



IST-2004-004475

DataMiningGrid

Data mining Tools and Services for Grid Computing Environments

Specific Targeted Research or Innovation Project

2.3.2.8 Grid-based Systems for Complex Problems Solving

D82(1): Description of User Groups

Due date of deliverable: M06 (28 February 2005)

Actual deliverable submission date: 3 April 2005

Start date of project: 1 September 2004

Duration: 24 months

University of Ulster

Revision: Submitted

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the Consortium (including the Commission Services)	
CO	Confidential, only for members of the Consortium (including the Commission Services)	

DATAMINING

qr10

**Deliverable D82(1):
Description of User
Groups**



DATA MINING TOOLS AND SERVICES FOR GRID COMPUTING ENVIRONMENTS

Deliverable D82(1): Description of User Groups

Responsible author(s): Terence Dörflinger
Co-author(s): Werner Dubitzky, Jürgen Franke, Nahum Korda,
Thomas Niessen, Gerd Paaß, Francois Perrevort,
Vlado Stankovski, Jernej Trnkoczy

Revision history

Deliverable administration and summary		
Project acronym: DataMiningGrid	ID: IST-2004-004475	
Document identifier:	DataMiningGrid-del-D82(1)	
Leading Partner: FHG		
Report version: 02		
Report preparation date: 28.2.2005		
Classification: Public		
Nature: Report		
Author(s) and contributors: Terence Dörflinger (FHG) in collaboration with all Partners		
Status:	-	Plan
	-	Draft
	-	Working
	-	Final
	X	Submitted
	-	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
25.10.04	Jernej Trnkoczy	Plan	'a report template will be defined that all reports will follow'
18.02.05	Terence Dörflinger	Final	First version of 'Description of User Groups'
03.04.05	Werner Dubitzky	Submitted	Final checks and minor corrections.

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.

Copyright

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

Citation

Terence Dörflinger (2005), Deliverable D82(1). DataMiningGrid Consortium, c/o University of Ljubljana, 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.

Executive Summary

A critical success factor for the DataMiningGrid project is to ensure that project outcomes meet the user demand and user requirements. This goal can only be achieved, if there exists detailed information on future user groups. But since the user groups that may exist after completion of the project are difficult to be known in advance, more or less accurate estimates have to be made.

In this document, a distinction will be made between *potential*, *targeted* and *actual* user groups. The potential user groups provide a general overview of possible user groups for the project outcomes. Within this broad range of users, specific user fields have to be picked that will be directly addressed by the use cases. The definition of these fields is referred to as *targeted user groups*. To ensure that targeted users will adapt the technologies developed by the project and become actual user groups, control mechanism will be applied and documented.

Table of Contents

Executive Summary	6
Table of contents	7
1 Introduction	9
1.1 Objectives of this document	9
1.2 Document amendment procedure.....	9
2 Objectives of the Description of User Groups.....	10
3 Description of User Groups	11
3.1 Types of user groups	11
3.2 Potential user groups.....	11
3.2.1 Enhancements to grid middleware.....	12
3.2.2 Grid data mining APIs.....	12
3.2.3 Data-mining applications.....	12
3.3 Targeted user groups	13
3.3.1 University of Ulster.....	13
3.3.2 Fraunhofer Institute for Autonomous Intelligent Systems.....	14
3.3.3 DaimlerChrysler.....	15
3.3.4 Israel Institute of Technology	15
3.3.5 University of Ljubljana	15
4 Controlling of User Groups	17
4.1 Controlling actions	17
4.1.1 Meetings.....	17
4.1.2 Surveys.....	17
4.2 Controlling consequences.....	17
4.2.1 Result adaptation.....	17

4.2.2 Redefinition of targeted user groups 18

5 Conclusions and Future Work..... 19

6 References..... 20

1 Introduction

1.1 Objectives of this document

The various results of the DataMiningGrid project are intended to meet the demand of certain user groups and their requirements. To ensure this, it is necessary to know relatively accurately, which the user groups there are going to be and what their demands and requirements are.

