PARSE.Insight

Deliverable D4.2
Draft Gap Analysis Report

<table>
<thead>
<tr>
<th>Project Number</th>
<th>223758</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title</td>
<td>PARSE.Insight. INSIGHT into issues of Permanent Access to the Records of Science in Europe</td>
</tr>
<tr>
<td>Title of Deliverable</td>
<td>Assessment of discrepancies between findings from selected surveys and future scenarios</td>
</tr>
<tr>
<td>Deliverable Number</td>
<td>D4.2</td>
</tr>
<tr>
<td>Contributing Work package</td>
<td>WP4: Gap Analysis</td>
</tr>
<tr>
<td>Deliverable Dissemination Level</td>
<td></td>
</tr>
<tr>
<td>Deliverable Nature</td>
<td>Report</td>
</tr>
<tr>
<td>Contractual Delivery Date</td>
<td>31st August 2009 (M17)</td>
</tr>
<tr>
<td>Actual Delivery Date</td>
<td>30st September 2009</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Moritz Gomm, Holger Brocks</td>
</tr>
</tbody>
</table>

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

In this report four main results are presented:

- The developed tool for visual gap analysis based on the surveys of WP3
- Domain-specific knowledge on the lead-user stakeholder-group\(^1\) of publishers
- The application of the framework and tool for the stakeholder-group of publishers
- Identified Gaps

The concept for IT support developed in D 4.1 was implemented as a software prototype enabling an interactive and visual analysis of gaps according to the developed framework based on the survey data. A simple interface allows for an intuitive drill-down into the data with real-time visual analysis feedback.

For the configuration of the tool and to later contextualise the results domain specific knowledge on the analysed stakeholder-group is necessary as described in D 4.1. Thus a lead-user stakeholder-group (“Publisher”) was selected for undergoing the gap analysis process. Within two expert workshops the domain was thoroughly explored delivering not only the required input for the tool but also valuable additional insights into the preservation issues of this stakeholder-group.

Using the domain specific knowledge the tool was configured and applied to the data of the publisher survey from WP3. This deliverable describes the results of this first testing of the framework and the tool.

The identified gaps of the lead-user stakeholder-group in general and specific Gaps in sub-areas are discussed. The next and final step is to validate the results of this draft report with the lead-users and feedback their input into the tool development.

The results deliver valuable insights on existing gaps within long-term preservation and a better understanding of their relevance and impact. The gap analysis tool offers a helpful utensil for discovering gaps, evaluate their impact and relevance, and start 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.

Keywords:
Visual Gap Analysis, Tool Support, Publisher, Front-End, Lead-User Stakeholder-group

---

\(^1\) In this deliverable we refer to the “community of publishers” as “stakeholder-group of publishers” to prevent confusion of the term “community” as it is used for a specific scientific discipline in the rest of the project.
### Contributors

<table>
<thead>
<tr>
<th>Person</th>
<th>Role</th>
<th>Partner</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moritz Gomm</td>
<td>Lead WP 4</td>
<td>FUH</td>
<td>Document owner and author</td>
</tr>
<tr>
<td>Björn Werkmann</td>
<td>Lead WP 4</td>
<td>FUH</td>
<td>Programming &amp; GUI Design and Tool Documentation</td>
</tr>
<tr>
<td>Holger Brocks</td>
<td>Lead WP 4</td>
<td>FUH</td>
<td>Support and Management</td>
</tr>
<tr>
<td>Matthias Hemmje</td>
<td>Lead WP 4</td>
<td>FUH</td>
<td>Supervision</td>
</tr>
<tr>
<td>Eefke Smit</td>
<td></td>
<td>STM</td>
<td>Ideas and Documentation</td>
</tr>
<tr>
<td>Tom Kuipers &amp; Jeffrey van der Hoeven</td>
<td></td>
<td>KB</td>
<td>Workshop and Data</td>
</tr>
</tbody>
</table>

### Document Approval

<table>
<thead>
<tr>
<th>Person</th>
<th>Role</th>
<th>Partner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matthias Hemmje</td>
<td>Lead WP 4</td>
<td>FUH</td>
</tr>
</tbody>
</table>

