



HEALTH-F4-2007-200754


[www.gen2phen.org](http://www.gen2phen.org)

## **D3.2 Development of High-Level Domain Model Version 1**

**WP3 – Standard data models and terminologies**


**V6.0  
Final Draft**

Lead beneficiary: EMBL  
Date: 13/02/2009  
Nature: Report  
Dissemination level: PU  
(Public)

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>	
	<b>Authors:</b> Tomasz Adamusiak, Juha Muiilu, Helen Parkinson	<b>Version:</b> v 6.0 Final draft	2/16

## TABLE OF CONTENTS

<b>DOCUMENT INFORMATION</b> .....	<b>3</b>
<b>DOCUMENT HISTORY</b> .....	<b>4</b>
<b>1. INTRODUCTION</b> .....	<b>4</b>
<b>2. DESCRIPTION OF WORK</b> .....	<b>6</b>
<b>3. DATA MODEL DEVELOPMENT AND EVALUATION SUMMARY</b> .....	<b>7</b>
<b>4. LSDB MODELLING</b> .....	<b>7</b>
4.1. FIGURE 1 - UML MODEL FOR LSDB RELATED DATA EXCHANGE TO ENSEMBL .....	9
4.2. FIGURE 2 - LRG UML MODEL .....	10
<b>5. DIAGNOSTIC LAB MODEL</b> .....	<b>11</b>
5.1. FIGURE 3 - DIAGNOSTIC LAB PROCESS DIAGRAM .....	11
5.2. FIGURE 4 - DIAGNOSTIC LAB DATA MODEL .....	12
<b>6. HIGH THROUGHPUT DATA MODEL DEVELOPMENT</b> .....	<b>12</b>
6.1. FIGURE 5 - GENERIC HTP DATA MODEL.....	13
6.2. FIGURE 6 - GWAS SUB DOMAIN MODEL .....	14
<b>7. MODEL VALIDATION</b> .....	<b>15</b>
<b>8. FUTURE PLANS</b> .....	<b>15</b>
8.1. SCOPE AND RANGE REQUIREMENTS OF SPECIALIZED DOMAIN MODELS. (D3.4) .....	15
8.2. A HIGH-LEVEL DOMAIN MODEL VERSION 2, WITH SAMPLE/PHENOTYPE FOCUS. (D3.5)	15
8.3. DERIVATION AND SPECIFICATION OF EXCHANGE FORMAT (D3.7).....	16
<b>REFERENCES</b> .....	<b>16</b>

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>		<b>Security:</b> PU
	<b>Authors:</b> Tomasz Adamusiak, Juha Muiilu, Helen Parkinson		<b>Version:</b> v 6.0 Final dfrac


## Document Information

<b>Grant Agreement Number</b>	HEALTH-F4-2007-200754	<b>Acronym</b>	GEN2PHEN
<b>Full title</b>	Genotype-To-Phenotype Databases: A Holistic Solution		
<b>Project URL</b>	<a href="http://www.gen2phen.org">http://www.gen2phen.org</a>		
<b>EU Project officer</b>	Frederick Marcus ( <a href="mailto:Frederick.Marcus@ec.europa.eu">Frederick.Marcus@ec.europa.eu</a> )		

<b>Deliverable</b>	<b>Number</b>	D3.2	<b>Title</b>	Development of High-Level Domain Model Version 1
<b>Work package</b>	<b>Number</b>	3	<b>Title</b>	WP3 – Standard data models and terminologies

<b>Delivery date</b>	<b>Contractual</b>	December 2008	<b>Actual</b>	February 2009
<b>Status</b>	Version 6.0		final <input checked="" type="checkbox"/>	
<b>Nature</b>	Report <input checked="" type="checkbox"/> Prototype <input type="checkbox"/> Other <input type="checkbox"/>			
<b>Dissemination Level</b>	Public <input checked="" type="checkbox"/> Confidential <input type="checkbox"/>			

<b>Authors (Partner)</b>	EMBL, UH.FGC		
<b>Responsible Author</b>	Helen Parkinson		<b>Email</b> parkinson@ebi.ac.uk
	<b>Partner</b>	EMBL-EBI	<b>Phone</b> +44 (0)1223 494 672

