Insight into digital preservation of research output in Europe

Inventory of Communities

Project Number	223758
Project Title	PARSE.Insight: INSIGHT into issues of Permanent Access to the Records of Science in Europe
Title of Deliverable	Inventory of Communities
Deliverable Number	D3.2
Contributing Work package	WP3: Community Insight
Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	31st May 2008 (M3)
Actual Delivery Date	13th May 2009 (M15) Updated: 9th June 2010
Author(s)	Beate Sturm, Sabine Schrimpf, Clemens Bergmann, Björn Werkmann, Jeffrey van der Hoeven

PARSE.Insight (INSIGHT into issues of Permanent Access to the Records of Science in Europe) is a two-year project co-funded by the European Commission under the Seventh Framework Programme (FP7), Research Infrastructures.

Abstract

The purpose of the PARSE.Insight deliverable D3.2 *Inventory of communities* is to identify the communities in long-term preservation of digital data that are of interest to PARSE.Insight, to describe their tasks and responsibilities and to enumerate the key players, which will be visualized in an interactive map. Furthermore, it points out how PARSE.Insight will reach the several actors in order to perform the general survey and the in-depth studies.

Keyword list

Actor, stakeholder, discipline, community, funder, policy maker, preservation, research data, publisher, researcher, target group, contacts database, interactive map

Contributors

Person	Role	Partner	Contribution
Beate Sturm	Lead task 3.2	MPDL	Document owner and author
Jeffrey van der Hoeven	Lead WP 3	KB	Co-author of section about actors and contacts database
Sabine Schrimpf	WP3 participant	DNB	Co-author of section about actors
Clemens Bergmann	WP3 participant	DNB	Design and development of Interactive Map
Björn Werkmann	WP3 participant	FUH	Development and supervisor of Interactive Map

Document Approval

Person	Role	Partner
Jeffrey van der Hoeven	WP3 leader	KB
Simon Lambert	Reviewer	STFC

Revision History

Issue	Author	Date	Description
0.1	Beate Sturm (UGOE)	12-11-2008	Initial version of D3.2
0.2	Tom Kuipers (KB)	20-03-2009	Internal review of the whole document; comments incorporated
0.3	Björn Werkmann (FUH)	02-04-2009	Internal review of chapter 4.2; comments incorporated
0.4	Jeffrey van der Hoeven (KB)	13-05-2009	Incorporated small changes, added illustration for research domains
1.0	Jeffrey van der Hoeven (KB)	09-06-2010	Included design of interactive map, created by Clemens Bergmann. Finalised document in new template.

I. Executive Summary

In order to provide insight into long-term preservation of digital research data in Europe, PARSE.Insight performed a general scan of the actors involved. Based on these actors we defined our strategy for gaining insight into these actors. This document presents a classification of the actors involved and offers a way to manage and visualise these actors in Europe.

PARSE.Insight has defined following groups of stakeholders (grouping all actors):

- **Researchers**: the stakeholder that is responsible for doing research (researchers, research managers all part of a research institute).
- **Data managers**: the stakeholder that manage research output. This could be done by profit or non-profit data archives, traditional memory institutions (libraries, archives and museums), and organisations doing research and development in preservation technology.
- **Funders/Policy makers:** the stakeholder that is concerned with creating policies for research and funding research.
- **Publishers**: the stakeholder that disseminates research output via books, journals or other means.

All stakeholders have their own interests, roles, rights and responsibilities and therefore need a special part in the infrastructure. In the subsequent deliverables of PARSE.Insight these stakeholders will be closely examined. PARSE.Insight will also learn about who the key players on digital preservation in Europe will be. We defined four criteria for an organisation to be a key player in its field:

- play an important role in preservation of digital research data;
- are representative of wider communities and their interests;
- invest in projects that deal with digital long-term preservation and;
- for this reason do pioneer work in preservation of digital research data.

To manage all information about the stakeholders, a Contacts Database has been created. This online database is available at: http://www.parse-insight.eu/contacts/index.php

The PARSE.Insight Interactive Map visualizes all information from the Contacts Database. This map is available at: http://www.parse-insight.eu/imap_intro.php

II. Table of Contents

1. Introduction

Scientific research[1] produces data and publications. Data curation and preservation is an ongoing and necessary process. For the moment, primary research data are preserved by research institutes and/or data archives; publications are preserved by the institutions themselves, libraries and publishers. This traditional division of tasks is currently undergoing changes with the emergence of 'going digital', open access publishing and a move towards a one-stop repository covering both primary research data and publications as well as the necessity to interlink them to make sure that their relations maintain persistent. In this report, the whole spectrum from primary research data to publications in digital form are addressed by the overall denominator *digital research data* or *scientific research data*.

The actors who create, use and preserve data as well as those who fund, manage and plan these actions have different levels of awareness of activities in the area. In some cases they are not aware of all similarities and differences in their work. There is a danger of duplication, inconsistent approaches or wasted effort. Therefore the aim of Work package 3 (Community insight) of PARSE.Insight is to identify who is doing what in the field of preservation of digital scientific information in Europe and to find out why this is done with the objective of giving the PARSE.Insight project a comprehensive overview of the state of the field.

1.1. Purpose of this document

In order to provide insight into the current state of the art in the e-infrastructure in long-term preservation of digital research data, PARSE.Insight performed a general scan of the actors involved and three in-depth case studies (questionnaires and interviews) within different scientific disciplines. Therefore, comprehensive insight into the relevant communities is needed to identify the actors and to define the target groups for both the general survey and the case studies. This document describes the classification of the actors involved and presents ways to manage and visualise these actors.

1.2. The structure of the document

This report is divided into two parts. The first part identifies the actors in long-term preservation of digital research data, describes how they are divided into categories and to which communities they could belong. Furthermore, it describes the actors' tasks and responsibilities within the area of long-term Preservation of digital research data and points out how the key players within these communities are defined. The second part illustrates how PARSE.Insight will try to reach the

[1] The use of the word 'science' is not intended to exclude the humanities.

key players and manage the contacts in order to perform the general survey and the in-depth case studies mentioned above.

The Description of Work (DoW) envisaged that this deliverable would also include an account of the methodology adopted for the survey and case studies. However, the project team considered that this would be better included in the corresponding deliverables (D3.3 *Case studies report*, D3.4 *Survey report* and D3.6 *Insight report*), and thus it does not appear here.

