

**DATAMINING**

**qr10**

**Deliverable D72 (PC1)  
Final Collaboration  
Plan**



# DATA MINING TOOLS AND SERVICES FOR GRID COMPUTING ENVIRONMENTS

Deliverable D72 (PC1) Final Collaboration Plan

**Responsible author(s):** Nahum Korda  
**Co-author(s):** Werner Dubitzky, Juergen Franke

## Revision history

| Deliverable administration and summary  |                            |
|---|----------------------------|
| <b>Project acronym:</b> DataMiningGrid  | <b>ID:</b> IST-2004-004475 |
| <b>Document identifier:</b> D72 (PC1) Final Collaboration Plan                  |                            |
| <b>Leading partner:</b> Technion – Israel Institute of Technology               |                            |
| <b>Report version:</b> 0.4  |                            |
| <b>Report preparation date:</b> 29.12.2004                                      |                            |
| <b>Classification:</b> Public   |                            |
| <b>Nature:</b> R  |                            |
| <b>Author(s) and contributors:</b> Nahum Korda, Werner Dubitzky, Juergen Franke |                            |
| <b>Status:</b>  | Plan                       |
|   | Draft                      |
|   | Working                    |
|   | Final                      |
| <b>X</b>  | <b>Submitted</b>           |
|   | Approved                   |

The DataMiningGrid © Consortium has addressed all comments received, making changes as necessary. Changes to this document are detailed in the change log table below.

| Date       | Edited by       | Status | Changes made   |
|------------|-----------------|--------|--|
| 23.12.2004 | Nahum Korda     | Draft  | First draft  |
| 28.12.2004 | Juergen Franke  | Draft  | Section 8 Training Activities                                      |
| 28.12.2004 | Werner Dubitzky | Final  | Joint Dissemination, Final check and minor editorial modifications |
|            |                 |        |  |

Notice that other documents may supersede this document. A list of latest public DataMiningGrid deliverables can be found at the DataMiningGrid webpage at [www.DataMiningGrid.org/dissemination](http://www.DataMiningGrid.org/dissemination).

## Copyright

This report is © DataMiningGrid Consortium 2004. Its duplication is allowed only in the integral form for anyone's personal use for the purposes of research or education.

## Citation

Nahum Korda, Werner Dubitzky, Juergen Franke, (2004). Deliverable D72 (PC1) Final Collaboration Plan. DataMiningGrid Consortium, [www.DataMiningGrid.org](http://www.DataMiningGrid.org).

## Acknowledgements

The work presented in this document has been conducted in the context of the EU Framework Programme VI project IST 2004 004475 DataMiningGrid. DataMiningGrid is a 24-month project that started on September 1st, 2004 and is funded by the European Commission as well as by the industrial partners. Their support is appreciated.

The partners in the project are University of Ulster (UU), Fraunhofer Institute for Autonomous Intelligent Systems (FHG), DaimlerChrysler (DC), Israel Institute of Technology (TECHNION) and University of Ljubljana (LJU). The content of this document is the result of extensive discussions within the DataMiningGrid© Consortium as a whole.

## More information

Public DataMiningGrid reports and other information pertaining to the project are available through DataMiningGrid public web site under [www.DataMiningGrid.org](http://www.DataMiningGrid.org).

## Executive summary

The DataMiningGrid project will closely follow the recommendations developed within the IST-FP6 Grid Projects Concertation initiative. In each of the proposed collaboration areas DataMiningGrid plans activities that are suitable to its duration and Consortium size. The DataMiningGrid Consortium recognizes that it will have to selectively focus its concertation activities. Due to its short project duration and small size, the Consortium cannot be expected to initiate and realize all of the envisioned comprehensive concertation activities. However, the members of the Consortium are eager to join emerging initiatives, and to contribute to these initiatives as much as possible within the existing time and resource constraints.

There three concrete concertation activities planned in the DataMiningGrid project: (1) initiation of joint discussion regarding the data mining requirements from next-generation data grids (leading possibly to joint publications), (2) organization of a joint mini-conference or workshop in collaboration with other European grid-oriented projects, and (3) issuing of a Project Impact Assessment Report at the end of the project (using the recommended project impact assessment indicators).