### Distribution

<table>
<thead>
<tr>
<th>Person</th>
<th>Role</th>
<th>Date</th>
<th>Partner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moritz Gomm</td>
<td>Lead WP 4</td>
<td>15. Sept. 2009</td>
<td>FUH</td>
</tr>
</tbody>
</table>

### Revision History

<table>
<thead>
<tr>
<th>Issue</th>
<th>Author</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
</table>
# Table of Contents

1 INTRODUCTION: PURPOSE AND SCOPE ......................................................................................................................... 5

2 THE GAP ANALYSIS FRAMEWORK ................................................................................................................................. 5
   2.1 FRAMEWORK DIMENSIONS ........................................................................................................................................ 5
   2.2 THE METHODOLOGY ..................................................................................................................................................... 6

3 SOFTWARE COMPONENTS AND ARCHITECTURE ................................................................................................. 7
   3.1 THE WORKFLOW .......................................................................................................................................................... 7
   3.2 SOFTWARE MODULES AND PROCESSES .................................................................................................................... 8
      3.2.1 Gap Analysis Tool .................................................................................................................................................. 8
      3.2.2 Gap Analysis Preparation Tool ............................................................................................................................. 9
      3.2.3 TreeModel ........................................................................................................................................................... 10
      3.2.4 SurveyAccess Module .......................................................................................................................................... 10
      3.2.5 Generator Module ................................................................................................................................................. 10
      3.2.6 FilterEngine Module .............................................................................................................................................. 10
      3.2.7 ComputationEngine Module .................................................................................................................................. 10
   3.3 TECHNOLOGY ............................................................................................................................................................. 10
   3.4 INFORMATION VISUALIZATION ................................................................................................................................... 10

4 THE LEAD-USER GROUP “PUBLISHERS” .................................................................................................................... 11
   4.1 INTRODUCTION ........................................................................................................................................................... 11
   4.2 WORKSHOPS TO GATHER DOMAIN SPECIFIC KNOWLEDGE .................................................................................. 12

5 APPLYING THE TOOL TO THE STAKEHOLDER-GROUP OF PUBLISHERS ............................................................... 14
   5.1 SETTING UP THE TOOL ................................................................................................................................................. 14
   5.2 GAP ANALYSIS ........................................................................................................................................................... 15
      5.2.1 Analysis of overall gaps in the stakeholder-group of publishers ............................................................................... 15
      5.2.2 Gaps: Submission of Research Data ....................................................................................................................... 16
      5.2.3 Gaps: Validation Process ........................................................................................................................................... 16
      5.2.4 Gaps: Preservation Policy .......................................................................................................................................... 17
      5.2.5 Gaps: Online Collaboratories .................................................................................................................................... 18
      5.2.6 Gaps: Disaster-Recovery .......................................................................................................................................... 18
      5.2.7 Gaps: Planning of Preservation Initiatives .............................................................................................................. 19
      5.2.8 Gaps: Economic Value of Data .................................................................................................................................. 20

6 SUMMARY AND NEXT SEPTS ........................................................................................................................................... 21

7 RELATED WORK ............................................................................................................................................................ 22

APPENDIX: FULL LIST OF FINDINGS OF THE LONDON WORKSHOP ................................................................. 23
1 Introduction: Purpose and Scope

The main objective of work package 4 is to identify gaps in the European e-infrastructure, including enabling technologies and corresponding interoperability models. Based on the findings of the draft roadmap and survey results, the discrepancy between the requirements from case studies and future scenarios (WP3) and the developing European research infrastructure is to be assessed in a systematic way.

The survey results from WP 3 reveal the status-quo in long-time preservation of digital data in a variety of countries and institutions. The survey itself was designed with knowledge about digital preservation and is thus based on a model of what exactly digital preservation is (PARSE insight report) and how it should be done (PARSE roadmap). A gap in this context is therefore the difference between the actual implementation of any relevant aspect of digital preservation and its objective requirement for a safe long-time preservation.

The objective of deliverable 4.2 is to document the implementation of the framework and concept into a software prototype and to demonstrate its functionality within a stakeholder-group. The tests result in documented gaps in the preservation landscape of this stakeholder-group.

2 The Gap Analysis Framework

2.1 Framework Dimensions

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-infrastructure landscape of the future.