2. Actors

2.1. Actors in long-term preservation of digital research data

Long-term preservation is a strategy for a trustworthy and persistent preservation of and access to digital data such that they remain accessible and usable for decades. Digital preservation begins before the digital record is created and continues for as long as the digital object needs to be retained. Hence it covers all activities concerning the data: creation, use/reuse and publication as well as preservation and storage. The actors who are involved in these steps, according to the data life cycle model, are:

- **Creators** of digital research data: researchers[2], industry
- **Users** of digital research data: researchers, industry
- **Providers** of digital research data: data archives, commercial providers, traditional memory institutions
- **Funders/Policy makers**
- **Funders** research funders, project funders, institutional funders
- **Developers** of preservation components and infrastructure
- **Publishers** of research publications and online databases, often interlinking these with original research data

Although the tasks in long-term preservation are not always well-defined – e.g. a researcher is often involved in preservation and storage as well – PARSE.Insight divides these actors into the following stakeholders:

- **Researchers**: the stakeholder that is responsible for doing research (researchers, research managers all part of a research institute).
- **Data managers**: the stakeholder that manage research output. This could be done by profit or non-profit data archives, traditional memory institutions (libraries, archives and museums), and organisations doing research and development in preservation technology.

[2] Note that the word ‘researcher’ refers to a scientist who generates or uses data in his/her field of endeavour. Of course there is research in digital preservation technology itself, but here we use the word ‘developer’ to avoid confusion.

- **Funders/Policy makers:** the stakeholder that is concerned with creating policies for research and funding research.
- **Publishers**: the stakeholder that disseminates research output via books, journals or other means.

All these stakeholders have their own interests, roles, rights and responsibilities and need a special infrastructure to perform their tasks. In the subsequent deliverables of PARSE.Insight these stakeholders will be closely examined. Each stakeholder is explained in more detail hereafter.

2.2. Stakeholder categories

2.2.1. Researchers

All researchers create, use and reuse research data. Research data is the key output of these activities. For different reasons many institutions are involved in the preservation and storage of these data (motivated for example by the high investment in research, that other researchers may need and replicate the research, data is routinely made available to reviewers, data is impossible to recreate).

Often researchers are the only ones who are responsible for preservation and storage of data. As creators of data, they are responsible for:[3]

- managing data for the duration of the project
- making the data available in a form that can be used by others
- using standards where possible
- complying with data policies
- disseminating their research work by writing articles and other publications

As users of third-party data, researchers are responsible for:

- adhering to any license and restrictions of use
- acknowledging data creators and curators
- proper citation of and reference to previous work and data as found in publications

[3] The description of roles and responsibilities refers to Liz Lyon, Dealing with Data: Roles, Rights, Responsibilities and Relationships, 2007.

- managing every derived data
- providing feedback to the research community and data archives

The research institutions are responsible for:

- setting internal data management policy
- managing data in the short term
- meeting standards for good practice
- advocacy and training
- possibly for the development of institutional repositories

In research, different communities exist, depending on which field of research they belong to. These fields are widely recognized categories of specialized expertise within research, but they are not always classified similarly[4],[5]. PARSE.Insight has chosen to base its classification on the way research is classified by the Dutch KNAW (Royal Netherlands Academy of Arts and Sciences) [6] . The following classification in main categories / disciplines has been made (see figure 1):

- Agriculture and Nutrition
- Behavioural Sciences
- Humanities
- Life Sciences
- Medicine
- Physical Sciences
- Socio-cultural Sciences
- Technology

