

PaN-data Europe

Deliverable D2.2

Common policy framework on analysis software

Grant Agreement Number	261537
Project Title	PaN-data Europe Strategic Working Group
Title of Deliverable	Common policy framework on analysis software
Deliverable Number	D2.2
Lead Beneficiary	ESRF
Deliverable Dissemination Level	Public
Deliverable Nature	Report
Contractual Delivery Date	31 Jan 2011 (Month 8)
Actual Delivery Date	4 Apr 2011

*The PaN-data Europe project is partly funded by the European Commission
under the 7th Framework Programme, Information Society Technologies, Research Infrastructures.*

PANDATA WP2 D2.2 – Data analysis software policy framework

Abstract

A common approach towards data analysis software would be extremely beneficial for the research infrastructures of the PaN-data common consortium. This document presents a policy which is recommended to facility management for common adoption. It covers development of software, preservation of software and access to software.

Keyword list

Software Policy, Open Source, GNU-GPL, sharing software, software maintenance

Document approval

Person	Role	Partner
Simon Lambert	Project coordination	STFC
	Approval of content	All partners

Revision history

Issue	Author(s)	Date	Description
0.1	Rudolf Dimper	01/03/2011	First draft
0.2	Juan Bicarregui	08/03/2011	Some additions
0.3	Rudolf Dimper	13/03/2011	Further editing
1.0	Rudolf Dimper	18/03/2011	Integrated comments from Partners
1.1	Rudolf Dimper	21/03/2011	Integration of further comments and corrections
1.2		04/04/2011	Ready for submission to EC

PANDATA WP2 D2.2 – Data analysis software policy framework

Table of contents

Page

Contents

1. DEFINITIONS	5
2. GENERAL PRINCIPLES.....	5
3. POLICY ON DEVELOPMENT OF ANALYSIS SOFTWARE IN PAN-DATA LABORATORIES	6
4. POLICY ON PRESERVATION OF DATA ANALYSIS SOFTWARE	7
5. POLICY ON DISCOVERY AND ACCESS TO DATA ANALYSIS SOFTWARE	7

PANDATA WP2 D2.2 – Data analysis software policy framework

Introduction

Data analysis software is at the heart of the scientific activity in our Research Infrastructures (RIs). A substantial amount of human resources are devoted in every RI to purchase or develop, install, and maintain data analysis software packages. The PaN-data consortium has realised early on that a common approach towards data analysis software would be extremely beneficial and could lead to substantial savings in human and financial resources, lead to better software, and help our scientists by finding similar software environments in the RIs.

The policy focuses on data analysis software but the same issues and reasoning are applicable to other domains like control system software, database applications, and office automation tools. It is left to the partner laboratories to discuss whether this policy shall be extended to those domains in the future.

Objectives

The objectives of the data analysis software policy can be summarised as follows:

- Share information on envisaged data analysis software developments, software development methods, and specifications, to create possible synergies between the partner labs,
- Maintain a table of open source and commercial data analysis software packages used in the partner laboratories to foster information exchange,
- Identify laboratories who will champion the maintenance of one or several software packages to save human resources for the benefit of the community,
- Make data analysis software packages and the associated expertise available to the partners,
- Increase reproducibility of data analysis,
- Exchange information on the purchasing conditions of popular commercial data analysis software packages.

Recommendation

The PaN-Data consortium strongly recommends to their respective facility managements to adopt and publish the data analysis software policy framework. Having an identical approach to the management of data analysis software will ease the life of scientists using more than one facility and add to the overall transparency of the scientific process.