The actual gap analysis is performed by FUH with support of all work package participants, with the individual foci determined by their respective competencies and backgrounds. The work was strongly supported by the project partners from STM and KB. FUH provides a visual Gap Analysis Tool which extends classical statistical analysis and allows explorative and experimental drill-down into the types and extent of gaps.

As described in detail in deliverable 4.1 the developed 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 form the gap analysis framework are visualised in Figure 1.
The gap analysis is based on the results from the survey in WP3 and defines a step-wise procedure to transform the survey data into the gap analysis framework (modelling the domain, defining target values for gaps, extracting data sets) and interactively and visually analysing the gaps (including reporting, documentation, evaluation and contextualisation).

### 2.2 The Methodology

The gap analysis is based on the results from the survey in WP3. The data within the gap analysis can be separated between information on the “Status-quo of long-time preservation” (e.g. Investment in digital preservation, etc.) and descriptive information on the participants (e.g. volume of data, size of company).

Following the methodology presented in deliverable 4.1 the gap analysis is performed in six steps (see figure 2).

1. **Modelling the Domain**: The data from the survey has to be structured according to the gap analysis framework, thus assigning each question to either of the gap categories awareness, knowledge, implementation or commitment. The resulting “domain tree” structures the *status-quo-information* in 1:n-relationships (e.g. "data types" as a branch and the different formats such as Office-Documents, .jpg etc. as its leafs).

2. **Determine Target Values**: The target values for the questions on digital preservation have to be defined by experts. If these values are reached, the specific gap is zero, i.e. there is no gap.\(^2\)

3. **Extract Data-Sets**: The entire data-set from the surveyed community is then transferred into the domain tree and the status-quo-information is presented according to a

\(^2\) Step 1 and 2 require experts on the specific domain “long-term preservation” and should thus be combined in one activity (e.g. a workshop).
specified set of selection criteria. For each item the gap is calculated as the average difference between the item values and the targets values from Step 3. Each brunch is then calculated. Each leaf, trunk and connecting line is coloured on a spectrum from green over yellow to red according to the calculated gaps (green = no or small gap, red = big gap).

**Step IV “Analyse the gaps”:** The expert user can now analyze and compare gaps by changing the selection criteria and thus drill in and out of the entire data set. Thus more general gaps can be separated from gaps in special areas (e.g. countries, industries etc.). The threshold for the colours can be varied by the user as well to adopt them to the specifics of the analysed data-sets.

**Step V “Reporting and Documentation”:** The user can produce reports or snap-shots from his settings to store, present and communicate them. He can also reload these settings later to continue the gap-analysis.

**Step VI “Evaluation and Contextualisation”:** The last step in the gap analyses process is the evaluation and discussion of results with experts from the analysed domain. This is best done in form of workshop. The goal is to contextualise the results, identify missing aspects (that haven´t been covered by the survey), discuss approaches and solutions for filling the gap, and estimating impact of both the gaps and means to close them.

In the following section the Software Components and Architecture of the Gap Analysis Tool are described.

### 3 Software Components and Architecture

Building on the methodology introduced in Section 2 this chapter describes the concrete workflow when using the GAT (Gap Analysis Tool), the software modules and data entities involved throughout the steps of the workflow process, as well as the technology that is used. The final section will address information visualization aspects.

#### 3.1 The Workflow

It is assumed that the survey is performed with SurveyMonkey or a similar tool and that the data sets comply with the format described in the [SurveyMonkey User Manual] section IV.3, Relational Database Format. The resulting data sets include the following information pertaining to a single survey:

1. Information about all participants
2. Structure of the questionnaire
3. All question items
4. Answer to each question for every participant

This information is pre-processed during the “Preparation Phase” phase using the GAPT (Gap Analysis Preparation Tool), an integral part of the overall system and central to the use of the GAT. After clean-up and conversion, the first step performed during this phase is to parse the survey structure to generate artifacts that support further transformation of survey data into the gap analysis framework. One result is an interactive Excel file that is created based on the survey structure, to support further categorization of survey questions that will lead to the hierarchical analysis framework for the GAT. The analyst will request this information from a small group of experts and consolidate their responses into a final tree structure representing the categorization. For this purpose a “Tree Definition Template” is generated, that can be filled in by the analyst, and that will be used by the GAT. Any customization of the computation model used by the GAT can be provided along with the tree definition, although the default model will be available without further additions. A possible use for the computation model is to assign weights to certain items of the questionnaire to increase or reduce their impact.

