PARSE.Insight

Deliverable D4.1

Specification of gap analysis schema and tool support

Project Number	223758
Project Title	PARSE.Insight. INSIGHT into issues of Permanent Access to the Records of Science in Europe
Title of Deliverable	Specification of gap analysis schema and tool support
Deliverable Number	D4.1
Contributing Work package	WP4: Gap Analysis
Deliverable Dissemination Level	
Deliverable Nature	Report
Contractual Delivery Date	31 November 2008 (M10)
Actual Delivery Date	11 May 2009
Author(s)	Moritz Gomm, Holger Brocks

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

Abstract

*This deliverable describes the development of the gap analysis schema and corresponding tool support. First a **gap analysis framework** was developed, thus eliciting and structuring the relevant dimensions and corresponding attributes for future e-infrastructures into a formal schema. Then a **stepwise procedure** was developed for conducting the gap analyses, providing methods and metrics to identify gaps within the European preservation infrastructure, although later broadened somewhat top include gaps in a broader science data infrastructure. Finally an appropriate tool support for the gap analysis was conceptualised. This included aspects of modelling, data management, data analysis and reporting.*

The developed gap analysis framework encompasses two dimensions: firstly the **life-cycle of scientific data** *including production, publication, storage and re-use of scientific data and secondly the types of gaps in the **diffusion of long-term preservation within scientific communities**. According to our framework 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.*

For performing community and industry specific gap analysis a concept for IT support of the stepwise procedure was developed, which helps to identify and visualizes gaps within the data gathered in WP 3.

Keyword list

Gap Analysis, Gap Analysis Framework, Tool Support, Gap Types, Front-End

Contributors

Person	Role	Partner	Contribution
Moritz Gomm	Lead WP 4	FUH	Document owner and author
Björn Werkmann	Lead WP 4	FUH	Programming & GUI Design
Holger Brocks	Lead WP 4	FUH	Concept Support and Management
Matthias Hemmje	Lead WP 4	FUH	Management and Supervision

Document Approval

Person	Role	Partner
Matthias Hemmje	Lead WP 4	FUH
David Giaretta	Coordinator	STFC

Distribution

Person	Role	Date	Partner
Moritz Gomm	Lead WP 4	11. May 2009	FUH

Revision History

Issue	Author	Date	Description
0.1	Moritz Gomm	01 May 2009	Initial draft Word document based on Wiki
1.0	Moritz Gomm	11 May 2009	Final version

Table of Contents

1 Introduction: Purpose and Scope

The main objective of work package 4 is to identify gaps in the European preservation e-infrastructure, including enabling technologies and corresponding interoperability models; this was later broadened to include other aspects of a broader European Science Data Infrastructure. Based on the findings of the draft roadmap (D2.1) 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. In a broader sense, a gap analysis determines "the space between where we are and where we want to be", and serves as a means to bridge that space.

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 mental model of what exactly digital preservation is and how it should be done. 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. To explore the types of gaps the following framework was developed. All partners were engaged in this task which is led by STFC.

2 Gap analysis framework

To develop a formal gap analysis framework the relevant dimensions and corresponding attributes had to be identified and structured. Based on this a stepwise, systematic procedure for assessing discrepancies between the requirements for permanent access and the actual European e-infrastructures landscape was then developed (see Chapter 3).

The actual gap analysis is performed by DNB with support of all work package participants, with the individual foci determined by their respective competencies and backgrounds. 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. All findings and interpretations of the gap analysis have been feedbacked to the relevant actors to validate and contextualize the results and further improve the methodology.

The developed gap analysis framework encompasses the life-cycle of scientific data and the diffusion of long-term preservation within scientific communities. Both dimensions are explained below.

