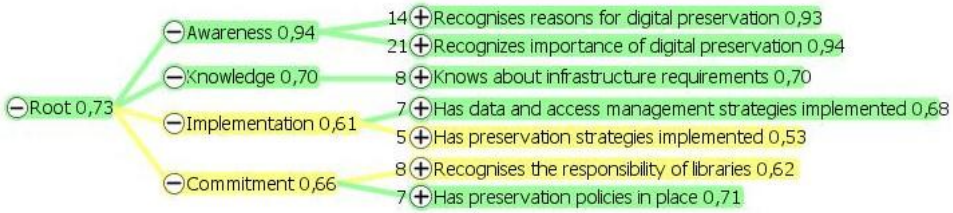
In contrast, the amount of data that a library currently stores seems to have a small but measurable impact on the gaps in preservation. The larger the amount of data that a library has to deal with, the smaller the gaps in the area of implementation and commitment are. There is a direct relation between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

4	"Please provide us with an estimate of the volume of stored digital data (excluding backups)"
100MB-1GB	

1GB-1TB	
1TB-1PB	

4.3.4 Preservation Strategies

It is conspicuous that a large gap is indicated in the area of implementation of preservation strategies in almost all pictures so far. 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.

5a	"Does your organisation have any of the following preservation strategies in place?"
Yes either: Emulation Migration Outsourced	

The gap values for the institutions, which indicated that they had either emulation or migration services in place or that they outsourced preservation to a third-party service were almost the same within the three categories, with only slightly different values within the category "outsourced to a third-party service", where the proper values are: knowledge: 0,63, implementation: 0,52, and commitment: 0,74. So they can all be subsumed within one picture.
Only minor commitment and implementation gaps are indicated for those groups.[2]
The commitment gap in the category "recognises the responsibility of libraries" cannot be explained as easily.

The gaps get even bigger when we look at those institutions that state explicitly that they do not have preservation strategies in place:

5b	"Does your organisation have any of the following preservation strategies in place?"

[2] The reason why there still is a minor gap in the sub category "has preservation strategies implemented", when we look at the answers of only those institutions that claim to have preservation strategies in place is due to the fact that most institutions have implemented only one or two of the potential strategies and have therefore not answered all possible questions. (E.g. Emulation: yes, but Migration: no, and Outsourced: no.)

Emulation - No	
Migration - - No	
Outsourced – No	

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. Furthermore there is bidirectional link between implementation and knowledge: On one hand commitment is needed on a corporate level to invest resources into digital preservation (implementation) and to set up policies. On the other hand commitment is also needed on an individual level to use the implemented systems and act according to the policies.[3]

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 to implement appropriate systems.

4.3.5 Training vs. Resources

Curiously enough, when we look at the libraries that think that more training is needed in order to guarantee preservation for access and use in the future, this very groups shows the smallest gap in the knowledge area:

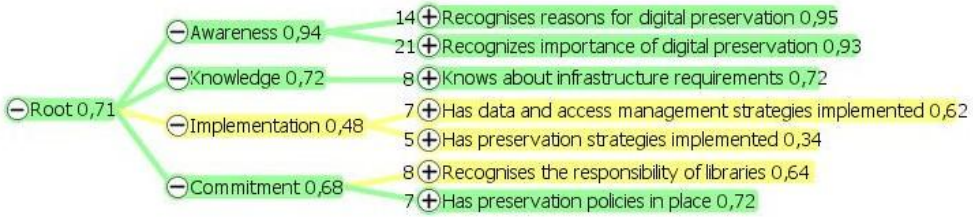
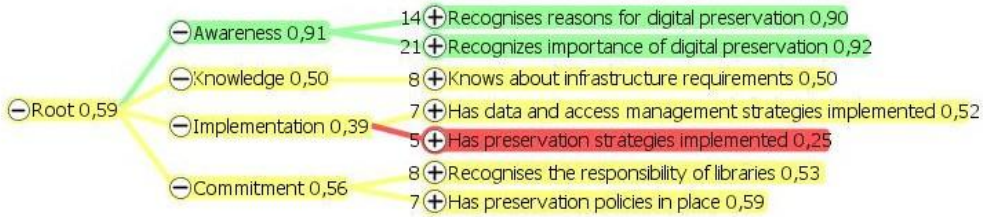
6	"Apart from an infrastructure, what do you think is needed to guarantee that valuable digital research data is preserved for access and use in the future?"
Training	

[3] See Figure 10 on page 10 for the dependencies between awareness, knowledge, implementation and commitment.

More resources	
More repositories/ archives	

4.3.6 Scalability

When we compare 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:

7	"Do you think your current infrastructure will scale with future requirements?"
Yes	
No	

All results were fed back to the experts of LIBER and asked if they missed anything. The feedback was positive, especially on the visualisation and the clear communication of gaps.

4.4 Summary of the LIBER Gap Analysis

The 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. Future study designs that want to make use of the Gap Analysis tool should

take this finding into account and exclude optional questions as far as possible from their 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 indicate 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 also show 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 good news for the libraries is there awareness for the importance of digital preservation. But serious issues amongst the libraries are the facts, that they have policies in place without successfully implementing them and that some of them show commitment but are missing the necessary knowledge to implement appropriate systems for digital preservation. Thus the majority of libraries seem willing but are yet unable to meet the preservation challenges. The gap analysis indicates a gap between well prepared and less prepared libraries and the less prepared libraries must be attentive that they do not fall behind.

4.5 Discussion of limitations

The gap analysis of the Libraries community made clear, that mandatory questions may influence the results. If respondents don´t answer questions at all that are indicators of awareness, knowledge, implementation and commitment the tool must expect, that there is a gap. These cases make it clear, that domain experts need to do the analysis who have thorough knowledge of the survey.

5 Application of the Gap Analysis to the User Group "Publishers"

The following is an update of the section from deliverable 4.2. After the workshop in Darmstadt it was argued, that a segmentation of publishers regarding size would be interesting. The results of this are described in section **Error! Reference source not found.**.

The publishers were selected as a pilot group for the gap analysis process because of the good statistical basis from WP3, the existing knowledge of the work package members in this domain and the willingness of a motivated group of lead-users from this stakeholder-group to contribute.

Before starting to search for gaps within the stakeholder-group results from the "Insight Report" (work package 4) were revisited and summarized under the perspective of the gap analysis framework:

- **Awareness:** The majority of publishers (small: 74%; big: 67%) are convinced that an infrastructure will help counter the threats to digital preservation. More than half of the publishers believe that illustrative material (59%; 61%) and data sets / auxiliary material (55%; 57%) should be preserved besides the article themselves.
- **Knowledge:** 57% of the large publishers and 23% of the small publishers outsource the preservation of their digital publications to a third **party service.** Most of the big publishers (78%) and every second small publisher (58%) believe the future will be dominated by a mixed business model, in which subscription-based and open access journals will both exist.