All subsequent steps in this phase are executed without user interaction and will create persistent GAT data structures, which, together with the “Tree Definition Template” comprise a complete GAT project. These data structures are described in detail in Section 3.2.
During the “Tool Support” phase, such a project is loaded into the GAT, the data is analysed, and results are documented and prepared for reporting, as described above in Section 2.2.

3.2 Software Modules and Processes

This section describes the software modules constituting both GAPT and GAT. First, a high level view of both GAPT and GAT are given, the description of the constituting modules follows.

3.2.1 Gap Analysis Tool

Figure 3 gives a high level overview of the components that comprise the GAT on a domain level. In the current implementation these components are executed within a single Java Virtual Machine. The View aspects of the application are implemented using Java Swing, the prefuse visualization library, and a JDIC (Java Desktop Integration Components) browser component for windows. The UI components of the FilterView are dynamically created by the GAPT, which also creates the necessary bindings to connect the view to the model and the FilterEngine, to be able to filter the survey data based on user interaction. The three views directly correspond to the three sections marked in Figure 3: Gap Analysis Tool.

![Figure 3: Gap Analysis Tool Domain Model](image-url)
3.2.2 **Gap Analysis Preparation Tool**

Figure 4 gives a high level overview of the components that comprise the GAPT and the process in which they are used, along with the most relevant data objects.
3.2.3 **TreeModel**

The TreeModel is a key component of the GAT, both on a conceptual and implementation level. On an implementation level it reflects the application state and is directly represented to the user in form of the TreeView. Conceptually, both data entities and ComputationModel are aligned around the TreeModel.

3.2.4 **SurveyAccess Module**

This module provides a level of abstraction between the application components and the storage mechanism used to store the Survey data sets. It provides the functionality that is necessary to efficiently execute the filter Operations performed by the FilterEngine.

3.2.5 **Generator Module**

The Generator module Generates code and data from survey data and user input. The following artefacts are created by this module:

1. Persistent Survey data structure
2. Persistent internal data structures
3. Tree
4. FilterSpecification and binding
5. User Interface template and instances
6. ComputationalModel Specification
7. Survey Categorization Excel File

3.2.6 **FilterEngine Module**

In response to a change of the FilterSpecification triggered by the user, this module accesses the SurveyAccess module, determines which questionnaires should still be considered for analysis and updates the TreeModel accordingly. This leads to value assignments only affecting leaf nodes of the tree, further computation is performed by the Computation Module.

3.2.7 **ComputationEngine Module**

This module computes the gap values. In a first step the leaf node values are assigned based on the current state as determined by the FilterEngine. In the following back propagation step, the inner nodes, and finally the root nodes are assigned a value. The computation is affected by the ComputationModel specification specified by the analyst.

3.3 Technology

As previously mentioned, the main applications are executed within a Java Virtual machine. One a source code level both Java and Groovy are used. The GAPT application takes advantage of existing runtime libraries such as iconv, that exist for most platforms, but currently only the Windows implementation is used.

The user interface is implemented as Swing components and uses the prefuse visualization toolkit. The GAPT modules that generate Excel files use Office-Automation mechanisms to access an existing installation of Excel. The SurveyAccess data storage mechanism for the GAT is built on top of an SQLite database binding for Java.

3.4 Information Visualization

The FilterView as well as the RangeView have been implemented according to the established tight-coupling technique [Shneiderman 92] to provide immediate feedback to the user. Using this mechanism the user is enabled to narrow down the choice of relevant information objects for a subsequent drill-down. The drill-down metaphor is based on the described tree visualization, that allows access to the information domain starting on a category level, down to the level of actual data items and detailed information about survey question items as depicted in Figure 5.
With respect to Focus-Context visualization [Card 99] to aid data analysis, the GAT employs regular zoom operations as well as degree of interest (DOI) based pruning techniques [Furnas 86] to collapse whole sub-trees with low DOI as depicted in Figure 6.

