INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

OpenMolGRID

Properties and priorities of the data for pharmaceutical and phytopharmaceutical compounds

Document identifier:	**OpenMolGRID-1-D1.3-0110-2-1-DataSpecification**
Date:	**18/09/2003**
Work package:	**WP1: Grid Data Warehousing of Molecular Structure – Property (Activity) Information**
Partner	**NEGRI**
Lead Partner	**UU**
Document status:	**APPROVED**
Deliverable identifier:	**D1.3**

Abstract: A general description of the database characteristic, abstract data type and metadata implementation for pharmaceutical and phytopharmaceutical compounds.

Delivery Slip

	Name	Partner	Date
From	Mose' Casalegno	NEGRI	18/09/2003
Verified by	WPM's	All	
Approved by	G.H.F Diercksen (TC) R. Ferenczi (QE)	OMC CGX	10/10/2003 07/04/2004

Document Log

Issue	Date	Comment	Author
1-0	13/09/2003	New section 4 added	Mose' Casalegno
2-0	18/09/2003	Templates updated	Mose' Casalegno, Damian McCourt

Document Change Log

Issue	Item	Reason for Change
	none	

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 1/Deliverables/ OpenMolGRID-1-D1.3-0110-2-1

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:	Franco ACCORDINO
Address:	Rue de la Loi, 200 (Office: Av. de Beaulieu 29, BU31 4/16) B-1160 Bruxelles – BELGIQUE
Phone:	+32/2/299 82 72
Fax:	+32/2/299 17 49
E-mail	franco.accordino@cec.eu.int
Project Coordinator:	Professor Dr Mati KARELSON
Address:	Department of Chemistry University of Tartu Jakobi Street 2 EE 2400 **Tartu** - ESTONIA
Phone:	+372 7 375 255
Fax:	+372 7 375 264
Mobil:	+372 5 021519
E-mail	mati@chem.ut.ee

Contents

1. Introduction

This document describes properties and specifications of the data for pharmaceutical and phytopharmaceutical compounds for the development of reliable QSPR/QSAR model. The aim of this document is to give overview of characteristics of data and model parameters applied in the development of reliable ecotoxic QSARs. The description of data types, available for the integration from various data sources, is presented in section 3.

2. Database characteristics

2.1. Overview

Data used in formulating QSARs should be reliable, of high quality, and reflect a well-defined and continuous toxic endpoint. Data for pharmaceutical compounds will be generated within OpenMolGrid, by the same laboratory, which is by far most preferable than using data obtained by different laboratories in different experimental conditions. Also in terms of ecotoxicity, such data should ideally be measured by a single protocol, even in the same laboratory and by the same workers. High-quality toxicity data will have lower experimental error associated with them. Such toxicity data typically come from standardised assays measured in a consistent manner with a clear and unambiguous endpoint. With ecotoxicity data, the quality is associated with the toxicity values that are accurate, consistent with other data for the same endpoint, and consistent with data for other endpoints. In the latter case, it is as important for data to be consistent between different endpoints as for the inconsistencies to be consistent. This latter issue is often overlooked and reliable data should demonstrate that if there is an inconsistency between potency for a given chemical, the data for other compounds of matching molecular structure should demonstrate an identical inconsistency.

Fortunately, many toxicity data are of high quality. However, there are also lower quality public data that may still be useful for the QSAR development but for which the resultant models will be less reliable. This fact is more apparent when data from a variety of sources and protocols are used in concert (i.e., the use of literature compilations or commercial databases).

In summary, essential or desirable features of the development of robust QSARs include using ecotoxicological data that are reliable, of superior quality, and signal a well-defined and continuous endpoint.

2.2. Database Description

This section presents public (freely available) databases, which meet the above listed requirements.

ECOTOX

Name of the database, owner, identification

The ECOTOXicology database (ECOTOX) is a source for locating single chemical toxicity data for aquatic life, terrestrial plants and wildlife of phytopharmaceutical and other pollutants. ECOTOX is a useful tool for examining impacts of chemicals on the environment. Peer-reviewed literature is the primary source of information encoded in the database. Pertinent information on the species, chemical, test methods, and results presented by the author(s) are abstracted and entered into the database. Another source of the test results is independently compiled data files provided by various United States and international government agencies. ECOTOX was created and is maintained by the U.S. EPA, Office of Research and Development (ORD), and the National Health and Environmental Effects Research Laboratory's (NHEERL's) Mid-Continent Ecology Division.

