MeMo Design

Design Status: Draft Release
Design Version: Version 2.2
Date: June 06, 2006
Author: Dr Irena Spasic

School of Chemistry
The University of Manchester
Manchester M60 1QD
United Kingdom

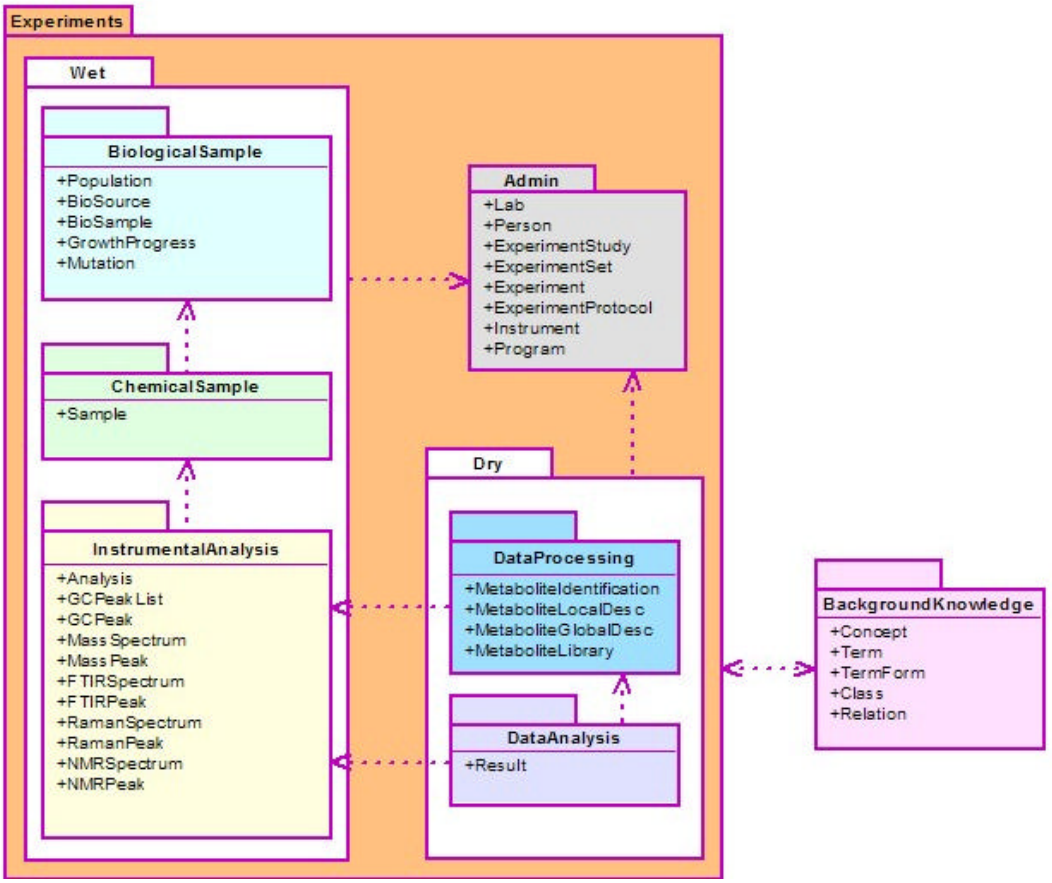
Table of Contents

1. Introduction

This document explains the design of schema for metabolomics modelling (MeMo) and its SQL implementation.

Biochemical data is nowadays being generated much faster than it could effectively be analysed. One possible approach to tackle this problem is to automate the process of hypothesis generation and experimentation. Another approach is to mine the data obtained as the results of wet experiments in order to produce conclusions automatically. In this project, we describe the general framework of the latter approach by means of a schema describing the structure of a relational database (RDB) used to store experimental data and background knowledge. Figure 1 represents the global components of the schema supporting this approach. We concentrate on a specific problem of metabolic footprinting as a strategy for functional genomics. However, the general schema is applicable to a wide range of metabolomics experiments. Namely, the schema is being developed so as to be extensible through the specialisation of abstract classes. In other words, the core of the schema consists of abstract classes which can be further specialised in order to embrace different types of experiments, results, organisms, etc.

Figure 1. The global components of the MeMo schema



2. The Architecture

Unified Modelling Language (UML) is used to represent the architecture of MeMo. Figure 1 describes the global components in MeMo represented by UML packages. MeMo describes the experimental framework for metabolomics experiment consisting of the following components:

1. **Admin**. Administrative data about an experimental framework (e.g. protocols, personnel , instruments).

2. **Biological Sample**. Description of a biological source (e.g. genotype, phenotype, biological sample).
3. **Chemical Sample**. Description of a sample prepared for instrumental analysis.
4. **Instrumental Analysis**. The results (e.g. mass spectrum) of analyzing a sample using an analytical instrument.
5. **Data Processing**. Identification of metabolites based on the results of instrumental analysis.
6. **Data Analysis**. Computational analysis of the analytical data and the associated meta-data.
7. **Background Knowledge**. The background knowledge explicitly stored in the database in a machine-readable form (e.g. gene classification, compound properties such as molecular weight, structure, reactions, etc.).
8. **XML Schemas**. (not shown in Figure 1) Data that describes the structure of textual fields that can store XML data.

Administrative data link related sets of experiments and provide contact and other information about the personnel performing the experiments.

The experimental framework for metabolomics (described by the package called *Wet Experiments* in Figure 1) can be divided into three main themes, including biological sample preparation, chemical sample preparation and instrumental analysis, performed in this order. There is a package in the schema corresponding to each experiment type. The output of *biologica sample* preparation is used as input for *chemical sample* preparation. The results of chemical sample preparation are in turn used in *instrumental analysis*. These relations are noted in the schema by directed dependencies between the packages involved.

As mentioned earlier, the metabolomics experiments in the post-genomic era often need to be extended beyond the traditional "wet" experimental framework. In order to process the vast amounts of data produced as a result of wet experiments, so-called *dry experiments* need to be performed *in silico* to extract knowledge. Therefore, a separate package is used to model dry experiments, which covers the processing of raw analytical data (*data processing*) and their computational analysis (*data analysis*). Dry experiments follow the analytical ones continuing the chain of wet experiments. Finally, a package named *Experiments* is introduced to group all experiment types, both wet and dry (Figure 1).

Further, experiments are generally performed in relation to some background knowledge, with the ultimate aim of enriching that knowledge. This knowledge is used to provide the biological context for the genetic strains being examined (e.g. information on specific organisms, gene to class mappings, etc.), to interpret the results of analytical experiments (e.g. compound properties such as molecular weight, structure, reactions, etc.), to support the reasoning process of data analysis (e.g. classification rules), etc. The relations between experiments and the background knowledge are modelled by an association between the corresponding packages in the schema.

Finally, XML Schemas are used to simplify the structure and the maintenance of the MeMo implementation as a RDB. Complex data that are aimed at browsing rather than querying are structured by XML and the entire XML document is stored in a single field, which improves the transparency of the RDB, but still leaves room for automatic processing of the stored data. This component is part of implementation, not the MeMo model.

