

# PPDG Internal Review of GDMP

## Introduction

This document is the report of the PPDG internal review of the GDMP project activity carried out on April 23, 2002. There is some material provided in advance by the GDMP team and this text is shown below in blue. Reviewers comments at the end are in Arial font.

## Project Activity: GDMP

**Project Lead:** Heinz Stockinger, CERN, European Organization for Nuclear Research  
CMS Experiment/Computing Group  
Bat. 40-3A-24, CH-1211 Geneva 23  
phone: +41-22-767-1608, fax: +41-22-767-8940  
<http://www.cern.ch/hst/>

**Current Team:** Heinz Stockinger, Shahzad Muzzafar, Flavia Donno, Aleksandr Konstantinov

**Attendees:** Ruth Pordes, Flavio Donno, Miron Livny, Alain Roy, Heinz Stockinger, Shahzad Muzzafar, Koen Holtman, Jennifer Schopf

**Reviewers:** Doug Olson, Arcot Rajasekar, executive team (Ruth, Miron)

**Date:** 23 April 2002

## Documents/URLs to be Read by the Reviewers:

(from GDMP team)

The main documentation of the GDMP project can be found at:

<http://cmsdoc.cern.ch/cms/grid/documentation.html> - In particular, the User Guide for GDMP 3.0

Important publications at conferences can be found at:

<http://cmsdoc.cern.ch/cms/grid/publications.html> - in detail:

Heinz Stockinger, Asad Samar, Bill Allcock, Ian Foster, Koen Holtman, Brian Tierney. File and Object Replication in Data Grids, 10th IEEE Symposium on High Performance and Distributed Computing (HPDC-10), San Francisco, California, August 7-9, 2001.

Asad Samar, Heinz Stockinger. Grid Data Management Pilot (GDMP): A Tool for Wide Area Replication, IASTED International Conference on Applied Informatics (AI2001), Innsbruck, Austria, February 2001.

Mehnaz Hafeez, Asad Samar, Heinz Stockinger. A Data Grid Prototype for DistributedData Production in CMS, VII International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2000), October 2000.

**Status of the Project:**

(from GDMP team)

GDMP version 3.0 is finished and now the project is in the deployment phase and ready to be used. GDMP 3.0 is supposed to be the final GDMP version of this kind. A next-generation version still needs to be designed and details need to be worked out.

**Plans for the Project:**

For the next 3 months:

Deploy GDMP in several testbeds in the US and in Europe as well as use it in production in CMS.

For PPDG Year 2:

A next generation version of GDMP will be based on web services. The timelines will be co-ordinated with EDG -WP2 replica manager and web-service development as well as with the Globus OGSA development.

**Questions for the Project (written responses submitted before the review are appreciated):**

- a) What are the deliverables of your project activity, how has the project met the deliverables to date, what effort has been contributing to the project?  
The main deliverables have been the several GDMP releases that are supposed to mainly satisfy the user requirements of PPDG and EDG. The latest deliverable is GDMP 3.0 including documentation for installation, configuration and usage.
- b) What is the deployment plan for your project activity and what is the state of that deployment?  
GDMP 3.0 will be deployed in the EDG testbed. There are also plans to use GDMP 3.0 in CMS (spring production) and Atlas (Magda)  
GDMP 2.1 is already in use in the EDG testbed and also partly in the CMS experiments. Earlier versions of GDMP were used successfully in the CMS production effort.
- c) Has the project benefited from being part of the PPDG work and if so how?  
From the PPDG side, Asad Samar and Shazhad Muzaffar have done major contributions to the GDMP code. James Amundson also contributed to some of the GDMP releases. The project benefits very much through this contribution. In addition, interactions with other PPDG projects and project members have positively contributed to some of the design solutions taken.

- d) Has the project been hindered by being part of the PPDG work and if so how?  
No.
- e) What collaborations does your project activity rely on and/or contribute to? Have these been of benefit or a hindrance?  
In addition to what is stated under c) the project also established several other contact points and interactions that contributed and will contribute to the success of PPDG, EDG and GriPhyN:  
- Strong interactions and collaborations have been built with the Globus project. In particular, Replica Catalogue and Replica Management design and development as well as information exchange as regards GridFTP.  
- Another collaboration has been built with LBL and in particular Arie Shoshani's group as regards SRM interaction. A trigger was a first interface of GDMP to HRM. Later, a collaboration between EDG and PPDG on SRMs has started and is ongoing. As a first result, a common Global Grid Forum document has been produced and presented.  
- A good collaboration has been established with SLAC and in particular Andrew Hanushevsky to work on a distributed replica catalogue system. The prototype work now is also one of the sources for the Globus-EDG-PPDG collaboration for Replica Catalogues.  
- A mutual document with Reagan Moore (SDSC) on comparisons of Data Grids has been done.
- f) What is your assessment of the potential for adapting the s/w from this project to other experiments?  
Several experiments have already expressed interest in using GDMP.
- g) What do you see as the future needs, deliverables and effort needed for the Project Activity?  
The main source required is personnel for support. Shahzad Muzzafar will officially leave PPDG by the end of April and thus GDMP experts are required at the PPDG side to support GDMP in the PPDG user community. Additional bug fixes of the current release will most likely be required, too.
- h) Is there anything PPDG should be doing more/differently to help with the project activity?  
A replacement for Shahzad Muzzafar is required – from Ruth: The replacement is Alain Roy and Al ? from University of Wisconsin, Condor Team.

