
ArMet sub-component design: GC-MS

Design Status: Draft Release

Version: Version 2.1

Date: 2nd of April, 2004

Author: Helen Fuell

Department of Computer Science
University of Wales
Aberystwyth
Ceredigion
SY23 3DB
UK

Copyright © 2004 by University of Wales, Aberystwyth

Table of Contents

1. Introduction	2
2. Datasets	2
3. The GC-MS Logical Data Model	2
3.1. Class Diagram.....	2
3.2. Entities	3
3.3. Relationships.....	4
3.4. Dependencies.....	4
3.5. Attributes	5
A. ArMet Data Types	8
Bibliography	10

1. Introduction

This document presents the design of a sub-component to the ArMet Metabolome Estimate component, [1], to support the processed results of metabolomics experiments performed using a GC-MS analytical instrument. The design is presented by way of a description of the types of datasets it aims to support and presentation of its logical data model.

2. Datasets

The four different analytical approaches that may be taken to perform a metabolomics experiment have been characterised in [2] as follows:

- **Targeted Analysis.** Detection and precise quantification of single or small set of target compounds within a metabolome sample.
- **Metabolite Profiling.** Detection and approximate quantification of a large set of target metabolites within a metabolome sample.
- **Metabolomics.** Detection, approximate quantification and tentative identification of as many of the compounds within a metabolome sample as possible.
- **Fingerprinting.** Generation of a signature for a metabolome sample without regard for the individual compounds that it contains.

A sub-component to support the processed results of a metabolome experiment carried out using a GC-MS analytical instrument should support:

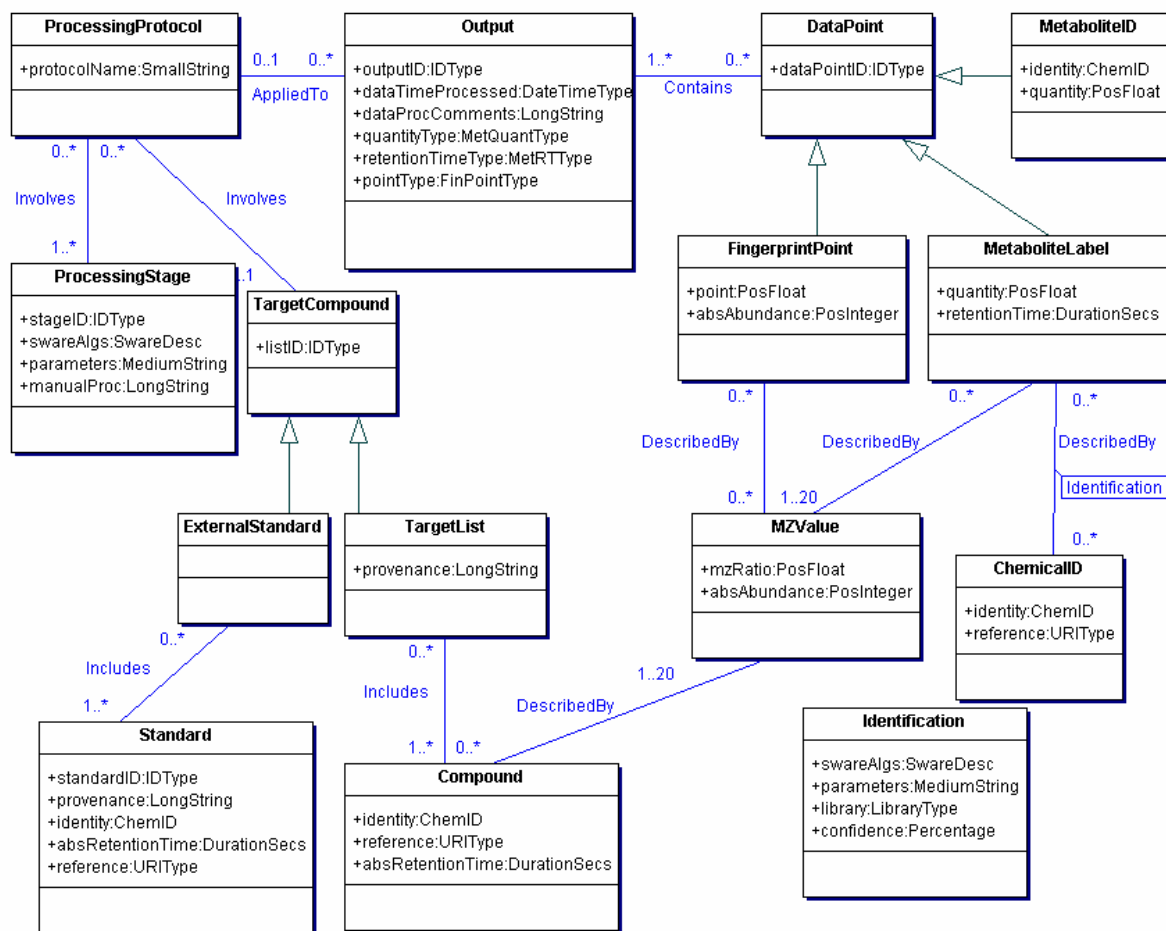
- Lists of identified metabolites together with associated data processing and target metabolite list metadata for targeted analysis and metabolite profiling experiments.
- Lists of detected metabolites each with one or more tentative identities and a representative mass spectrum together with data processing metadata for metabolomics experiments.
- Lists of datapoints within either summed mass spectra, TIC chromatograms or TIC chromatograms with associated mass spectra for fingerprinting experiments. Fingerprinting datasets may optionally be associated with data processing metadata.