## Table of contents

|  |    |
|--|----|
| Executive summary .....                                    | 5  |
| Table of contents .....                                    | 6  |
| 1 Concertation .....                                       | 7  |
| 2 Exploitation of Synergies / Technical Concertation ..... | 8  |
| 3 Joint Forums for Exchange and Dissemination .....        | 10 |
| 4 Coordination of Standardisation Efforts .....            | 11 |
| 5 Repository of Reference Implementations .....            | 12 |
| 6 Collaboration on Research Inventories and Roadmaps ..... | 13 |
| 6.1 Collaboration on Research Inventories .....            | 13 |
| 6.2 Collaboration on Roadmaps .....                        | 14 |
| 7 Indicators and Impact Assessment .....                   | 15 |
| 8 Training Activities .....                                | 17 |
| 9 Conclusions and Future Work.....                         | 18 |
| 10 References.....   | 19 |
| 11 Acronyms .....  | 20 |

## 1 Concertation

The DataMiningGrid will focus its concertation activities mainly on the potential technology exchange that will proceed in the following steps:

First, **scouting and intelligence gathering regarding the technologies that are planned to be researched and developed in other European grid-related projects.** This effort must go beyond the follow-up of the publicly available information, since most of the really interesting information is in the projects' Descriptions of Work, which are not publicly available, and in internal discussions. Accordingly, personal contact must be sought with all relevant projects in order to better understand if there are potential duplications of efforts between them, and to avoid potential technological competition.

Second, **decision regarding questions on which of the parallel European projects are relevant for collaboration, and a clear definition of the possible outcomes of such collaboration.** The direct collaboration was established with the intelliGrid and OntoGrid projects in the field of the semantic grid standards, as well as with the Provenance project by providing the requirements for the mechanisms that allow information generated and managed within a grid infrastructure to be proven and trusted.

Third, **establishing a joint workplan for collaboration.** Such a workplan may factually include joint research and development efforts, but also mutual testing, validation and evaluation of the project results, integration of the other projects' results into the research and development, etc. It is, however, mandatory that the joint workplans do not interfere with the original individual projects' workplans<sup>1</sup>.

Fourth, **joint evaluation of the collaboration results.** There exists a project impact assessment criterion suggested by the IST-FP6 Grid Projects Concertation initiative (see [below](#)) that requires such evaluation: "The cooperation among initiatives in the member states and EC initiatives has increased."

Fifth, **planning of the future collaboration activities that outlive the actual projects' lifetimes.** The ideal outcome would be, of course, joint exploitation and commercialization of the collaboration results. However, joint proposals for possible future EC initiatives or continuing projects based on bilateral cooperation between the member states could be also expected.

---

<sup>1</sup> But if they do interfere, it is currently unclear whether there are mechanisms that could resolve this problem.

## 2 Exploitation of Synergies / Technical Concertation

The IST-FP6 Grid Projects Concertation initiative set the following eight topics as the technological foci for cooperation between the projects:

- T1 Grid Architecture
- T2 Mobile Grid
- T3 Distributed Collaborations
- T4 Monitoring and Fabric Management
- T5 Data Management
- T6 Trust and Security
- T7 Semantic Grid
- T8 Business Services and Workflow

As STREP, with relatively short duration and small number of participants, the DataMiningGrid project cannot become a major contributor in most, if not all, of these topics. Moreover, the concertation efforts within the project will be effective only if they are strictly focused on a limited number of activities.

Since data mining critically depends on sophisticated data access mechanisms, the technological focus *T5 Data Management* seems to be the most suitable direction for concentrating the collaboration efforts within the DataMiningGrid project.

Currently, the most influential project in this area is Enabling Grids for E-Science in Europe (EGEE) [1], which evolved from the European DataGrid (EDG) [2]. Within EGEE, the Large Hadron Collider (LHC) Computer Group (LCG) [3] maintains currently the largest data grid worldwide. A small testbed integrated with this data grid, called Gilda [4], is available to other European projects courtesy of the Istituto Nazionale di Fisica Nucleare (INFN – The Italian National Institute for Nuclear Physics) [5] that is responsible for its maintenance. Gilda offers an exquisite opportunity for testing grid applications, as well as for demonstration and training activities.

Collaboration with EGEE in general, and specifically with the Gilda testbed, can be sought at various levels, and it may also include joint dissemination and training activities. Nevertheless, the most important aspect of the collaboration with EGEE and other European projects involved in *T5 Data Management* activities is a **joint discussion of the specific requirements from the next-generation data grids that would allow effective data mining on the grid**. We expect that the DataMiningGrid project, despite its relatively small size as a STREP, can make a significant impact on the future evolution of data grids by making the European and the global data grid community aware of specific data mining requirements.