- **Implementation:** Preservation is in place for 96 % of journals (small: 56%; big: 89%). A majority of the publishers (small: 68%; big: 71%) have no arrangements for the preservation of research data (yet). 71 % of large publishers and 58 % of small publishers allow authors to submit underlying digital research data, together with their manuscripts, to the journal, which comes available for free upon publication of the article.
- **Commitment:** The majority of the large publishers (89%) have a **policy** for the preservation of digital publications in place, as opposed to 56% of the small publishers.

To better understand the specific gaps and their relevance in these dimensions three workshops were organised, bringing together experts of the stakeholder-group of publishers which cover different roles in the publishing value-chain (see Figure 8).

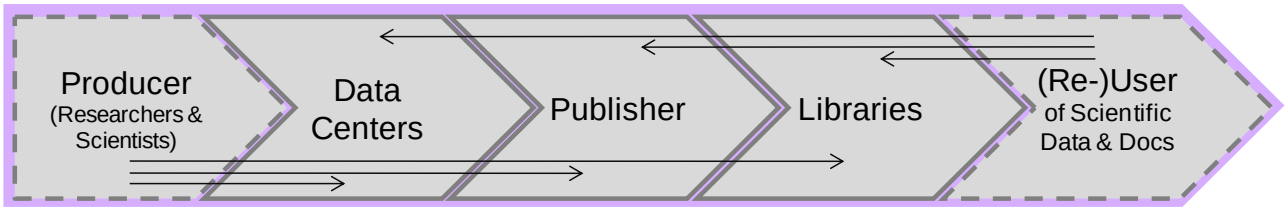


Figure 8: The value chain in the publishing community

Within the workshops the experts discussed the issues of preservation and assigned them to the four gap dimensions:[4]

Awareness:

- Misconceptions about digital preservation exist mainly among small publishers. There is confusion between related but distinct activities such as archiving, digitizing backfiles, persistent accessibility, etc.
- The main common objective is to ensure a sustainable and persistent linking system between datasets and all publications referring to them and vice versa. Also to ensure that those who wish to re-use datasets, do not take them out of context and are able to interpret them properly by consulting all relevant publications about the original research.
- For digital preservation, enthusiasm exists to come to common standards and common practices, establish persistent and interoperable identifiers, agree policies and support linking systems for datasets and publications. The publishers community can be typified as a 'Coalition of the Willing'.
- Establishing an infrastructure and business processes is a complex, dynamic, and evolutionary objective and should first and foremost use all stepping stones that are already available (such as DOI's to link from and to publications).

Knowledge:

- Common language (e.g. „Archiving" is not the same as „Preservation"), understanding the concepts.
- Usage of Meta-Data – publishers are very eager to share metadata standards between them so that datasets can be properly handled together with manuscripts for official publications. The common standard among publishers is the DOI.
- Best-Practices (who does what; ongoing efforts) are often the best way to achieve common standards and practices.
- Maturity of sharing and preserving research data is different for each discipline. To predict where new demands for datasets will occur and how to create a proper infrastructure for that, it could help to look back for change agents and triggers in the more developed areas like chemistry, genetics, life science.
- Clear roles and responsibilities: Who-archives-what-and-where-and-with-whom; it would be good if clear policies, common standards and common practices were being promoted.

[4] See the appendix of deliverable 4.2 for the full list of findings from the workshop.

Implementation:

- New infrastructure needs to be interoperable with what exists, e.g. with established initiatives such as DOI's, TIB-initiative, ACAP, Crossref, Portico, Clockss.
- Lack of established standards ("minimal sets of requirements"); publishers are more than willing to help establish these.
- International approach and Internet based (Multimedia, "mashed-up data")
- Linkage in both directions Data ⇔ Publication (basic: Link to web page of author/institute)
- Technology required (e.g. Storage space and costs, digital tool-sets)
- How to cite data? („micro citation") How to link to them? Central information point (Where can I find more about this?) What is the "audit trail"? (life cycle of publication)

Commitment:

- Policies und incentives are needed, probably mostly so for the researchers to deposit and share research data (e.g. getting credits for data, linking, citation).
- Policies on certification of trusted repository and deposit place so that people know where to go. Boundary conditions: not too many, not too much fragmentation, eg one main one per continent.
- Central archiving and linking should become a prerequisite to obtain research funding
- Find out the reasons why so many researchers do not want to share data (competitiveness, proving wrong, discipline specific sharing culture, willingness to share later, efforts)
- Publishers are committed to support linkage between datasets and publications (discover the "Undiscovered value of data"), but datasets are not 'owned' by publishers.
- There is no publisher-specific policy for data handling. In most cases the editor decides, while in general most publishers are offering facilities to submit datasets with manuscripts.
- Infrastructures can be established faster if promoted and supported by an "Alliance of the willing" (cooperation of lead users and big players)
- Setting basics for Certification, Compliance, Codes of Conduct ("Minimal Requirements")

It was agreed, that data preservation is a complex and extremely discipline-specific issue thus confirming the approach of the PARSE.Insight project and the concept of basing the gap analysis tool support on a profound understanding and interaction with the community being analysed.

5.1 Setting up the Tool

With the domain specific knowledge from the workshops the survey from WP3 was transformed into the tree structure as described in section 3.1 above. A total of 60 items were identified and grouped, as shown in the following table:

Dimension	Sub-Categories (Questions)
Awareness (20)	Offers Data Upload and Access (8a, 10c, 10d, -38d, 64a) Verifies Data (12_1, 12_2, 13b, 13c, 33c) Sees Data Storage Need (16e, 16f, -16g, 18b, 28d, 30e, 31d, 35a) Preserves Data (21d, 33b)
Knowledge (11)	General Knowledge(-21f, -22f, -45c) Technology Knowledge (22a, 22b, 22c, 22d, 22f, 28e, 30f, 31e)
Implementation (19)	Offers Upload, Validation and Collaboration (10b, 13a, -18a, -38d, -42c, 18b, 44a, -44c, 45a, 46a) Preserving Strategies & Operations (21a, 21b, 21c, 21d, -21e, -23c, 26a, 26b, -26c)
Commitment (10)	Policies (15a, -16g, 20a, 22a, 23a) Willingness to Pay (25a, 27d, 28c, 30d)

	Willing to share data (47a)