More information on the ECOTOX database is available at:

ECOTOX Support Staff

U.S. Environmental Protection Agency

Office of Research and Development

National Health and Environmental Effects Research Laboratory

Mid-Continent Ecology Division (MED)

6201 Congdon Boulevard

Duluth, Minnesota 55804

Telephone: 218-529-5225

Fax: 218-529-5003

E-mail: ecotox.support@epa.gov

Background information

The development of the ECOTOX was started in 1995, and in March of 1996, it was released to governmental users through a telnet access procedure. In February, 2000, ECOTOX was released as a web based interface system. The ECOTOXicology database (ECOTOX) is a source for locating the toxicity data for single chemicals from three U.S. Environmental Protection Agency (U.S. EPA) ecological effects databases; AQUIRE, TERRETOX, and PHYTOTOX. Aquatic data in AQUIRE are limited to test organisms that are exclusively aquatic (saltwater and freshwater). Species that are associated with the water but do not have gills, such as ducks and geese, are included in the terrestrial database. Amphibians are included in both AQUIRE and TERRETOX databases, with the life stages that exist exclusively in the water (e.g., tadpole) located in AQUIRE and the terrestrial life-stage (e.g., adult) in TERRETOX. Bacteria and viruses are not included in the ECOTOX database. TERRETOX is the terrestrial animal database. Its primary focus is wildlife species but when data gaps exist for a particular chemical, data for domestic species are included. PHYTOTOX is a terrestrial plant database. In the development of PHYTOTOX, the tests represent mostly agricultural chemicals and predominantly agricultural species.

Characteristics regarding software

The ECOTOX database can be accessed by using a web browser software via the Internet at http://www.epa/gov/ecotox. The more detailed information regarding field data definitions can be obtained from the coding guidelines for the aquatic database (Aquatic Coding Guidelines) and the terrestrial database (Terrestrial Coding Guidelines). ECOTOX (ECOTOXicology Database System) is a comprehensive computer-based system that provides single chemical toxic effect data for aquatic life, terrestrial plants, and terrestrial wildlife.

Characteristics regarding data coverage

Database searches can be conducted using either a Quick Query or an Advanced Query menu. The Quick Query supports searches on habitat, taxonomic kingdom, species common or Scientific name, Chemical Abstract Service Registry number, chemical name, observed effect group and publication year. The Advanced Query menu includes all options under Quick Query, and enables you to focus on more specific criteria such as study site type (e.g., laboratory, field), exposure media (e.g., freshwater, soil), route of chemical exposure (e.g., oral, diet), and statistically-derived endpoints (e.g., LD50, NOEL). The search results can be downloaded either as an ASCII delimited file format, which can be transferred into a database or spreadsheet, or into a browser-viewable report format.

Basics of the database

The resources of the data come from several databases:

- ECOTOX
- AQUIRE
- TERRETOX
- PHYTOTOX

The ECOTOXicology database (ECOTOX) is a source for locating single chemical toxicity data from three U.S. Environmental Protection Agency (U.S. EPA) ecological effects databases; AQUIRE, TERRETOX, and PHYTOTOX. The AQUIRE and TERRETOX databases contain information on lethal, sublethal and residue effects. The AQUIRE database includes toxic effects data on all aquatic species including plants and animals and freshwater and saltwater species. TERRETOX is the terrestrial animal database. It's primary focus is wildlife species but the database does include information on domestic species. PHYTOTOX is a terrestrial plant database that includes lethal and sublethal toxic effects data.

Characteristics regarding output

Data obtained from compiled data files must meet the minimum data requirements and quality assurance guidelines defined for each ECOTOX database component. The key data fields that must be included are: test chemical name, test organism, test duration, effect, and effect concentration or application rate. Documentation describing the test methods must be provided within the publication. If tests are missing key parameters, the data are rejected. No effort is made to locate unreported data (e.g., authors are not contacted, citations referring to methods used are not obtained). During the incorporation of an electronic data file, a quality assurance check of the CAS number, species scientific name, and reference citation is carried out.

Planned or expected further development

Plans exist to extend on data coverage.