3. The GC-MS Logical Data Model

3.1. Class Diagram

The class diagram for the GC-MS data model is depicted in Figure 1.

Figure 1. The GC-MS Data Model



3.2. Entities

The GC-MS data model contains the following entities:

Table 1. GC-MS Entities

Name	Description
ProcessingProtocol	A data processing protocol
ProcessingStage	A stage or step in a data processing protocol
TargetCompound	A list of target compounds for either a targeted analysis dataset or a metabolite profiling dataset
ExternalStandard	A TargetCompound that is for a targeted analysis dataset
Standard	An entry in a list of target compounds for a targeted analysis dataset
TargetList	A TargetCompound that is for a metabolite profiling dataset
Compound	An entry in a list of target compounds for a metabolite profiling dataset

Name	Description
Output	A set of processed results from a metabolomics experiment performed using a GC-MS analytical instrument
DataPoint	A data point in a set of processed results
MetaboliteID	A DataPoint in a targeted analysis or metabolite profiling dataset
MetaboliteLabel	A DataPoint in a metabolomics dataset
ChemicalID	A chemical identity for a data point in a metabolomics dataset
Identification	The provenance of a chemical identity for a data point in a metabolomics dataset
FingerprintPoint	A DataPoint in a fingerprinting dataset
MZValue	A spectral point from the mass spectrum for either a Compound in a metabolite profiling reference list, a data point in a fingerprinting dataset or a data point in a metabolomics dataset

3.3. Relationships

There are the following requirements and constraints on relationships:

- **TargetCompound:ProcessingProtocol.** Processing protocols that involve a TargetCompound should only be associated with datasets with MetaboliteID data points.
- **ExternalStandard:TargetCompound/TargetList:TargetCompound.** There is a mandatory, or constraint on the specialisation of a TargetCompound into an ExternalStandard or a TargetList.
- **ProcessingProtocol:Output.** Datasets with MetaboliteID or MetaboliteLabel data points must be associated with a ProcessingProtocol.
- **Output:DataPoint.** All DataPoints associated with the same Output should be of the same type, i.e. MetaboliteID, MetaboliteLabel or FingerprintPoint.
- **MetaboliteID:DataPoint/MetaboliteLabel:DataPoint/FingerprintPoint:DataPoint.** There is a mandatory, or constraint on the specialisation of a DataPoint into a MetaboliteID, MetaboliteLabel or FingerprintPoint.
- **MetaboliteLabel:DataPoint/FingerprintPoint:DataPoint.** A datapoint in a metabolomics dataset is defined by a MetaboliteLabel and its associated MZValues. Likewise a point in a fingerprinting dataset is defined by a FingerprintPoint and any associated MZValues.