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>	
	<b>Authors:</b> Tomasz Adamusiak, Juha Muilu, Helen Parkinson	<b>Version:</b> v 6.0 Final dfrac	4/16

## Document History

Name	Date	Version	Description
T. Adamusiak	01/12/2008	1.0	Draft
H. Parkinson	23/01/2008	2.0	Revision
T. Adamusiak	28/01/2008	3.0	Revision
J. Muilu	28/01/2008	4.0	Revision
H. Parkinson	29/01/2009	5.0	Final draft
A. Ahonen-Bishopp, J.L. Oliveira, S. Heath	11/02/2009	6.0	Formal review

## Definitions

- Partners of the GEN2PHEN Consortium are referred to herein according to the following codes:

**ULEIC** – University of Leicester (UK) – Coordinator

**EMBL** – European Molecular Biology Laboratory (Germany) – Beneficiary

**FIMIM** – Fundació IMIM (Spain) – Beneficiary

**LUMC** – Leiden University Medical Center (Netherlands) – Beneficiary

**INSERM** – Institut National de la Santé et de la Recherche Médicale (France) – Beneficiary

**KI** – Karolinska Institutet (Sweden) – Beneficiary

**FORTH** – Foundation for Research and Tecnology Hellas (Greece) – Beneficiary

**CEA** – Commissariat à l’Energie Atomique (France) – Beneficiary

**EMC** – Erasmus Universitair Medisch Centrum Rotterdam (Netherlands) – Beneficiary

**UH.FGC** – Helsingin Yliopisto (Finland) – Beneficiary


**UAVR** – Universidade de Aveiro (Portugal) – Beneficiary

**UWC** – University of the Western Cape (South Africa) – Beneficiary

**CSIR** – Council of Scientific and Industrial Research (India) – Beneficiary

**SIB** – Swiss Institute of Bioinformatics (Switzerland) – Beneficiary

**UNIMAN** – The University of Manchester (UK) – Beneficiary

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>	
	<b>Authors:</b> Tomasz Adamusiak, Juha Muilu, Helen Parkinson	<b>Version:</b> v 6.0 Final draft	5/16


**BIOBASE** – BioBase GmbH. (Germany) – Beneficiary

**deCODE** – Islensk Erfoagreining EH (Iceland) – Beneficiary

**PHENO** – Phenosystems S.A. (Belgium) – Beneficiary

**BCP** – Biocomputing Platforms Ltd. Oy (Finland) – Beneficiary

- **Grant Agreement:** The agreement signed between the beneficiaries and the European Commission for the undertaking of the GEN2PHEN project (HEALTH-200754).
- **Project:** The sum of all activities carried out in the framework of the Grant Agreement by the Consortium.
- **Work plan:** Schedule of tasks, deliverables, efforts, dates and responsibilities corresponding to the work to be carried out for the GEN2PHEN project, as specified in Annex I to the Grant Agreement.
- **Consortium:** The GEN2PHEN Consortium, conformed by the above-mentioned legal entities.
- **Consortium agreement:** agreement concluded amongst GEN2PHEN participants for the implementation of the Grant Agreement. Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the Grant Agreement.

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>	
	<b>Authors:</b> Tomasz Adamusiak, Juha Muiilu, Helen Parkinson	<b>Version:</b> v 6.0 Final draft	6/16

## 1. INTRODUCTION

Work package 3 ‘Standard data models and terminologies’ provides domain standards to develop GEN2PHEN specific architecture, facilitate data exchange, and integrate data across existing and emerging resources. This work package is focused on providing standards to act as the foundation for much of the database development activities of other work packages.

The work package objectives include the rapid development of a standard data model(s) capable of representing the minimum agreed content standard (as determined by WP2) and a derived data exchange format. Data models developed in coordination with WP3 will have several uses in GEN2PHEN: data from pre-existing databases will be mapped to generate data in a derived data exchange format, thus offering a flexible solution for integrating and exchanging existing and new data. In this respect, data model development is a necessary prerequisite, initially separated from implementation details.