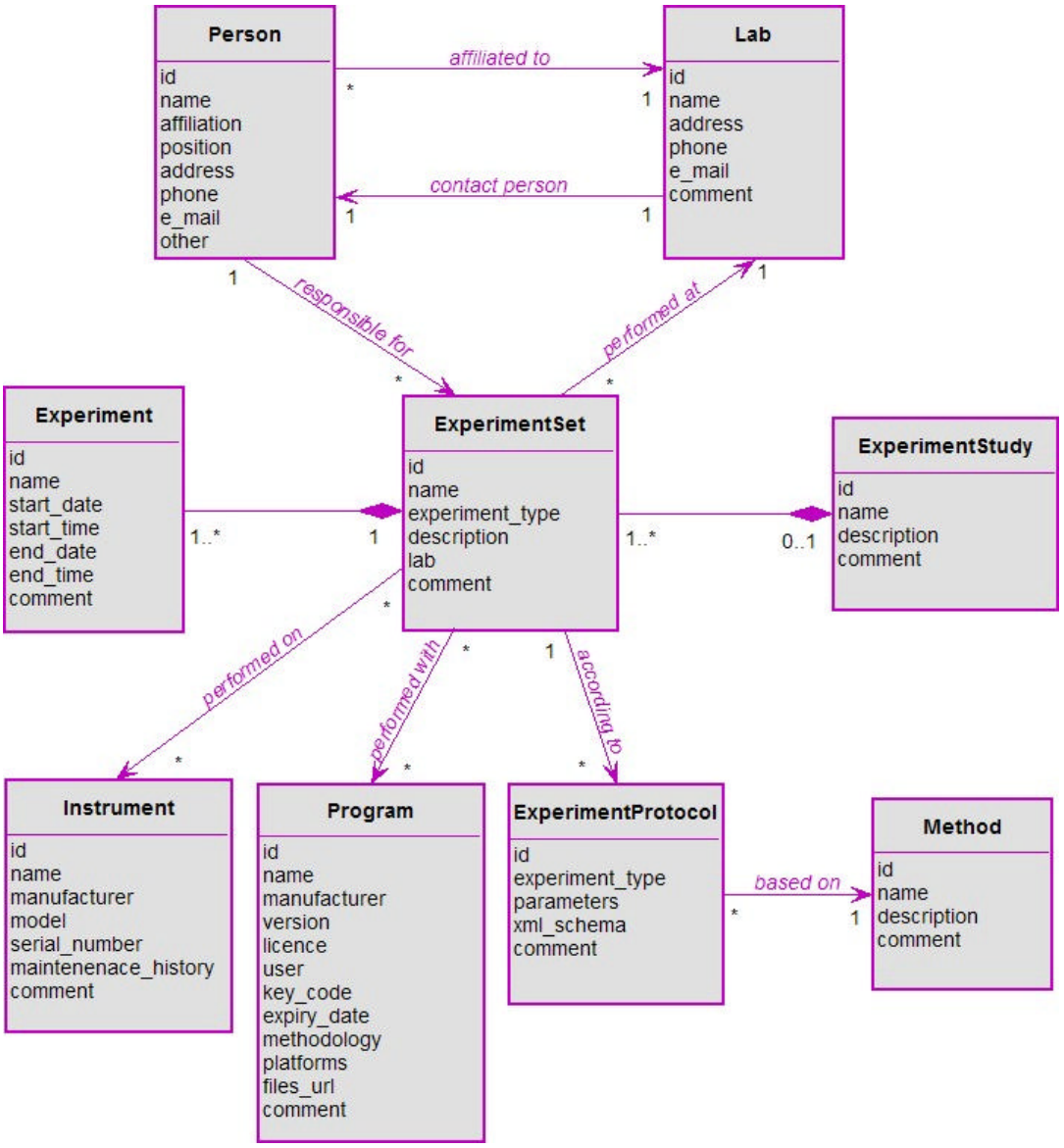
3. The Components

Each of the packages described in Section 2 consists of classes, which are used to represent the key entities in the corresponding components. This section describes each component separately.

3.1. Admin

Class Diagram. The class diagram for the Admin package is depicted in Figure 3.1.1.

Figure 3.1.1. The Admin Component



Entities. The Admin component consists of the following entities:

Table 3.1.1. Admin Component Entities

Entity	Description	Occurrence
Person	A member of lab personnel, usually the one responsible for conducting an experiment	There will be a Person entity for each person conducting any kind of experiment
Lab	A lab at which experiments can be performed	There will be a Lab entity for each lab conducting an experiment
ExperimentProtocol	Specification of an *experimental setup* used to perform a set of related experiments	There will be an ExperimentProtocol entity for each setting used to perform at least one set of experiments

ExperimentSet	A set of related experiments (of the same type, e.g. a set of biological sample preparation experiments). A set of experiments is typically performed under the same conditions. Grouping such experiments under set removes the need to specify these conditions repetitively for each individual experiment.	There will be an ExperimentSet entity for each set of related experiments
Experiment	Specification of a *single experiment*. It is used to represent all three types of wet experiments (i.e. biological sample preparation, chemical sample preparation and instrumental analysis) as well as dry experiment (i.e. computational data analysis).	There will be an Experiment entity for each experiment performed
ExperimentStudy	A summary of the experimental study, which links biological sample preparation, chemical sample preparation and instrumental analysis	There will be an ExperimentStudy entity for each study
Method	A global *method* upon which an experiment protocol, analytical instrument or software is based (e.g. metabolic footprinting for biological sample preparation, GCMS for instrumental analysis, GP for data analysis, etc.)	There will be a Method entity for each global method that can be used in an experiment
Instrument	A particular *analytical instrument* (as an instance, not as a type!)	There will be an Instrument entity for each analytical instrument used or available
Program	A particular *software package* that can be used for computational data analysis	There will be a Program entity for each software package used or available for data analysis

Dependencies. The Admin component is dependent upon the XML Schemas component. These and internal dependencies are described by the following relationships between entities:

Table 3.1.2. Admin Component Dependencies

Entities	Multiplicity	Relationship
Experiment : ExperimentSet	1..* : 1	PartOf
ExperimentSet : ExperimentStudy	1..* : 1	PartOf
Person : ExperimentSet	1 : *	ResponsibleFor
Person : Lab	* : 1	AffiliatedTo
Lab : Person	1 : 1	ContactPerson
ExperimentSet : ExperimentProtocol	* : 1	AccordingTo
ExperimentProtocol : Method	* : 1	BasedOn
ExperimentSet : Instrument	* : *	PerformedOn
ExperimentSet : Program	* : *	PerformedWith
ExperimentProtocol : XMLSchema	* : *	StructuredBy

Attributes. The entities and relations from the Admin component are described below through their attributes.

Table 3.1.3. Person

Attribute	Description	Domain & Characteristics			
		Type	R[1]	PK	FK
ID	An identifier for the person conducting an experiment	VARCHAR(50)	✓	✓	
name	The name of the person	VARCHAR(50)	✓	✓	
affiliation	A reference to the institute to which the person is affiliated	VARCHAR(50)			✓
position	His/her position at the institution where the experiment is performed	VARCHAR(50)			
address	His/her work address	VARCHAR(200)			
phone	His/her work telephone number	VARCHAR(50)			
e_mail	His/her e-mail address	VARCHAR(50)	✓		
other	Additional information about the person	TEXT			

Table 3.1.4. Lab

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the lab at which the experiments were performed	VARCHAR(50)	✓	✓	
name	The name of the lab	VARCHAR(200)	✓		
address	The address of the lab	VARCHAR(200)			
phone	A telephone number at the lab	VARCHAR(50)			
e_mail	An e-mail address for the lab	VARCHAR(50)			
person_ID	A reference to a contact person	VARCHAR(50)			✓
comment	Comments about the lab	TEXT			

Table 3.1.5. Experiment Protocol

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the experiment protocol	VARCHAR(50)	✓	✓	
experiment_type	Type of the experiments in the set: ***BS*** (biological sample prep), ***CS*** (chemical sample prep), ***IA*** (instrumental analysis) or ***DA*** (data analysis)	VARCHAR(10)	✓		✓

[1] **R**, **PK** and **FK** are abbreviations for Required, Primary Key and Foreign Key respectively.

method	A reference to a method upon which the protocol is based (e.g. metabolic footprinting for biological sample preparation, GCMS for instrumental analysis, GP for data analysis, etc)	VARCHAR(50)	✓		✓
parameters	A particular choice of parameter values used in the protocol (e.g. instrument or program settings)	TEXT			
xml_schema	A reference to an XML schema (if applicable) according to which the parameters field (above) is structured into an XML document	VARCHAR(50)			✓
comment	Comments about the experiment protocol	TEXT			

Table 3.1.6. Experiment Set

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the set of experiments	VARCHAR(50)	✓	✓	
name	The name used for the particular set of experiments	VARCHAR(50)	✓		
description	A description of the experiment set including its objectives and the hypothesis behind it	TEXT			
lab_ID	A reference to the laboratory in which the experiment was conducted	VARCHAR(50)	✓		✓
protocol_ID	A reference to the protocol followed in the experiment	VARCHAR(50)	✓		✓
conductor_ID	The user ID for the person conducting the set of experiments	VARCHAR(50)	✓		✓
comment	Comments about the set of experiments	TEXT			