4 The Lead-User Group “Publishers”

To “model the domain” in a structure computable by the IT-tool thorough knowledge on the specifics of the domain is required. The selection of a lead-user stakeholder-group and the preliminary steps to configure the tool are described below.

4.1 Introduction

For an in-depth analysis of the framework and to further reflect and contextualise the findings the stakeholder-group of publishers was selected. Reasons were 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.

The first step is to gain domain-specific knowledge in order to model the domain (“domain tree”). The data is taken from the survey from WP3. Before starting to search for gaps within the stakeholder-group it is helpful to revisit the status-quo of this community (reflected in the “PARSE insight report”) under the perspective of the gap analysis framework.

The main findings from the Insight Report can be brought into the context and dimensions 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 hybrid business model, in which subscription-based and open access journals will both exist.
Implementation:

- 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.

In order to better understand the specific gaps and their relevance in each of these dimensions a workshop was organised, bringing together experts of the stakeholder-group of publishers which cover different roles in the publishing value-chain. This community encompasses publishers, libraries, data-centers and in a broader sense includes the producers and (re-)users of scientific data and publications as well (see Figure 7).

![Figure 8: The value chain in the publishing community](image)

4.2 Workshops to gather domain specific knowledge

Preparation

To prepare the workshop and identify the adequate setting for it a preliminary meeting was held on the 10th of February 2008 in Darmstadt. The participants defined the broader goals, both in terms of the planned workshop as well as the overall benefit for the publishers.

The discussion delivered the following eight preliminary results for existing challenges:

1. finding a common language
2. adjusting developments across value chain
3. coping with copyright and charging issues for scientific data
4. Investigate access-side of publishers
5. establish policies for preservation
6. combining of Open Access and Publishing
7. defining common service requirements (formats, processes, standards,...)
8. integrating multimedia data in Publishing

The findings would serve as the input and a starting point for discussion in the workshop with the lead-users.

Workshop

The purpose of the workshop is twofold:

- Understand the publishers’ domain regarding long-term preservation in order to transform the survey and its results from WP3 into the data-structure of the gap analysis
- Identify gaps within the publishers to validate the results from the gap analysis tool and to collect gaps which have not been covered by the survey

---

3 Participants in Darmstadt: Holger Brocks (FUH), Eefke Smit (STM), Gertraud Griepke (Springer), Sabine Schrimpf (DNB), Matthias Hemmje (FUH), Gerald Jäschke (Globit), Moritz Gomm (FUH)
The workshop was held on the 21st of April at the Centre for e-Research at King’s College in London (UK). 15 people from 13 institutions took part in the four hour session. The preliminary results from the PARSE.insight Report were presented and discussed. Afterwards the Gap Analysis Framework was introduced and a group discussion emerged on the basis of the findings from the preparation meeting.

The group discussed the issues and assigned them into the four gap dimensions. It was defined, that an “awareness gap” arises, when the existence of a problem or solution is not known; a “knowledge gap” arises if existing solutions are not known; an “implementation gap” arises, if existing solutions are known, but not installed; a “commitment gap” exists if the implemented solutions are not used or avoided.

The following gaps and challenges derived from them were identified:

**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. 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 following the publications from 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.
- Building an infrastructure and business processes is a complex, moving, and evolutionary objective.

**Knowledge:**
- Common language (e.g. „Archiving“ is not the same as „Preservation“)
- 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.
- Best-Practices (who does what; ongoing efforts) is 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.
- 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.

**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)

---

4 Participants in London: Nicholas Everett (Taylor and Francis), Jonathan Clark (Elsevier), Toni Tracy (Portico), Heike Klingebiel (Springer), Richard Kidd (RSC), Amanda Ward (Nature), Norman Paskin (IDF), Eefke Smit (STM), Alfred Kranstedt (DNB), Tom Kuipers (KB), Gerald Jäschke (Globit), Neil Beagrie (Charles Beagrie), Matthias Hemmje (FUH), Moritz Gomm (FUH), Holger Brocks (FUH)

5 See the appendix for the full list of findings from the workshop.
• 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).
• 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 (cooperative setting), IPR Issues, ACAP (How to use data)
• Setting basics for Certification, Compliance, Codes of Conduct (“Minimal Requirements”)

Other Issues
• The “Value of Data” was also discussed controversially: What is the Value? Who determines it and when? At what time does it occur? When is it worth the investment of preservation? When and only if the project has resulted in a publication?
• This also led to the question of business models and business cases for data preservation: How to create a Business Case? Who would be the “entrepreneur”? Is it necessary to invest in Data to make it available and therefore valuable?

