INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

OpenMolGRID

SPECIFICATION OF THE DATABASE ACCESS TOOL FOR THE OPENMOLGRID DATA WAREHOUSE

Contract Reference:	**IST-2001-37238**
Document identifier:	**OpenMolGRID-1-D1.1f-0106-2-1-DBATMOLDW**
Date:	**08/01/2004**
Work package:	**WP1: Grid Data Warehousing of Molecular Structure – Property (Activity) Information**
Partner:	**UU, FZJ**
Lead Partner:	**UU**
Document status:	**APPROVED**
Classification:	**PUBLIC**
Deliverable identifier:	**D1.1f**

Abstract: This document is to describe the input and output formats for the OpenMolGRID data warehouse (MOLDW).

Delivery Slip

	Name	Partner	Date
From	Damian McCourt	UU	15/09/03
Verified by	WPM	All	10/10/03
Approved by	G.H.F.Diercksen (TC) R.Ferenczi (QE)	OMC CGX	22/10/03 25/10/03

Document Log

Issue	Date	Comment	Author
0-0	05/09/03	Initial version – extracted from D1.1	Damian McCourt
1-0	10/09/03	Submitted To WPM	Damian McCourt
2-0	15/09/03	Submitted for Authorisation	Damian McCourt
2-1	08/01/04	Updated due to the change of the document template (version1.3)	Jean Jing

Document Change Record

Issue	Item	Reason for Change
2-0	Document Status	Document Approved
2-1	Document Template Change	The standard template of the document is changed

Files

Files in this section relate to actual storage locations on the BSCW server located at https://hermes.chem.ut.ee/bscw/bscw.cgi. The URL below describes the location on BSCW from the root OpenMolGRID directory

Software Products	User files / URL
Word 2000/XP	OpenMolGRID/Workpackage 1Deliverables/ OpenMolGRID-1-D1.1f-0106-2-1-DBATMOLDW

Project information

Project acronym:	OpenMolGRID
Project full title:	Open Computing GRID for Molecular Science and Engineering
Proposal/Contract no.:	IST-2001-37238
European Commission:	
Project Officer:	Annalisa BOGLIOLO
Address:	European Commission - DG Information Society F2 - Grids for Complex Problem Solving B-1049 Brussels Belgium
Office	BU31 4/79
Phone:	+32 2 295 8131
Fax:	+32 2 299 1749
E-mail	annalisa.bogliolo@cec.eu.int
Project Coordinator:	Mathilde ROMBERG
Address:	Forschungszentrum Jülich GmbH ZAM D-52425 Jülich Germany
Phone:	+49 2461 61 3703
Fax:	+49 2461 61 6656
E-mail	m.romberg@fz-juelich.de

Contents

1. Introduction

1.1. Purpose and Scope

The purpose of this document is to describe the input and output formats for the OpenMolGRID data warehouse database access tool, referred to as DBAT_MOLDW throughout the remainder of this document. The OpenMolGRID data warehouse will be referred to as MOLDW throughout the remainder of the document. The details contained in this document are complimentary to the specification of the general approach to database access described in D4.1a [1] and D4.1 c [2].

1.2. Overview

In order to enable the use of MOLDW it is essential that the access mechanism is defined. This document describes access and describes the expected inputs and outputs of MOLDW.

1.3. Document Structure

In addition to this section the document contains the following sections:

- Section 2 – A general description of access to MOLDW
- Section 3 – A description of the Input expected by MOLDW
- Section 4 – A description of the expected output from MOLDW

1.4. Terminology

UT	University of Tartu, Estonia
UU	University of Ulster, UK
FZJ	Research Centre Juelich, Germany
Negri	Mario Negri Institute, Italy
CGX	ComGenex, Hungary
MOLDW	OpenMolGRID Data Warehouse
UNICORE	Uniform Interface to Computer Resources
XML	Extensible Markup Language
SQL	Structured Query Language
DBAT	Database Access Tool
DBAT_MOLDW	The DBAT for MOLDW

2. Access to MOLDW

Access to MOLDW will be realised as generic database access with the development of a tool called DBAT_MOLDW. This tool will be used to provide external access to MOLDW. It will have a command line interface. The tool will have the following syntax:

```
dbat_moldw infile outfile
```

In this sample `infile` relates to the query being sent to MOLDW and this is specified in section 3 of this document. `Outfile` relates to the response from MOLDW and is specified in section 4 of this document. UNICORE will initiate the execution of this tool using the syntax outlined above. Generic database access via UNICORE is described in deliverable D4.1c [2].