For that purpose DataMiningGrid will offer to co-chair the Data Access Workgroup and ensure that data mining requirements play a significant role in the workgroup's agenda. As a result of this discussion, towards the end of the project, DataMiningGrid will seek to co-author a position paper specifying the data mining requirements from the data grids.

In 2004 a similar effort led to the foundation of the Grid Information Retrieval Workgroup (GIR-WG) [6] within the GGF. The GIR-WG recently published GIR Requirements and GIR Architecture recommendations. A complementary contribution on grid data mining requirements and architecture would certainly be considered a great achievement of the DataMiningGrid project. Accordingly, possible collaboration with GIR-WG towards extending their activity into the area of grid data mining will be sought within the project's lifetime.

### 3 Joint Forums for Exchange and Dissemination

For STREP the greatest benefit of the collaboration between projects comes arguably from joint dissemination efforts that significantly lower the costs and organizational strain, and guarantee an improved visibility. Accordingly, DataMiningGrid plans to organize a mini-conference or symposium in collaboration with other related European projects sometime in fall 2005. The project partners have already been working on establishing a wider community in the area of the project's research by successfully organizing an international Workshop [7] on grid and data mining technologies at the prestigious IEEE International Conference on Data Mining [10] (Brighton, UK, November 2004). Other related, community-building activities where members of the Consortium are actively involved include the International Symposium on Knowledge Exploration in Life Science Informatics: KELSI 2004 [9] (Milano, Italy, 25-26 November 2004), 3rd International Workshop on Biomedical Computation on the Grid: BioGrid 2005 [8] (hosted with 5th IEEE/ACM International Symposium on Cluster Computing and the Grid) (Cardiff, UK, 9-12 May 2005), and a special issue on HealthGrid: Toward Collaborative and On-Demand Healthcare in Journal of Clinical Monitoring and Computing, Kluwer. As mentioned above, in the section dedicated to the concertation efforts, the scouting and intelligence gathering regarding the technologies planned to be researched and developed in other European projects can reveal which other projects could be approached with the proposal for organizing this joint event.

DataMiningGrid will also actively seek to join events organized by larger, leading grid-related projects and forums in order to ensure its appropriate representation, thus achieving improved visibility.

## 4 Coordination of Standardisation Efforts

The DataMiningGrid project is actively participating in the European GRID Standards Co-ordination Group. The objectives of this group were defined as:

1. Define GRID standards to focus on.
2. Regularly track and review all GRID standards candidates across all the projects.
3. Identify and foster synergies, dependencies and opportunities for collaboration between the GRID standards candidates.
4. Actively identify and promote opportunities for resource sharing between projects, specifically when it comes to participation in the activities standards bodies.
5. Promote SME participation.

In the light of the criticism expressed on behalf of the Commission it is expected that these objectives will be modified, and that a more practical and focused approach will be endorsed. However, the DataMiningGrid project will extend its cooperation in order to support the future activity of this group.

As STREP, with relatively short duration and small number of participants, the DataMiningGrid project cannot hope to bear a significant impact on the standardisation activities related to the grid technologies. The project is, nevertheless, coordinating its activities with COPRAS [11], another Consortium put together with the objective to improve the interface between research and standardisation efforts. COPRAS is run by the three officially recognized European Standards Organizations: CEN, CENELEC and ETSI, together with the Open Group and the W3C. The overall strategic objectives of COPRAS are to support and encourage the IST-FP6 projects to partake in the standardization activities in Europe and worldwide, and to generally increase awareness of existing standardisation activities among the researchers and developers.

## 5 Repository of Reference Implementations

Integrated in its official website, the DataMiningGrid will maintain a page containing links to all relevant open source software that will be used, or referred to, within the project. This approach is certainly preferable to maintaining yet another mirror-repository of the already existing repositories (that are in fact kept on the fast-downloading servers preferred for downloading). This page will include information regarding the relevancy of the technologies to the DataMiningGrid project. It will also include links to other specialized inventories and repositories.

Nevertheless, the software that will be developed within the project and released under the open source license will be exposed to the public in a dedicated repository. This dedicated repository will be also integrated into the official project website.

Whenever necessary, a Control Versions Systems (CVS) [12] will be established between the relevant partners either within the project, or even between collaborating projects.