2.1 Life-cycle of Data

The first and *domain-specific dimension* of the gap analysis framework encompasses the life-cycle of scientific data form creation to publishing. This dimension is represented in the survey and segmented as follows:

- Creation and use of data (by producer)
- Data re-use (by others)
- Data preservation (by any institution)
- Publishing (by a publisher)

2.2 Diffusion of the concept of "long-term preservation of data"

The second and *task-specific dimension* of the gap analysis framework covers the levels in the diffusion of a (new) concept such as "long-term preservation of data" within and across scientific communities. Four levels were identified:

Awareness – The stakeholders need to be aware, that long-term preservation is a relevant issue for them in terms of advantages from it (chances) or threats of not using it (risks).

Knowledge – If stakeholders are aware of the issue it is a matter of knowing about the issue in terms of available infrastructures (technologies, standards, methods, etc.).

Implementation – The infrastructure and its components are only of value, if they are properly implemented within the institutions and communities.

Commitment – The infrastructure has to be used by the individuals within the institutions and communities and not circumvented to be effective.

The two orthogonal dimension form the gap analysis framework are visualised in Figure 1.

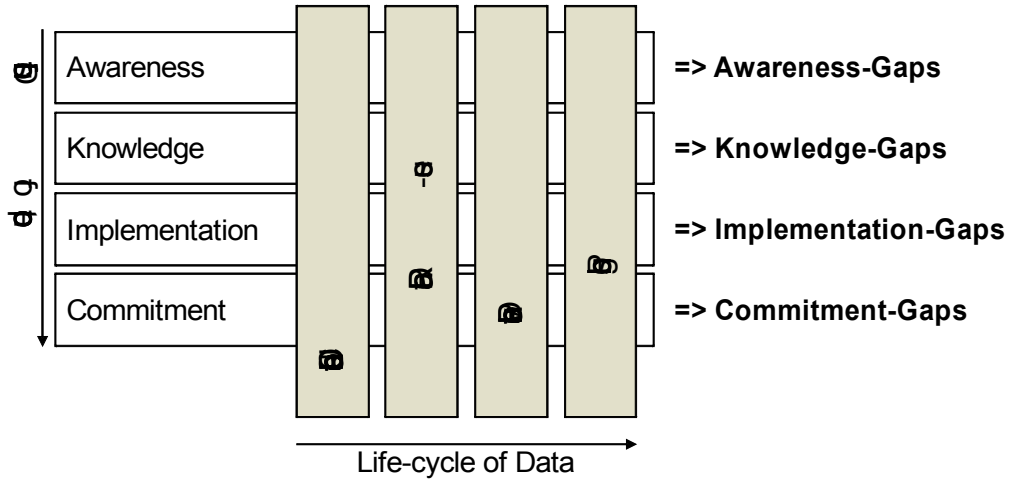

Figure 1: Gap analysis framework

2.3 Gap Analyses Methodology and Process

The gap analysis is based on the results from the survey in WP3. The surveys form a extensive data set, that allows to systematically analyze and „measure" the gaps in long-term preservation within different communities.

The data within the gap analysis can be separated into two different types:

- Status-quo of long-time preservation (e.g. Investment in digital preservation, etc.)
- Descriptive information (e.g. volume of data, size of company, country)

In order to analyse the gaps within the framework some extraction and transformation of the survey data is required. The gaps will then be analyzed using mathematical and visual algorithms. Finally conclusions from the gap analysis can be drawn by the project consortium with feedback and discussion with experts within the communities. For that a reporting function is planned for the tool.

The developed gap analysis process encompasses six steps:

- Step I: Modelling the domain
- Step II: Define targets values for gaps
- Step III: extract data sets
- Step IV: analyse the gaps
- Step V: Reporting and documentation

- Step VI: Evaluation and Contextualisation

Each step is described below.

2.3.1 Step I: 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. Furthermore the results from the questions in the survey have to be differentiated between *status-quo information* (revealing information about gaps, e.g. "what type of gap?") and descriptive information (serving as selection criteria in the analysis, e.g. "who has these gaps?").