3. Query Input Specification

This section of the document will describe the input expected by MOLDW. This format is general and will be the same for all access mechanisms including UNICORE, which is the only one currently envisaged.

3.1. General Information

Input to the OMGDW will be an XML file containing one or more queries and user authentication. Currently no further details are required.

3.2. Sample

The format for the input specification based on the information described above is listed in **Figure 1**.

```
<?xml version="1.0"?>
<input>
     <query>SELECT * FROM toxicity WHERE OMG_CAS=12345</query>
     <username>SampleUserName</username>
     <password>SomePassword</password>
</input>
```

Figure 1: Sample Input for MOLDW

This sample assumes the existence of a table called "toxicity" that contains a field "OMG_CAS".

3.3. Tag Description

Each tag that exists in the sample input is described in **Table 1**, except the `<?xml version="1.0"?>` tag which must appear in every XML document.

Table 1: A Description of Input Tags

Tag	Description
input	The root element. Encapsulates the user request.
query	Contains information about the query. This will be in SQL format. More than one query element can be specified.
username	The name of the user carrying out request.
password	The password of the user.

3.4. Notes

There are some important points to note about this format:

- When username and password are not specified, the database access tool tries to log on anonymously.
- Special characters in the query and/or username and password have to be encoded properly, since an XML parser is used to read this file. For example, the character "<" has to be encoded as "<"
- Access to metadata will occur in the same way as access to operational data. Metadata can therefore be queried.
- Currently it is not envisaged that MOLDW will require password authentication, as MOLDW will trust UNICORE to ensure secure access.

4. Query Output Specification

This section will describe the output provided by MOLDW. This format is general and will be the same for each request received.

4.1. General Information

Output from MOLDW will be written in XML. When a query is executed, the result can be visualized as a table. This will be written row by row to the output file and will be accompanied by the necessary information to reconstruct the table in the form of a document type definition (DTD). The DTD for DBAT_MOLDW is the same as that for general database access as described in D4.1c [2], but is inserted here to help describe specific MOLDW output.

4.2. Document Type Definition

```
<!ELEMENT info (#PCDATA)>
<!ELEMENT status (#PCDATA)>
<!ELEMENT no_of_columns (#PCDATA)>
<!ELEMENT no (#PCDATA)>
<!ELEMENT label (#PCDATA)>
<!ELEMENT typeSQL (#PCDATA)>
<!ELEMENT typeJDBC (#PCDATA)>
<!ELEMENT typeOMG (#PCDATA)>
<!ELEMENT column (no,label,typeOMG?,typeSQL?,typeJDBC?)>
<!ELEMENT column_info (no_of_columns,column*)>
<!ELEMENT value (#PCDATA)>
<!ELEMENT row (value+)>
<!ELEMENT results (column_info,row*)>
<!ELEMENT dbat_output (info?,status?,results)>
```

Figure 2: Document Type Definition for MOLDW output

4.3. Sample

This section will provide an example of the output that can be expected from MOLDW. The sample is based on **Table 2**.

Table 2: Sample MOLDW output viewed as a table

OMG_CAS	chemical_name	mol_weight
302170	chloral hydrate	149.404
312856	sodium lactate	112.061
357573	brucine	394.469

NB! This table was generated using information from websites [2] and [3].