## 6 Collaboration on Research Inventories and Roadmaps

DataMiningGrid joined the efforts coordinated by the CoreGRID project. Objectives and activities of this task comprise:

- The continuation of the GRIDSTART inventory,
- Collaboration with GridCoord which will work on the inventory of national Grid initiatives, and
- The creation of a European-wide roadmap build on the activities of CoreGRID and other projects activities in this area

During the Concertation Event it was decided to split the work in two separate tracks: the collaboration on research inventory, and the collaboration on roadmaps.

### 6.1 Collaboration on Research Inventories

The objective of this task will be the provision of a thematical European Research Inventory. This work must be closely related to the activities within the Technical Groups within the collaboration task on Exploitation of Synergies, and the technical concertation, as well as to the work planned in the collaboration group on Impact Assessment and the Collaboration on Joint Repositories. Synchronization is required primerily in the production of whitepapers as part of the technical concertation. For example, the attempts to produce a state-of-the-art overview on the semantic web technologies need not to be duplicated in this activity. In addition, in order to support the Impact Assessment activity it is required to provide the baseline that is applied to the metrics that are supposed to measure the progress.

The initial baseline of this activity is surely the work performed within the FP project GRIDSTART resulting in their inventory document.

Two major work items that are to be addressed in the first phase are proposed:

- Overview and Assessment of Web Services Standards, Identified interoperability problems on specification level, and level of support of this specification in toolkits, and
- Middleware components available from FP5 projects, National Initiatives and commodity-of-the-shelf (COTS) technologies organized around major building blocks.

The DataMiningGrid project will extend its cooperation in order to support these activities, and provide the requested material.

## 6.2 Collaboration on Roadmaps

Objectives of the European Grid Roadmap Group were defined as following:

- Create a Europe-wide roadmap build on the activities of CoreGRID and other projects' activities in this area, and
- Increase the efficiency of collaboration between the GRID projects.

The work of this group will have a strong relation to the outcome of the roadmap activities taking place in each of the projects.

Depending on the type of a project partners from different projects play partly different roles in the concertation activities. The DataMiningGrid project, as a STREP will contribute the view and the perspectives of its specific niche, and will provide the project's roadmap to be integrated into a general, Europe-wide roadmap for grid-related activities.

## 7 Indicators and Impact Assessment

In order to measure the impact on the strategic objectives of the work programme with respect to grid computing, the DataMiningGrid adopts both the evaluation criteria and the indicators that were developed within the Collaboration Task 6 of the IST-FP6 Grid Projects Concertation initiative [13]. The following are the evaluation criteria that have been established:

- The cooperation among initiatives in the member states and EC initiatives has increased.
- The influence of members of the ERA in the global grid community has increased.
- Interaction between academic and industrial partners has increased.
- Degree of grid penetration in *complex problem solving* and new application areas.
- Grid-enabled collaboration within business communities (like distributed supply chains) has increased.
- The capability and functionality of next generation grid toolkits and middleware has been increased.
- New grid generic toolkits and middleware makes the application of grid technology possible and easier.

Each evaluation criterion has further indicators:

- The cooperation among initiatives in the member states and EC initiatives has increased.  
Indicator 1. Number of steps taken and actions initiated.  
Indicator 2. ERA Coordination.
- The influence of members of the ERA in the global grid community has increased.  
Indicator 3. Degree of involvement in Grid Standards Community (GGF and OASIS).  
Indicator 4. Downloads of European Grid Middleware.
- Interaction between academic and industrial partners has increased.  
Indicator 5. Participation of commercial representatives in European grid events.  
Indicator 6. Number of commercial products including grid technology from European Projects has increased.
- Degree of grid penetration in Complex Problem Solving and new application areas.  
Indicator 7. Percentage and absolute number of applications using grid technology.

Indicator 8. Number of grid enabled computing resources, which are made available by vendors and academia.

- Grid-enabled collaboration within business communities (like distributed supply chains) has increased.

Indicator 9. Number of grid enabled analysis methods, problem solving environments and workflow tools in selected application areas, which are relevant for industrial usage.

- The capability and functionality of next generation grid toolkits and middleware has been increased.

Indicator 10. Number of new products resulting from the GRID IST projects.

- New grid generic toolkits and middleware makes the application of grid technology possible and easier.

Indicator 11. Evaluation reports by external users of the new products, tools and environments

Applying these indicators to the project results, the DataMiningGrid will issue a Project Impact Assessment Report at the end of the project.