3.4. Dependencies

The GC-MS data model is dependent upon the Admin component and the Instrumental Analysis component of ArMet. These dependencies may be described by way of the following relationships between classes:

Table 2. GC-MS Data Model ArMet Dependencies

Entities	Multiplicity	Relationship
User:Output	1..1:0..*	Processes
Run:Output	1..1:0..*	Produces

3.5. Attributes

The data items for the GC-MS data model are described below. Their data types (as given in Figure 1) are described in Appendix A.

Table 3. ProcessingProtocol Attributes

Attribute	Description
protocolName	A unique user given name for the data processing protocol.(Required/Primary Key)

Table 4. ProcessingStage Attributes

Attribute	Description
stageID	A unique identifier for the stage. (Required/Primary Key)
swareAlgs	An item of software or an algorithm for performing a stage of data processing. (Optional)
parameters	The parameters to the software or algorithm. (Optional)
manualProc	A description of the manual processing performed upon the results of the software/algorithms or used in place of the software/algorithms. (Optional)

Additional requirements and constraints. The following additional requirements and constraints exist for the ProcessingStage entity attributes.

- The five letter entity identifier that makes up part of stageID should be "stage".
- Values must be stored for either swareAlgs or manualProc or both.

Table 5. TargetCompound Attributes

Attribute	Description
listID	A unique identifier for the list of target compounds. (Required/Primary Key)

Additional requirements and constraints. The following additional requirements and constraints exist for the TargetCompound entity attributes:

- The five letter entity identifier that makes up part of listID should be "tlist".

Table 6. Standard Attributes

Attribute	Description
standardID	A unique identifier for the external standard. (Required/Primary Key)
provenance	A description of the generation of the standard. (Required)
identity	The chemical identity of the standard. (Required)

Attribute	Description
absRetentionTime	The absolute retention time of the standard. (Required)
reference	A reference to further information on the standard. (Required)

Additional requirements and constraints. The following additional requirements and constraints exist for the Standard entity attributes:

- The five letter entity identifier that makes up part of standardID should be "exstd".

Table 7. TargetList Attributes

Attribute	Description
provenance	A description of how the metabolite profiling reference list was created. (Required)

Table 8. Compound Attributes

Attribute	Description
identity	The chemical identity of the compound. (Required/Partial Primary Key)
reference	A reference to further information on the compound. (Optional)
absRetentionTime	The absolute retention time of the compound. (Required/Partial Primary Key)

Table 9. Output Attributes

Attribute	Description
outputID	A unique identifier for the dataset. (Required/Primary Key)
dateTimeProcessed	The date and time at which the dataset was produced. (Required)
dataProcComments	Any comments on the processing made by the data processor. (Optional)
quantityType	An indicator of absolute or relative metabolite quantities in MetaboliteID and MetaboliteLabel datasets. (Optional)
retentionTimeType	An indicator of absolute, relative or index values for retention time in MetaboliteLabel datasets. (Optional)
pointType	An indicator of the type of fingerprinting dataset. (Optional)

Additional requirements and constraints. The following additional requirements and constraints exist for the Output attributes:

- The five letter entity identifier that makes up part of outputID should be "dtast".
- The value of dateTimeProcessed must follow the values for dateTime in all associated runs.
- The quantityType attribute must only contain a value if the dataset comprises MetaboliteID or MetaboliteLabel data points.
- The retentionTimeType attribute must only contain a value if the dataset comprises MetaboliteLabel data points.
- The pointType attribute must only contain a value if the dataset comprises FingerprintPoint data points.

Table 10. DataPoint Attributes

Attribute	Description
dataPointID	A unique identifier for the data point. (Required/Primary Key)

Additional requirements and constraints. The following additional requirements and constraints exist for the DataPoint attributes:

- The five letter entity identifier that makes up part of dataPointID should be "datap".