This document is the second of the three deliverables of the 'Dissemination, Awareness and Exploitation' workpackage of the DataMiningGrid project. It identifies potential user groups and characterizes the targeted user groups of the particular outcomes of different Partners in order to provide customized solutions for the user.

The purpose of the description of user groups is to provide a formal planning document for customizing project results and ensure that user needs will be addressed. It is vital for success of the project that all Partners have a clear understanding of targeted users and realistically estimate whether their results meet user demands and requirements. This Plan will be revised and updated upon changes in knowledge about user demands and triggers changes in applications design.

The description of user groups builds on the Technical Annex [Annex04] and compliments the Dissemination Plan [Dissemination05] and the Exploitation Plan [Exploitation05] to further highlight how user groups or results change for the DataMiningGrid project. The aim is to provide a clear design strategy for all Partners and members of the Management Board.

1.2 Document amendment procedure

The Description of User Groups is Deliverable D82(1) of WP8. It will be updated at month 6 and 18.

2 Objectives of the Description of User Groups

Having detailed knowledge about the user groups of the DataMiningGrid project outcomes is necessary for a variety of tasks.

To have clear picture of targeted user groups, it is most important to:

- Ensure that project outcomes meet the user demand.

Moreover it affects other important project aspects such as:

- Specification of requirement;
- Effective dissemination;
- Effective exploitation.

To ensure that user demands will be matched, the following actions have to be carried out:

- Identifying potential user groups for project results,
- Identifying targeted user groups for project results,
- Determining whether targeted user groups get addressed,
- Controlling whether user targeted groups change throughout development.

The identification of potential user groups focuses on the general overview of possible users. It provides valuable information for the definition of markets that are subject to penetration in the exploitation. The targeted user groups are that part of the potential user groups that have been decided to particular address by the specific project results.

For the overall success of the DataMiningGrid project it is indispensable that the identified targeted user groups will actually benefit from the developed results. Hence, control mechanisms have to be established, which, while the project progresses, have to be evaluate on a regular basis whether the demands of the targeted user group still correspond to the anticipated results of the project. Possible hazards and risks are changes in user groups that, in the worst case, make the anticipated project outcomes irrelevant. In this case, either the requirements for software developments need to be adjusted, or the targeted user group has to be redefined. The latter case involves considerations on whether the project results can be sufficiently exploited within the changed targeted user group, as well as changes to the dissemination strategy to address the desired user.

3 Description of User Groups

3.1 Types of user groups

A user group is a group of individuals with common interests in a technology or an application. In the scope of the DataMiningGrid project user groups will be referred to as the *group of individuals that have use for one or more specific results, which will be developed* in the project. We distinguish between groups of targeted users, potential users and actual users:

- Potential user groups: All user groups that potentially have a demand for or interest in the project's outcomes,
- Targeted user groups: User groups that are intentionally and deliberately addressed by the project results,
- Actual user groups: User groups that actually will use the project results after its completion.

While it is fairly easy to determine a diffuse class of potential users and to define a group of targeted users, it is impossible to know in advance of the actual users. Therefore, the goal must be, to implement a mechanism to estimate the actual users in advance. This estimation should then match with the definition of targeted user groups in order to avoid acceptance problems.

3.2 Potential user groups

A very important aspect that highly influences the usage of project outcomes and therefore the identification of user groups is the question of the intellectual property rights (IPR). The IPR concerning newly developed software are defined in the Consortium Agreement with respect to three different categories of project results as follows:

- Enhancements to Grid Middleware – will remain open source,
- New Generic Grid data mining APIs - will be joint property of all parties and shall not be available as open source for third parties. Each party shall be granted access rights to this jointly developed knowledge to the extent that:
 - No exclusive, transferable nor limited licenses of any form will be granted to third parties without prior written consent by all parties. A detailed licensing policy will be agreed by all parties, and specified in a form of an Annex to the Consortium Agreement,
 - Each Party is entitled to use the APIs under the Consortium Agreement for future EU-funded projects and for research and technology development projects in cooperation with industrial and/or academic Partners,
- New Grid-based data mining Algorithms and Tools - will be regarded as knowledge and become property of the individual parties or groups of parties that have developed them.

The IPR structure of project results is applicable to determine potential user groups. For the three different categories of results, different user groups can be identified.