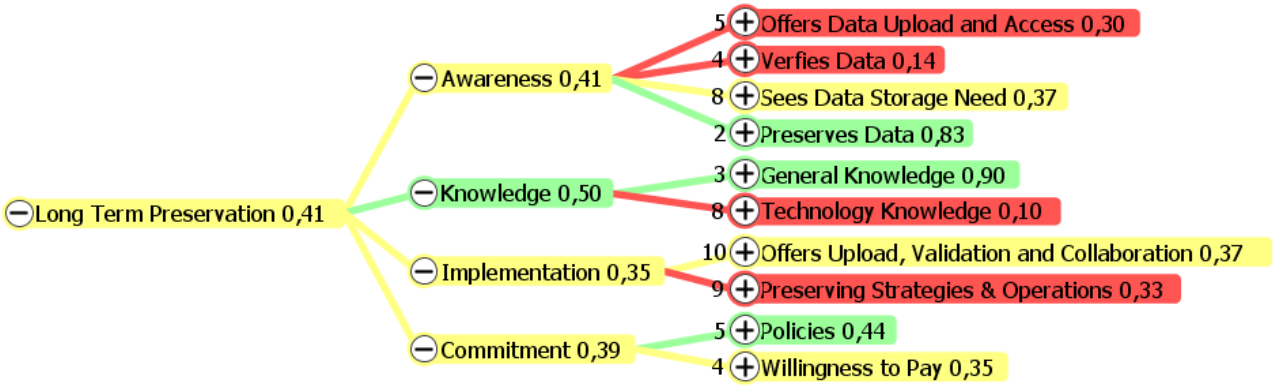
The table also shows that the dimensions of the framework are not reflected equally in the survey with 33% relating to awareness, 32% to Implementation, 18% to knowledge, and 17% to commitment. It has to be kept in mind that the survey had to be designed before the framework was developed and that designing the questionnaires specifically for the Gap Analysis would allow a much more balanced coverage of the four dimensions.

The data from the publisher survey was then extracted as a .CSV-file from the platform "survey-monkey.com" and loaded into the Gap Analysis Tool. At this point the tool is ready for visual analysis and drill-down into the survey data.

5.2 Gap Analysis

5.2.1 Analysis of Overall Gaps

Feeding the tool with the data of the 185 datasets from the survey gives the following results from the Gap Analysis Tool for the Publishers:



The visualization indicates at a glance, that the knowledge[5] about long term preservation (green colour) of the entire sample seems to be high (due to some "general knowledge" about the issue) and that awareness as well as implementation is on a moderate level (yellow colour). The highest gap lies in the commitment to actually "live" and promote preservation (red colour). Particularly the willingness to share data among the publishers seems to be low.

The tool now allows for a further drill-down into the data-sets, selecting those publishers who have answered to specific items.

- **Awareness** for issues of long-term preservation exists due to the fact, that publishers either preserve data (especially articles) or at least see the need of storing data. But only few publisher already offer or plan facilities for researchers to upload and access data or to verify the data. This is due to the fact, that submission of data sets is a relatively new category.
- The group of publishers has general **knowledge** about preservation but have only little technology knowledge which is a prerequisite for a thorough implementation of preservation systems. This is due to the fact, that most publishers outsource preservation and therefore don´t need a high degree of technical knowledge.
- The **implementation** gaps lie in installing preservation strategies and operations, though a moderate level of online collaboration and data upload and validation processes is implemented within the community of publisher. Again this is also related to the high degree of outsourcing in preservation.
- In the **Commitment** dimension the high number of publishers having preservation policies in place is very positive. The overall gap is still just moderate since the willingness to pay

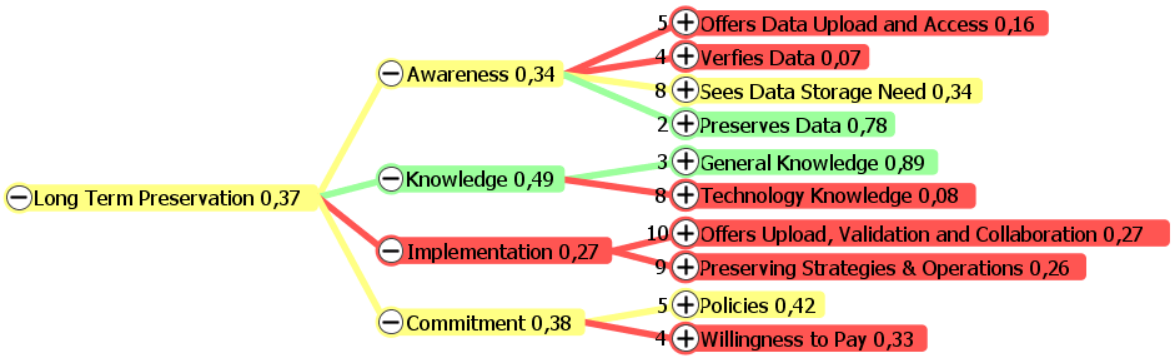
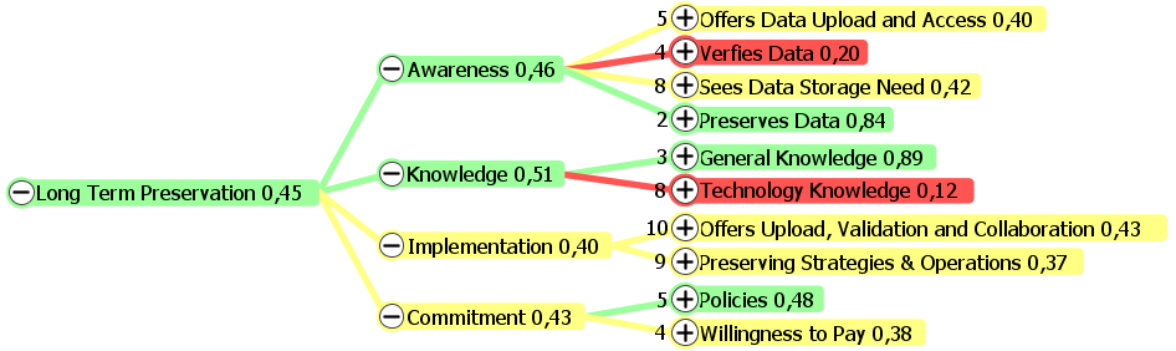
[5] Within the survey there were only very few questions on "knowledge" with many items, thus the knowledge dimension can´t be measured as well as the other dimensions.

for preservation is not as high.

All questions were analysed using the gap analysis tool. Interesting findings are shown and discussed below.