The 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). The relevant *descriptive information* in the survey serves as the selection criteria in order to drill into and analyze gaps for special groups or subsets (e.g. countries, research institutions, SME etc.).

2.3.2 Step II: Define targets values for gaps

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. The target values are implicitly part of the mental model underlying the survey. They can be easily determined by answering the survey from a viewpoint of some who has full awareness, knowledge, implementation and commitment of longterm-preservation.

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 or a meta-survey).

2.3.3 Step III: extract data sets

The entire data-set from a surveyed community (e.g. "publishers") is now transfered into the domain tree and the status-quo-information are presented according to a 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).

2.3.4 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.

2.3.5 Step V: Reporting and documentation

The user can produce reports or snap-shots from his/her settings to store, present and communicate them. (S)He can also reload these settings later to continue the gap-analysis.

2.3.6 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. This will be linked with developments on the Roadmap (WP2), particularly the legal, social and organisational aspects, as well as the results of WP6 (Sustainability).

3 Concept of the IT Tool Support for the Gap Analysis

FUH is developing an appropriate IT support for conducting the gap analyzes following the process described in Section 2. The concept of the tool is described here using a mock-up of the frontend, which is currently being implemented (see Figure 2).

The aim of the IT tool is to easily visualize and drill down into the survey data in order to identify and analyze those gaps for the formulation of an effective roadmap and possible impact analysis.