NTP

Name of the database, owner, identification

Federal and State Regulatory Agencies use the NTP study data in considering the need for regulation of specific chemicals to protect human health. The National Toxicology Program (NTP) coordinates toxicological testing programs within the Department of Health and Human Services (DHHS), strengthen the science base in toxicology: It develops and validates improved testing methods and provides information about potentially toxic chemicals to health regulatory and research agencies, the scientific and medical communities, and the public.

More information on the NTP database is available at:

NTP Webmaster E-mail: ntpwm@niehs.nih.gov

Central Data Management (CDM)
P.O. Box 12233, MD EC-03
Research Triangle Park, NC 27709
Telephone: (919) 541-3419
E-mail: cdm@niehs.nih.gov

NTP Liaison and Scientific Review Office
P.O. Box 12233, MD A3-01
Research Triangle Park, NC 27709
Telephone: (919) 541-0530
E-mail: liaison@starbase.niehs.nih.gov

Background information

More than 80,000 chemicals are registered for use in commerce in the United States, and an estimated 2,000 new compounds are introduced annually for use in everyday items such as foods, personal care products, prescription drugs, household cleaners, and lawn care products. The effects of many of these chemicals on human health are unknown, yet people and our environment may be exposed to them during the manufacture, distribution, use, and disposal or as pollutants in our air, water, or soil. Although relatively few chemicals are thought to pose a significant risk to human health, the safeguarding of public health depends on identifying the effects of these chemicals and the levels of exposure at which they may become hazardous to humans. The National Toxicology Program (NTP) has been established in 1978 by the Department of Health and Human Services (DHHS) to coordinate toxicological testing programs within the Department. The NTP is an interagency program consisting of relevant toxicology activities of the National Institutes of Health's National Institute of Environmental Health Sciences (**NIH/NIEHS**), the Centers for Disease Control and Prevention's National Institute for Occupational Safety and Health (CDC/ NIOSH), and the Food and Drug Administration's National Center for Toxicological Research (FDA/NCTR). The NIH's National Cancer Institute (**NIH/NCI**) was a charter agency; however, the NCI Carcinogenesis Bioassay

Program was transferred to the NIEHS in 1981. The NCI remains active in the Program through membership on the NTP Executive Committee. NTP's mission is to evaluate agents of public health concern by developing and applying tools of modern toxicology and molecular biology.

Characteristics regarding software

The NTP database is accessible using web browser software via the Internet at http://ntp-server.niehs.nih.gov/.

Basics of the database

The National Institute of Environmental Health (NIEHS) Sciences Division of Extramural Research and Training (DERT) in collaboration with the National Toxicology Program (NTP) has initiated a new grants program which funds investigator-initiated research to provide data to aid in defining the mechanism of action of agents under study by the NTP. This new program uses the NIH R03 Small Grant mechanism that encourages investigator-initiated hypothesis-driven investigations on animals/tissues/cells from animals undergoing the NTP 2-year cancer bioassay or shorter toxicological characterisations. The National Toxicology Program (NTP) conducts toxicity/carcinogenesis studies on agents suspected of posing hazards to human health. Chemical-related study information is submitted to NIEHS and is archived and maintained in a central location (Central Files) so that all study information can be monitored and tracked efficiently. Currently, more than 800 chemical studies are on file. NTP Information is routinely provided to industry and the public on an as requested basis.

Characteristics regarding data coverage

Currently, the data coverage of the NTP includes mostly the compound identification (CAS number, chemical formula, etc.), physico-chemical properties (solubilities, volatility, stability, etc.), toxicity data (LD50 on several species, carcinogenicity, mutation data, teratogenicity, etc.) and other data.

Planned or expected further development

The NTP plans extend the data coverage and web-enabled source facilities. Moreover, several annual meeting on specific subjects (Endocrine Disruptors, Phathalates, etc.) will be organised.

2.3. Pharmaceutical

Some typical databases of the toxicity of chemicals have been discussed above. Many similar databases exist for other chemicals and endpoints. However, typically they contain less data for a series of reasons. For instance, data collection on chronic toxicity data are less abundant, due to the higher costs and longer time needed to get this data.

The case of pharmacological data is quite peculiar. The economic value of this information means that in most of the cases the data are private. Pharmaceutical companies spend large amount of money to build up huge libraries of chemicals and biological/pharmacological properties. This information is not publicly available as it is the source of a successive screening to develop potential new drugs. Anyhow, the structure of these databases is conceptually very similar to those we already discussed, even though they can be much simpler and contain less details for a single experiment.