[4] http://en.wikipedia.org/wiki/Fields_of_science

[5] http://www.sciencemadesimple.com/science-definition.html

[6] http://www.knaw.nl/

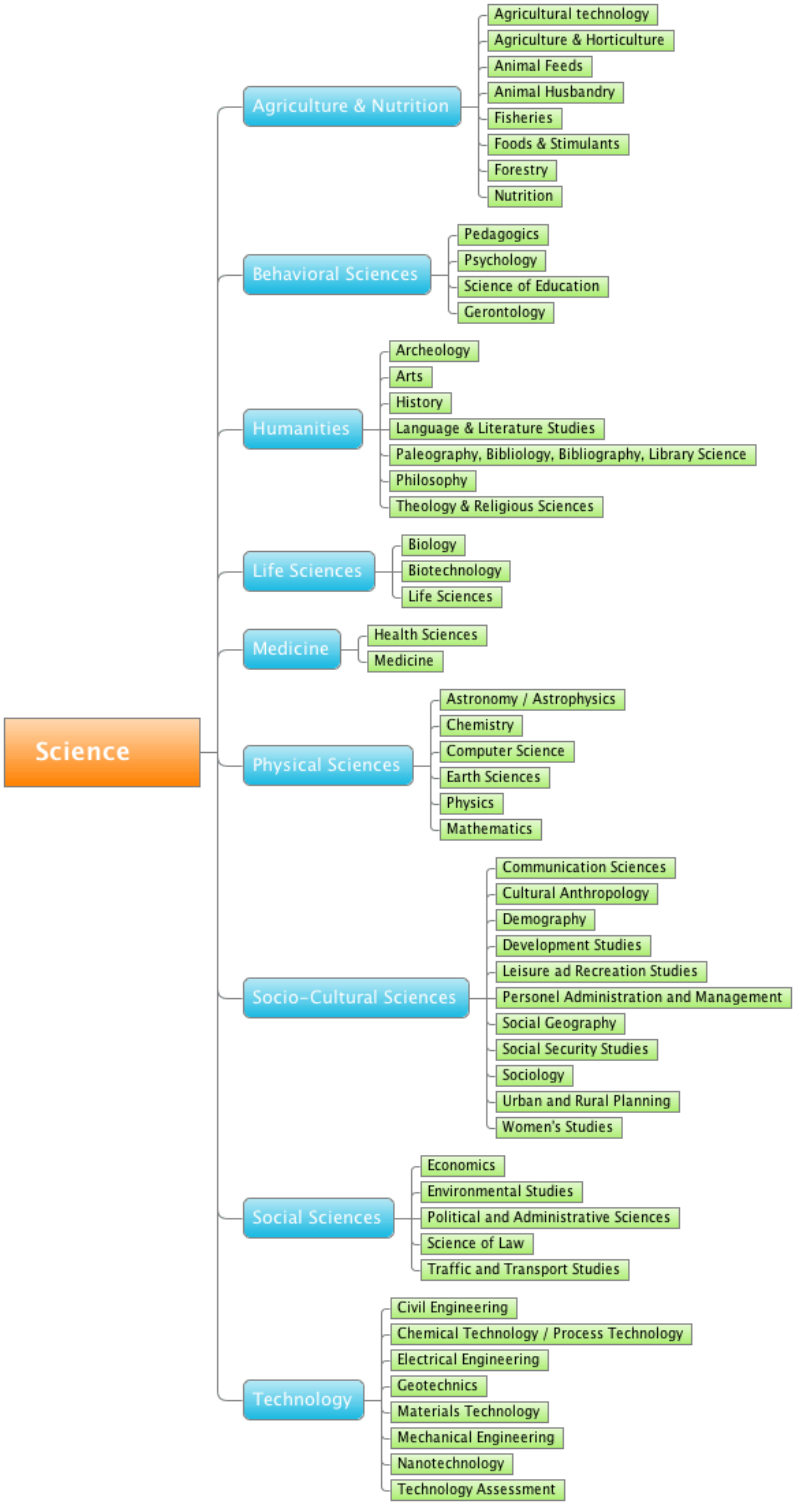


Figure 1: decomposition of research domains in Europe

2.2.2. Data managers

The category “data managers” covers profit and non-profit data archives, traditional memory institutions (libraries, archives & museums) as well as research and development in preservation technology itself.

Data archives collect data, make them accessible to researchers, and preserve them for future use. Hence, data archives are centres of expertise in data acquisition, preservation, dissemination and promotion. Data archives are setup as (cross-) institutional, national or disciplinary archive and are part of a network of national and international organisations with whom they maintain contact in order to keep knowledge current regarding new data collections, information management techniques and new computing technology. There are several national and international data archives in Europe[7].

Memory institutions follow a similar workflow as data archives but have a broader focus. They collect all kind of information of humanity's culture, make that accessible to researchers and are responsible for long-term preservation. Archives, museums and libraries are paradigmatic examples of traditional memory institutions. Most of them preserve digital collections as well[8]. These collections cover both born digital documents and digitized material.

Data archives and memory institutions have to manage data for the long term. In this context, they are responsible for:

- Identification, validation and selection
- performing preservation watch, planning and actions to ensure data remains understandable
- supporting ingest and metadata capture
- protecting the rights of data creators
- providing access
- supporting re-use via special tools
- participation in the development of standards
- training

[7] The PARSE.Insight interactive map in digital preservation shows some of the stakeholders within research in Europe, see http://www.parse-insight.eu/imap_intro.php

[8] PARSE.Insight survey report, section 7.3 *preservation – state of affairs*, pag. 42 - 43

When digital information is archived for a longer period of time (e.g. more than twenty years), several problems may arise, for instance:

- the medium on which data is stored becomes obsolete and deteriorate
- the format in which the data is stored becomes obsolete
- the old software that can display the data doesn't run any more on any of the current hardware and operating system platforms

Hence, apart from keeping digital information available, information must also be kept accessible. For this reason developers of preservation components and infrastructure also partly belong to the category *data managers*. They develop hard- and software which is needed for long-term preservation of digital data.

2.2.3. Funders/Policy makers

The category "funders" comprises organisations that are involved in the process of management, funding and establishing policies in long-term preservation of digital research data. Funders provide capital (funds) for a person, project, or a private or public institution. Funds can be allocated for short-term as well as for medium and long-term purposes.

The main actors in the funding of European research are the European Commission and national governments. Both actors fund projects doing research in long-term preservation. Most national research funding agencies in Europe are associated in the European Science Foundation (ESF)[9], which promotes scientific collaboration between countries. Within the European Union, the European Research Council (ERC) is the major research funder. The national governments are represented by agencies amongst which the available research funds are divided. These agencies either have a general character or they are discipline specific. In addition, a small part of scientific research is funded by non-profit foundations and private institutions. In contrast to the United States, this part of research funding is negligible since it accounts for less than one percent of total research funding[10].

Next to funding, policies can be understood as political, management, financial, and administrative mechanisms, arranged to reach explicit goals. They form deliberate plans of action to guide decisions as well as to achieve rational outcomes. Many funding bodies, data management organisations and research institutions, no matter the scientific discipline, have policies in place concerning accessibility of research data (e.g. the Safeguarding Good Scientific

[9] ESF members and observers, available at: http://www.esf.org/research-areas/marine-sciences/about-us/members-and-observers.html

[10] http://cordis.europa.eu/erawatch/index.cfm?fuseaction=ri.content&topicID=4&countryCode=EU

Practice[11] (1998) by the German Research Foundation (DFG), the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities[12] (2003) signed by several international scientific organisations and the Declaration on Access to Research Data from Public Funding[13] by the Organisation for Economic Co-operation and Development (OECD)). However, policies concerning long-term digital preservation are far less often available and seem to be in an early stage of development.

In this, funding organisations are important actors in long-term preservation of digital research data. They can set the scene for new guidelines and regulations regarding digital preservation. In this field, they have the following responsibilities:

- to consider wider policy perspectives
- to develop policies, either in co-operation with other stakeholders or by themselves
- to monitor and enforce policies
- to act as advocate for data curation and fund expert advisory services

2.2.4. Publishers

The category "publishers" covers publishers of academic books and journals and is regarded to be the most important stakeholder for disseminating the results of research.

With the advent of digital information systems and the internet, the scope of publishing has expanded to include digital resources, such as the digital versions of books and periodicals. Publishing is undergoing major changes, emerging from the transition of print to digital format. Currently, publishers are experimenting with interlinking publications with original research data, and integrating new kinds of digital research output into publications. Most publishers have preservation arrangements for all their publications in place, largely outsourced to third party services from National Libraries (for example KB) or specialised (non profit) organisations (like Portico). In most cases, supplementary material that is submitted by the author together with the manuscript receives the same preservation treatment as the scientific article. Publisher do understand the importance of preservation but do not regard preservation of data as an activity in their own area of expertise.

Responsibilities of publishers in the context of preservation are to:

[11] http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/self_regulation_98.pdf.

[12] http://oa.mpg.de/openaccess-berlin/berlindeclaration.html.

[13] http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html.

- organise the dissemination of research output
- manage all publications under their copyright
- ensure sustainable business model for this form of research dissemination
- acknowledge data creators and curators
- adhere to any licenses and restrictions of use
- protect the rights of data creators
- support ingest and metadata capture as well as re-use via special tools
- arrange for digital preservation of the publications

2.3. Key players

After having defined the stakeholders in long-term preservation of research data, PARSE.Insight has to identify the key players in Europe. We defined four criteria for an organisation to be a key player in this field:

- play an important role in preservation of digital research data;
- are representative of wider communities and their interests;
- invest in projects that deal with digital long-term preservation and;
- for this reason do pioneer work in preservation of digital research data.