## Reviewers Comments

### Findings:

1. GDMP has been and continues to be a successful collaboration of PPDG and EDG WP2.
2. GDMP version 3.0 delivered and installed in the EDG testbed and in VDT and is being used in the CMS production testbed.
3. Work to integrate GDMP with MAGDA in ATLAS is starting, as part of DataTAG effort. This effectively makes the PPDG involvement with GDMP a cross-cut activity since it no longer is focused only on CMS.
4. Other experiments (ALICE, LHCb) have expressed interest and will use GDMP in the EDG testbed.
5. Shahzad Muzzafar has been a strong contributor to the GDMP team and is leaving. Alain Roy will be replacing Shahzad as the PPDG contributor to GDMP at a level of 0.5 FTE. The exact details of Alain's participation are yet to be defined.
6. The roadmap for the next version of GDMP (V4) is still being defined but it will likely be significantly different than GDMP V3 while retaining a publish/subscribe model for data replication. EDG-WP2 will be deploying a Replica Manager and Replica Location Services independently of GDMP V3. GDMP V4 is likely to be implemented in a web services framework.
7. GDMP V3 is expected to work effectively up to a level of 100K files.
8. GDMP V3 can be used with HRM.
9. GDMP V3 can "publish" files to the Replica Catalog but this is not required for GDMP operation. In the EDG testbed the Resource Broker gets file information via the Replica Catalog.
10. Heinz is very concerned that issues of GDMP user support get worked out. He suggests that experiments have a GDMP expert in house to handle problems as a first line of support and the GMDP team gets contacted for problems that can not be handled by that person.
11. Difficulties with administration and installation of certificates for using GSI has been a cause of some problems with GDMP installations. It is probably useful to emphasize that people configuring GDMP need to follow very closely in the Globus installation guide.

### Comments / Concerns:

1. The degradation of performance in GDMP V3 at a level of about 100K files is likely an issue for production use in CMS and STAR. It is expected to be technically feasible to improve this but it will be a question of manpower and priorities.
2. There is not a clear timeline for integration of GDMP and RLS. A timeline should be developed.

3. A lot of the functionality of GDMP for data transfer/replication is being replaced in WP2/Globus with Replica Manager and Replica Location Services. How does this impact the role of GDMP in PPDG?
4. Because of the significant differences of GDMP V3 and V4 it will be helpful to have guidance (guidelines?) to determine if other experiments should deploy GDMP V3, such as for STAR.
5. Operational problems in GDMP V3, such as the deletion problem, are being addressed and will continue to need effort while GDMP V3 is deployed.
6. The problems with proper administration of X.509 certificates indicates a need that may not be just a GDMP need.
7. The 0.5 FTE replacement for Shahzad may not be enough, considering the additional use GDMP is getting and the split between developing V4 and supporting V3. Perhaps some CMS support can be found in DataTAG similar to the ATLAS-DataTAG work with GDMP.

### **Recommendations:**

1. Encourage the use of the EDG and VDT bug tracking systems when reporting problems with GDMP.
2. Pre-installation tests/checklist to verify proper conditions (Globus, etc.) for installation of GDMP (may be taken care of in packaging with EDG, VDT?)
3. The issue of support for GDMP should be clarified and publicized. The roles of the GDMP team, EDG, VDT team, and mutual expectations with experiments should be described.
4. STAR should try to use GDMP (via VDT) along with HRM for the BNL-LBNL data replication.
5. The plans for GDMP V4 and WP2 are in being developed, it seems mostly within WP2. There should be significant PPDG involvement in developing these plans.
6. Given the evolution of GDMP to V4 as a publish/subscribe application on top of a separate Replica Manager and Replica Location Services, PPDG should consider if the role of collaboration with EDG on GDMP should stay focused on GDMP or if it should include more of the WP2 activities.
7. VDT team (Miron, et al) should discuss with SRM team (Arie, et al.) the possibility of packaging HRM in VDT. At least, HRM should be packaged for use with GDMP from VDT.