Table 11. MetaboliteID Attributes

Attribute	Description
identity	The chemical identity of the metabolite. (Required)
quantity	The quantity of the metabolite detected. (Required)

Additional requirements and constraints. The following additional requirements and constraints exist for the MetaboliteID attributes:

- The value for identity should be present in the associated metabolite profiling reference list.

Table 12. FingerprintPoint Attributes

Attribute	Description
point	Either a scan number/retention time or an MZ value. (Required)
absAbundance	The abundance of ions measured for the point. (Required)

Table 13. MZValue Attributes

Attribute	Description
mzRatio	The mzRatio of a mass spectral point. (Required/Partial Primary Key)
absAbundance	The abundance of ions measured for a particular spectral point. (Required/Partial Primary Key)

Table 14. MetaboliteLabel Attributes

Attribute	Description
quantity	The quantity of the metabolite detected. (Required)
retentionTime	The retention time of the metabolite. (Required)

Table 15. ChemicalID Attributes

Attribute	Description
identity	A chemical identity. (Required/Partial Primary Key)
reference	A reference to further information on the chemical. (Required/Partial Primary Key)

Table 16. Identification Attributes

Attribute	Description
swareAlgs	The software or algorithm used to derive the chemical identity. (Required)
parameters	The parameters to the software or algorithm used to derive the chemical identity. (Optional)
library	The library used by the software or algorithm to lookup the chemical identity. (Optional)
confidence	The percentage confidence that the identity is the correct one. (Required)

A. ArMet Data Types

Table A-1. ArMet Data Types

Name	Format and Size	Allowable Values
SmallString	Non-empty sequence of up to 50 Unicode characters	
IDType	Non-empty sequence of up to 43 Unicode characters: A five letter entity identifier followed by a 38 digit unique number	
MediumString	Non-empty sequence of up to 100 Unicode characters	
LongString	Non-empty sequence	

Name	Format and Size	Allowable Values
	of up to 200 Unicode characters	
DurationSecs	A duration value that conforms to ISO 8601	
URIType	Non-empty sequence of up to 200 US-ASCII characters conforming to the Network Working Group RFC 2396	
DateTimeType	A date and time value that conforms to ISO 8601	
MetQuantType	Non-empty sequence of 8 Unicode characters	"relative", "absolute"
MetRTType	Non-empty sequence of up to 8 Unicode characters	"relative", "absolute", "indices"
FinPointType	Non-empty sequence of up to 3 Unicode characters	"TIC", "3-D", "MS"
FloatType	A number that conforms to the W3C Recommendation for XML Schema Datatypes lexical representation for floating point numbers: http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/	
PosFloat	See FloatType above	Positive values only
IntegerType	A number that conforms to the W3C Recommendation for XML Schema Datatypes lexical representation for integer numbers: http://www.w3.org/TR/2001/REC-xmlschema-2-20010502/	
PosInteger	See IntegerType above	Positive values only
Percentage	See FloatType above	Values in the range 0.0 to 100.0

Complex Types. The following types are complex and described using class diagrams

Figure A-1. SwareDesc



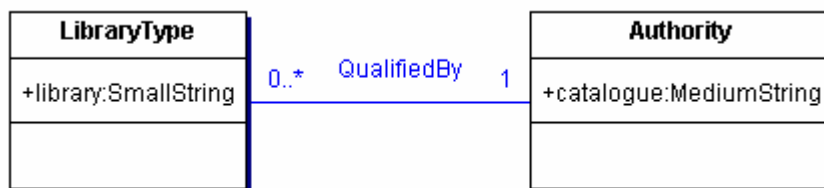
Allowable Values. Values from documented ontologies or controlled vocabularies of software items/algorithms for data processing.

Figure A-2. ChemID



Allowable Values. Values from documented ontologies of controlled vocabularies of chemical identities.

Figure A-3. LibraryType



Allowable Values. Values from documented ontologies or controlled vocabularies of chemical libraries.

Bibliography

- [1] Helen Fuell, *ArMet design*, Technical Report, University of Wales, Aberystwyth, June 2004.
- [2] Oliver Fiehn, *Metabolomics - the link between genotypes and phenotypes*, Plant Molecular Biology, **16**, 155-171, 2002.