



## PARSE.Insight Newsletter Nr. 1 (July 2009)

### 1. Introduction

This Newsletter summarizes the first year's goals and achievements of PARSE.Insight (<http://www.parse-insight.eu/>), a two-year project co-funded by the European Union under the Seventh Framework Programme. It is concerned with the preservation of digital information in science, from primary data through analysis to the final publications resulting from the research. Its aim is to develop a roadmap and recommendations for developing the e-infrastructure in order to maintain the long-term accessibility and usability of scientific digital information in Europe. The project conducts surveys and in-depth case studies of different scientific disciplines and stakeholders and it will base its results on these findings, as well as knowledge of ongoing developments.

PARSE.Insight is closely linked to the Alliance for Permanent Access to the Records of Science (<http://www.alliancepermanentaccess.org/index.php?id=1>). The output from the project is intended to guide the European Commission's strategy about research infrastructure.

### 2. Stage of Work

On June 2, 2009 PARSE.Insight had its first annual review by reviewers of the European Commission in Brussels. Looking back on the past year, the project showed good progress in defining a draft roadmap for building the e-science infrastructure in European. Furthermore, an extensive survey is in progress for identifying Europe's current position on digital preservation regarding research. Also, specifications have been defined for performing a gap analysis between the roadmap and the current state of affair and development of standards on which an audit and certification process for digital repositories can be based. The reviewers were pleased with the first year's goals and achievements that are available under [http://www.parse-insight.eu/review\\_year1.php](http://www.parse-insight.eu/review_year1.php).

### 3. Gaining Insight

In the first year of project the main emphasis of the project has been surveying communities with an interest in digital preservation to build up insight, and developing a draft roadmap for the e-infrastructure. General surveys and case studies have been performed, interviews are going to deliver more detailed results and representative examples of best practices in coming months. The survey and interview results will be used as a basis for plugging the gaps in the European e-infrastructure with regard to the long-term usability of raw scientific data. The general surveys were aimed at distinct groups of stakeholders: researchers, funders, publishers and data archivists. The surveys employed online questionnaires constructed to obtain information about the knowledge, attitudes, practices and desires of the stakeholders

with respect to digital preservation. The communities were contacted using a wide range of mailing lists and other publicity.

For more information about the stakeholders see [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D3-2\\_InventoryOfCommunities\\_final.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D3-2_InventoryOfCommunities_final.pdf). Information about the survey platform is available under [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D3-1\\_SurveyAndForumPlatforms\\_final.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D3-1_SurveyAndForumPlatforms_final.pdf).

### *General Survey*

One of the ideas underlying the surveys was threats to preservation eventualities that could lead to the loss of digital resources or the inability to understand them. For the survey seven threats were formulated, ranging from the loss of data to the inability to "read" preserved data because of the unavailability of appropriate hard- and software. Between 60% and 80% of the respondents indicate that all threats are recognized as either "Important" or "Very Important", with about half supporting the need for an international preservation infrastructure. Another clear message is that researchers would like to (re-)use data from both their own and other disciplines, and it is suggested that this is likely to produce more and better science. However more than 50% report that they have wished to access digital research data gathered by other researchers which turned out to be unavailable.

### *Case Studies*

Three case studies are being conducted: in high energy physics, earth observation and social sciences/humanities (psycholinguistics and book studies). They have a different motivation from the general survey, aiming to investigate more deeply and more narrowly the characteristics of certain communities.

High-Energy Physics (HEP) aims to understand the way our universe works. HEP experiments are run by hundreds to thousands of scientists and produce a literal data deluge. There is a growing concern about the fate of these irreproducible data beyond the lifetime of experimental facilities since these are virtually not understandable without expert knowledge and also because there is an alarming lack of resources for their preservation.

CERN, as part of the PARSE.Insight consortium, ran a large-scale survey of the attitudes and concerns of HEP scientists on this burning issue. The survey received 1,200 detailed responses (about 5% of all practitioners) reflecting the worldwide character of the discipline and cover all age groups, from PhD students to professors. An overwhelming majority of practitioners perceives the issue of long-term preservation as "very important" or "crucial", the most quoted reasons being: independent verifications of results and the possible re-use of old data in the light of future ideas. Many respondents feel that important scientific data have been lost for want of a concerted preservation effort. On the positive side, even though the high complexity of HEP data makes a re-analysis of "raw" data impossible without insider knowledge, about 45% of the respondents would like to see raw data preserved.

Respondents to the survey felt that HEP data should be stored at an independent platform, and this should happen early in the life of an experiment. Such a platform should be easy to use, free of charge and Open Access, with data completely documented. Against this positive

attitude, it is somewhat sobering that the effort to be deployed into data preservation is perceived as enormous and practitioners mostly think their organizations will not be able to make that investment. Other concerns include: the possible inflation of incorrect results due to the potential of unregulated re-use of badly documented data, the un-fair sharing of credits and responsibilities, data security and integrity, and long-term accessibility.

Several thousands of free-text answers to the survey are currently being analyzed which contain much more information on the intricacy of data preservation in HEP. First results of the HEP case study are available at <http://arxiv.org/abs/0906.0485>.

The large amount of new Earth Observation (EO) missions coming in the next years and the increased demands from the user community, bring a challenge for EO satellite operators, Space Agencies and EO data providers to manage the data and to offer the access to the different products as coherently and easily as possible. With the increasing interest on global change monitoring, the use and exploitation of EO data has been increasing systematically and many users request time-series of data spanning 20 years and more calling for a need to preserve the EO data without time constraints and keep them accessible. On the other hand:

- Current EO data preservation approaches are still mostly limited to the satellite lifetime and few years after.
- More and more EO missions data can be called 'historic' and more and more operators are faced with the decision if and how to preserve these data.
- The data volumes are increasing dramatically.
- EO data preservation policies, if existing at all, are different for each EO mission, each operator or Agency.
- In one sentence, the "preservation and easy access of the whole European EO space dataset cannot be guaranteed".