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 Applying the Tool to the Stakeholder-group of Publishers

5.1 Setting up the Tool

Using the domain specific knowledge the survey from WP3 was transformed into the tree structure according to step 1 and 2 of the Gap Analysis Process (“Modelling the domain”). Thus every item from the survey was evaluated regarding its information on each of the four gap dimensions (awareness, knowledge, implementation and commitment).

A total of 60 items were identified and grouped. For example a positive answer on the question “Are you planning to develop a validation process for data submission?” was taken as an indication for “awareness” and can be grouped along with other items under the sub-category “verifies data”.

The following table lists the different sub-categories (in brackets the questions from the survey related to them).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Sub-Categories (Questions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Awareness (20)</td>
<td>Offers Data Upload and Access (8a, 10c, 10d, -38d, 64a)</td>
</tr>
<tr>
<td></td>
<td>Verifies Data (12_1, 12_2, 13b, 13c, 33c)</td>
</tr>
<tr>
<td></td>
<td>Sees Data Storage Need (16e, 16f, -16g, 18b, 28d, 30e, 31d, 35a)</td>
</tr>
<tr>
<td></td>
<td>Preserves Data (21d, 33b)</td>
</tr>
<tr>
<td>Knowledge (11)</td>
<td>General Knowledge(-21f, -22f, -45c)</td>
</tr>
<tr>
<td></td>
<td>Technology Knowledge (22a, 22b, 22c, 22d, 22f, 28e, 30f, 31e)</td>
</tr>
</tbody>
</table>
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 (step 3). 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 in the stakeholder-group of publishers

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

The visualization indicates at a glance, that the knowledge 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 publisher either preserve data 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.

- 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.

- 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.
• **Commitment** is the lowest of the four dimensions. Even though policies are established to some extend the willingness to pay for long-term preservation is low and there is even less willingness to share data.

All questions were analysed using the gap analysis tool. A selection of very interesting Findings are shown and discussed below.

### 5.2.2 Gaps: Submission of Research Data
For example the following visualisation shows all publishers who allow “authors to submit their underlying digital research data with their publication”:

![Diagram showing gaps in submission of research data]

<table>
<thead>
<tr>
<th>&quot;Can authors submit their underlying digital research data with their publication to you?&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;No&quot; (44)</td>
</tr>
<tr>
<td>![Diagram of &quot;No&quot; responses with gaps and values]</td>
</tr>
<tr>
<td>&quot;Yes&quot; (106)</td>
</tr>
<tr>
<td>![Diagram of &quot;Yes&quot; responses with gaps and values]</td>
</tr>
</tbody>
</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 Gaps: 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:

<table>
<thead>
<tr>
<th>&quot;Are you planning to develop a validation process for data submission?&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

---

Page 16 of 24
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, partly implementation) the late-comers still have great awareness, implementation and commitment gaps and also the knowledge is still on a moderate level.

5.2.4 Gaps: 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:
The publishers who don’t have a policy for preservation 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 and obviously also the little need for preserving data.

Those companies who have a policy in place also have much higher values in the awareness and implementation dimension.

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 **Gaps: Online Collaboratories**

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

![Diagram](image)

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. The direct benefit in customer satisfaction seems to be a good lever, than as can be seen above, the commitment itself isn’t so high.