Table 3.1.7. Experiment

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the experiment	VARCHAR(50)	✓	✓	
name	The name used to describe the particular experiment	VARCHAR(200)			
set_ID	A reference to the set of which the experiment is a part	VARCHAR(50)	✓		✓
start_date	The date on which the experiment started	DATE	✓		
start_time	The time at which the experiment started	TIME			
end_date	The date on which the experiment ended	DATE	✓		
end_time	The time at which the experiment ended	TIME			
comment	Comments about the experiment	TEXT			

Table 3.1.8. Experiment Study

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the experiment study, which links biological, sample preparation and analytical experiments	VARCHAR(50)	✓	✓	
name	The name used for the particular experiment study	VARCHAR(50)	✓		
description	A description of the experiment study including its objectives and the hypothesis behind it	TEXT			
bio	A reference to a set of biological experiments	VARCHAR(50)	✓		✓
sample_preparation	A reference to a set of sample preparation experiments	VARCHAR(50)	✓		✓
analysis	A reference to a set of analytical experiments	VARCHAR(50)	✓		✓
comment	Comments about the experiment study	TEXT			

Table 3.1.9. Method

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the method	VARCHAR(50)	✓	✓	
name	The name used for the particular method (e.g. metabolic footprinting, Gas Chromatography / Mass Spectrometry, genetic programming, etc.)	VARCHAR(50)	✓		
description	A description of the method	TEXT			
comment	Comments about the method	TEXT			

Table 3.1.10. Instrument

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the analytical instrument	VARCHAR(50)	✓	✓	
name	The name used for the particular instrument	VARCHAR(50)	✓		
manufacturer	The name of the instrument's manufacturer	VARCHAR(50)			
model	The model of the instrument	VARCHAR(50)			
serial_number	The serial number of the instrument	VARCHAR(50)			
maintenance_history	A description of the maintenance history for the instrument	TEXT			
comment	Comments about the instrument	TEXT			

Table 3.1.11. InstrumentUsed

Attribute	Description	Domain & Characteristics

		Type	R	PK	FK
set_ID	A reference to an experiment set	VARCHAR(50)	✓	✓	✓
instrument_ID	A reference to an instrument used in the experiment set	VARCHAR(50)	✓	✓	✓

Table 3.1.12. Program

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the software	VARCHAR(50)	✓	✓	
name	The name used for the particular software	VARCHAR(50)	✓		
manufacturer	The name of the manufacturer of the software	VARCHAR(50)			
version	The version of the software	VARCHAR(50)			
licence	The type of licence (single or multiple user, open source, ...)	VARCHAR(50)			
user	User(s) that holds the licence	VARCHAR(50)			✓
key_code	The key code used to activate the software	VARCHAR(50)			
expiry_date	The licence expiry date	DATE			
methodology	A reference to the methodology implemented by the software	VARCHAR(50)	✓		✓
platforms	The platforms on which the software runs	VARCHAR(200)			
files_url	The location of the installation files	VARCHAR(200)			
comment	Comments about the software	TEXT			

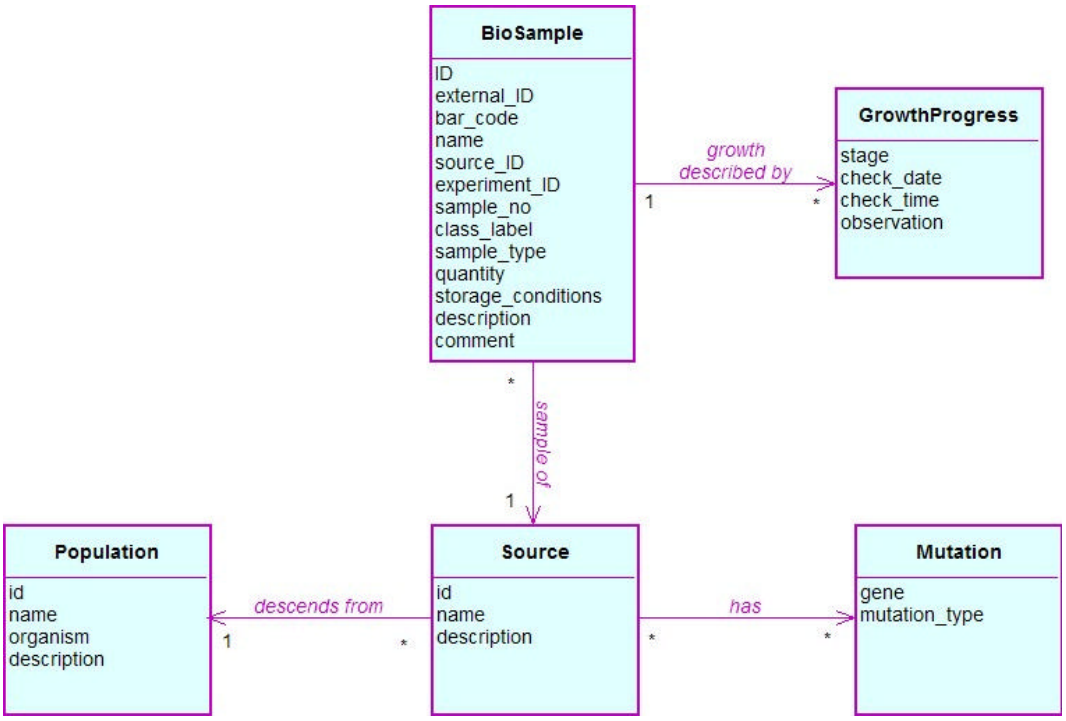
Table 3.1.13. ProgramUsed

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
experiment_ID	A reference to an experiment *set*	VARCHAR(50)	✓	✓	✓
program_ID	A reference to a program used in the experiment set	VARCHAR(50)	✓	✓	✓

3.2. Biological Sample

Class Diagram. The class diagram for the Biological Sample package is depicted in Figure 3.2.1.

Figure 3.2.1. The Biological Sample Component



Entities. The Bio component consists of the following entities:

Table 3.2.1. Bio Entities

Entity	Description	Occurrence
Population	Description of a particular collection of biological sources (e.g. wild type for various mutant strains, human patients with a specific problem such as diabetes, etc.)	There will be a Population entity for each collection of biological sources
Source	Description of a specific biological source (e.g. mutant yeast strain, patient)	There will be a Source entity for each strain used during the project
BioSample	Description of a biological sample (e.g. metabolic footprint for yeast, blood serum for humans) taken from the biological source.	There will be a BioSample entity for each biological sample of each biological source
Mutation	A gene mutated in a specific yeast strain	There will be a Mutation entity for each mutated gene in a specific strain
GrowthProgress	Records the observations of *growth progress* throughout a number of specified stages	There will be a GrowthProgress entity for each growth experiment

Dependencies. The Bio component is dependent upon the Admin component and the Background Knowledge[2] component. These and internal dependencies are described by the following relationships between entities:

[2] The Background Knowledge component is not implemented and it depends on the resources available and specific implementation choices. Question marks (?) are used to denote a potential foreign key that could link a

Table 3.2.2. Bio Component Dependencies

Entities	Multiplicity	Relationship
Source : Population	* : 1	BelongsTo
Source : Gene	* : *	Mutation
BioSample : Source	1..* : 1	SampleOf
BioSample : Experiment	* : 1	ProducedBy
Source : GrowthProgress	1 : *	GrowthDescribedBy

Attributes. The entities from the Bio component are described below through their attributes.