3.2.1 Enhancements to grid middleware

The Consortium will produce concrete enhancements of the existing grid software (like e.g. Globus [Globus04], Condor [Condor04]) and will make these software improvements publicly available (e.g., the data-mining-based monitoring of grid software).

These grid software enhancements are initially intended for use by the project Partners, supporting or enabling the development of data-mining algorithms, tools and applications. Besides the internal usage, the potential user group is the general public. Since the enhancements will be open source they are potentially available for anyone interested.

The DataMiningGrid project encourages the use and adoption of grid software enhancements by the grid community and EU grid research projects as pointed out in the Final Collaboration Plan [Collab04].

3.2.2 Grid data mining APIs

On a generic application level, precisely defined data-mining interfaces will be generated for data access and management, data mining analysis, and text-mining and ontology learning services. Just as the enhancements to grid middleware, the data mining interfaces are initially intended for use by the project Partners, to enable standardized data access, data mining analysis, text mining and ontology learning services on the grid. Two potential user groups can be identified:

- Software companies and vendors (IBM, Oracle, SAS, SPSS, OLAP database producers, etc.), which can integrate these interfaces to improve their products and services, in particular performance (e.g. for large data amounts) and the compatibility,
- Data-mining technology users – developers of data mining services (data analysis service companies, public facilities, etc.) can solve most urgent data-mining problems (e.g. computation power and time) by using the interfaces for distributed computation.

3.2.3 Data-mining applications

Data-mining applications benefit from the utilization of the generic interfaces as well as from the development of new state-of-the-art algorithms and tools for highly specialized end users. The Consortium agreed on the creation of a joint venture that addresses the goal of producing different application-specific portals by customizing the produced generic interfaces.

The potential user group for data-mining applications is the entire business market that has a demand for data-mining solutions.

The market can be divided in:

- The private sector, consisting of industry and service companies which are using data-mining technologies in the field of data services, telecommunication, energy, civil engineering, nutrition, medicine and biomedicine, publishing etc. and
- The public sector, including all public facilities, which are using data-mining applications such as medical departments, statistical, economic or ecologic offices or ministries and departments from security, aerospace etc.

3.3 Targeted user groups

While the potential user groups offer a general overview of which users are reachable, the targeted user groups describe in detail, which of user is to be addressed. This user group is highly dependent on the specific usage of the certain project results and shall therefore be described for the individual use case contributions by the Partners.

The following descriptions of targeted user groups are the estimated actual users and reflect the current knowledge about the user groups at this moment. This knowledge has been generated by interviewing the targeted user groups.

3.3.1 University of Ulster

The expected result of the University of Ulster’s (UU) contribution towards the DataMiningGrid project consists of the following items:

- A specification for the access of data services for the other components in DataMiningGrid project (making efficient use of existing technologies and standards),
- To implement and distribute whatever resources are required to provide viable access and pre-processing methods to support the data mining activities of the project,
- To implement two demonstrator tasks that demonstrate how the DataMiningGrid project can be used to support research activities.

The University of Ulster’s (UU) contribution to the DataMiningGrid project is initially intended for use by the other project Partners. They will implement the data access components of their software according to the specifications provided by the UU. Ultimately, the target user groups are researchers wishing to use data-mining techniques on large amounts of data within grid-computing environments. The demonstrator components developed at UU will be of interest to wider bioinformatics community. User groups associated with UU’s requirements and applications are depicted in Table 1.

Table 1. UU user groups

Identifier	Targeted User Group
------------	---------------------

Data service access specification	<ul style="list-style-type: none"> • DataMiningGrid project Partners • Researchers wishing to use data mining techniques on large amounts of data within grid computing environments
Implementation and distribution of resources	<ul style="list-style-type: none"> • DataMiningGrid project Partners
UU demonstrator tasks	<ul style="list-style-type: none"> • Researchers wishing to use data mining techniques on large amounts of data within grid computing environments • Bioinformatics Community

3.3.2 Fraunhofer Institute for Autonomous Intelligent Systems

The Fraunhofer Institute for Autonomous Intelligent Systems (FHG), as a data-mining research and consulting center, will use the technology to offer new services. FHG will provide:

- A grid-enabled version of the Weka [Weka04] toolkit that will provide several advantages compared to the standalone application such as execution on any suitable machine in the grid, without the need of installing Weka on all these machines.