The potential population of actors is very large. To focus its efforts most effectively, the project decided to focus on the following actors:

- researchers/research institution(s) who/which create significant amounts of digital research data, do RTD in this field and participate in significant national and international projects concerning long-term preservation; or who represent particular types of discipline with their own needs and priorities
- preservation institutions which preserve a significant amount of digital research data and do RTD in this field and participate in significant national and international projects concerning long-term preservation
- funder/policy maker, who provide significant support for scientific/ academic research and digital long-term preservation and respectively develop policies

- publishers who publish electronic journals as well as other digital documents with a significant importance for the scientific communities, and who do RTD in this field, and who participate in significant national and international projects concerning long-term preservation

These key players will be managed via a contacts database and visualized in an interactive map. Both will be described in the following sections.

3. Finding and managing contacts

3.1. Find contacts

It is important to note that there is a distinction between two classes of contacts being assembled by the PARSE.Insight project. First, there are the aforementioned *key players*, whose activities must be taken into account in creating the roadmap and performing the gap analysis. In many cases, the PARSE.Insight project team already has a good awareness of these activities.

Second, there are the *general targets for surveys*, individual persons assembled from a wide range of sources and to be approached en masse to obtain a large amount of data as evidence for the development of the roadmap. Individually, these people will belong to communities identified as key players (for example, the High Energy Physics community), but they are treated as individuals contributing to the formulation of the larger picture.

To identify the key players, PARSE.Insight can benefit from the knowledge of its project partners who belong to the categories *researchers*, *preservation actors* or *publishers*. Via the Alliance for Permanent Access, the project benefits from the contacts of the European Science Foundation (ESF), which covers the category *funding/policy*.

Furthermore, the project gathers information about the people and institutions who/which fill out the questionnaire. In this way, the project will review its coverage and, where possible, fill the gaps.

3.2. Manage contacts

During the insight phase of the project, a lot of new contacts and knowledge will be gained. Although the deeper understanding of the answers to the survey and interview questions offer an insight in the current state of digital preservation in Europe, the information about the organisations involved is of equal importance. It tells us which type of organisations deal with digital preservation in a certain way. Also, geographic spread of preservation awareness is interesting. Therefore, PARSE.Insight created an online *Contacts Database*[14]. The database is only available for internal use, because the project is not allowed to circulate contact information beyond the project context.

3.3. Contacts database

The next sections explain the requirements, design and implementation of the contacts database.

[14] http://www.parse-insight.eu/contacts (authentication required)

3.3.1. Functional requirements

The database can maintain information about organisations and persons related to those organisations.

For each organisation, it should be possible to record:

- the name of the organisation
- the acronym
- a URL of the website
- the city where it is located
- the country where it is situated
- its core R&D activity (if applicable)
- if it is a national organisation or not
- if it is a profit / not-for-profit organisation
- associated people working for that organisation
- associated category (research, data management, publisher, funder)
- associated research communities (if applicable)
- additional notes

An organisation can only belong to one country and category, but may be associated with multiple communities and multiple persons working for that organisation.

The following scientific communities should be distinguished:

- Agriculture and Nutrition
- Behavioural Sciences
- Humanities
- Life Sciences
- Medicine
- Physical Sciences

- Socio-cultural sciences
- Technology

Based on this factual information, several representations of the data should be created. Basic views supported by the Contacts Database are browse and topic-search functionalities. Moreover, graphical representations are possible as well. This will be covered in the next chapter (see 4. Interactive Map).

A web-based interface is needed to give access to this database. This interface should meet the following requirements:

- online accessible
- authentication (login functionality)
- distinction between organisations and persons
- browsing through both organisations and persons
- search function
- possibility to add, modify and view contact info of an organisation
- possibility to add, modify and view contact info of a person
- allow selection of information based on criteria communities, countries, and categories.

3.3.2. Data model

To support the structure of the contacts database, a data model has been developed. Figure 3 displays the data model of the database offering the following tables:

Organisation: for each organisation the database keeps track of all elements inside the box. Each organisation has a relation with a country, category and community.

Country: states a list of the 27 EU member states. The list also includes an option 'pan-European,' since not all organisations are fixed to one country.

Category: denotes the nature of the organisation in such a way that the information can be used to target the surveys later on. At least the database distinguishes the following levels:

- Funding/policy
- Research

- Preservation
- Publisher

Community: states to which kind of research area the organisation is related, if appropriate. Although many classifications can be made, we choose the classification of the Dutch KNAW (as discussed previously).

Person: states all necessary contact information for each person. It is related to a country as well and to one or more organisations.