Table 3.2.3. Population

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the specific population	VARCHAR(50)	✓	✓	?
name	The name used to refer to the particular population	VARCHAR(200)	✓		
organism	The name of the corresponding organism	VARCHAR(50)	✓		✓
description	A description of the observable structure, function or behaviour of the given population	TEXT			
comment	Comments about the population	TEXT			

Table 3.2.4. Source

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the specific biological source	VARCHAR(50)	✓	✓	
name	The name used to refer to the particular biological source	VARCHAR(200)	✓		
population_ID	A reference to the belonging population	VARCHAR(50)	✓		✓
description	A description of the observable structure, function or behaviour of the given biological source	TEXT			
comment	Comments about the biological source	TEXT			

Table 3.2.5. BioSample

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for a biological sample of a biological source	VARCHAR(50)	✓	✓	
external_ID	An identifier for the biological sample as given by the supplier	VARCHAR(50)			

specific attribute to an entity from the Background Knowledge component (e.g. to find information about a mutated gene).

bar_code	Internally used bar code assigned to the given sample	VARCHAR(200)			
name	The name used to refer to the particular biological sample	VARCHAR(200)			
source_ID	A reference to the source from which the sample was taken	VARCHAR(50)	✓		✓
experiment_ID	A reference to the biological (i.e. growth) experiment used to obtain the given sample	VARCHAR(50)	✓		✓
sample_no	An internally used number assigned to the sample (usually reflects the order of processing)	INT			
class_label	A (externally given) label indicating a class to which the sample belongs	VARCHAR(50)			
sample_type	The type of sample where applicable (e.g. plasma, serum, etc.)	VARCHAR(50)			
quantity	The amount of the sample (in millilitres)	INT			
storage_conditions	The conditions at which the sample is stored	TEXT			
description	A description of the observable structure, function or behaviour of the given sample	TEXT			
comment	Comments about the biological sample	TEXT			

Table 3.2.6. Mutation

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
source_ID	An identifier for the biological source	VARCHAR(50)	✓	✓	✓
gene_ID	An identifier for the mutated gene	VARCHAR(10)	✓	✓	?
mutation_type	The way in which the gene has been mutated	VARCHAR(50)	✓		✓

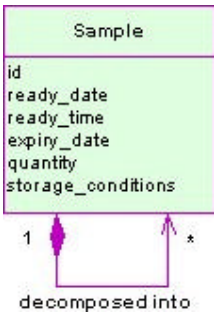
Table 3.2.7. Growth Progress

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
biosample_ID	An identifier for the biological sample (i.e. strain replicate) being grown	VARCHAR(50)	✓	✓	✓
stage	The stage at which the sample was observed	INT	✓	✓	
check_date	The date on which the sample was observed	DATE			
check_time	The time at which the sample was observed	TIME			
observation	A brief description of the observable structure, function or behaviour of the given sample (strain) at the given stage	TEXT	✓		

3.3. Chemical Sample

Class Diagram. The class diagram for the Chemical Sample package is depicted in Figure 3.3.1.

Figure 3.3.1. The Chemical Sample Component



Entities. The Chemical Sample component consists of the following entities:

Table 3.3.1. Sample Preparation Component Entities

Entity	Description	Occurrence
Sample	A chemical *sample* produced from the biological sample as a result of a sample preparation experiment.	There will be a number of sample entities for each sample preparation experiment based on a specific *biological* sample

Dependencies. The Chemical Sample component is dependent upon the Admin component and the Biological Sample component. These and internal dependencies are described by the following relationships between entities:

Table 3.3.2. Sample Preparation Component Dependencies

Entities	Multiplicity	Relationship
Experiment : Sample	1 : *	Produces
Sample : BioSample	1 : 1	SampleOf
Sample : Sample	1 : *	DecomposedInto

Attributes. The entities from the Chemical Sample component are described below through their attributes.

Table 3.3.3. Sample

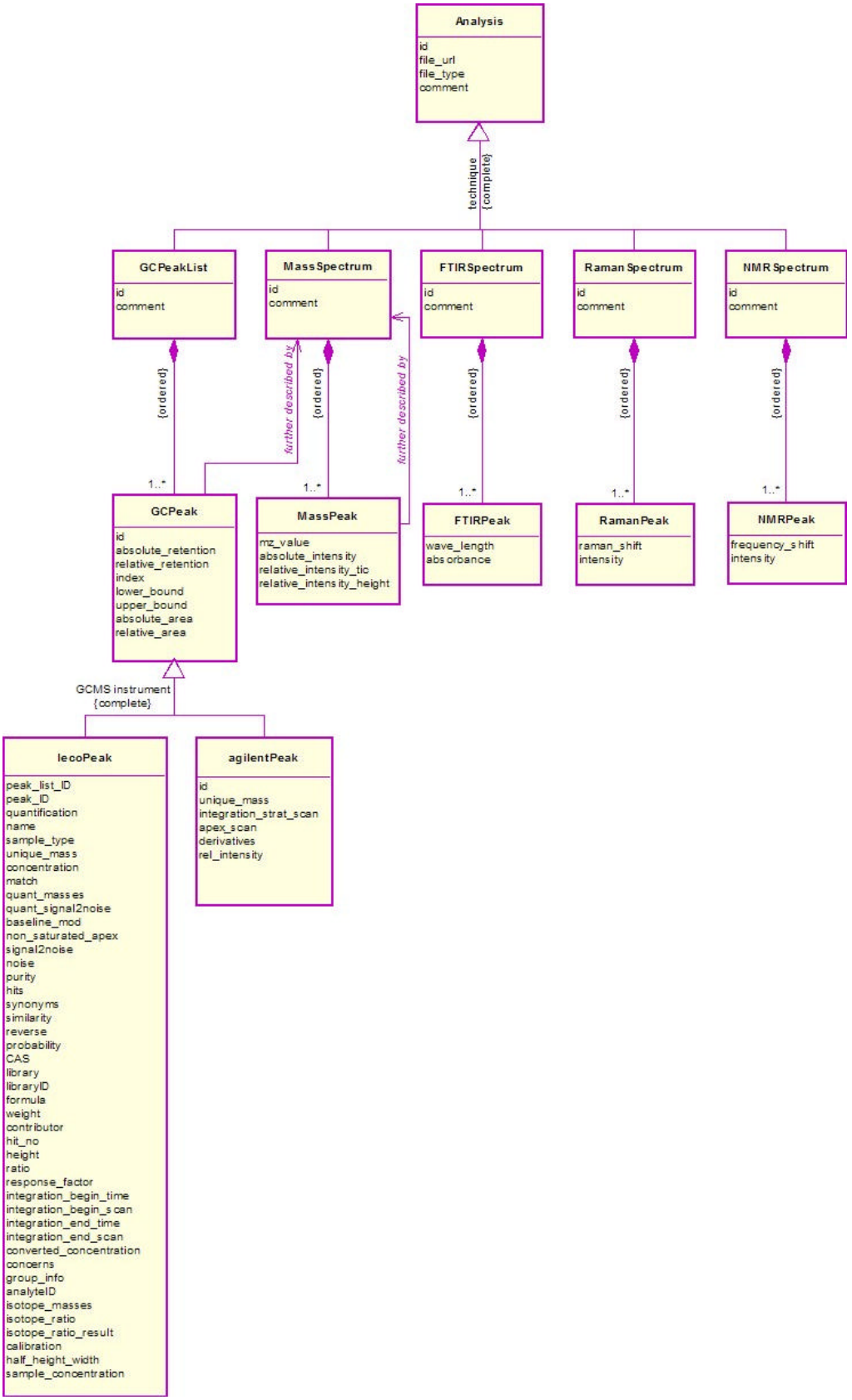
Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the chemical sample prepared for instrumental analysis	VARCHAR(50)	✓	✓	
name	The name used to refer to the chemical sample	VARCHAR(200)			
biosample_ID	A reference to the biological sample from which the chemical sample was prepared	VARCHAR(50)	✓		✓
experiment_ID	A reference to the chemical sample preparation experiment in which the sample was prepared	VARCHAR(50)	✓		✓

ready_date	The date on which the chemical sample was made "ready to use"	DATE			
ready_time	The time at which the chemical sample was made "ready to use"	TIME			
expiry_date	The chemical sample's "best before" date	DATE	✓		
quantity	The amount of the chemical sample (in millilitres)	INT	✓		
storage_conditions	The conditions at which the chemical sample is stored	TEXT			
part_of	A reference to the chemical sample of which the given chemical sample is a part	VARCHAR(50)			✓
comment	Comments about the chemical sample	TEXT			

3.4. Analysis

Class Diagram. The class diagram for the Analysis package is depicted in Figure 3.4.1.