5.2.6 **Gaps: 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. Every second responding company had such a system in place:

![Diagram](image)
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 Gaps: 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?”
**5.2.8 Gaps: 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?”**

<table>
<thead>
<tr>
<th>“very important” (112)</th>
<th>“&lt; 1 year” (51)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Long Term Preservation 0.35</strong></td>
<td><strong>Long Term Preservation 0.43</strong></td>
</tr>
<tr>
<td><strong>Commitment 0.25</strong></td>
<td><strong>Commitment 0.45</strong></td>
</tr>
<tr>
<td><strong>Implementation 0.34</strong></td>
<td><strong>Implementation 0.40</strong></td>
</tr>
<tr>
<td><strong>Knowledge 0.49</strong></td>
<td><strong>Knowledge 0.45</strong></td>
</tr>
<tr>
<td><strong>Awareness 0.32</strong></td>
<td><strong>Awareness 0.42</strong></td>
</tr>
<tr>
<td>5 Offers Data Upload and Access 0.28</td>
<td>5 Offers Data Upload and Access 0.36</td>
</tr>
<tr>
<td>5 Verifies Data 0.22</td>
<td>5 Verifies Data 0.31</td>
</tr>
<tr>
<td>8 Sees Data Storage Need 0.32</td>
<td>8 Sees Data Storage Need 0.42</td>
</tr>
<tr>
<td>2 Preserves Data 0.40</td>
<td>2 Preserves Data 0.61</td>
</tr>
<tr>
<td>3 General Knowledge 0.89</td>
<td>3 General Knowledge 0.78</td>
</tr>
<tr>
<td>6 Technology Knowledge 0.08</td>
<td>6 Technology Knowledge 0.12</td>
</tr>
<tr>
<td>10 Offers Upload, Validation and Collaboration 0.36</td>
<td>10 Offers Upload, Validation and Collaboration 0.43</td>
</tr>
<tr>
<td>9 Preserving Strategies and Operations 0.32</td>
<td>9 Preserving Strategies and Operations 0.37</td>
</tr>
<tr>
<td>5 Policies 0.41</td>
<td>5 Policies 0.61</td>
</tr>
<tr>
<td>4 Willingness to Pay 0.29</td>
<td>4 Willingness to Pay 0.47</td>
</tr>
<tr>
<td>1 Willing to share data 0.04</td>
<td>1 Willing to share data 0.22</td>
</tr>
</tbody>
</table>

Those companies who have not planned any further steps have great gaps in awareness, implementation and commitment, whilst those with short term plans have high values in all dimensions except in “implementation”. This is very consistent with the question on further implementation.
6 Summary and Next Septs

The tool proved to give much deeper insight in the gaps within the stakeholder-group of publishers and the implications of having gaps in the four dimensions of the framework and how they relate to each other. The focus on the publishers does not limit the usage of the gap analysis on other stakeholders or communities within and outside of the PARSE.Insight project.

Awareness, implementation and commitment gaps were measured very well in the survey. Unfortunately there wasn’t enough space to cover sufficient aspects in the knowledge dimension within the questionnaire in WP3. Thus the interpretation of the results regarding “knowledge gaps” in the stakeholder-group of publisher is limited.

The analysis show that different gaps of different magnitude exist in each of the analysis sub groups of publishers. Within task 4.4 these findings will be fed back to the publishers using follow-up expert workshops. The aim is not only to further validate the tool and contextualise the gap results but also to estimate the impacts of these gaps and to find means of closing them.

The results deliver valuable insights on existing gaps within long-term preservation and a better understanding of their relevance and impact. The gap analysis tool offers a helpful utensil for discovering gaps, evaluate their impact and relevance, and start a discussion within and across communities on how to close them. The tool and methodology will be available for other communities and stakeholder-groups for in depth analysis of domain-specific gaps.

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.
7 Related Work

Following a list of selected related work, that has been used as inspiration for the gap analysis framework.

- **The prefuse visualization toolkit**: [http://www.prefuse.org/](http://www.prefuse.org/)
- **SQLite**: [http://www.sqlite.org/](http://www.sqlite.org/)

A list of related projects was gathered by the workshop members:

- Crossref: [http://www.crossref.org/](http://www.crossref.org/)
- IDF  [www.doi.org](http://www.doi.org)
- ACAP [www.the-acap.org](http://www.the-acap.org)
- JISC [www.jisc.ac.uk](http://www.jisc.ac.uk)
- Cambridge Crystallographic Datacentre  [www.ccdc.cam.ac.uk/](http://www.ccdc.cam.ac.uk/)
- Editeur [www.editeur.org](http://www.editeur.org)
- Handle system: [www.handle.net](http://www.handle.net)
Appendix: Full List of Findings of the London Workshop

**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.

- Most if not all large and mid-size publishers preserve publications in their final version (Version of Record) but cannot guarantee properly the preservation of datasets that authors submit together with manuscripts. Also, datasets are regarded per se to be ‘Open access’ and not copyrightable (Brussels Declaration).