A specific pharmaceutical database is available at: http://www.medscape.com/px/urlinfo. This database contains the toxicity data of drugs. Within OMG we will use data achieved by the planned experiments. The data format for QSAR models will be the same chosen for ecotoxicity.

3. Abstract Data types and Metadata

The integration of data from various sources requires a level of abstraction. Categories have to be identified in which the data from the different sources fit into. In technical terms, a data type needs to be defined so that it can be used to integrate several data sources. Each component of the data type will be described. This will provide the first level of metadata.

3.1. Identification of Chemicals

Chemicals can be identified in a number of ways. Having analysed the way in which users currently search for information, CAS number is found to be very important. For this abstract data type, the CAS will be used as the unique identifier. However, users often carry out search based on the chemical name. It is therefore important to store a list of names by which a chemical is known. All chemicals also have an associated chemical formula. This describes the composition of the chemicals in terms of the number and the type of atoms. Such information is readily available. The molecular weight is another piece of information that is often available. Given this information it is possible to relate this to a given CAS number.

Name	Type	Description
CAS (unique identifier)	Integer	The Chemical Abstract Service (CAS) number. This is an integer without hyphens or spaces
Molecular Weight	Float	The Atomic Mass Unit (amu) associated with the CAS.
Chemical Formula	String	The chemical formula associated with the CAS
Chemical Names	List of Strings	The names by which the chemical is known

3.2. Toxicity

Toxicity always relates to some chemical. Since each chemical is associated with a CAS number, its toxicity can be also associated with this CAS number. There are many pieces of information associated with toxicity. However, not all of these are always available and they do not necessarily portray important information. After consultation with daily users of toxicity data, the following information was shown to be important:

Name	Type	Description
CAS (unique identifier)	Integer	The Chemical Abstract Service (CAS) number. This is an integer without hyphens or spaces
Target Species Names	List of Strings	A list of names associated with the target species
End Point Type	String	The type of toxicity measure used in this protocol e.g. LC_50, LD_50.
Dose Metric	Float	The dose of chemical studied (for aquatic toxicity the compound is in the tank, and the fish exposed)measured in milligrams per kilogram (mg/kg) or milligrams per litre (mg/l).
Dose Metric Units	String	The units associated with the Metric dose
Dose Mol	Float	The dose of chemical studied measured in millimols per kilogram (mmol/kg) or millimols per litre (mmol/l).

Dose Mol Units	String	The units associated with the Mol dose
Exposure Time	Float	The amount of time the target species was exposed to the chemical. This is measured in hours.
Protocol Details	Text	General information associated with the protocol. There is no particular format associated with this.
Mode of Action USA	String	The Mode of Action of the chemical. These are classes based on the EPA (Duluth, USA) standard. These are as follows: • Non Polar Narcosis (Base Line Narcosis) • Polar Narcosis (Narcosis II) • Narcosis III (Ester/ Acrylate compounds) • Oxidative Phosphorilation uncoupling • Respiratory inhibition • Electrophile and proelectrophile reactivity • Acetilcholinesterase inhibition • Central nervous system seizure responses
Mode of Action EU	String	The mechanism of action of the chemical. These are classes based on the EU (Netherlands) standard. These classes are as follows: • Non Polar Narcosis • Polar Narcosis • Reactive • Receptor Mediated
Author	String	The author who reported the toxicity measure.
Year	Integer	The year in which the author reported the toxicity measure.
Database	String	The database from which the toxicity measure was obtained.

3.3. Carcinogenicity

Carcinogenicity is a special case of toxicity. It is concerned mainly with whether or not a chemical is carcinogenic to humans. Carcinogenicity always relates to a CAS number.

Name	Type	Description
CAS (unique identifier)	Integer	The Chemical Abstract Service (CAS) number. This is an integer without hyphens or spaces

Carcinogenicity	String	Classification of carcinogenicity according to classes proposed by the World Health Organisation (International Agency for Research on Cancer-IARC). These classes are as follows: 1 – Carcinogenic to human 2a – Probably Carcinogenic to human 2b – Possibly Carcinogenic to human 3 – Unknown 4 – Non-carcinogenic to human

3.4. Structure

All chemicals have an associated molecular structure. There are various ways in which this structure can be represented. It is not possible to store all of these representations. It is useful however to be able to store the most common representations, if available. These specifications are listed as follows:

Name	Type	Description
CAS (unique identifier)	Integer	The Chemical Abstract Service (CAS) number. This is an integer without hyphens or spaces
SMILES	String	A particular format for storing structure. (Simplified Molecular Input Line Entry System)
2D Molfile	Text	A text file representing the 2D structure of the chemical.
3D	Text	The 3D structure associated with the CAS. The realisation of this is unknown at present.