Figure 3.4.1. The Analysis Component



Entities. The Analysis component consists of the following entities:

Table 3.4.1. Analysis Component Entities

Entity	Description	Occurrence
Analysis	The result of an analytical experiment	There will be an Analysis entity for each analytical experiment performed
GCPeakList	A *GC peak list* produced as an output of a GC run (corresponds to a minimal common output set for all GC instruments)	There will be a gcPeakList entity for each sample analysed on a GC instrument
GCPeak	A *GC peak* as part of a GC peak list	There will be a gcPeak entity for each peak in a GC peak list
LecoPeak	Additional information supplied for a GC peak produced by a *Leco* instrument	There will be a lecoPeak entity for each GC peak from the GC peak list produced by a Leco instrument
AgilentPeak	Additional information supplied for a GC peak produced by an *Agilent* instrument	There will be an AgilentPeak entity for each GC peak from the GC peak list produced by an Agilent instrument
MassSpectrum	A mass spectrum produced by a MS (or hyphenated MS) instrument	There will be a massSpectrum entity for each mass spectrum produced as an output of MS (or hyphenated MS) run or tandem MS
MSPeak	A *mass peak* produced as part of a mass spectrum	There will be a msPeak entity for each peak in a mass spectrum
FTIRSpectrum	An *FT-IR spectrum* produced as an output of a FT-IR run (corresponds to a minimal common output set for all FT-IR instruments)	There will be a ftirSpectrum entity for each spectrum in a sample analysed on an FT-IR instrument
FTIRPeak	An *FT-IR peak* as part of an FT-IR spectrum	There will be an ftirPoint entity for each peak in an FTIR spectrum
RamanSpectrum	A *Raman spectrum* produced as an output of a Raman run (corresponds to a minimal common output set for all Raman instruments)	There will be a RamanSpectrum entity for each spectrum in a sample analysed on a Raman instrument
RamanPeak	A *Raman peak* as part of a Raman spectrum	There will be a ramanPeak entity for each peak in a Raman spectrum
NMRSpectrum	An *NMR spectrum* produced as an output of a NMR run (corresponds to a minimal common output set for all NMR instruments)	There will be an NMRSpectrum entity for each spectrum in a sample analysed on an NMR instrument
NMRPeak	An *NMR peak* as part of an NMR spectrum	There will be an NMRPeak entity for each peak in an NMR spectrum

Dependencies. The Analysis component is dependent upon the Admin component. These and internal dependencies are described by the following relationships between entities:

Table 3.4.2. Analysis Component Dependencies

Entities	Multiplicity	Relationship
Experiment : Analysis	1 : 0..1	ResultsIn
Analysis : GCPeakList	1 : 0..1	FormattedAs
GCPeakList : GCPeak	1 : 1..*	OrderedListOf
GCPeak : LecoPeak	1 : 0..1	AdditionallyDescribedBy
GCPeak : AgilentPeak	1 : 0..1	AdditionallyDescribedBy
GCPeak : MassSpectrum	1 : 0..1	DescribedBy
Analysis : MassSpectrum	1 : 0..1	FormattedAs
MassSpectrum : MassPeak	1 : 1..*	OrderedListOf
Analysis : FTIRSpectrum	1 : 0..1	FormattedAs
FTIRSpectrum : FTIRPeak	1 : 1..*	OrderedListOf
Analysis : RamanSpectrum	1 : 0..1	FormattedAs
RamanSpectrum : RamanPeak	1 : 1..*	OrderedListOf
Analysis : NMRSpectrum	1 : 0..1	FormattedAs
NMRSpectrum : NMRPeak	1 : 1..*	OrderedListOf

Attributes. The entities from the Analysis component are described below through their attributes.

Table 3.4.3. Analysis

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the results of analysing a sample	VARCHAR(50)	✓	✓	✓
experiment_ID	A reference to the corresponding analytical experiment	VARCHAR(50)	✓		✓
sample_ID	A reference to the sample analysed	VARCHAR(50)	✓		✓
name	A name used to refer to specific analysis	VARCHAR(50)			
software_name	The name of the software used to process the raw data	VARCHAR(50)			
software_ver	The version of the software used to process the raw data	VARCHAR(50)			
file_url	A reference to the raw data file produced as a result of an analytical experiment	VARCHAR(50)			
file_type	Type of the raw data file	VARCHAR(50)			
comment	Comments about the results	TEXT			

Table 3.4.4. GC Peak List

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the peak list produced by a GC-MS instrument and a reference to the general description of experimental results	VARCHAR(50)	✓	✓	✓
comment	Comments about the GC-MS peak list	TEXT			

Table 3.4.5. GC Peak

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
peak_list_ID	An identifier for the GC-MS peak list of which the GC-MS peak is a member	VARCHAR(50)	✓	✓	✓
absolute_retention	The absolute retention time for the peak in seconds	FLOAT	✓	✓	
relative_retention	The relative retention time for the peak in seconds	FLOAT			
retention_index	A retention time index for the peak	FLOAT			
retention_lower_bound	The lower bound of the retention time window in which the peak falls	FLOAT			
retention_upper_bound	The upper bound of the retention time window in which the peak falls	FLOAT			
absolute_area	The absolute area of the peak	INT	✓		
relative_area	The relative area of the peak	FLOAT			

Table 3.4.6. Mass Spectrum

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the mass spectrum (automatically generated primary key to facilitate referencing to individual mass spectra)	VARCHAR(50)	✓	✓	
analysis_ID	A reference to the results of the corresponding analytical experiment (note: coincides with GC peak list ID, when GC is used)	VARCHAR(50)	✓		✓
GC_peak_ID	A reference to a GC peak belonging to a peak list produced by the given experiment (experiment_ID)	VARCHAR(50)			✓

tree_number	The number denoting the position of the spectrum in the hierarchy of mass spectrums (for *tandem* mass spectrometry)	VARCHAR(200)	✓		
comment	Comments about the mass spectrum	TEXT			

Table 3.4.7. MS Peak

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ms_ID	An identifier for the mass spectrum of which the MS peak is a member	VARCHAR(50)	✓	✓	✓
mz_value	A mass-to-charge value	INT	✓	✓	
absolute_intensity	The absolute intensity registered for the given mass-to-charge value	INT	✓		
relative_intensity_tic	The intensity relative to the total ion current (TIC)	FLOAT			
relative_intensity_height	The intensity relative to the maximal peak height	FLOAT			

Table 3.4.8. Leco Peak

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
peak_list_ID	An identifier for the GC-MS peak list of which the GC-MS peak is a member	VARCHAR(50)	✓	✓	✓
peak_ID	An identifier for the GC-MS peak obtained from a Leco GC-MS instrument	INT	✓	✓	✓
quantification	The calibration/analyte name used for the quantification for this peak	VARCHAR(100)			
name	The name assigned to a peak. It can come from the library search, the quantification, auto naming from the software, or from the operator	VARCHAR(100)			
sample_type	The category the sample is in	VARCHAR(100)			
unique_mass	The mass value chosen by the instrument as the best peak based on purity calculations for the peak	INT			
concentration	The concentration calculated for the corresponding analyte using the indicated calibration	FLOAT			
match	A numerical description [0..999] of the match between the analyte and the referent mass spectrum. The higher the number, the better the match	INT			

quant_masses	The m/z value used for quantification	VARCHAR(200)			
quant_signal2noise	The peak height of the quantified mass divided by the calculated noise	FLOAT			
baseline_mod	Indicator of manual modification of the analyte's baseline	BOOLEAN			
non_saturated_apex	If the peak is saturated, this is the location of the closest point to the apex that is not saturated	FLOAT			
signal2noise	Baseline corrected peak height of the unique mass divided by the noise	FLOAT			
noise	The variation of the signal as calculated from the analyte peaks	FLOAT			
purity	A unitless value that helps to describe how much unique_mass coelutes with other compounds	FLOAT			
hits	The library hit(s) names found for the peak	VARCHAR(200)			
synonyms	Synonym names for the compound corresponding to the best hit	VARCHAR(200)			
similarity	A numerical description [0..999] (defined by the NIST search algorithm) of how well the library hit matches the peak using all masses. The higher the value the better the fit	INT			
reverse	A value [0..999] defined by the NIST search algorithm describing how well the library hit matched the peak using only the masses present in the NIST DB mass spectrum. The higher the value the better the fit	INT			
probability	An estimation of the likelihood that a matching compound has been found	INT			
CAS	The CAS (Chemical Abstracts Service) registry number	VARCHAR(100)			
library	The name of the library used to identify the analyte	VARCHAR(100)			
libraryID	The NIST identification number of the matching compound for the selected hit	INT			
formula	The chemical formula of the identified analyte based on the library search (hit)	VARCHAR(100)			
weight	The molecular weight of the matching compound	INT			
contributor	The library contributor of the matching compound	VARCHAR(50)			
hit_no	The number of the selected hit	INT			