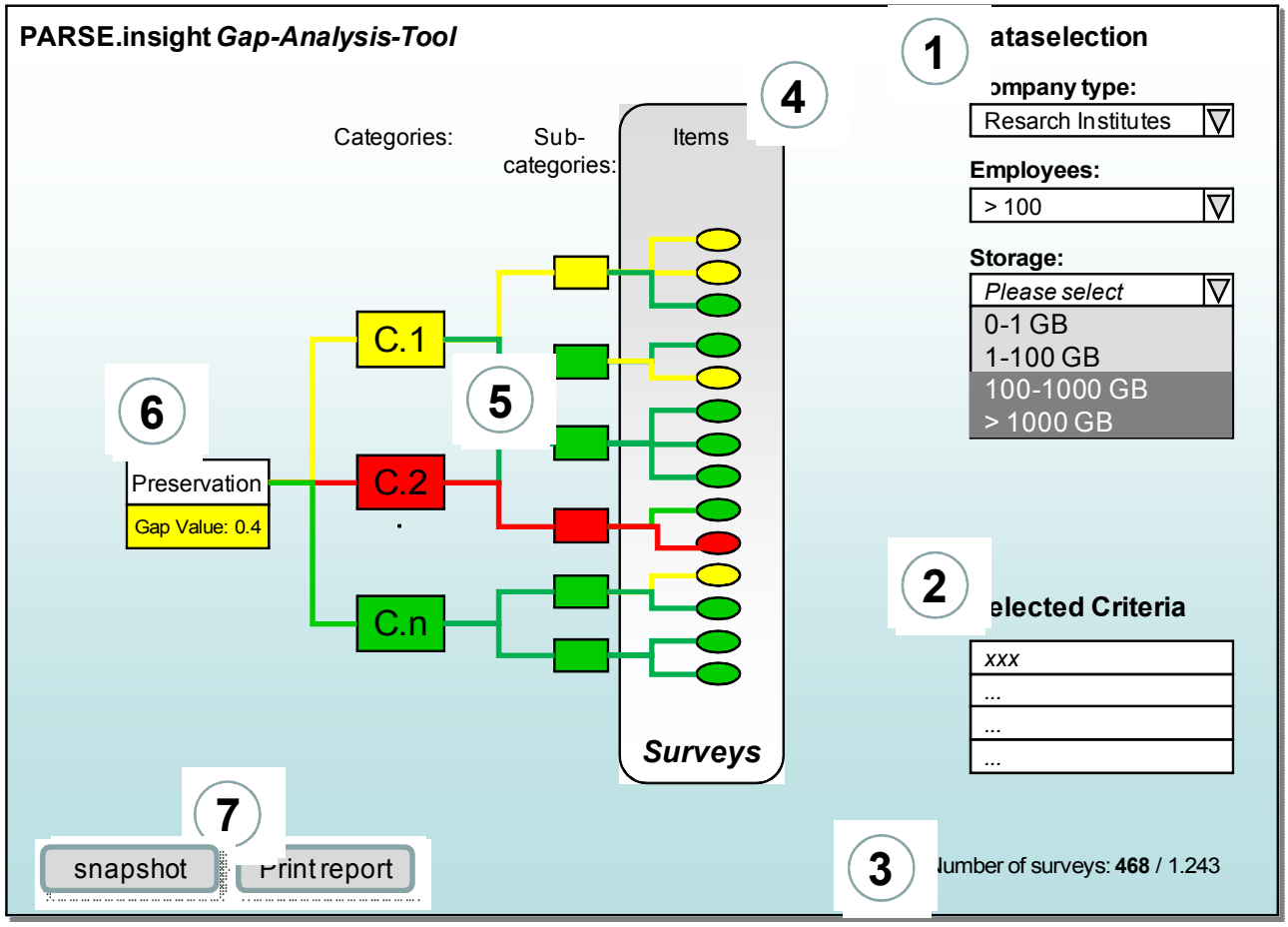

Figure 2: Prototype Front-end for Visual Gap-Analysis

The following seven areas in the front-end are described below:

1. In the “Data selection” all the items of the survey are shown and can be selected by the user. The tool then updates the data set accordingly (e.g. “show me all surveys of institutions without a long-term preservation policy”).
2. All currently selected criteria from (1) are listed here.
3. The resulting set of surveys are shown here together with the total set.

4. All surveys, that match the selection criteria are visualised in the tree diagram. According to the average calculates values for the data set the colours are set (green = no or small gap, yellow = medium gap, red = big gap).
5. The values are also calculated for the entire hierarchy to the root of the tree, showing at one glance, where the gaps are (in the example in the category C.2 e.g. “knowledge about long-term preservation).
6. The root shows the total calculated gap value for the selected data set.
7. For saving, circulating and presenting the results, snapshots can be made and stored. Furthermore brief reports can be produced.

4 Related Work

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

- The **Research Data Strategy Working Group** has completed a gap analysis of research data stewardship in Canada. Using the data lifecycle as a framework, the report examines Canada's current state versus an 'ideal state' based on existing international best practices across 10 indicators. The indicators include: policies, funding, roles and responsibilities, standards, data repositories, skills and training, accessibility, and preservation. Link: http://data-donnees.gc.ca/eng/news/gap_analysis.html
- A research and development gap analysis conducted by **NASA**: The objective was to update and extend a previously produced research and development gap analysis to identify research gaps in the new AATT ATM/OPSCON 2002. It provides an assessment of gaps in the research applications needed to achieve the enhancements to the NAS services described in the AATT ATM/OPSCON - 2002 Update. Link: http://vams.arc.nasa.gov/activities/opscons_archive/gaprpt/gap_toc.html
- **eGovRTD2020** gap analysis: The goal was to assess the differences between today and possible future outlooks for eGovernment research. This includes investigation of the future scenarios in respect to the current research taking place and eliciting research gaps to be addressed and measures to be taken to implement the future scenarios. Link: http://www.egovrtd2020.org/navigation/work_packages/wp3_gap_analysis
- **e-Science Gap Analysis** funded by the National e-Science Centre: This report analyses the current state of Grid and Cyberinfrastructure technology with respect to their use in e-Science, based largely on personal interviews with UK e-Scientists augmented with discussions with some European and US Grid experts. The report goes further than just identifying the gaps, it builds a possible development program, which aims to deliver by 2006 Grid technologies supporting the e-Science functionalities identified in the report. Link: http://www.nesc.ac.uk/technical_papers/UKeS-2003-01/