## 2. DESCRIPTION OF WORK

The focus of the GEN2PHEN High-Level Domain Model Version 1 development process is:

- To understand the process of object modelling in the biomedical domain
- To share previous experience of object modelling in the biomedical domain
- To evaluate relevant public domain models
- To harmonize and unify specific domain models
- To support GEN2PHEN use cases
- To develop a sub-domain object model for LSDBs (Locus Specific Databases) from which to derive and validate data exchange formats
- To develop a sub-domain object model for high throughput data

The first GEN2PHEN Modelling Workshop (Hinxton 9-11 April 2008) was hosted by EMBL with objectives to share previous experience of object modelling in the biomedical domain and to begin the evaluation of relevant public domain models and prioritise use cases required by GEN2PHEN partners. The results of this workshop are summarised in this report, complete proceedings are appended in the Appendix 1.

Subsequent work was continued during the second GEN2PHEN Modelling Workshop (Helsinki 19-22.1.2009) hosted by UH.FGC. Use cases were gathered and models were developed and minimum content standards to be used in exchanging data between partners were discussed and used as the basis for building and evaluating sub-domain models. See Appendix 2 for workshop proceedings.

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>	
	<b>Authors:</b> Tomasz Adamusiak, Juha Muiilu, Helen Parkinson	<b>Version:</b> v 6.0 Final draft	7/16

### 3. Data model development and evaluation summary

Several public data models<sup>1</sup> currently exist for genotype-phenotype association data and these were evaluated for relevance, domain coverage compared to existing resources, ease of use, and complexity during the First Modelling Workshop. The most relevant model in this domain to date is the OMG (Object Management Group) standard model – PAGE-OM (Phenotype and Genotype Experiment <http://www.pageom.org>). PAGE-OM is a complete reference model that describes genotype data on the level of an individual and of a cohort summary statistics. It also represents LSDB-type data, phenotype data, and supports some legacy technology use cases. PAGE-OM is very detailed and is useful as a reference model; meaning that GEN2PHEN specific models can be aligned to it and it can be used as a meta-mapping model for mapping external data representations. As GEN2PHEN partners already have legacy implementations, the focus on modelling within the project to date has been dedicated to supporting data exchange between existing applications, and therefore simpler use case specific models, from which to derive new exchange formats and map existing ones, have been produced in the first year.


Modelling efforts in the Consortium have been separated into two domains: LSDBs including diagnostic lab requirements, and representation of data for UHT (ultra high-throughput technologies) used in e.g. GWAS (Genome Wide Association Studies) using arrays, HTP (high-throughput) sequencing, etc. This separation is further enforced by the fundamental differences in the needs of these two fields: LSDBs gather information about rare diseases, and focus on precise loci in the genome, whilst GWAS typically gather data about very common diseases, and explore variations in the whole genome. Furthermore it allows us to build on the work of WP2 and WP3 in developing minimum standards for LSDB data exchange, and the LRG (Locus Reference Genomic Sequence) format specification for standard reference sequences needed to consistently report reference sequences. Two core data models and associated sub-domain models have therefore been produced and will be used in the next phase of modelling for format development for data exchange purposes.

### 4. LSDB modelling

An LSDB data model was produced during the Second Modelling Workshop based on the discussion of ‘LSDB minimal content’ document developed by LUMC Partner and HGVS (Human Genome Variation Society <http://www.hgvs.org>). This model was evaluated for use cases related to data exchange between central repositories such as Ensembl (<http://www.ensembl.org>) and LSDBs in light of the minimum requirements document. The discussion on the ‘LSDB minimal content’ document is recorded on the GEN2PHEN project website and is available from

<http://askja.gene.le.ac.uk/drupal5/content/lsdb-minimal-requirements>

<sup>1</sup> Some of the data models have been documented at [www.schemalet.org](http://www.schemalet.org), which is an experimental wiki site for documenting use case specific data models.