height	The distance between horizontal lines passing through the maximum and baseline of the peak	FLOAT			
ratio	ratio = As / Ais, where As = the response (are/height) of the analyte to be measured, Ais = the response of the internal standard	FLOAT			
response_factor	The measure of the mass spectral response of the analyte's wrt its concentration: RF = AS x Cis / (Ais x Cs), where As = the response (are/height) of the analyte being measured, Cis = the concentration of the internal standard, Ais = the response of the internal standard, Cs = the concentration of the analyte being measured	FLOAT			
integration_begin_time	The retention time at which peak integration started	FLOAT			
integration_begin_scan	The number of the scan at which peak integration started	INT			
integration_end_time	The retention time at which peak integration ended	FLOAT			
integration_end_scan	The number of the scan at which peak integration ended	INT			
converted_concentration	The concentration multiplied by the conversion factor from the calibration	FLOAT			
concerns	Definition of what makes the peak a problem (II = isomer interference, SM = saturated mass, QM = low quantification mass match, LM = low library much)	CHAR(2)			
group_info	The name given for that quantification of the peak.	VARCHAR(100)			
analyteID	The user-given name for the compound	VARCHAR(100)			
isotope_masses	Isotope masses entered in the calibration table	VARCHAR(200)			
isotope_ratio	The ratio between the expected isotope ratio and the calculated isotope ratio	FLOAT			
isotope_ratio_result	Displays if the isotope ratio passed or failed according to the parameters set in isotope ratio tolerance in the calibration table	VARCHAR(100)			
calibration	The calibration name used for the quantification of the peak	VARCHAR(100)			
half_height_width	The number of spectrum wide that the peak is at approximately half the height of the peak	FLOAT			
sample_concentration	The amount of analyte in the sample before any dilution	FLOAT			

Table 3.4.9. Agilent Peak

Attribute	Description	Domain & Characteristics

		Type	R	PK	FK
peak_list_ID	An identifier for the GC-MS peak list of which the GC-MS peak is a member	VARCHAR(50)	✓	✓	✓
peak_ID	An identifier for the GC-MS peak obtained from an Agilent GC-MS instrument	INT	✓	✓	✓
mz_unique	The mass-to-charge value chosen by the instrument as the best peak based on purity calculations for the peak	INT			
integration_start_scan	The number of the scan at which peak integration started	INT			
apex_scan	The scan number at the apex of the peak	INT			
derivatives	The derivatives of the peak compound	VARCHAR(200)			
relative_intensity	The relative intensity of the unique ion excluding system peaks in the calculation	FLOAT			

Table 3.4.10. FTIR Spectrum

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the FT-IR spectrum produced by a FT-IR instrument and a reference to the results of the corresponding analytical experiment	VARCHAR(50)	✓	✓	✓
comment	Comments about the FT-IR spectrum	TEXT			

Table 3.4.11. FTIR Peak

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
peak_list_ID	An identifier for the FT-IR spectrum of which the FT-IR peak is a member	VARCHAR(50)	✓	✓	✓
wave_length	A wave length value	FLOAT	✓	✓	
absorbance	The absorbance registered for the given wave length	FLOAT	✓		

Table 3.4.12. Raman Spectrum

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the Raman spectrum produced by a Raman instrument and a reference to the results of the corresponding analytical experiment	VARCHAR(50)	✓	✓	✓
comment	Comments about the Raman spectrum	TEXT			

Table 3.4.13. Raman Peak

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
peak_list_ID	An identifier for the Raman spectrum of which the Raman peak is a member	VARCHAR(50)	✓	✓	✓
raman_shift	A Raman shift value	FLOAT	✓	✓	
intensity	The intensity registered for the given Raman shift	FLOAT	✓		

Table 3.4.14. NMR Spectrum

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the NMR spectrum produced by an NMR instrument and a reference to the results of the corresponding analytical experiment	VARCHAR(50)	✓	✓	✓
comment	Comments about the NMR spectrum	TEXT			

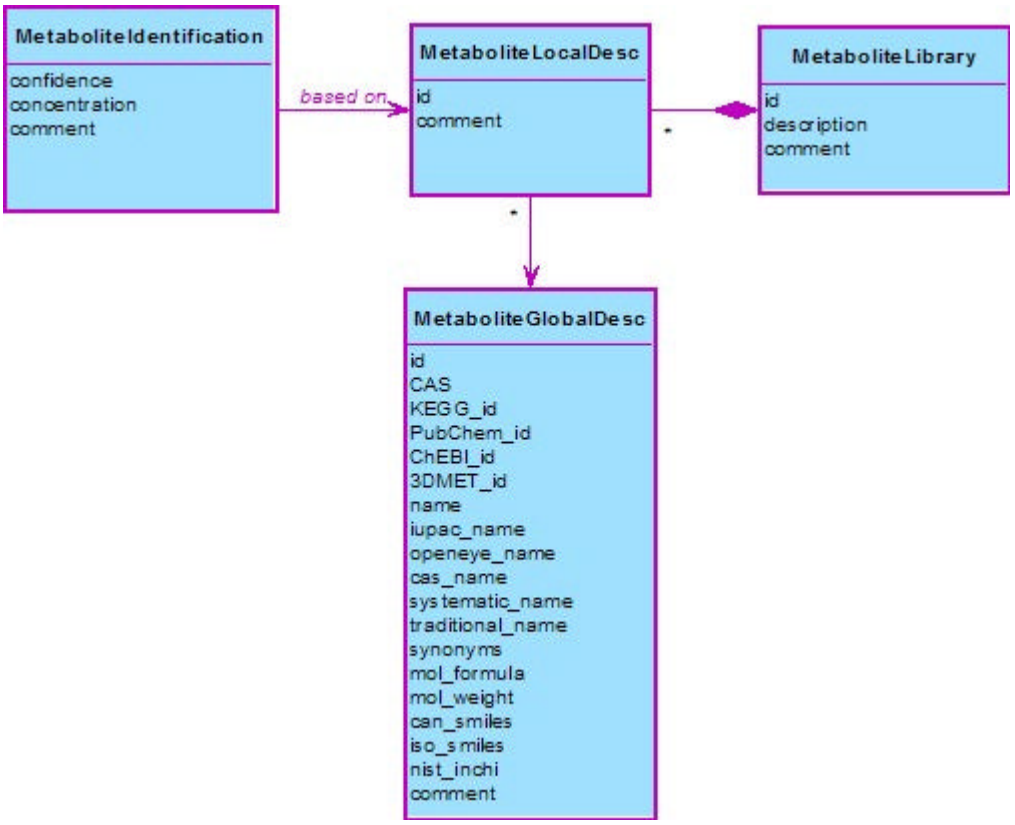
Table 3.4.15. NMR Peak

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
peak_list_ID	An identifier for the NMR spectrum of which the NMR peak is a member	VARCHAR(50)	✓	✓	✓
frequency_shift	A frequency shift value	FLOAT	✓	✓	
intensity	The intensity registered for the given frequency shift	FLOAT	✓		

3.5. Data Processing

Class Diagram. The class diagram for the Data Processing package is depicted in Figure 3.5.1.

Figure 3.5.1. The Data Processing Component



Entities. The Data Processing component consists of the following entities:

Table 3.5.1. Data Processing Component Entities

Entity	Description	Occurrence
MetaboliteIdentification	Identified metabolite based on the processing of data produced by instrumental analysis	There will be a MetaboliteIdentification entity for each metabolite identified in the sample analysed
MetaboliteLibrary	Each instrumental analysis produced as a characterisation of a given metabolite is used to create a new item in a MetaboliteLibrary, which is a collection of all metabolites considered in a specific experimental study.	There will be a MetaboliteLibrary entity for each collection of metabolites of interest for one or more experimental studies.
MetaboliteLocalDesc	A local description of a given metabolite represents a single item in a metabolite library. It refers to a characterisation of a metabolite using a specific analytical technique on a specific instrument under the conditions specific for the experimental study conducted.	There will be a MetaboliteLocalDesc entity for each metabolite of interest. Note that there may be multiple local descriptions for a single metabolite, one for each machine replicate.
MetaboliteGlobalDesc	Each local description of a metabolite is further mapped to its global description (MetaboliteGlobalDesc), where the general information about the metabolite is stored: its identifiers in other resources, its names, molecular formula and weight, SMILES strings, etc.	There will be a MetaboliteGlobalDesc entity for each metabolite of interest. Multiple local description of a single metabolite will be mapped to a single global description even when they belong to different libraries.