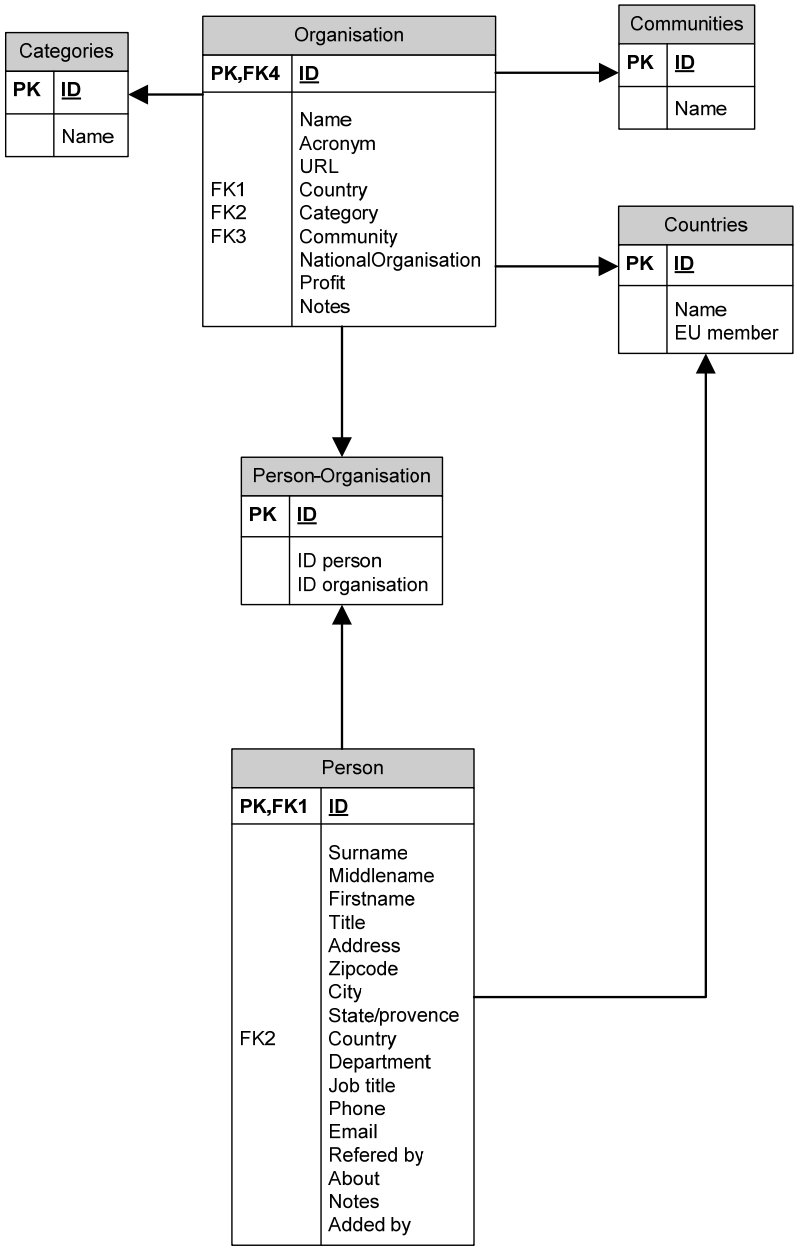


Figure 2: data model of contacts database

3.3.3. Usage & maintenance

The Contacts Database can be accessed via:

http://www.parse-insight.eu/contacts/index.php

Authentication is required for this. For gaining access, please contact a PARSE.Insight member.

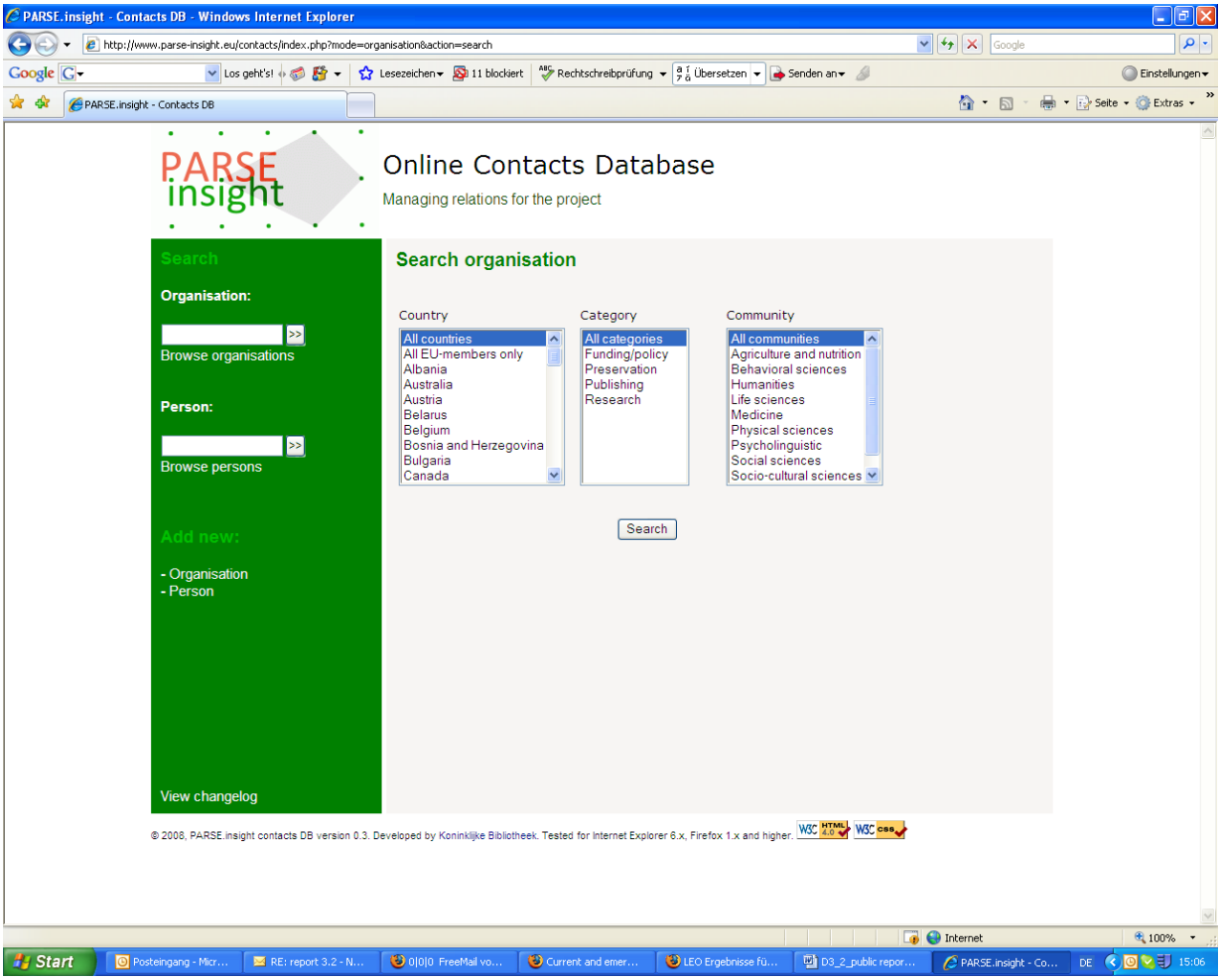


Figure 3: screenshot of the browsing functionality of the PARSE.Insight contacts database

Once logged in to the Contacts Database any user is allowed to update information on persons and organisations. It is also possible to add new organisations and persons and link them to each other. Deletion of entries is not possible. This can only be done by changes in the MySQL database directly.

4. Interactive Map

With the contacts database information about people and organisations can be kept and browsed. However, in advance to this functionality, PARSE.Insight want to offer a more visual view on the data. Therefore, an interactive map of key players depicting their R&D activities in digital preservation related technologies have been created.

4.1. Change with respect to the DoW

According to the DoW, this deliverable has to be delivered before having analysed the general survey and the case studies results. However, during the project meeting in The Hague (15/16 December 2008) the project members decided to postpone the delivery of the interactive map and the searchable database and not to deliver it together with the report (around April 2009), because it was considered more useful to create both the interactive map and the searchable database after having analysed the results of the general survey and the case studies. This provides the opportunity to use the data from the general survey and the case studies to picture both the interactive map and fill the database.

4.2. Requirements

Based on the output of the general survey and the case studies, PARSE.Insight identifies Europe's key players within the categories *Funding/Policy*, *Preservation*, *Publishing* and *Research*. The purpose of the interactive map is to visualize this information.