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>		<b>Security:</b> PU
	<b>Authors:</b> Tomasz Adamusiak, Juha Muilu, Helen Parkinson		<b>Version:</b> v 6.0 Final draft

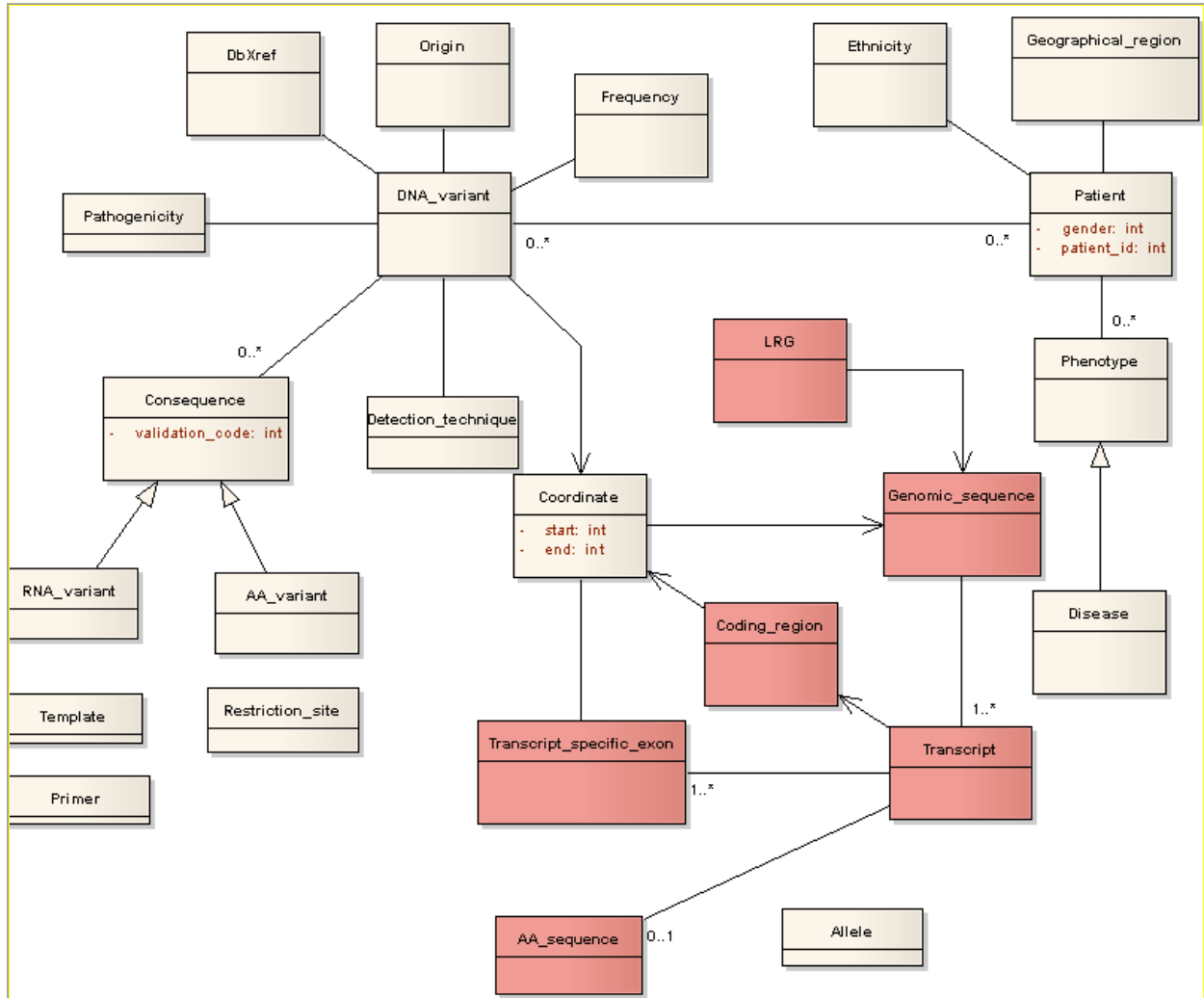
Use cases for this model were discussed by GEN2PHEN Partners with specific interest in this domain:

- Data exchange between different LSDBs e.g. data for the same gene(s) or disease(s)
- Data exchange/access between LSDBs and a central repository e.g. Ensembl
- Using an LRG in the context of Ensembl and LSDB data exchange

These use cases were used to derive the UML model shown in Figure 1. As use cases in this domain require the use of an LRG standard (for which a specification has been produced already) a UML model has been created for the LRG. Subsequently respective use cases and the LRG model were aligned. The LRG model is shown in Figure 2. Some minor modifications were suggested to the LRG schema during the Helsinki workshop based on additional use cases, and these will be discussed with the NCBI who are also developing this standard. The additional use cases are documented below:

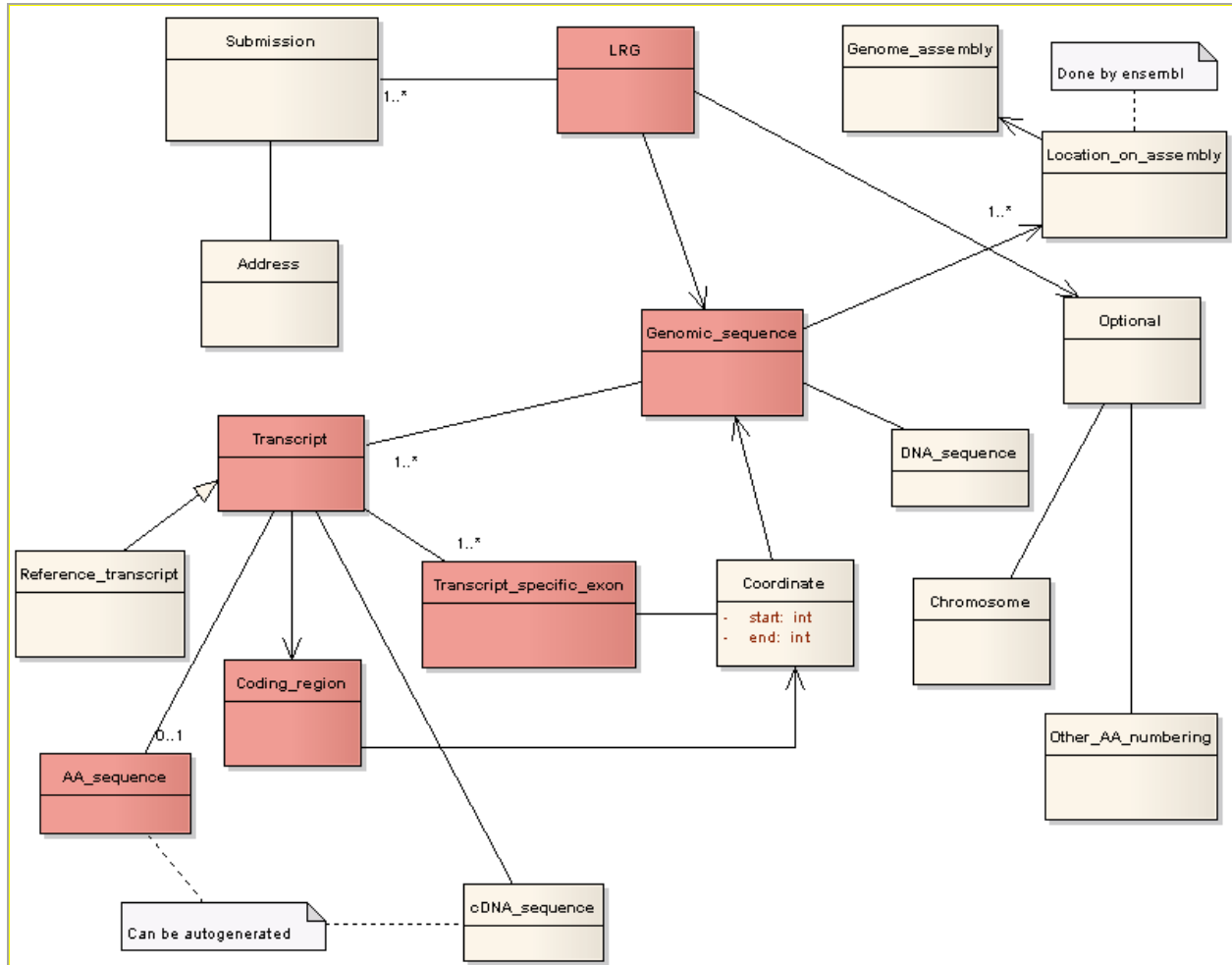
- There are cases where LSDB infrastructure exists, but no active curation is taking place. Therefore an ability to comment in LRG records is needed to express this information.
- Cross referencing of LSDBs from LRG records can be one LRG to many LSDBs. In this case a way to resolve the LSDB names, ids and URIs is needed.
- LRG standard may be versioned, hence the XML format needs a version attribute for downstream parsers.
- URI formatting recommendations are needed for, LSDBs references. Also a cardinality change may be needed as some LSDBs may have multiple URIs for the same gene.
- cDNA coordinates are commonly used by LSDBs and diagnostic labs. These can be calculated and included in the LRG format by default.
- Multiple database cross references are needed in case of e.g. Genbank and Uniprot. For example <protein id> </protein id> could be used for multiple databases and so an ability to cite multiple databases consistently is needed.
- There should be a possibility to handle multiple transcripts for a single LRG. There should be at least one ‘idealized transcript’, which can be used to represent the most common or longest splice form. The LRG author is responsible for assigning the “idealized transcript”. Each transcript has its own exon numbering system (transcript specific exon).
- The LRG is gene centric. Other functional elements like non-coding RNAs and transcription factors should be investigated as well.