Given that 'OMG_CAS' is an integer, 'chemical_name' is a string and 'mol_weight' is a real number, the result file shown in **Figure 3** would be produced.

```
<?xml version=”1.0”?>
<dbat_output>
     <status_info>request successful
     </status_info>
     <status>0
     </status>
     <results>
          <column_info>
               <no_of_cols>3</no_of_cols>
               <column>
                    <col_no>1</col_no>
                    <label>OMG_CAS</label>
                    <typeOMG >LONG</typeOMG>
                    <typeSQL>LONG</typeSQL>
                    <typeJDBC>4</typeJDBC>
               </column>
               <column>
                    <col_no>2</col_no>
                    <label>chemical_name</label>
                    <typeOMG>VARCHAR</typeOMG>
                    <typeSQL>VARCHAR</typeSQL>
                    <typeJDBC>12</typeJDBC>
               </column>
               <column>
                    <col_no>3</col_no>
                    <label>mol_weight</label>
                    <typeOMG>FLOAT</typeOMG>
                    <typeSQL>FLOAT</typeSQL>
                    <typeJDBC>6</typeJDBC>
               </column>
          </column_info>
          <row>
               <value>302170</value>
               <value>chloral hydrate</value>
               <value>149.404</value>
          </row>
          <row>
               <value>312856</value>
               <value>sodium lactate</value>
               <value>112.061</value>
          </row>
          <row>
               <value>357573</value>
               <value>brucine</value>
               <value>394.469</value>
          </row>
     </results>
</dbat_output>
```

Figure 3: Sample MOLDW Output

4.4. Tag Description

Table 3 describes each tag contained in the sample output except the `<?xml version="1.0"?>` tag which must appear in every XML document.

Table 3: Tag Description for MOLDW Output

Tag	Description
dbat_output	The root element. Encapsulates the response to the user request.
Status	This tag is used to indicate the status of the request. It will be a system-generated number whose possible values must be defined.
status_info	This tag is used to give a textual description of the status of the request. It is intended that the user easily understand this information.
results	This is a wrapper tag that encapsulates all information relating to the result of a user query.
column_info	This is a wrapper tag that encapsulates all information associated with the output "table"
no_of_cols	This relates to the number of columns contained in the output "table"
column	This is a wrapper class for the information associated with one column. This is metadata. More than one column tag can appear in the XML output
col_no	This is the associated column number. It is an element contained within the parent element column.
label	This is the text that identifies the name or label of the column in the data warehouse. It is an element contained within the parent element column.
typeOMG	This indicates the type of data stored in this column in relation to its representation within OpenMolGRID. This allows user defined types to be used within the project. It is an element contained within the parent element column.
typeSQL	This indicates the type of data stored in this column within the database. It is an element contained within the parent element column.
typeJDBC	This indicates the type of data stored in the column in relation to how JDBC can read this data. JDBC types are integers relating to actual types. It is an element contained within the parent element column.
row	This is the wrapper for a row in the output table. This is where the "real" data appears. There are multiple row tags within an output file.
value	The value of each column for a particular row is contained within this tag. The number of value tags within a row is dependent on the number of columns in a row.

4.5. Notes

There are some important notes about this format:

- The output XML format will not take advantage of XML concepts such as schemas.
- All information (operational data and metadata) required to reconstruct the result from the XML file is included, such as column type information by using the DTD.
- Information relating to the success or failure of the query will be available as status information.
- Special characters in the response have to be encoded properly, since an XML parser is will be used to read the output. For example, the character "<" has to be encoded as "<"

Term/Abbreviation	Meaning
UT	University of Tartu, Estonia
UU	University of Ulster, UK
FZJ	Research Centre Juelich, Germany
Negri	Mario Negri Institute, Italy
CGX	ComGenex, Hungary
MOLDW	OpenMolGRID Data Warehouse
UNICORE	Uniform Interface to Computer Resources
XML	Extensible Markup Language
SQL	Structured Query Language
DBAT	Database Access Tool
DBAT_MOLDW	The DBAT for MOLDW

5. References

[1] M. Romberg and B. Schuller, "Specification of the generic user interface for database access," https://hermes.chem.ut.ee/bscw/bscw.cgi/d6202/OpenMolGRID-4-D4.1a-0101-2-0, 15/09/03.

[2] B. Schuller and M. Romberg, "Specification of Database Access Interface," https://hermes.chem.ut.ee/bscw/bscw.cgi/d6357/OpenMolGRID-4-D4.1c-0103-2-0, 15/09/03.