3.5. Chemical Descriptors

Chemical descriptors are also associated with a CAS number. The type of descriptors generated for a given CAS is largely dependent on the process by which they were created. This means that the values produced are experimental and require updating on a regular basis. As the data warehouse is a read-only information repository, chemical descriptors must be stored outside the data warehouse. They can, however, be easily integrated as they are associated with a CAS number.

Name	Type	Description
CAS (unique identifier)	Integer	The Chemical Abstract Service (CAS) number. This is an integer without hyphens or spaces
DES_1	Float	A descriptor associated with the CAS
DES_2	Float	A descriptor associated with the CAS
…		
DES_n	Float	A descriptor associated with the CAS

3.6. ADME Related Properties and Descriptors

The pharmaceutical and biotech industry together with their investors have long accepted the fact that from every 6 chemical entities that enter human clinical trials only one emerges as a marketable drug product. This ratio of attrition carries a significant financial burden on the industry because the average cost of bringing drugs to the market is estimated to be between $500-800 million. More importantly,

about 75 % of this cost is spent during the development of failed drug candidates. A statistical analysis of regulatory agencies database indicates that about 28 % of the failed drug candidates are dropped for lack of efficacy, 5 % for market reasons, 20 % due to excessive toxicity and 37 % for poor pharmacokinetic properties such as adsorption, distribution metabolism and excretion (ADME). It is therefore obvious that the reduction of the rate of the failed drug candidates could result in substantial savings in the development cost of drugs, because the leads terminated late in the development process divert hundreds of millions of dollars from commercially more viable compounds. The challenge now is to fail fast and eliminate the high-risk compounds early in the process while selecting those that are most likely to survive the rigorous development process. Advances in technology are making it possible to integrate information into the early drug discovery process.

The physicochemical properties characterize the compounds in absorption, distribution and partly in excretion (ADE). During the last two decades some of these properties were believed to be the most important in ADE, like acidity/basicity, lipophilicity, hydrogen bond donor/acceptor capabilities and molecular size-related properties.

There are different parameters for the representation of these properties, like pK_a, log$P$ and log$D$, hydrogen bond donor count (HBA), hydrogen bond acceptor count (HBA), polar surface area (PSA) and molar refractivity (MR). All of these parameters are to model the behavior of the compound in solutions and in crossing different barriers. Traditionally, most of these parameters have been measured, but due to the increased capacity of chemical synthesis the prediction tools have become more important.

Comparing the properties of known drugs and the results of the screens at Pfizer, Lipinski et al. realized that the number of the hits was increasing with the lipophilicity and the molecular weight of the compounds, which – above a certain limit – causes reduced intestinal absorption. They also found that the number of OH and NH groups (possible hydrogen bond donor groups), as well as N and O atoms (possible hydrogen bond acceptor groups) are increased in Pfizer compounds, and finally decided to set up rules that help to filter out molecules that possibly have bad absorption behavior. According to these rules, a compound tends to be poorly absorbed if at least two of the following four criteria are true:

- the molecule contains more than 5 hydrogen-bond donors (defined as NH or OH groups)
- the molecule contains more than 10 hydrogen-bond acceptors (defined as O or N atoms, including those forming part of hydrogen-bond donors)
- the molecular weight is greater than 500
- ClogP > 5.0 or MlogP > 4.15

This is the so-called Rule-of-Five, where the fifth rule is an exception:

- Compounds that are substrates for biological transporters are exception

There are additional descriptors that can be used together with Rule-of –Five:

- Number of fused rings: < 5 (i.e. capability of intercalation into DNA molecules)
- Number of Rotatable bonds: < 8 (i.e. increased conformational flexibility)

It is important that if the ADME related descriptors and properties (listed in the Table below) are available, they will be incorporated into the Data Repository, so the Data Warehouse shall be able to deal with handling these properties. These are not requisite parameters for the development of new compounds, but based on them, different rules (e.g. Rule-of-Five) can be applied for selection, the drug candidates can be scored by the system, and elaborated structures can be developed. The introduction of these parameters will increase the scientific value and appreciation of the OMG system, and will contribute to the easier exploitation of the system in the pharmaceutical industry.