#### 4.1. Figure 1 - UML model for LSDB related data exchange to Ensembl



This model was aligned to the draft of the ‘LSDB minimal content’ document (produced by LUMC) by cross referencing the minimal information elements to the UML objects. The LRG specification is important for the data exchange among Ensembl and LSDBs and the model supporting the schema is therefore shown separately in Figure 2. These two models have been aligned and future changes will be made to the LRG model based on future versions of the schema (see use cases above).


#### 4.2. Figure 2 - LRG UML model



Both models are also available in Enterprise Architect format from

<http://askja.gene.le.ac.uk/drupal5/filemanager/active?fid=47>

Classes shared between the two diagrams are denoted in red. The LRG is an integral part of data flow between the LSBDs and Ensembl and all variants will be coded in context of LRG. DNA\_Variants are positioned and annotated in context of LRG. Exon assignments are related to a specific transcript, which is referenced by using a transcript identifier (this parametric association is not shown in model) and these have unique numbering. DNA\_Variant has zero or many consequences on the amino acid and/or RNA level. This is coded using standard HGVS nomenclature, which is based on the amino acid or RNA sequence stored into the LRG object. DNA\_Variant can be associated to zero or many Patients and each Patient can have zero or more DNA\_Variants.

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>	
	<b>Authors:</b> Tomasz Adamusiak, Juha Muilu, Helen Parkinson	<b>Version:</b> v 6.0 Final dfrac	11/16

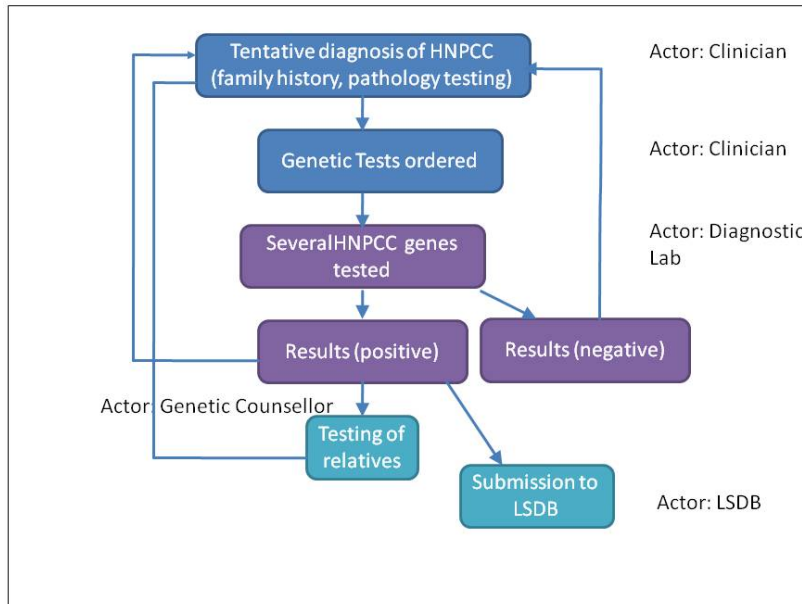
Patient has information on Phenotype (Disease), Ethnicity and Geographical\_region. Annotation should be based on specific vocabulary (e.g. ISO standard for Geographical\_location and Ethnicity).