The main function of the interactive map is the possibility to search within the entries and give a visual representation of the results. The application allows a search for entries using different criteria, e.g. a search:

- within all categories and communities;
- within all categories within a special community;
- within all categories within a selection of countries.

The results should be made easy to understand, so that one has a quick overview of the outcome for the combination of criteria that was chosen.

4.3. Design & Implementation

This section explains the design and implementation of the Interactive Map. It first tries to explain how the different parts work together and then describes each part in greater detail. The

first part should be easy to read an understand even by people that are not trained in any of the used technology. The second part is mainly for people that already worked with the corresponding programming languages and techniques.

4.3.1. Graphical User Interface (GUI)

The purpose of the interactive map is to display the location and details of the actors and stakeholders. To do so, we chose to build a geographical overview of stakeholders that can be selected via menu options. The user can filter which kind of organisations are shown. He/She can filter the attributes “country”, “category” and “community”. Only organisations that match the given criteria are shown on the map. Figure 4 shows the user interface of the Interactive Map.

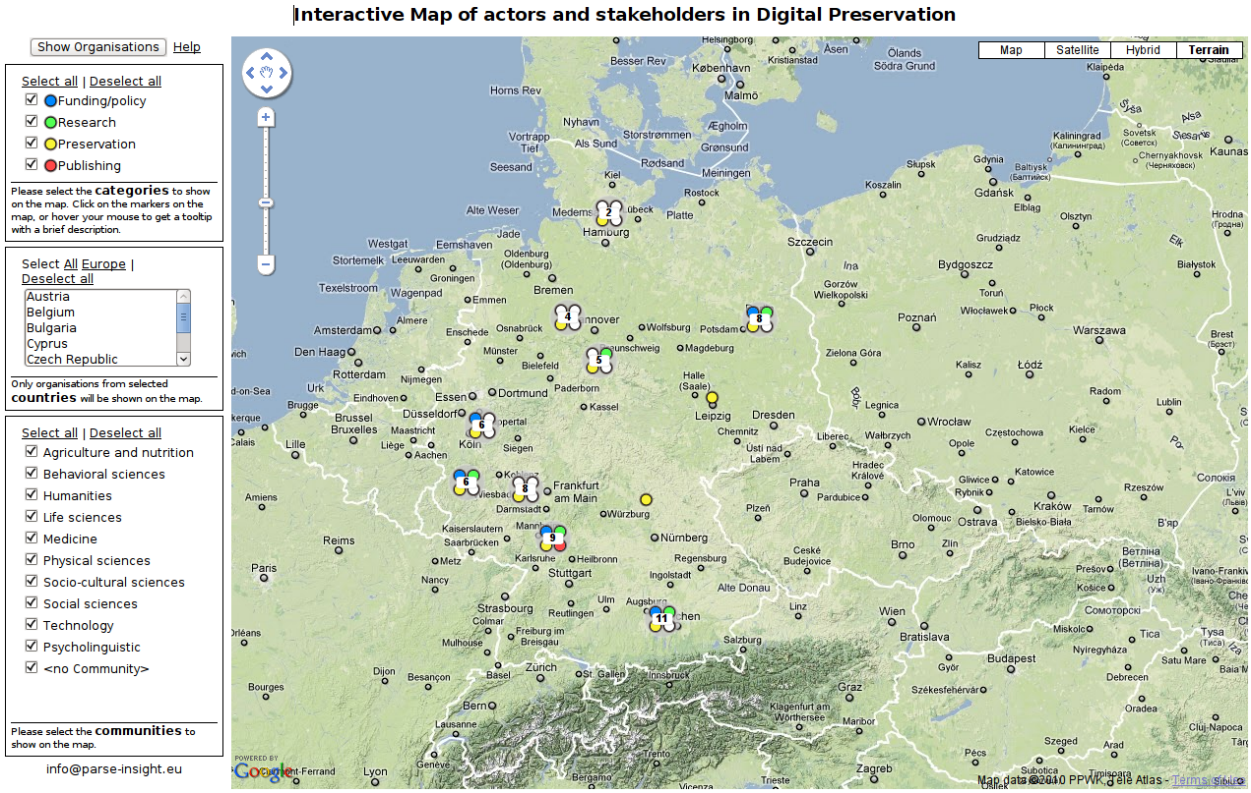


Figure 4: Graphical User Interface of Interactive Map

As you can see in the above screenshot the user interface is divided in two areas. The left column gives you the possibility to filter the organisations. The right area of the user interface is the actual map and displays the organisations.

PARSE.Insight uses the Google Maps API[15] to create the interactive map. Google Maps API provides the opportunity to embed Google Maps in web pages with JavaScript, so that the

[15] http://code.google.com/intl/en/apis/maps/

interactive map will be available on the website of PARSE.Insight. The API provides a number of utilities for manipulating maps and adding content to the map through a variety of services.

Google Maps API provides the opportunity of specifying fields (attributes) to qualify entries of the map, e.g. categories like *Funding/Policy* or communities like *Humanities*. Results can be colour-coded for the respective attributes. Thus, it is possible to visualize the results of the project's work by using the terms applied in the project's reports and the searchable database.

The server that runs the PARSE.Insight website (both the public website (see D7.1, 2.2) and the internal contacts database (see 3.2)) supports PHP and a MySQL database. Both the Contacts Database and the Interactive Map use PHP for dynamically creating pages; the contacts database uses the MySQL database for storing all contacts and organisations. The HTML pages for the interactive map are dynamically generated from the data in the contacts database, so that it is possible to cope with large amounts of data.

4.3.2. Overall structure

The webpage showing the map is a static HTML page including a number of Javascript files. These scripts load the content from the server and display it on the map. In this text this will be referred to as *client*.

The other side of the application is the *server* part. It is called over the URL http://www.parse-imap.eu/imap/server/server.php. The client asks this part of the application for information (e.g. the list of countries). The server then connects to the database, collects this information and sends it back to the client.

The following information is transferred to the client in this form:

- organisations with Position, Name, Country, Community(s), Category(s)
- List of Communities
- List of Categories
- List of Countries
- List of Country groups (e.g. Europe)

This information is used by the client to filter the organisations when the user requests this.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119[16].

[16] RFC 2119, available at: http://www.ietf.org/rfc/rfc2119.txt

4.3.3. Client

As mentioned the client is a static HTML page. It is written in XHTML 1.1. This page refers to a list of Javascript files that provide the main User interface.