5.2.2 Submission of Research Data

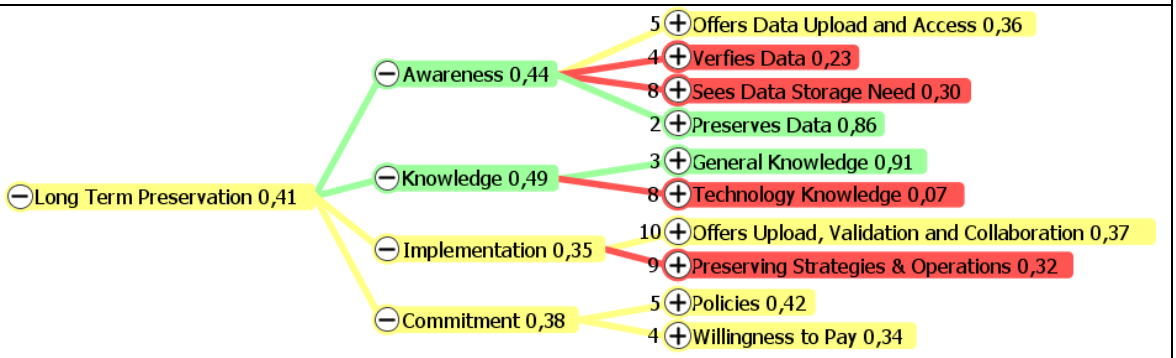
The following visualisation shows all publishers who allow "authors to submit their underlying digital research data with their publication":

<table>
<tr><td colspan="2">"Can authors submit their underlying digital research data with their publication to you?"</td></tr>
<tr><td>"No"
(44)</td><td> 
</td></tr>
<tr><td>"Yes"
(106)</td><td> 
</td></tr>
</table>

The difference between the group of publishers who answered "Yes" and "No" is very obvious: The "Yes"-group has higher awareness and implementation levels whilst these values for the "No"-group are much lower than the publishers average.

5.2.3 Validation Process

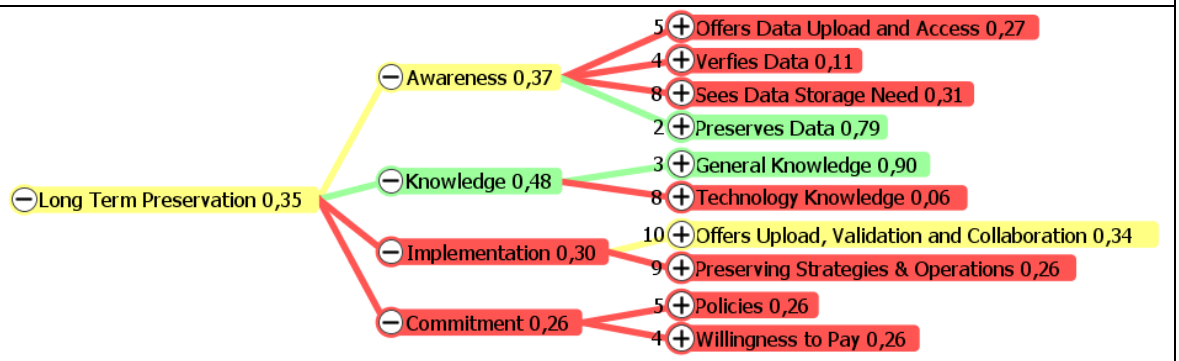
The scope of developing a validation process for data submission also seems to have a strong impact on the gaps in preservation as can be seen in the following table:

"Are you planning to develop a validation process for data submission?"	
"No" (14)	
"In less than 3 years" (12)	

Whilst the group of publishers who will develop a validation process within the next three years seem to have done quite some of their homework (awareness, knowledge, much higher levels of implementation) the late-comers still have great awareness, implementation and commitment gaps.

5.2.4 Preservation Policy

More and more companies give themselves a "policy" for preservation and it is interesting to have a closer look if this also leads to higher levels in the dimensions of the Gap Analysis Framework:

<table>
<tr><td colspan="2">"Does your organisation have a policy for preservation of digital publications?"</td></tr>
<tr><td>"No" or "Don´t know" (87)</td><td> 
</td></tr>
<tr><td>"Yes" (99)</td><td> 
</td></tr>
</table>

The publishers who don´t have a policy for preservation (or don´t know of it) have very little commitment as can seen in the figure above. Interestingly those people who "don´t know" if they have a policy score the least. One reason of the low commitment is the little awareness for the necessity of preservation.

Those companies who have a policy in place also have much higher values in all dimensions. The high values in "knowledge" have to be looked at carefully since only very few questions with a total of 11 items gave hints about the knowledge in the companies.

5.2.5 Online Collaboratories

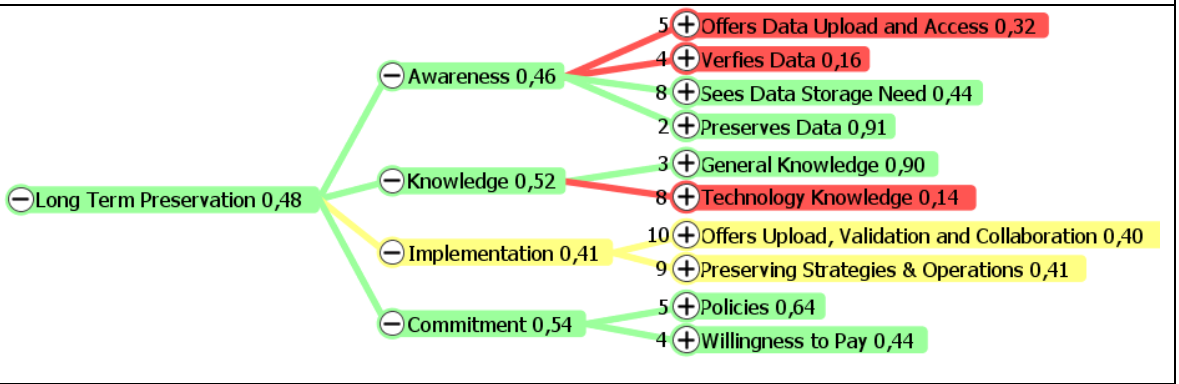
Some companies use preservation in the context of services for their customers, such as those offering "online collaboratories":

"We offer online collaboratories and we preserve them (for the long term)"	
(14)	

Obviously offering value added services such as "online collaboratories" that require long term preservation is a good context to build up knowledge about preservation and implementing appropriate measures.