## 8 Training Activities

Members of the DataMiningGrid Consortium took part in the Collaboration Session C2 held at the European Grid Technology Days 2004, IST-FP6 Grid Projects Concertation Meeting (16-17 September, 2004, Brussels). This session was related to the tasks *Joint Forums for Exchange and Dissemination* and *Training*.

As a result of the session there was an agreement regarding the possibility of having a single web portal for training activities.

A focus group lead by the representatives for training activities of the CoreGrid (Rosa M. Badia), NextGrid (Malcolm Atkinson), Akogrimo (Victor Villagr ) and OntoGRID (Mike Wooldridge) projects is organized in order to decide what should be the next steps. In order to formulate an initial proposal, a telephone conference was held end of October 2004.

Furthermore, all project representatives were asked to provide a contact person who will represent their projects in the Training Workgroup.

In the meantime, a BSCW (Basic Support for Cooperative Work) server [14] was setup by the Fraunhofer-Institut in Sankt Augustin (participant in the Simdat Project) specifically for the purpose of coordinating all training activities.

Additionally a document (D.CG.04 Draft Collaboration Plan for Task 7: Training activities (CPC1-T7)) was composed by the members of the focus group in cooperation with members from each FP6 Grid based project. This document was delivered at the 31<sup>st</sup> December 2004 to the commission.

Currently the DataMiningGrid Consortium is not planning to independently initiate joint training activities, but rather to join and support selected activities proposed by the Training Workgroup.

## 9 Conclusions and Future Work

The following concertation activities are planned within DataMiningGrid project:

1. Scouting other EC initiatives for the purpose of technology exchange, and joint development, testing, validating and evaluating.
2. Initiating joint discussion regarding the specific requirements from the next-generation data grids that would allow effective data mining on the grid. This effort will be focused within the workgroup handling the Data Access technological focus. Possibly, joint publication of a position paper on data mining on the grid may be initiated.
3. Organizing a joint mini-conference or seminar in collaboration with other European grid-oriented projects.
4. Issuing a Project Impact Assessment Report at the end of the project. The report will use the recommended project impact assessment indicators, and follow the recommendations of the IST-FP6 Grid Projects Concertation initiative.

Joining and supporting selected training activities proposed by the IST-FP6 Grid Projects Concertation initiative.

## 10 References

- [1] EGEE website, <http://public.eu-egee.org/>.
- [2] The DataGrid website, <http://eu-datagrid.web.cern.ch/eu-datagrid/>.
- [3] LHC Computer Group website, <http://lcg.web.cern.ch/LCG/>.
- [4] Gilda testbed website, <https://gilda.ct.infn.it/testbed.html>.
- [5] Istituto Nazionale di Fisica Nucleare website, <http://www.infn.it/>.
- [6] The Grid Information Retrieval Working Group website, <https://forge.gridforum.org/projects/gir-wg>.
- [7] The DMGrid Workshop website, <http://www.cs.technion.ac.il/~ranw/dmgrid/>.
- [8] BioGrid 2005, at <http://www.cse.uconn.edu/~huang/BioGrid-05/>.
- [9] KELSI 2004, at <http://research.bioinformatics.ulster.ac.uk/kelsi2004/>.
- [10] The ICDM website, <http://icdm04.cs.uni-dortmund.de/>.
- [11] The CORPAS website, <http://www.w3.org/2004/copras/>.
- [12] Control Version Systems website, <https://www.cvshome.org/>.
- [13] C.-A. Thole (Editor), "Collaboration Task 6: Indicators and Impact Assessment" (Outline of a Collaboration Task among GRID IST Projects), <http://www.nextgrid.org/events/>.
- [14] BSCW website, <http://bscw.scai.fraunhofer.de/bscw/bscw.cgi/0/1296>

## 11 Acronyms

|        |  |
|--------|--|
| COPRAS | Cooperation Platform for Research and Standards  |
| CVS    | Control Version System   |
| EDG    | European DataGrid  |
| EGEE   | Enabling Grids for E-Science in Europe   |
| ERA    | European Research Areas  |
| GGF    | Global Grid Forum  |
| GIR-WG | Grid Information Retrieval Workgroup   |
| INFN   | Istituto Nazionale di Fisica Nucleare (The Italian National Institute for Nuclear Physics) |
| LCG    | LHC Computer Group   |
| LHC    | Large Hadron Collider  |