Both technically and economically, long term preservation of environmental data is a subject that calls for harmonization, cooperation and sharing among all the data owners for the benefit of the user community. To respond to the urgent need for a coordinated and coherent approach for the long term preservation of the existing European EO space data, ESA has setup a dedicated working group, the European Long Term Data Preservation initiative (LTDP) that was approved by the ESA Ministerial in 2008.

In this context and as partner of the European Union's FP7 PARSE.Insight project, ESA issued a survey to better understand and evaluate the needs and requirements of the Earth Science user community on long term preservation of historical environmental data. Responsiveness to this survey and the interest raised by data users have been much higher than expected, and the result will be considered by the European LTDP initiative and used as basis for a set of initiatives to safeguard this valuable digital material over time and to ensure that it remains accessible, usable and understandable in the future.

The social sciences and humanities case study is divided into two sub case studies: psycholinguistics and book studies. The Psycholinguistic survey has been composed of different parts for researchers and data managers. Because of the small size of the community, the Book Studies survey did not make this distinction. Early analysis shows, that on the one hand, there is an interest in concerns of long-term preservation of scientific data and the awareness that efforts are needed to shape the e-infrastructure for these data. On

the other hand, respondents show a great uncertainty concerning the existence of present e-infrastructure components and how to use them.

#### **4. Roadmap**

The draft roadmap is another major achievement of the first year of the project. Using the survey results as a foundation, the roadmap characterizes what is meant by an e-infrastructure for science data and proposes components to build it. The draft roadmap provides an overview and initial details of a number of specific components, both technical and non-technical, which would be needed to supplement existing and already planned infrastructures for scientific data. The infrastructure components are aimed at bridging the gaps between islands of functionality, developed for particular purposes, often by other European projects. Thus the infrastructure components are intended to play a general, unifying role in scientific data. While developed in the context of a Europe-wide infrastructure, there would be great advantages for these types of infrastructure components to be available much more widely. The draft roadmap is available at [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D2-1\\_DraftRoadmap\\_v1-1\\_final.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D2-1_DraftRoadmap_v1-1_final.pdf).

#### **5. Gap Analysis**

The gap analysis compares the roadmap with the inventory of existing and planned capabilities, to focus resources where they are most needed in order to develop the full preservation e-infrastructure required. For this a gap analysis framework was developed, differentiating four gap types relating to the different stages of diffusion of the concept of “long-term preservation of scientific data”. Then different technologies were assessed to implement an IT support of the gap analysis including data management, analysis and experimentation. Two communities - “publishers” and “libraries” - are chosen for the first application of the tool. The so obtained validation results will support the dissemination into other sectors. For further information see [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D4-1\\_SpecOfGapAnalysis\\_final.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D4-1_SpecOfGapAnalysis_final.pdf).

#### **6. Sustainability**

In the sustainability and evaluation work, the focus has been on the progress towards an international standard for audit and certification of digital repositories. A workshop was held in the US at which excellent progress was made on the draft standard, which is now close to submission to the ISO process. The workshop report is available under [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D6-1\\_SustainabilityWorkshopReport\\_final.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D6-1_SustainabilityWorkshopReport_final.pdf).

#### **7. EU Consultation Meeting**

In November 2008 a concertation meeting on sustainable e-infrastructures organized by the European Commission took place in Lyon, France. Dr. David Giaretta of STFC, the PARSE.Insight coordinator, acted as rapporteur for the data track. The report is available at [http://www.beliefproject.org/docs/6th-eConcertation%20Meeting-Report-Final.pdf/at\\_download/file](http://www.beliefproject.org/docs/6th-eConcertation%20Meeting-Report-Final.pdf/at_download/file). This meeting, at which many important stakeholders were present, was an opportunity to expose and discuss aspects of the PARSE.Insight roadmap. In

consultation with the European Commission, the project team saw an opportunity to reorient the project within the wider context of science data infrastructure, in which preservation is considered as part of a bigger picture of preservation, reuse and (open) access, rather than in isolation.

## **8. How to Proceed?**

During the next year, PARSE.Insight will revise its roadmap, which will influence the agenda of development in the science data infrastructure for the coming years. The roadmap will be complemented by an understanding of the gaps with respect to the current situation. Several workshops are being planned to engage important stakeholders in the remaining work of the project.

By the end of the project an important base of data will have been assembled concerning the attitudes and practices of a wide range of scientific communities concerning digital preservation and science data infrastructure. This will provide an excellent body of evidence for policy makers, strategists and funders.

Have we piqued your interest? You will find all results on the project's website: <http://www.parse-insight.eu>.

Kind regards on behalf of the PARSE.Insight-team,

Ada Beate Sturm

Niedersächsische Staats- und Universitätsbibliothek Göttingen- Goettingen State and University Library,  
Germany (SUB) - R & D, Papendiek 14, D-37073 Göttingen  
Mail: [sturm@sub.uni-goettingen.de](mailto:sturm@sub.uni-goettingen.de)  
URL: <http://www.sub.uni-goettingen.de/>

Max Planck Digital Library (MPDL) - R & D  
Amalienstr. 33, D-80799 München  
[www.mpd.l.mpg.de](http://www.mpd.l.mpg.de)