5.2.6 Disaster-Recovery

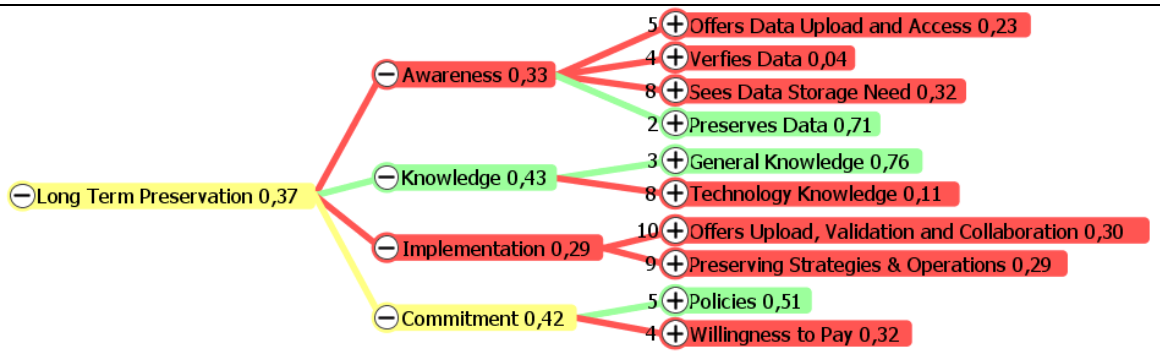
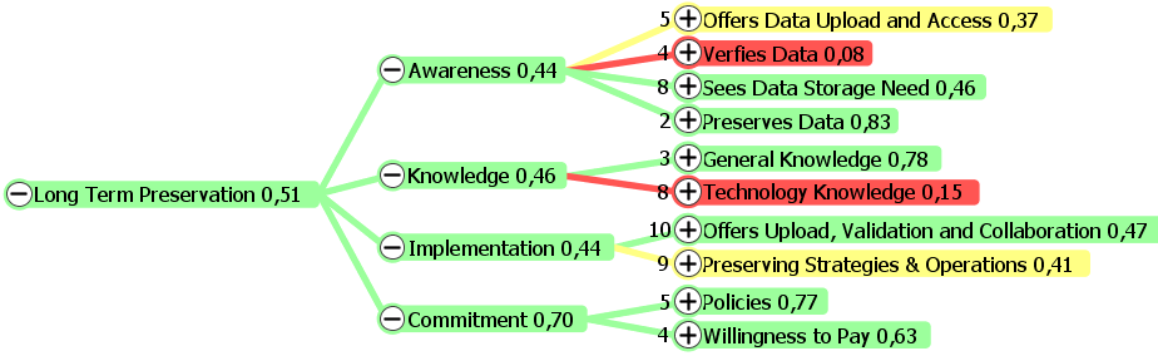
One of the very basic preservation issues is the protection against disasters of all kind which could destroy the entire basis of business. More than 40 % of the responding companies have such a system in place:

"Does your organisation have a disaster recovery policy for its digital content?"	
"No" or "Don´t know" (107)	
"Yes" (79)	

Coming to terms with very basic requirements such as "disaster recovery" seems to not only have a positive effect on awareness and knowledge but also on the commitment and implementation towards preservation.

5.2.7 Planning of Preservation Initiatives

Some questions in the survey asked for the timeline for preservation initiatives, such as the following question:

"If your organisation has not yet systematically organised long-term preservation, when do you expect to do this?"	
"1-3 years" (7)	
"< 1 year" (6)	

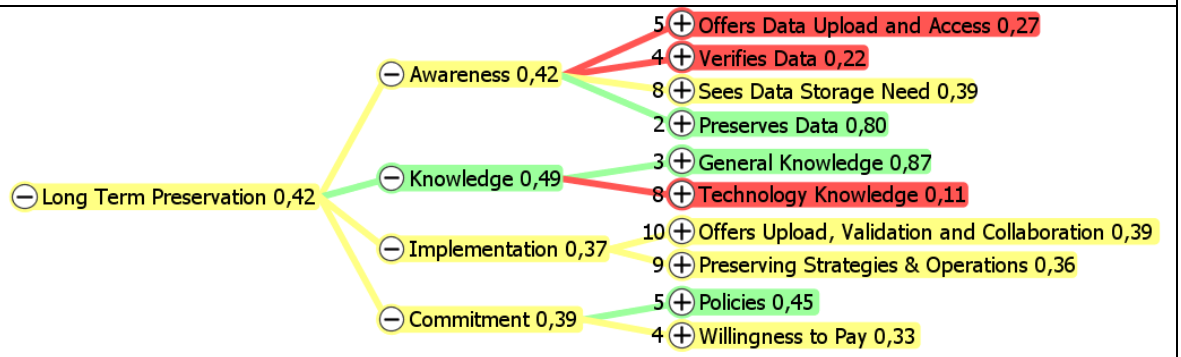
Those companies who have planned further steps in the future have higher gaps in awareness, implementation and commitment, whilst those with existing short term plans have high values in all dimensions except in "implementation". This is very consistent with the question on further implementation.

5.2.8 Economic Value of Data

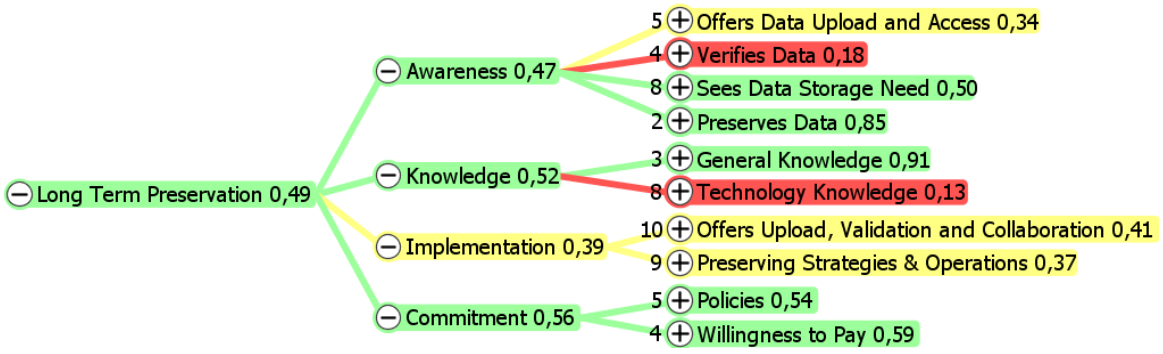
One strong driver for investing in preservation is the economic value which lies in it. The following question indicates that finding sustainable business models for offering scientific data can be a high motivator:

"How important is the reason 'digital data potentially has economic value' for preservation?"

