

DATAMINING

GRID

**Deliverable D71:(PC1)
Draft Collaboration Plan**

DATA MINING TOOLS AND SERVICES FOR GRID COMPUTING ENVIRONMENTS

D71: (PC1) Draft Collaboration Plan

Responsible author:	Nahum Korda
Co-authors:	Assaf Schuster, Vlado Stankovski, Werner Dubitzky

Report Version:	0.3
Report Preparation Date:	31.10.2004
Classification:	Public
Status:	Draft
Nature:	Report

Revision history:

Deliverable Administration & Summary	
Project acronym : DataMiningGrid	ID: IST-2004-004475
Deliverable number & name: D71: (PC1) Draft Collaboration Plan	
Version : 0.3	Date: 31 October 2004
Authors & contributions: Nahum Korda, Assaf Schuster, Vlado Stankovski, Werner Dubitzky	
Compiled by: Nahum Korda, Technion	Classification: Public
Short description: Collaboration plan	
Status:	<input checked="" type="checkbox"/> Draft (issued for contributions)
	<input checked="" type="checkbox"/> Working (issued for comments)
	<input type="checkbox"/> Final (issued for approval)
	<input checked="" type="checkbox"/> Approved (by document QA)
	<input checked="" type="checkbox"/> Cleared for submission to the EC (by Quality Manager)
	<input type="checkbox"/> Accepted (by the EC)

Status of this document:

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

Comments are welcome at www.DataMiningGrid.org/comments and general discussion of related technology is welcome at www.DataMiningGrid.org/comments.

The DataMiningGrid Consortium maintains a list of any patent disclosures related to this work.

This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current DataMiningGrid publications and the latest revision of this deliverable can be found at the project website under www.DataMiningGrid.org/disseminaton.

Change log table			
Date	Author	Status	Changes
31.10.2004	NK, AS, VS, WD	Draft	First complete draft.

Copyright

This report is property of the DataMiningGrid Consortium 2004. Its duplication is restricted to the personal use within the Consortium, funding agency and project reviewers.

Citation

Nahum Korda, Assaf Schuster, Vlado Stankovski, Werner Dubitzky, *Deliverable D71 – First Draft*, The DataMiningGrid Consortium, www.DataMiningGrid.org, 2004.

Acknowledgements

The work presented in this document has been conducted in the context of the project IST 2004 004475 DataMiningGrid. DataMiningGrid is a 24-month project that started on 1st September 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 the University of Ljubljana (LJU), the University of Ulster (UU), the Fraunhofer Institute for Autonomous Intelligent Systems (FHG), DaimlerChrysler (DC) and the Israel Institute of Technology (TECHNION). The content of this document is the result of extensive discussions within the DataMiningGrid Consortium.

More information

Public DataMiningGrid reports are available through DataMiningGrid public website at 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 the 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
1 CONCERTATION	7
2 EXPLOITATION OF SYNERGIES / TECHNICAL CONCERTATION.....	9
3 JOINT FORUMS FOR EXCHANGE AND DISSEMINATION	11
4 COORDINATION OF STANDARDISATION EFFORTS	12
5 REPOSITORY OF REFERENCE IMPLEMENTATIONS.....	13
6 COLLABORATION ON RESEARCH INVENTORIES AND ROADMAPS.....	14
7 INDICATORS AND IMPACT ASSESSMENT	15
8 TRAINING ACTIVITIES.....	17
9 SUMMARY	18
10 REFERENCES.....	19

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. A good example why such intelligence is required for successful collaboration is to be found in a potential opportunity emerging from the fact that the Enabling Grids for E-Science in Europe (EGEE) [1] project is apparently considering replacing the existing Resource Broker component (developed within the European DataGrid project [2]) with Condor-G [3] in their next software release. Since, on the other hand, DataMiningGrid intends to develop a monitoring application that efficiently parses distributed Condor logs, collaboration between the projects would be welcome from both sides, and the unnecessary duplication of efforts could be completely avoided.

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.** To go back to the previous example, a good collaboration objective could be that the DataMiningGrid monitoring technology is included in the final EGEE software release. Such objective could save unnecessary efforts by EGEE and at the same time efficiently promote the results of the DataMiningGrid project.

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' workplans¹.

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."

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



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.

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) [4] maintains currently the largest data grid worldwide. A small testbed integrated with this data grid, called Gilda [5], is available to other European projects courtesy of the Istituto Nazionale di Fisica Nucleare (INFN – The Italian National Institute for Nuclear Physics) [6] 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.

Last year a similar effort led to the foundation of the Grid Information Retrieval Workgroup (GIR-WG) [7] within the GGF. The GIR-WR 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 [8] on grid and data mining technologies at the prestigious IEEE International Conference on Data Mining [9] (Brighton, UK, November 2004). 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

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 should, nevertheless, coordinate its activities with COPRAS [10], a 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.

As mentioned above in the section dedicated to the concertation activities, the DataMiningGrid could attempt to seek collaboration with the GGF GIR-WG towards extending their activity into the area of grid data mining. Alternatively, the project could establish new workgroup within GGF dedicated specifically to data mining on in grid computing environments (complementing the GIR-WG). This possibility will be investigated during the project's lifetime.

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) [11] will be established between the relevant partners either within the project, or even between collaborating projects.

6 Collaboration on Research Inventories and Roadmaps

There are several inventories presenting grid research initiatives in Europe and worldwide. The DataMiningGrid project must ensure that it is enlisted on all such inventories and websites that are typically used to search collaboration partners.

For example, the GRIDSTART initiative sponsored by the European Commission maintains a website that serves as a portal to a diverse collection of activities and initiatives taking place not only within the European Union but elsewhere, most notably in the United States. By being enlisted in the GRIDSTART and similar portals, the DataMiningGrid will achieve excellent visibility.

Some DataMiningGrid Partners have been involved in IST-FP5 roadmap projects, and will continue that work as part of the Consortium's concertation 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 [12]. 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 (engineering and pharma), 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 Badia) and NextGrid (Malcolm Atkinson) projects will be organized in order to decide what should be the next steps. In order to formulate an initial proposal, a telephone conference is expected to be held sometime in 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 was setup by the High Performance Computing Centre in Stuttgart (participant in the Akogrimo Project) specifically for the purpose of coordinating all training activities.

A first draft of a document that was created during the meeting lists the preliminary ideas of collaboration on training activities such as:

- Joint Conference on Grid Technologies
- Concertation Meeting
- CoreGRID quarterly newsletter
- IST Grid Research Web Portal

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 Summary

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.
5. 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] The Condor Project website, <http://www.cs.wisc.edu/condor/condorg/>.
- [4] LHC Computer Group website, <http://lcg.web.cern.ch/LCG/>.
- [5] Gilda testbet webiste, <https://gilda.ct.infn.it/testbed.html>.
- [6] Istituto Nazionale di Fisica Nucleare website, <http://www.infn.it/>.
- [7] The Grid Information Retrieval Working Group website, <https://forge.gridforum.org/projects/gir-wg>.
- [8] The DMGrid Workshop website, <http://www.cs.technion.ac.il/~ranw/dmgrid/>.
- [9] The ICDM webiste, <http://icdm04.cs.uni-dortmund.de/>.
- [10] The CORPAS website, <http://www.w3.org/2004/copras/>.
- [11] Control Version Systems website, <https://www.cvshome.org/>.
- [12] 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/>.
- [13] BSCW website, <http://bscw.fit.fraunhofer.de/>.