Name	Type	Description
CAS (unique identifier)	Integer	The Chemical Abstract Service (CAS) number. This is an integer without hyphens or spaces
1st acidic pK_a	Float	Strongest acidic dissociation constant
2nd acidic pK_a	Float	2nd strongest acidic dissociation constant
3rd acidic pK_a	Float	3rd strongest acidic dissociation constant
4th acidic pK_a	Float	4th strongest acidic dissociation constant
1st basic pK_a	Float	Strongest basic dissociation constant
2nd basic pK_a	Float	2nd strongest basic dissociation constant
3rd basic pK_a	Float	3rd strongest basic dissociation constant
4th basic pK_a	Float	4th strongest basic dissociation constant
Log*P*	Float	Octanol/water partition coefficient of neutral form
Log*D at pH=0*	Float	Octanol/water distribution coefficient at pH=0
Log*D at pH=1*	Float	Octanol/water distribution coefficient at pH=1
Log*D at pH=2*	Float	Octanol/water distribution coefficient at pH=2
Log*D at pH=3*	Float	Octanol/water distribution coefficient at pH=3
Log*D at pH=4*	Float	Octanol/water distribution coefficient at pH=4
Log*D at pH=5*	Float	Octanol/water distribution coefficient at pH=5
Log*D at pH=6*	Float	Octanol/water distribution coefficient at pH=6
Log*D at pH=7*	Float	Octanol/water distribution coefficient at pH=7
Log*D at pH=7.4*	Float	Octanol/water distribution coefficient at pH=7.4
Log*D at pH=8*	Float	Octanol/water distribution coefficient at pH=8
Log*D at pH=9*	Float	Octanol/water distribution coefficient at pH=9
Log*D at pH=10*	Float	Octanol/water distribution coefficient at pH=10
Log*D at pH=11*	Float	Octanol/water distribution coefficient at pH=11
Log*D at pH=12*	Float	Octanol/water distribution coefficient at pH=12
Log*D at pH=13*	Float	Octanol/water distribution coefficient at pH=13
Log*D at pH=14*	Float	Octanol/water distribution coefficient at pH=14
MR	Float	Molar refractivity
HBD	Float	Number of Hydrogen-bond donors
HBA	Float	Number of Hydrogen-bond acceptors
PSA	Float	Polar Surface Area
TPSA	Float	Topological Polar Surface Area

3.7. Physico-Chemical Properties (PCP)

Every chemical (identified by CAS number) has set of unique physicochemical data that is readily available in databases (also for instance in NTP) and can be used in QSAR analysis as property or descriptor values. Following physicochemical data should also be downloaded from NTP database and be accessible through the DW.

Name	Type	Description
CAS (unique identifier)	Integer	The Chemical Abstract Service (CAS) number. This is an integer without hyphens or spaces
PCP Name	String	List of PCP-s: Specific Gravity Density Melting Point Boiling Point Vapor Pressure Vapor Density Decomposition Temperature Solubility Flash Point Autoignition Temperature Ph value Viscosity Refractive Index Octanol Water Partition Coefficient Partition Coefficient Other PCP
PCP Value	Float	Value measured for particular PCP
PCP Unit	String	Unit of measured PCP
PCP Method	String	Method of measurement
PCP Temp	Float	Condition of the measurement, temperature. List of PCP-s where it is applicable: Specific Gravity Density Vapor Pressure Solubility Viscosity Refractive Index
PCP Temp Unit	String	Condition of the measurement, unit for temperature
PCP Pressure	Float	Condition of the measurement, pressure. List of PCP-s where it is applicable: Boiling Point From current list applies only on Boiling Point
PCP Pressure Unit	String	Condition of the measurement, unit for pressure
PCP Solvent	String	Condition of the measurement, solvent. List of PCP-s where it is applicable: Solubility Applies only for solubility measurements
PCP Concentration	Float	Condition of the measurement, concentration. List of PCP-s where it is applicable:

		Ph value Applies only for Ph measurements
PCP Concentration Unit	String	Condition of the measurement, unit for concentration
PCP Reference	Text	Literature source for respective PCP

Note on Material Safety Data Sheets (MSDS-s)

MDSD-s contain full information about chemicals, including physicochemical properties and toxicological data. Unfortunately toxicological data is not consistent and verified like in NTP or ECOTOX. *Physicochemical properties in MSDS can complement the physicochemical information obtained from the NTP database and can be one of* ***data sources for the DW in future***. Right now most of the MSDS (also NTP database follows the general setup for MSDS) are provided by chemical manufactures as standardized text or html files.