"very un-important" (15)



"very important" (26)

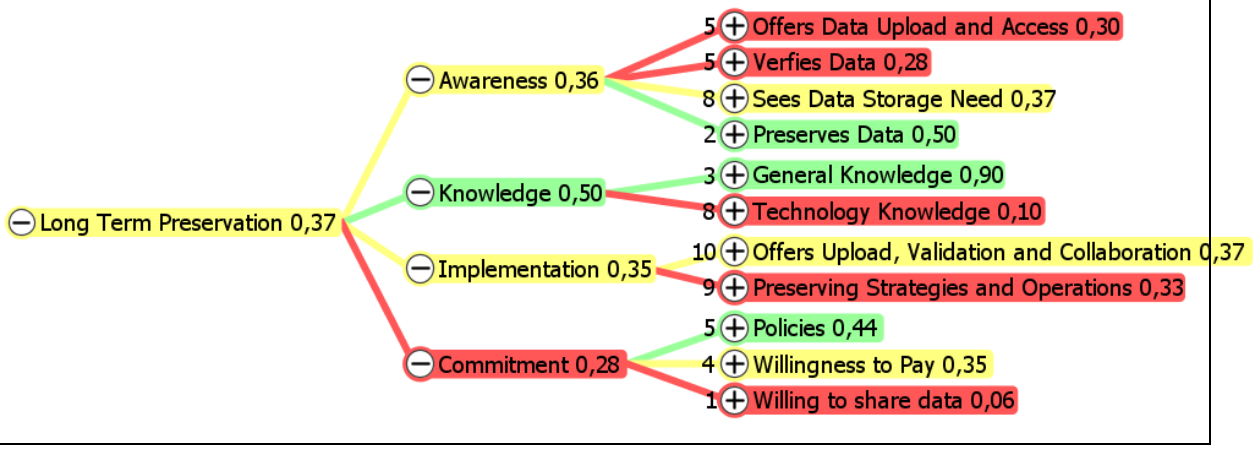


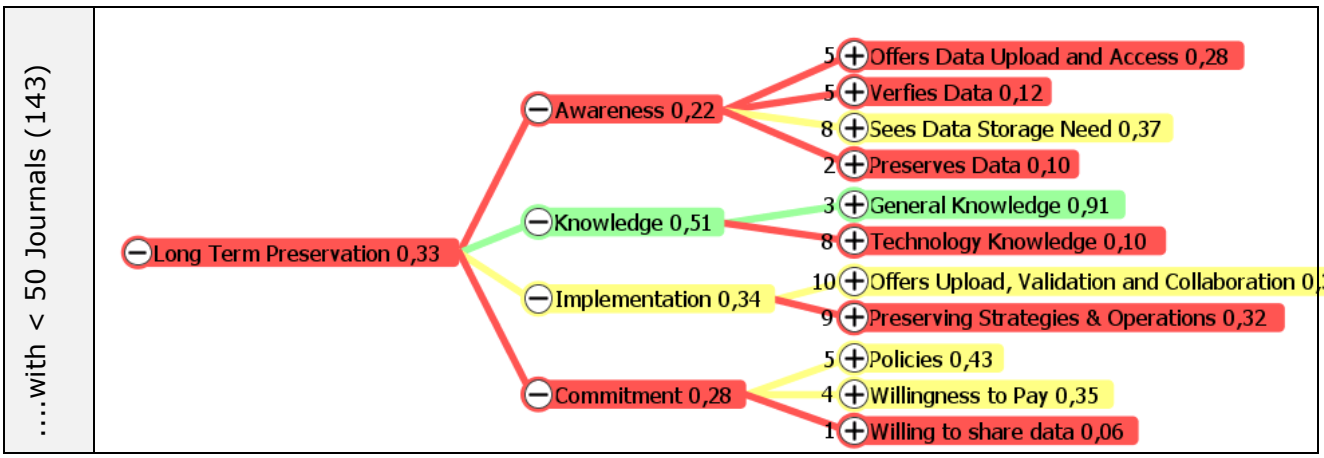
5.2.9 Size of Publisher

After the presentation of the gap analysis results a request for another in depth analysis arose from the expert panel. It was argued, that there should be great differences between "big" and "small" publishers due to their organisational structure, the differences in financial and human resources and their need for preservation. As a measure for the "size" of the publishers the number of journals was expected to be the best factor.

To verify the hypothesis a snapshot of all publishers with less than 50 Journals (143) was produced and compared to the entire data-set (185). The results are shown below:

All Publishers (185)





There are very minor differences in the "knowledge" and "implementation" dimensions but more significant differences in the "awareness" dimensions, indicating that publishers with fewer journals (< 50) are less aware of preservation issues than the "bigger publishers".

The results and findings were fed back to the experts. The difference confirms what the experts from the publisher group have expected: Smaller publishers are often less focused on longer term issues such as preservation and hence have a lower awareness. Also, their IT departments will be smaller (or absent) so that there is no or only little support from that side.

Interestingly the experts expected a similar difference on the other dimensions, in fact they expected a higher score for the bigger publishers on 'implementation' and on 'technology knowledge'. The experts suggested, that the fact that they outsource most of their preservation leads to rather shallow strategic, technical and operational knowledge. This reasonable explanation for the lack of difference was called an "outsourcing effect".

5.3 Discussion of limitations

The gap analysis shows some discrepancies to the insight report due to the different approach of analysis as well as the combination and aggregation of multiple questions. The tool uses the answers from the survey without "knowing" details of the context of the respondents such as their role, their qualification etc. For example most of the variances between the gap analysis result and the perception of experts can be explained by a bias regarding the size of participating companies. Since all answers are counted equally in the tool small publishers with only a few publications have a larger-than-life influence on the overall analysis result compared to their importance in the overall number of publications. It is this sort of insight that is gained through the objective and quantitative approach of the gap analysis.

Another cause for discrepancies between gap analysis results and workshop discussions among the experts is related to the "outsourcing effect": Since many publishers outsource digital preservation to third parties, they may no longer need profound knowledge and extensive implementations. This is not necessarily a cause for trouble since preservation is taken care of by other competent players. Also such information that goes beyond the survey can only be seen by experts and not by the tool itself. But again the gap analysis was a fruitful starting point for a discussion that led to new insights by comparing the straight forward survey results to the multi-dimensional gap analysis results.

The explanation from the experts for the relatively high gaps in the implementation dimension is the emphasis on preservation of "datasets" in the survey. Whilst the preservation of journal articles etc. is in many aspects not a problem for publishers the preservation of corresponding research data is a relatively new phenomena. Thus the "state-of-the-art" of digital preservation amongst publishers is higher than the gap analysis results imply.

Overall it has to be kept in mind, that the tool can only deliver results in a quality that matches the quality of the input. Bias is not only produced by selecting and formulating the survey

questions but of course also by the structuring and weighing of the different survey items by the experts while they build the hierarchical information. The great advantage of the tool is the direct and interactive visualization of the analysis results which stimulates discussion among the experts and thus produces new insights as shown above.