PANDATA WP2 D2.2 – Data analysis software policy framework

1. Definitions

For the purpose of this policy document:

- 1.1. The term **software license** is a legal instrument governing the usage or redistribution of software. A typical software license grants an end-user permission to use one or more copies of software in ways where such a use would otherwise constitute copyright infringement of the software owner's exclusive rights.
- 1.2. The term **open-source software (OSS)** refers to software that is available in source code form and for which the source code and certain other rights normally reserved for copyright holders are provided under a software license that permits users to study, change, improve and distribute the software.
- 1.3. The term **free software license** is a software license which grants recipients rights to modify and redistribute the software, which would otherwise be prohibited by copyright law.
- 1.4. The term **GNU General Public License (or simply GPL)** refers to the most widely used free software license. For more details of what the GPL defines, please refer to: http://en.wikipedia.org/wiki/GNU_General_Public_License

2. General Principles

- 2.1. This data analysis software policy pertains to the design, ownership, and usage of data analysis software.
- 2.2. It is not expected that the data analysis software policy will induce additional cost, but on the contrary it is expected that synergies will ultimately lead to significant savings of human resources.
- 2.3. Acceptance of this policy is a condition to participate in the software sharing activity. The PaN-data consortium will endeavour to extend this activity beyond the partner RIs.
- 2.4. The WP6 of the PaN-data Europe FP7 project makes provision for the establishment of a list of software packages developed, used, and maintained in the partner laboratories. This list is the basic document for the collaborative efforts on data analysis software within the consortium.
- 2.5. Software packages shall be preserved along with their computing environment (operating system, libraries) as long as possible, and for at least 3 years (the typical duration of a PhD thesis), allowing to establish a link to the metadata information of the processed data and allowing to reproduce data analysis runs during this time span.

PANDATA WP2 D2.2 – Data analysis software policy framework

- 2.6. The PaN-data partners will strive to exchange information on purchasing conditions of commercial data analysis software packages.
- 2.7. The implementation of this software policy shall be driven and monitored by the respective IT groups/departments of the RIs.

3. Policy on development of analysis software in PaN-data laboratories

- 3.1. Partners will make every effort to communicate between themselves their intention to develop new software before the actual development work starts. Whenever possible specifications shall be shared to allow for feedback and to avoid parallel software developments.
- 3.2. The PaN-data RIs endeavour to license newly developed in-house data analysis software under the GNU-GPL.
- 3.3. The PaN-data RIs will strive to share software development guidelines, expertise on modern programming languages and expertise on modern software development tools. It is expected that, as a consensus on programming languages and tools emerges, this will lead to better sharing of competences and programming methods.
- 3.4. Data analysis code should be broken down into independent modules with well defined APIs and workflows.
- 3.5. No major data analysis software package should be developed by a single person. Whenever possible developments shall be shared between the PaN-data partners, such that at least two laboratories participate in the overall design of a package.
- 3.6. Every newly developed data analysis software package should be compatible with the HDF5/Nexus data format.
- 3.7. Every newly developed data analysis software package should be extensively documented (code documentation as well as end-user documentation), in particular for the installation on various platforms and for its use in interactive and/or batch/script mode.
- 3.8. Every newly developed data analysis software package should be complemented by an automatic test suite based on a well defined test data set.
- 3.9. Partners undertake to avoid 'branching' of software developments unless there are good technical reasons for doing so, in order to avoid diverging software packages. Before a decision is taken to develop a new branch of an existing software package, a technical discussion between the stakeholders shall take place.

PANDATA WP2 D2.2 – Data analysis software policy framework

4. Policy on Preservation of data analysis software

- 4.1. Software used in the analysis of data shall be preserved in an executable state whilst it is listed as available in the catalogue of analysis software and for a period of at least one year after it is removed from that catalogue.
- 4.2. For software developed at the participating laboratories source code shall be preserved ad infinitum or until such time as an agreed and deliberate decision is made to dispose of it.
- 4.3. Each of the PaN-data laboratories shall engage in the long-term maintenance and support on as many platforms as possible of at least one data analysis software package for the benefit of the other partners.

5. Policy on discovery and access to data analysis software

- 5.1. The PaN-data RIs endeavour to make their in-house developed data analysis source code available and communicate its whereabouts to the others. Version control software will be used to manage the source code and associated materials.
- 5.2. The PaN-data RIs will maintain a catalogue of data analysis software packages and exchange that information with the partners.
- 5.3. The PaN-data RIs will make data analysis software packages and the associated expertise available to the partners on a best effort basis.