Dependencies. The Data Processing component is dependent upon the Analysis component. These dependencies are described by the following relationships between entities:

Table 3.5.2. Data Processing Component Dependencies

Entities	Multiplicity	Relationship
Analysis : MetaboliteIdentification	1..* : *	Implies
GCPeak : MetaboliteIdentification	0..1 : 0..1	Implies
MetaboliteIdentification : MetaboliteLocalDesc	0..1 : 1	BasedOn
MetaboliteLocalDesc : MetaboliteLibrary	1..* : 1	PartOf
MetaboliteLocalDesc : MetaboliteGlobalDesc	* : 1	CharacterisationOf

Attributes. The entities from the Data Processing component are described below through their attributes.

Table 3.5.3. Metabolite Identification

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for a metabolite identified in the sample analysed	VARCHAR(50)	✓	✓	
analysis_ID	A reference to the results of an analytical experiment which are processed	VARCHAR(50)		✓	✓
CAS	The CAS (Chemical Abstracts Service) registry number	VARCHAR(100)		✓	
confidence	A numerical description of the confidence with which the metabolite is identified	FLOAT			
concentration	The concentration calculated for the identified metabolite	FLOAT			

Table 3.5.4. Metabolite Library

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for a library of metabolites characterised by the results of their instrumental analysis	VARCHAR(50)	✓	✓	
description	An description of a library of metabolites characterised by the results of their instrumental analysis	TEXT			
comment	Additional comments about the library	TEXT			

Table 3.5.5. Metabolite Global Description

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	A local identifier for the metabolite	VARCHAR(50)	✓	✓	
CAS	The CAS (Chemical Abstracts Service) registry number for the given metabolite	VARCHAR(100)			
KEGG_ID	A link to an external database. The unique identifier in the KEGG COMPOUND database, which is a database of chemical compound structures that are relevant to biochemical reactions and pathways	CHAR(6)			
PubChem_SID	A link to an external database. PubChem Substance ID (SID) is the non-zero unsigned integer PubChem accession ID for a deposited substance	INT			
ChEBI_ID	A link to an external database. The unique identifier in the Chemical Entities of Biological Interest (ChEBI) database of molecular entities focused on 'small' chemical compounds	INT			
3DMET_ID	A link to an external database. The unique identifier in the 3DMET database of three-dimensional structures of natural metabolites	CHAR(6)			
name	Metabolite's name	VARCHAR(500)			
IUPAC_name	Calculated IUPAC preferred name variant, based on the 2005 standard, for the metabolite using the OpenEye, Inc.'s Ogham implementation and the IUPAC-style setting (see PubChem Substance documentation)	VARCHAR(500)			
OpenEye_name	Calculated IUPAC acceptable name variant, using standards earlier than 2005, for the metabolite using the OpenEye, Inc.'s Ogham implementation and the OpenEye-style setting (see PubChem Substance documentation)	VARCHAR(500)			
CAS_name	Calculated IUPAC name variant, typically calculated by other IUPAC naming applications, for the metabolite using the OpenEye, Inc.'s Ogham implementation and the Cas-style setting (see PubChem Substance documentation)	VARCHAR(500)			

systematic_name	Calculated IUPAC name variant, using a more systematic naming approach that attempts to be predictive of where the IUPAC naming conventions are headed, for the metabolite using the OpenEye, Inc.'s Ogham implementation and the Systematic-style setting (see PubChem Substance documentation)	VARCHAR(500)			
traditional_name	Calculated IUPAC name variant, using a more traditional name, for the metabolite using the OpenEye, Inc.'s Ogham implementation and the Traditional-style setting (see PubChem Substance documentation)	VARCHAR(500)			
synonyms	Other names of the given metabolite separated by the comma	TEXT			
mol_formula	Molecular formula for the metabolite	VARCHAR(500)			
mol_weight	Molecular weight for the metabolite	FLOAT			
can_smiles	Calculated canonical SMILES for the metabolite using the OpenEye, Inc.'s OEChem implementation (see PubChem Substance documentation)	VARCHAR(500)			
iso_smiles	Calculated isomeric SMILES for the metabolite using the OpenEye, Inc.'s OEChem implementation (see PubChem Substance documentation)	VARCHAR(500)			
NIST_INChI	Calculated INChI string for the metabolite using the NIST INChI implementation (see PubChem Substance documentation)	VARCHAR(500)			
comment	Additional comments about the metabolite	TEXT			

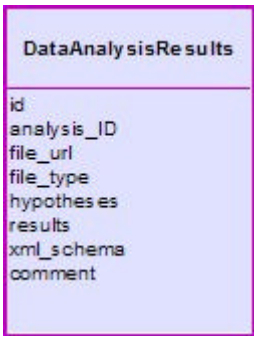
Table 3.5.6. Metabolite Local Description

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the local description of the metabolite	VARCHAR(50)	✓	✓	✓
metabolite_ID	A reference the global description of the metabolite	VARCHAR(50)	✓		✓
library_ID	A reference the local library of local metabolite descriptions to which the given description belongs	VARCHAR(50)	✓		✓
comment	Additional comments about the local description of the metabolite	TEXT			

3.6. Data Analysis

Class Diagram. The class diagram for the Data Analysis package is depicted in Figure 3.6.1.

Figure 3.6.1. The Data Analysis Component



Entities. The Data Analysis component consists of the following entities:

Table 3.5.1. Data Analysis Component Entities

Entity	Description	Occurrence
DataAnalysis Results	The result of computational data analysis	There will be a DataAnalysis Results entity for each data analysis experiment performed

Dependencies. The Data Analysis component is dependent upon the Admin, Analysis and XML Schemas components. These dependencies are described by the following relationships between entities:

Table 3.6.2. Data Analysis Component Dependencies

Entities	Multiplicity	Relationship
Experiment : DataAnalysis Results	1 : 0..1	Produces
ExperimentSet : DataAnalysis Results	1 : 1	InputOf

Attributes. The entities from the Data Analysis component are described below through their attributes.

Table 3.6.3. Data Analysis Results

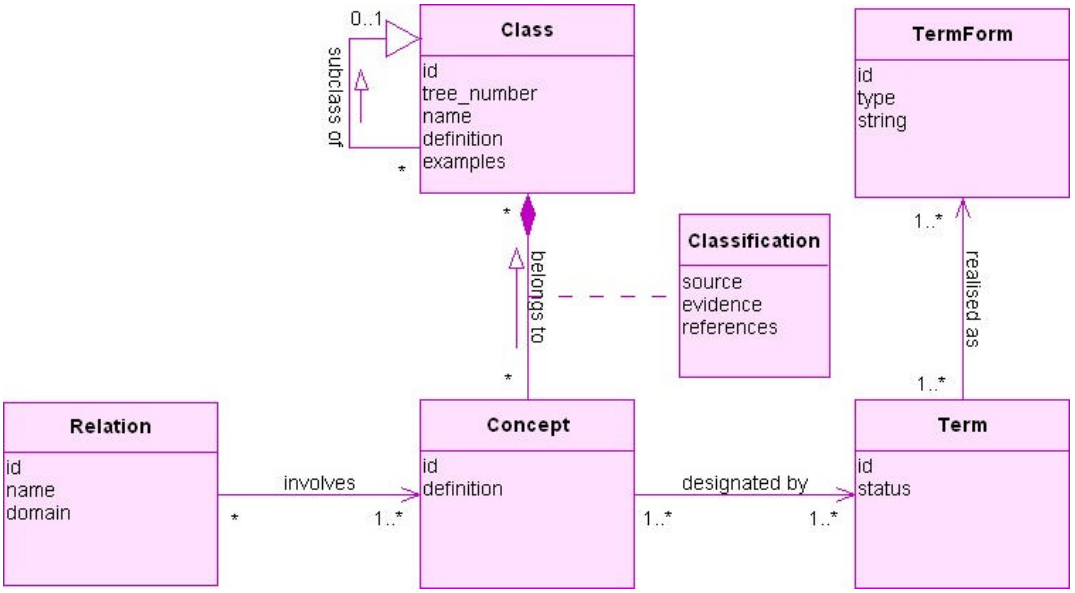
Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the results of a data analysis experiment	VARCHAR(50)	✓	✓	✓
analysis_ID	A reference to the results of an analytical experiment *set* that are processed within the data analysis experiment	VARCHAR(50)	✓		✓
file_url	A reference to the output file	VARCHAR(50)			
file_type	Type of the output file	VARCHAR(50)			

hypotheses	Hypotheses generated as a results of the data analysis experiment	TEXT			
results	The results of the data analysis experiment	TEXT			
xml_schema	A reference to an XML schema (if applicable or used) according to which the results field (above) is structured into an XML document	VARCHAR(50)			
comment	Comments about the results	TEXT			