6 Implications for the roadmap

6.1 Refinement of the insight report findings

The tool proved to give deeper insight in the gaps within the stakeholder-group of publishers and scientific libraries and the implications of having gaps in the four dimensions of the framework and how they relate to each other.

Deliverable 4.2 showed that different gaps of different magnitude exist in each of the analysis sub groups of publishers. Within the feedback workshops with the experts these findings were validated and the gap analysis framework and tool proved its value for assessing. Some of the results were different from the workshop discussion and lead to further discussions on the state-of the art of digital preservation in the corresponding communities. While the gap analysis provided a better understanding and a revisiting of assumptions (e.g. for the libraries) the perception on the state-of the art in digital preservation in the publishers workshops were in some cases more positive than the results in the gap analysis. This can be explained, with the fact that the workshop participants were more positive than the survey respondents, since the workshop participants are much more engaged in preservation activities than the average survey respondent.

The gap analysis tool offers a helpful approach for discovering gaps, evaluating their impact and relevance, and starting a discussion within and across communities on how to close them. The **roadmap (D 2.2)** and the **insight report (D 3.6)** list the major threats in preservation (see Figure 9). The gap analysis supports the findings of the insight report but sheds more light on the aspects of "awareness" and "knowledge" on preservation. It is clear, that in order to tackle the threats a lot of "implementation" of technologies and processes as well as political, financial, and personnel "commitment" is necessary. But the gap analysis broadened the perspective taking "awareness" and "knowledge" as important prerequisites into account as well. As the gap analysis results show, there are still a lot of gaps in these two dimensions, suggesting, that building an European preservation infrastructures regarding "implementation" needs more than technologies and in terms of "commitment" more is needed than a "lip service policy". Much more attention needs to be given to enlighten people and institutions on the threats, challenges and opportunities of preservation to raise awareness and promote knowledge.

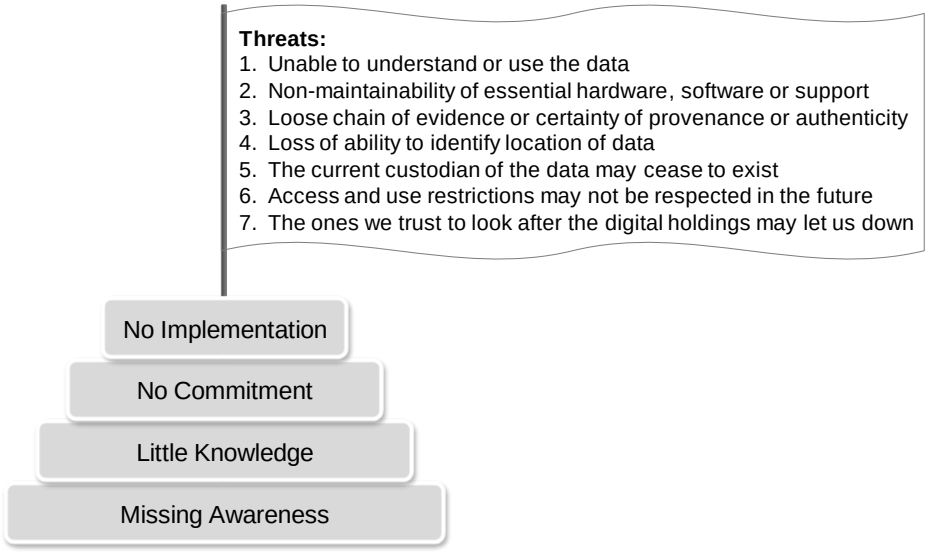


Figure 9: Threats in preservation

Main Objectives for further initiative thus encompass all four dimensions:

- **Foster awareness:** inform about threats of not-preserving and explain the benefits of digital preservation;
- **Foster knowledge:** identify and assess the requirements, means and channels for spreading research results, tools and services offerings;
- **Foster implementation:** push forward the preservation activities from the current fragmented situation characterized by isolated experiences promoted by "Pioneers" and "Early Adopters" into established – pervasive - practices, benefiting large numbers of "Mainstream" users and stakeholders.
- **Foster commitment:** use the valuable information created by PARSE.Insight to promote dialogue and alignment between preservation stakeholders, facilitating implementation of digital preservation strategies, practices and operations.

To close the gaps an interdisciplinary exchange of best practices should be promoted by bringing together stakeholders from different domains (e.g. scientific communities but also others, such as innovative technical industries) to raise awareness, generate and share knowledge and support publishers, libraries, data centres but also scientists in preserving their digital data.

The Gap Analysis Tool should be further improved to gain more valuable insights into existing and future gaps in digital preservation. The Gap Analysis Framework proved its value in PARSE.Insight and is ready to be used by other researches, practitioners, policy makers and all involved in securing the permanent access to the record of sciences in Europe (see deliverable 4.4).

As said before the most important aspect to keep in mind for further studies is to design the survey along the dimensions of the gap analysis framework to best support the methodology. The tool can only be as precise as the input data from the survey. To cover the reality of a complex issue such as digital preservation the most adequate and relevant questions on awareness, knowledge, implementation and commitment have to be asked. Further improvements on the software should concentrate on giving the experts more freedom in specifying the calculation schemas such as individual weights within and between the dimensions and to take into account mandatory questions. For PARSE.Insight the aim was to develop the basic software and make the user interface and the methodology as easy to understand and use as possible. The trade-off between functionality and complexity for the user should be kept in mind when developing the software and the methodology further.

6.2 Dependencies between awareness, knowledge, implementation and knowledge

The gap analysis framework provides a valuable tool to structure the "hot spots" of digital preservation and sheds light on the interdependencies of awareness, knowledge, implementation and commitment on digital preservation which add to the insight report and the roadmap. Based on the results of the gap analysis the relations and dependencies between the four dimensions can be refined as follows:

a) Implementation requires some degree of knowledge

b) Knowledge hardly exists without awareness.

c) Commitment requires awareness and can be strengthened by knowledge.

d) Commitment can exist without implementation which must be considered as a "lip service".

e) Systems can be in place (= implementation) without being used, if the commitment of using them is missing or if no policy actions exist around it (a set of rules and common practices as well as peer pressure).

f) Commitment can be endorsed by policy actions, eg at corporate level or by governments/funding bodies and on a personal level (willingness to use the

implemented systems, understanding the greater good to the community)

Theses dependencies are visualized in Figure 10.