Targeted users for the grid-enabled version of the Weka toolkit are researchers as well as data-mining experts in industry that have a demand for parallel, distributed data mining. Moreover, this project result will enable present and future grid-related EU-projects to integrate and enhance distributed data mining facilities.

- A collection of text-mining modules, as well as the associated knowledge how to develop solutions with these modules. The text-mining modules can be combined into complex workflows using the common workflow editor. In this way, various pre-processing, clustering, and classification steps may be chained and distributed over the grid.

Targeted users are people that administrate large text repositories, e.g., document collections, service records, customer contact records or administrative records, which require efficient text-mining systems. In addition, the ability to have the raw text databases, method repositories and compute servers at different locations is a possible advantage for users, who want to control the location of their data. User groups associated with FHG's requirements and applications are depicted in Table 2.

Table 2. FHG user groups

Identifier	Targeted User Group
Grid-enabled version of Weka	<ul style="list-style-type: none"> • Researchers with demand for parallel distributed data mining • Data mining experts in the industry that have a demand for parallel distributed data mining • Present and future grid related EU-projects

Collection of text-mining modules

- Administrators of large text repositories, who require efficient text mining systems

3.3.3 DaimlerChrysler

DaimlerChrysler (DC) has a direct need to use the applications that are developed in the scope of the DataMiningGrid project in its in-house processes related to quality management and customer relationship management, especially to compute large amounts and distributed, mostly confidential data.

At DC the targeted users are internal departments working on quality and security enhancement. Additional target user groups are all developers in the company, which should be supported by better abilities of finding related textual information.

3.3.4 Israel Institute of Technology

Israel Institute of Technology (TECH) intends to develop generic monitoring capabilities, which may then be deployed as part of Net-batch, Condor, DataGrid [DataGrid04], and other grid systems. TECH's project results consist of:

- Mechanism for rapid deployment of data-mining applications on the grid that could significantly cut down time-to-market of complex data-mining applications.

Typical users are IT personnel and executives utilizing data mining.

- Log mining tool designed to detect various failure points and allow improved configuration and administration of the grid infrastructure.

Typical users are Grid administrators. User groups associated with TECH's requirements and applications are depicted in Table 3.

Table 3. TECH user groups

Identifier	Targeted User Group
Mechanism for rapid deployment of data-mining applications	<ul style="list-style-type: none"> • IT personnel and executives utilizing data mining
Log mining tool	<ul style="list-style-type: none"> • Grid administrators

3.3.5 University of Ljubljana

The University of Ljubljana's (LJU) is a research and education institution. In this sense, the immediate result that LJU expects is knowledge that shall be exploited to improve teaching activities. Moreover, LJU expects to contribute to the delivery of specific services in three distinct areas:

- Services that could be used in connection to standard digital libraries, which could improve the ways how digital libraries are used. This

specifically refers to certain text-mining and ontology-learning services that we intend to implement.

Typical users are researchers that need to access digital libraries around the world and want to discover useful information in these libraries. The technology shall help towards automation of the whole process. It shall provide easier semantic search possibilities for the researchers.

- Services that could be used for ecosystem modeling that take into account collected data as well as previous knowledge to generate new knowledge.

Typical users are researchers and other users that need to combine mathematical models of natural phenomena together with experimental data. This approach would not be possible or at least very difficult without the DataMiningGrid technology. The computational grid resources are used to generate quality equations (i.e., models) of the natural phenomena. In the DataMiningGrid project, this shall be demonstrated in a use case of modeling of the Lake Bled ecosystem.

- Services that could be used in the medical field, when conducting research studies on greater geographic areas. In this case, the developed data services shall take care about the security, privacy, and governance of the investigated data, therefore enabling on the fly medical analysis of much larger and distributed databases than it has been possible in the past.