3.7. Background Knowledge

Class Diagram. The class diagram for the Background Knowledge package is depicted in Figure 3.7.1.

Figure 3.7.1. The Background Knowledge Component



Note. The given schema suggests a global structure that can be used to accomodate the background knowledge. The extent and the organisation of this component depend on the actual implementation. For example, if one of the goals is functional classification of genes, then the Concept entity would be used to identify a gene, Term entity to store its name and aliases, while the classification hierearchy should incorporate that of gene functions. Also note that the background knowledge need not be part of the database per se, that is – it can provide the appropriate links within external resources (e.g. MIPS, SGD, GO, etc.).

Entities. The Background Knowledge component may consist of the following entities:

Table 3.7.1. Background Knowledge Component Entities

Entity	Description
Class	An abstraction of a set of related concepts, i.e. a description of concepts characterised by a uniform set of attributes
Concept	A domain-specific concept (e.g. a gene, protein, etc.)
Term	A name used to refer to a domain-specific concept

TermForm	Textual realisation of a term
Relation	A domain-specific relation (e.g. interaction)

Dependencies. The Background Knowledge component has no dependencies upon other components. Internal dependencies are described by the following relationships between entities:

Table 3.7.2. Background Knowledge Component Dependencies

Entities	Multiplicity	Relationship
Class : Class	* : 0..1	SubclassOf
Concept : Class	* : *	BelongsTo
Concept : Term	1..* : 1..*	DesignatedBy
Term : TermForm	1..* : 1..*	RealisedAs
Relation : Concept	* : 1..*	Involves

Attributes. The entities and relations from the Admin component are described below through their attributes.

Table 3.7.3. Class

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for a class of homogenous concepts	VARCHAR(50)	✓	✓	
tree_number	The number denoting the position of the class in the class hierarchy	VARCHAR(200)			
name	The name of the class	VARCHAR(200)			
definition	A short description of the class	TEXT			
examples	Several concepts from the given class	TEXT			

Table 3.7.4. Concept

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An internal identifier for a domain-specific concept	VARCHAR(50)	✓	✓	
definition	A short description of the given concept	TEXT			

Table 3.7.5. ConceptToClass

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
concept_ID	An identifier for a domain-specific concept	VARCHAR(50)	✓	✓	✓

class_ID	A class to which the concept belongs	TEXT	✓	✓	✓

Table 3.7.6. Term

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for a term denoting a domain-specific concept	VARCHAR(50)	✓	✓	✓
status	The status of the given term, i.e. if it is a standard name (S) or an alias (A)	CHAR(1)	✓		
concept_ID	A reference to the concept denoted by the given term	VARCHAR(50)	✓		✓

Table 3.7.7. TermForm

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for a term form	VARCHAR(50)	✓	✓	
term_ID	A reference to a term whose form is given	VARCHAR(50)	✓		✓
type	The type of a form (e.g. plural)	VARCHAR(50)	✓		
string	The actual term form	VARCHAR(200)	✓		

Table 3.7.8. Relation

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for a domain specific relation (in general, not as an instance)	VARCHAR(50)	✓	✓	✓
name	The name of the relation	VARCHAR(50)	✓		
domain	The arity of the relations and the types of concepts involved	TEXT			

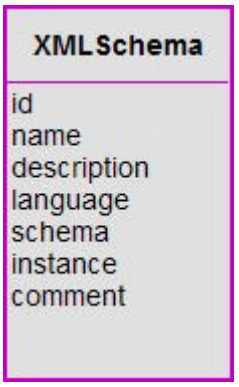
Table 3.7.8. RelationInstance

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for an instance of a domain specific relation	VARCHAR(50)	✓	✓	✓
concept_ID	A reference to a concept involved in the relation	VARCHAR(50)	✓	✓	✓

3.8. XML Schemas

Class Diagram. The class diagram for the XML Schemas component is depicted in 3.7.1.

Figure 3.8.1. The XML Schemas Component



Entities. The XML Schems component consists of the following entities:

Table 3.8.1. XML Schemas Component Entities

Entity	Description	Occurrence
xmlSchema	Description of an XML schema according to which some of the fields in other packages can be structured	There will be at least one xmlSchema entity for each field that can be further structured

Dependencies. The XML Schemas component has no dependencies on other components.

Attributes. The entities from the XML Schemas component are described below through their attributes.

Table 3.8.2. XML Schema

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
ID	An identifier for the XML schema	VARCHAR(50)	✓	✓	
name	The name of the XML schema	VARCHAR(50)	✓		
description	A description of the entity modelled by the schema	TEXT			
language	Language used to write the XML schema (e.g. XML Schema language, DTD)	VARCHAR(50)			
schema	The actual XML schema	TEXT	✓		
instance	An instance XML document illustrating the use of the given XML schema	TEXT			
comment	Comments about the XML schema	TEXT			

4. Implementation Notes

To facilitate the maintenance of the database each table is given a prefix corresponding to the component to which it belongs. The following prefixes are in use:

Table 4.1. Table Name Prefixes

Prefix	Component
ADM	Admin
BIO	Biological Sample
YST	Yeast Meta Data (subcomponent of Bio)
HMD	Human Meta Data (subcomponent of Bio)
SMP	Chemical Sample
ANL	Instrumental Analysis
DPR	Data Processing
DAN	Data Analysis
BK	Background Knowledge
XML	XML Schema
VOC	Controlled Vocabularies

5. Controlled Vocabularies

Controlled vocabularies are used to tag certain units of information, so that they are referred to in a consistent manner and therefore more easily retrieved by a search.

Table 5.1 A vocabulary table

Attribute	Description	Domain & Characteristics			
		Type	R	PK	FK
entry	An tag used for an entity, usually a short-hand annotation (e.g. GCMS)	VARCHAR(50)	✓	✓	
name	The full name of an entity (e.g. Gas Chromatography / Mass Spectrometry)	VARCHAR(50)	✓		
definition	A short description of an entity	TEXT			

Vocabularies. The following vocabulary tables are used:

Table 5.2. Vocabulary tables

Table	Description	Sample entries
Ethnicity	Ethnical background	White, Mixed, Indian
Experiment_type	Type of the experiments, e.g. biological sample prep, chemical sample prep, instrumental analysis, etc.	A, SP
Gender	Gender	M, F
Method	A global method upon which an experiment protocol, analytical instrument or software	GCMS, FTIR, MS

Table	Description	Sample entries
	is based (e.g. metabolic footprinting for biological sample preparation, GCMS for instrumental analysis, GP for data analysis, etc.)	
Mutation_type	Type of a gene mutation	knock out, knock in
Organism	Organism names	*Saccharomyces cerevisiae*

Dependencies. The following dependencies involve the vocabulary tables:

Table 5.3. Vocabulary uses

Table	Attribute	Vocabulary
ADM_ExperimentProtocol	method	Method
ADM_Program	methodology	Method
BIO_Mutation	mutation_type	Mutation_type
BIO_Population	organism	Organism
HMD_General	gender	Gender
HMD_General	ethnicity	Ethnicity