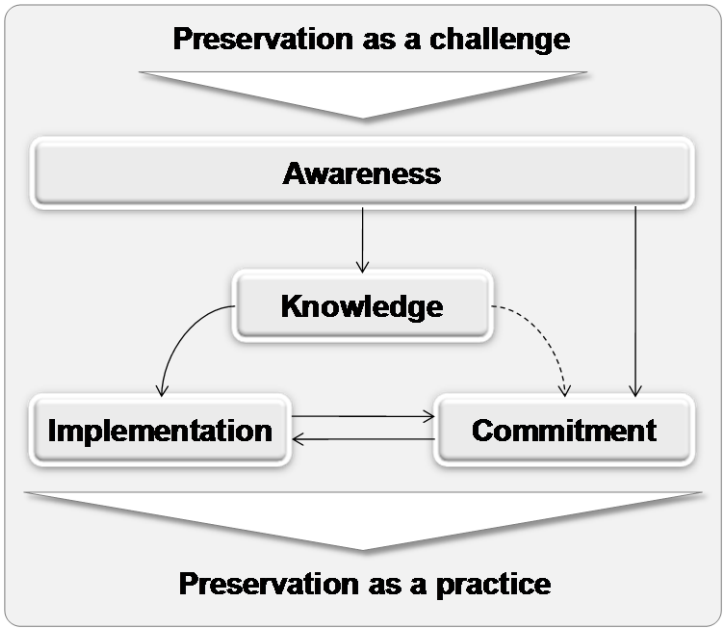


Figure 10: Relations and dependencies of the dimensions

6.3 A common vision for preserving scientific data

The findings from the Gap Analysis Workshops and the application of the tool in the community of publishers and scientific libraries made clear, that a common vision is required for the stakeholders in the scientific publication value chain. The vision cannot be directly derived from the gap analysis finding but from the debate on the findings and on ideas how the gaps could be closed. It was developed from the group of publishers and summarises the findings in form of a user story.[6] This vision can thus be seen as a draft of the preservation environment where all existing gaps are closed exemplified in the publishing and preservation value chain.

The vision in a nutshell:

For many researchers it is important to have good links from datasets to their journal articles, and vice versa, it takes a lot of concern away that their datasets might be re-used out of context. The best metadata on the datasets are after all in their own articles. Publishers are most interested in getting the infrastructure in place to ensure all of this! While publishers have the preservation of journals and their articles well in place (with 96 % of all journal publications covered), the datasets are a new phenomenon and need increasing attention. Libraries and data-centres are needed to ensure the seamless usage of such an infrastructure.

A Scenario:

After several years of hard work and gathering tremendous amounts of raw data, a researcher gets her paper accepted for top-journal AAA – the ground breaking conclusions of her research stem from a new analysis methodology that she developed herself to allow for proper processing of the data. Her paper extensively describes the data set as well as the methodology, in terms of its richness as well as its shortcomings. She decides to make the data available for sharing with other researchers via the subject specific data repository at the National Library in her home country or at an international research institute and adds the software code of the analysis methodology to it. But she has the following concerns:

- She wants anyone looking at the data or at the software to first consult her extensive research paper in journal AAA, to avoid misuse of the data or taking conclusions out of

[6] The text in this paragraph was kindly provided by Eefke Smit from STM and is also printed in the roadmap document.

context.

- She wants any researcher wanting to re-use these data to be aware of the special methodology needed for the proper analysis and how that was applied in the software (hence: to read the paper)
- She wants her dataset to be properly cited when re-used, with a pertinent and persistent link to where the dataset resides, idem for the software code and of course with a link to her original research paper in journal AAA
- She wants the data to be available as a separately citable item, counted in all significant citation scores
- She wants to be sure that any reader of her research paper in journal AAA can easily link to the datasets and the related software.

Her concerns are directly linked digital preservation since reuse is impossible if preservation is not in place.

Next steps:

- Safe and secure data repositories where datasets can be deposited and accessed for reuse, where subject-specific specifications are required, organised per scientific domain.
- EU-wide registry system of persistent identifiers for any deposited datasets, example: TIB Hannover who uses DOI's for datasets (for registration and resolution)
- Publishers include these persistent identifiers in their DTD's and ensure persistent links from published articles to these datasets, e.g. via Crossref
- The data repositories ensure links from datasets to any and all publications that have appeared about them, example: via Crossref
- Extension of the present Citation Systems to include datasets as separately citable items
- Idem for the analysis software belonging to the datasets.
- If time lags and/or embargoes are important for research data sets, this should be enabled by differentiated access and re-use systems, e.g. by using ACAP.

Final Destination:

Research articles always provide links to available datasets and/or the related analysis software, based on persistent identifiers that fit the publication-, linking and citation schemes of publishers (e.g. DOI's). Publishers of research journals encourage their authors to submit and/or deposit their data together with these persistent identifiers and encourage authors to properly cite anyone else's datasets via such identifiers. Publishers adapt their metadata systems for indexing and their DTD's to ensure persistent links and a common citation schemes to datasets (micro-citations) and arrange for free and unhindered access to such datasets. The standards for this will be adopted internationally, not just for the EU and build on existing building blocks (such as DOI's).

- Safe and sustainable dataset repositories ensure the registry and resolution of such international identifiers and provide links from datasets to all related research literature that has appeared about this (e.g. via the forward linking system of crossref.org).
- Research funders define requirements for depositing datasets from research funded by them, including proper identifiers and links to and from research articles.

References

- SurveyMonkey User Manual: http://s3.amazonaws.com/SurveyMonkeyFiles/UserManual.pdf
- The prefuse visualization toolkit: http://www.prefuse.org/
- SQLite: http://www.sqlite.org/
- 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
- CARD, S., MACKINLAY, J. and SHNEIDERMAN, B. 1999. Readings in Information Visualization: Using Vision to Think. Morgan-Kaufmann
- FURNAS, G. 1986. Generalized Fisheye Views. Proceedings of CHI'86: ACM Conference on Human Factors in Computing Systems, Boston, MA, ACM Press, 16-23
- Good Scientific Practice (1998) by the German Research Foundation (DFG) http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf.
- Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003): http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.
- Declaration on Access to Research Data from Public Funding by the Organisation for Economic Co-operation and Development (OECD) http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html.
- TIB Hannover: http://www.tib-hannover.de/en/the-tib/doi-registration-agency/
- Crossref: http://www.crossref.org/
- IDF www.doi.org
- ACAP www.the-acap.org
- JISC www.jisc.ac.uk
- Cambridge Crystallographic Datacentre www.ccdc.cam.ac.uk/
- Editeur www.editeur.org
- Portico
- CLOCKSS/LOCKSS
- e-depot
- Handle system: www.handle.net