Typical users are researchers in medical centers and health monitoring authorities are the main users of the technology. The technology will enable them to access geographically distributed medical databases and conduct 'on-the-fly' studies. In the DataMiningGrid project, the demonstration is based on a use case dealing with iodine deficiency of children entering high school where data are collected in several regional databases. User groups associated with TECH's requirements and applications are depicted in Table 4.

Table 4. LJU user groups

Identifier	Targeted User Group
Standard Digital Libraries	<ul style="list-style-type: none"> • Researchers that need to access digital libraries
Ecosystem modeling	<ul style="list-style-type: none"> • Researchers and other users that need to combine mathematical models of natural phenomena together with experimental data
Process medical data in distributed databases	<ul style="list-style-type: none"> • Researchers in medical centers and health monitoring authorities

4 Controlling of User Groups

While efforts have been made to precisely predict user groups, there is no guarantee, that these user groups will accept the results of the DataMiningGrid project and ultimately profit of the applications. It should be prevented that the targeted user groups do not turn into actual user groups after the end of the DataMiningGrid project. Since it is essential for the success of the project that the targeted users are accurately estimated, controlling mechanisms have to be established.

4.1 Controlling actions

4.1.1 Meetings

An effective measure to avoid unforeseen outcomes is to schedule meetings with representatives of the target user group and representatives of the relevant project Partners on a regular basis. Starting after month 6 these meeting are planned to happen every three months. Future versions of the Description of User Groups document [UserGroups05(2)] will list all held meetings in a table structure.

4.1.2 Surveys

To better find out about user needs, as well as to adjust the range of the targeted user groups it is planned to arrange surveys. The next version of the Description of User Groups document will list all surveys in a table structure.

4.2 Controlling consequences

Basically there are two ways how to deal with problems that arise on the inadequate definition of targeted user groups.

- Adapt project results to the demand of the targeted user group,
- Redefine the targeted user groups and apply these changes to other relevant project aspects, e.g., dissemination, exploitation etc.

4.2.1 Result adaptation

The first possibility is to change the progress in work and attend to the user group demand. This means that possibly new unforeseen changes to software, (i.e. grid enhancements, grid data-mining APIs or data-mining algorithms and applications) have to be applied. In most cases, this means that software developments have to be adjusted, in other cases this can even mean to abandon a development without substitution. Nevertheless, these changes in requirements will be comprehensible justified and accurately documented. The next version of the Description of User Groups document will list all applied changes in a table structure.

4.2.2 Redefinition of targeted user groups

A redefinition of the targeted user groups is an intervention in the project, which can possibly jeopardize (commercial) success. Therefore, it should be thoughtfully done and only if there are convincing reasons to do so. Furthermore, the redefinition of target users will affect the dissemination, exploitation and the collaboration in the DataMiningGrid project. Changes to targeted user groups will be listed in a table structure in the next version of the Description of User Groups document.

5 Conclusions and Future Work

Deriving from the general estimation of potential user groups for the results of the project, it is necessary to identify and define target user groups, which are forecasted to be the actual users. There have been efforts made to identify target user groups on the basis of use cases. While the project is still in an early stage, these initial descriptions of user groups are somewhat vague and need to be more refined in the future. Hence, it is planned to provide a more detailed targeted user groups description in the upcoming version of the Description of User Groups deliverable. Furthermore, controlling mechanisms have to be applied to ensure that user groups will be addressed by the project results. Controlling actions and consequences – starting with Month 6 – will be carried out and reported in the next version of this document.

6 References

- [Annex04] DataMiningGrid Annex I – ‘Description of Work’
- [Collab04] Deliverable D72 Final Collaboration Plan
- [Condor04] The Condor Project at <http://www.cs.wisc.edu/condor/>
- [DataGrid04] The DataGrid Project at <http://eu-datagrid.web.cern.ch/eu-datagrid/>
- [Dissemination05] Deliverable D81(1) Dissemination Plan
- [Exploitation05] Deliverable D83(1) Exploitation Plan
- [Globus04] The Globus Alliance at <http://www.globus.org/>
- [UserGroups05(2)] Deliverable D82(2) Description of User Groups
- [Weka04] Weka at <http://www.cs.waikato.ac.nz/ml/weka/>