Initialisation

The client starts onLoad of the HTML page by calling the function `initialize()` which is defined in the `functions.js` file. This function first asks the Server for important information that is used to create the User interface. This includes the List of communities, countries, categories and country groups. While the client waits for the result from the server it starts the Google map.

When the results from the server arrive, the function `init_stage2()` is called. This function uses the information to build the filter boxes on the left side. Every box is created by one class from the `filter.js` file which represents one Filter. The fields within the box are created by the `getInput()` function of the corresponding filter. The detailed Interface for a filter can be found in section filter.

After the UI is build the client loads the organisations from the Server.

When these arrive the organisations are saved to the array organisations and the function `init_stage3()` is called. This function enables the search button and thereby concludes the initialisation process.

Searching

Whenever the user clicks the “Show organisations” button the following steps are taken.

First the `prepareFilter()` function of each filter is called. This function checks the current settings of the filter and return true if everything is OK. If the user made a invalid selection for a the corresponding filter it returns false and displays a warning to the user.

If and only if all `prepareFilter()` functions returned true the client walks though all organisations and passes them to the filter(organisation:array): bool function of each filter. If only one filter function return false for this organisation it is not displayed. If all filter return true and thereby state that the organisation should not be filtered out the organisation is put into the list of organisations to show.

After this all organisations that passed the filter process are displayed on the map. All other organisations are removed from the map.

Filters

The following UML diagram shows the interface for a filter.

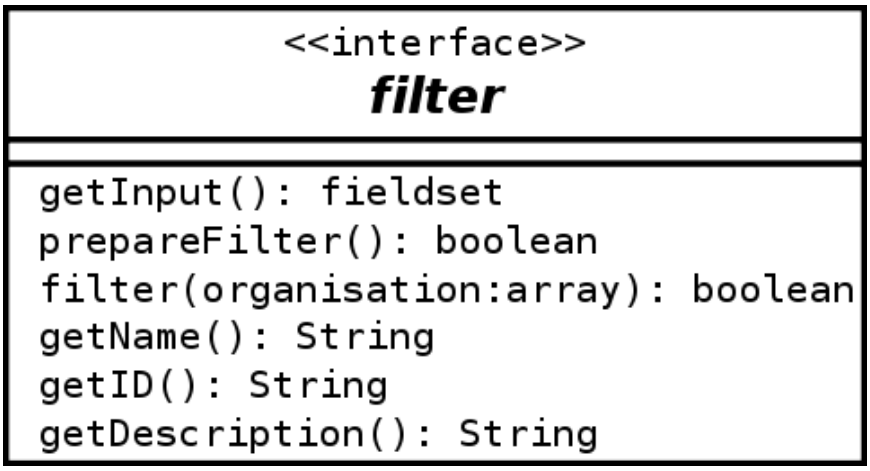


Figure 5: UML filter

Each filter must extend the `filter()` Class.

The `getInput()` function MUST return a fieldset HTML element with all fields that should be displayed to the user. The filter itself is responsible for taking action if something should happen when these change. Also the filter itself is responsible for saving some kind of reference (e.g. a variable) to each field so it can read the user input afterwards if this is necessary.

The `prepareFilter()` function is called before the actual filter process. The filter MUST, at this point in time, check it's input fields for invalid inputs. If any input is invalid this function MUST return false. It also SHOULD display a warning to the user. Otherwise the function MUST return true.

The execution proceeds if each `prepareFilter()` function returns true;

After that each filters `filter(organisation: array): boolean` function is called for each organisation. The argument is an array that describes one organisation. The format of this array is the same as the return value of the `get_orgs` server function. Refer to section **Fout! Verwijzingsbron niet gevonden.** for details about this format. The function MUST return false if the given organisation is not a valid organisation according to the filter options and should be filtered out. Otherwise the function MUST return true.

If any of the filters returns false for a given organisation that organisation is filtered out. It only stays in the list if all filters return true;

`getName()` MUST return a User-readable Name for the filter.

`getID()` MUST return a unique ID for this filter. This ID MUST NOT contain any symbols that are invalid characters for CSS identifiers (e.g. whitespaces or umlauts).

`getDescription()` MUST return a textural description about what this filter does and how it should be used. It SHOULD be not that long because it is displayed below the filter and would

otherwise break the layout of the page.

The filters are registered in the function `initfilter()` in `functions.js`.

4.3.4. Server

The server is written in PHP and normally runs on the same server from which the client page is served. But this is not required. It could also run on any other server as long as the client code can access its URL.

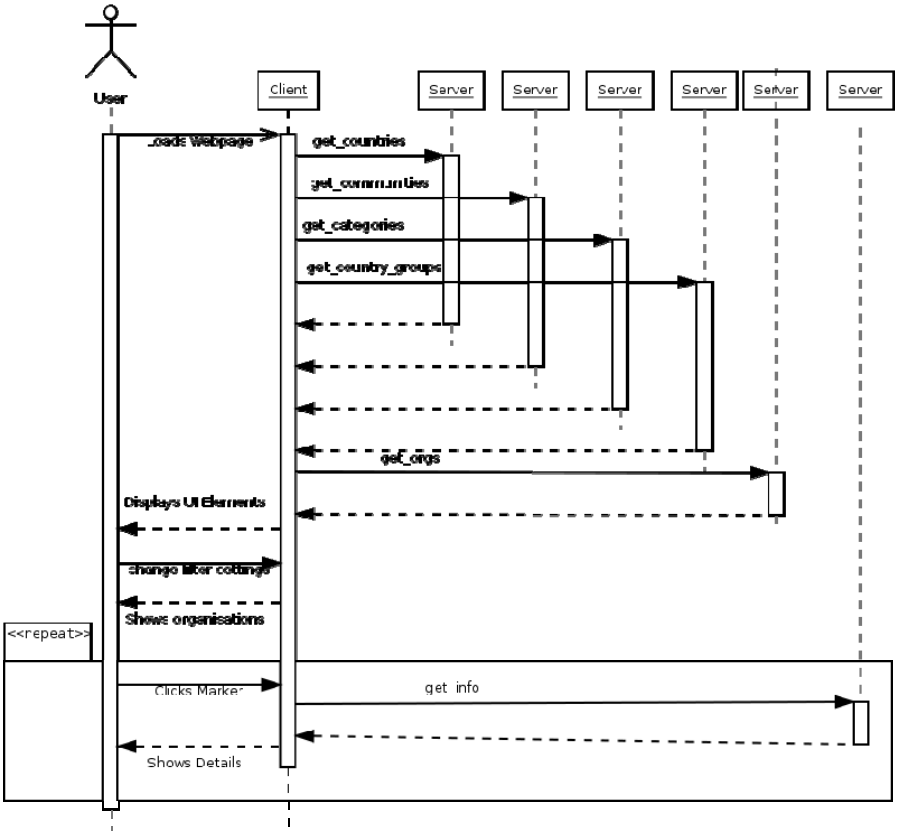


Figure 6: Client-server interaction

The above illustration shows the normal interaction between the client and the server. The server is called by the client via HTTP GET or POST and returns a result in the JSON[17] format.