- The main common objective is to ensure a sustainable and persistent linking system between datasets and all publications referring to them. 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 following the publications from the original research.

- Publishers have a successful track record in undertaking such initiatives jointly and in a wider context; see for example the Crossref linking system, the support for DOI’s via the IDF, sharing and making available DTD’s, etc. For digital preservation, a similar 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.

- Building an infrastructure and business processes is a complex, moving, and evolutionary objective – but within STM most parties are used to positively working together on matters like this.

**Knowledge:**

- Common language (e.g. „Archiving“ is not the same as „Preservation“) – see above under small publishers.

- 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.

- Best-Practices (who does what; ongoing efforts) – often the best way to achieve common standards and practices.

- Maturity of sharing and preserving research data is different for each discipline; publishers get most requests for dataset submissions in the area of genetics, earth sciences, oceanography, meteorological sciences, computational sciences in general and chemistry. In genetics and in chemistry several databases and a common nomenclature exist to which can be referred and linked. To predict better 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.

- Clear roles and responsibilities: TIB Hannover is a good example of a situation in which responsibilities have been made clear and have been divided well over the publishing chain. Researchers submit datasets of their research at TIB, where they get a DOI, publishers use these DOI’s for linking to the publications. Researchers will also want to be able to define access and re-use terms concerning their datasets, perhaps ACAP-protocols can help here.

- TIB is now entering a European partnership with ETH Zurich, BL, INIST (France), TIC (DK), TU Delft (NL) to set up a similar registration agency at the European level, also based on DOI’s. This clearly will help answer the question: Who-archives-what-and-where-and-with-whom; it would be good if clear policies, common standards and common practices were being promoted.

**Implementation:**

- Build on the existing building blocks: make sure the new infrastructure is interoperable
with what exists. Eg make use of what has been established with DOI’s, TIB-initiative, ACAP, Crossref, Portico, Clockss.

- International approach needed; STM publishers work internationally.
- Internet based (Multimedia, “mashed-up data”) – not per se web-based.
- Lack of established standards (a minimal sets of requirements would be helpful); publishers are more than willing to help establish these.
- Linkage in both directions Data ⇨ Publication, (Handles, DOI) MPDL, TIB, Data-Centers (Common Meta-Data), back links, Persistent Identifiers, Author ID’s.
- Technology required (e.g. Storage space and costs, digital tool-sets)
- Preservation of publications is well addressed, not so with preservation of data, How to cite data? (“micro citation”) How to link to them?
- Central information point (Where can I find more about this?)
- “audit trail” (life cycle of publication)
- Role of supplements and annexes with datasets to publications(not peer reviewed, indexed by title)
- Very basic: Link to web page of author/institute
- At the point of publication the data is often not indexed and stored yet (so no linkage in article possible)

**Commitment:**

- Policies und incentives are needed, probably mostly so for the researchers to deposit and share research data.
- Central archiving and linking should become a prerequisite for researchers to obtain research funding
- Legal or moral obligations are one thing, the best incentive for individual researchers is getting credits for data, linking to official publications will help, but also a common way of citing datasets.
- If time-lags and embargoes are important (researchers willing to share data only later or upon/after publication of the research results), the policies should take this into account. Find out the reasons why so many researchers do not want to share data (competitiveness, proving wrong, discipline specific sharing culture, Maybe willing to share later?, efforts)
- A multiple preservation strategy is preferred with diversification in approaches – not put all eggs in one basket.
- Publishers are very committed to support and help deploy a system for linking between datasets and publications to help discover the “Undiscovered value of data”, but datasets are not ‘owned’ by publishers, it belongs to the author or the research institute, so some policy for good cooperation is needed here.
- At present, 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.
- International standards are important; Publishers cannot work with a complex variety of national or European requirements and geographical fragmentation (interoperability).
- Infrastructures can be established faster if promoted and supported by an Alliance of the willing (cooperative setting), IPR Issues, ACAP (How to use data)
- The infrastructure should set the basics for Certification, Compliance, Codes of Conduct and a Minimal Set of Requirements