Most of the MSDS-s are freely downloadable from Internet and the best resources of them include up to 1'000'000 records [i]. There is also an initiative taken by US, Department of Defense to develop **XML format** for the MSDS-s [ii, iii]. The format is defined but unfortunately not in everyday use.

Example1: Vermont Safety Information Resources, Inc. MSDS collection

MSDS collection [iv] provided by Vermont Safety Information Resources, Inc. [v] can be searched over the Web and is also available for downolad [vi] in 145 MB tar file with approximately 180'000 MSDS records. Unfortunately this data source contains also mixtures, not only pure chemicals.

Example2: Fischer Scientific and Sigma-Aldrich MSDS collections.

Better alternatives would be online chemical catalogues by Fischer Scientific [vii] and Sigma-Aldrich [viii], but they require registration. Both resources include only chemicals. This is their advantage.

4. Integration of Data and Metadata in the MOLDW: its usefulness through the entire project.

The integration of data sources to derive metadata is fundamental to the success of the whole OpenMolGRID project. Up to now, gathering information through web-based databases was carried out manually by the user, with a considerable waste of time, money, and resources. The OpenMolGRID Data Warehouse, MOLDW, offers a simple and powerful way to select the molecules of interest trough a user-friendly interface, gathering all web-accessible information into a unique data source. This will considerably speed-up the search for a specific compound, immediately providing its structure, chemical-physical properties, toxicological-related properties, and, if available, chemical descriptors. Since the MOLDW represents the active input/output interface between the user and the UNICORE system, its usefulness goes beyond the objectives of **WP1**, to many work packages. In the following we report work packages and deliverables it affects:

WP2 (D2.2, D2.3): the information gathered by MOLDW can be used as input to predict environmental, toxicological and pharmacological end-points by querying the QSAR/QSPR software packages integrated within the OpenMolGRID system, such as MLR, PCA, ANN, etc. In this case, *Structures* and *Chemical Descriptors* represent the main input needed to build such a model. MOLDW offers the possibility to organize the initial data, and to pre-process the chemical descriptors, giving them the most suitable format for further calculations.

WP3(D3.2, D3.5): MOLDW will play a fundamental role in addressing the objectives stated in **WP3.** In this case, both *Structures* and *PCP* provide input for the molecular structure generation. *Structures* will serve as input to build a structural fragment library. Once the fragments are generated, it would be desirable to save them in the same format the molecular structures have. For this reason, the choice made for saving molecular structures radically change the way the user might manage molecular fragments.

WP5(D5.1, D5.5): test and validation of OpenMolGRID performance on real life chemical and pharmaceutical applications strongly depends on the flexibility of the data warehouse system. A huge variety of tasks will be subject to a performance investigation: from GPCR activity to human fibroblast toxicology. It is therefore necessary for the user interface to be enough flexible to manage environmental, toxicological and pharmaceutical queries as well.

WP6(D6.1): Results achieved by using OpenMolGRID will be initially visualized as output in a format compatible with that of the input files. In order to guarantee the I/O compatibility, and readily exploit the results so far obtained, input and output should share the same format. This requirement has been taken into account in developing the MOLDW repository.

The successful implementation of MOLDW will have a noticeable influence on the overall OpenMolGRID performance. As we have already seen, the objectives proposed in **WP2**, **WP3**, **WP5**, and **WP6**, could not be efficiently addressed without integrating Data and Metadata appropriately.

5. References

i http://www.ilpi.com/msds/

ii https://www.denix.osd.mil/denix/Public/Library/MSDS/HMIS/hmis.html

iii http://www.esohxml.org/

iv http://siri.org/msds/index.php

v http://siri.org/index.php

vi http://siri.org/msds/load.html

vii https://www1.fishersci.com/index.jsp

viii http://www.sigmaaldrich.com/