The function that the server performs is determined by the func parameter.

Currently the server knows of the following functions:

`get_countries`

This function returns an Array of countries. Each entry of this array consists of a associative

[17] http://www.json.org/

array with the entries “ID” and “name” that describe the numeric ID and the name of the country.

`get_country_groups`

This function returns an Array of country groups. A country group is something like “Europe” that is shown as selection field over the country list.

`get_communities`

This function returns an Array of communities. Each entry of this array consists of a associative array with the entries “ID” and “name” that describe the numeric ID and the name of the community.

`get_categories`

This function returns an Array of categories. Each entry of this array consists of a associative array with the entries “ID”, “name” and “color” that describe the numeric ID, the name of the category and the colour in which it should be displayed on the map. The later is currently ignored because the images that are shown on the map are static.

`get_info`

This function needs the additional parameter “id” in addition to the “func” parameter. The function returns detailed information about the organisation with the given ID in form of a associative Array with the following entries:

- ID
- city
- name
- acronym
- url
- country
- category
- communities
- national
- profit
- notes

`get_orgs`

This function returns an Array of organisations. Each entry of this array consists of a associative array with the following entries:

- Id
- name
- country
- category
- communities
- lat
- lng

4.3.5. Database

In addition to the already existing databases the Interactive Map introduces two tables and one view to the database. The first table is `organisation_position`. This table holds longitude (lon) and latitude (lat) of the position for each organisation.

The table `position_cache` saves known positions for a given city in a country. This reduces the number of requests against the google geocoding server to determine the position of a organisation.

The last addition to the database is a view called `organisation_with_positions` which just combines all information in the organisation table and the position for this organisation. This view is mainly there to ease the access to all information about a organisation.

4.4. Usage & maintenance

4.4.1. Accessing the map

The Map can be found under the URL http://www.parse-imap.eu/imap.

For detailed usage information please refer to the help page[18]. The same help information is shown in the appendix.

4.4.2. Importing new data

Every time new data is inserted into the database the database must be informed about this

[18] Help page is available at: http://178.63.1.204/imap/help.php

change and the position of each new Organisation must be generated.

The maintenance tasks are available over the admin interface which can be found under http://parse-insight.eu/imap/admin/

To import new data one visits the page *Generate Positions for all organisations*[19]. This page generates the position data and shows some debug information along the way. This process uses the city and country of each organisation in combination with the google geocoding service to generate a position of each organisation. Therefore each organisation MUST have at least these two information to be shown on the map and an error message will be displayed to the user during the import process if this is not the case. This process will include all positions to the database that it could geocode. It might be needed to repeat this process if anything went wrong. The script will not try to recode the already geocoded positions.

After this you should visit the webpage *Print statistical information about the database*[20]. This page shows what went wrong with the import process and gives tips how these problems can be solved. If the import process worked without any problems this page shows a green status message otherwise this message is in red colour.

[19] Generation page available at: http://parse-insight.eu/imap/admin/generate_pos.php?debug_level=10

[20] Print page available at: http://parse-insight.eu/imap/admin/data_statistics.php

5. Glossary

DFG	German Research Foundation
ESF	European Science Foundation
KNAW	Koninklijke Nederlandse Academie van Wetenschappen
MPG	Max Planck Society
OECD	Organisation for Economic Co-operation and Development
PARSE	Permanent Access to the Records of Science in Europe
RTD	Research Technology & Development

6. Appendix A: manual for Interactive Map

Interactive Map of Actors and Stakeholders in Digital Preservation

The purpose of the interactive map is to visualise actors and stakeholders in digital preservation. The application is based on a searchable database of key players and R&D activities in digital preservation related technologies, created by the PARSE.Insight team.

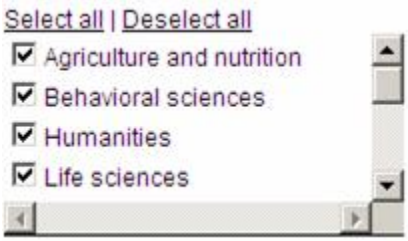
How does it work?

When you hit “Show organisations”, all organisations of the database are displayed in the context of a geographic map. All organisations are correctly assigned to cities, but not to particular street addresses. Caution: When you zoom into a specific city, the organisations are distributed randomly over the available streets.

Each organisation belongs to a country, category and community. You can specify which organisations are displayed by selecting or deselecting items in each of those categories. You can, for example, look for all Social Science (community) publishers (category) in EU member states (country), or for all research organisations (category) in the UK (country) which fund activities in the Life Sciences (community) etc.

Select one or more categories.

Select a county or the option “all EU-members”.



Select one or more communities.

Example: Representation of the selection “Preservation organisations in Germany (all communities):

Each organisation is marked in the map by a little coloured dot:

The colours code the four categories:

When there is more than one organisation in a city, an aggregate icon is displayed. It indicates the numbers of organisations behind it and the categories represented (marked with the colour code). There are always four bubbles within an aggregator icon. If a category is not represented by the aggregator icon, the respective bubble remains white. In this case, the colour code represents three organisations from the categories “Funding/policy” (blue) and “Preservation” (yellow):

A mouseover tooltip provides a brief description of the organisation or the organisations represented by an aggregator icon:

A mouse click on an individual dot reveals information about the organisation displayed:

Selection Options "Country"

You can pick from a list of 58 states. Since not all organisations are fixed to one country, you have also the option to pick "Europe" or "worldwide".

Selection Options "Category"

We distinguish the following levels: 1. Funding/policy 2. Research 3. Preservation (e.g. libraries, archives, data centers) 4. Publisher

Selection Options "Community"

The community states in which kind of research area the organisation is related. Some organizations, like national libraries, are not assigned to any specific community. In this case, the option "no community" is applicaple:

- Agriculture and nutrition
- Behavioral sciences
- Humanities
- Life sciences
- Medicine
- Physical sciences
- Socio-cultural sciences
- Social sciences
- Technology
- Psycholinguistic
- <no Community>