## 5. Diagnostic lab model

Diagnostic Labs are typically embedded in healthcare systems and are both users of locus specific databases and potential suppliers of data to LSBDs. The WP1 Pilot project on Lynch Disease was examined to derive a typical data flow diagram (Figure 3) and a data model (Figure 4). The following use cases were considered when designing the model:

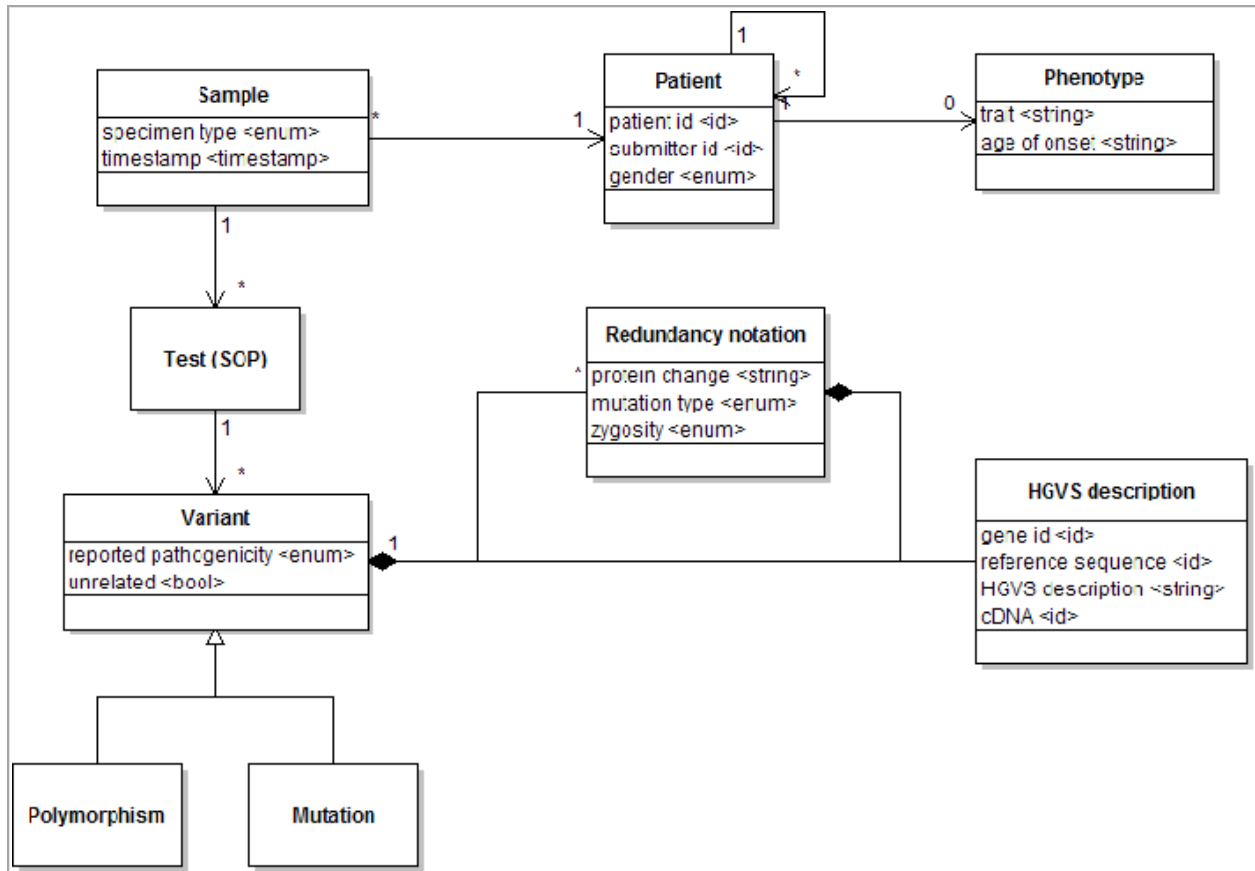
- Data exchange between diagnostic labs
- Data representation inside a diagnostic lab
- Data access from an LSBDB by a diagnostic lab
- Data access from a central repository e.g. Ensembl by a diagnostic lab

### 5.1. Figure 3 - Diagnostic lab process diagram



The diagnostic process flow highlights the importance of the clinical staff, the referral to the diagnostic lab and the interaction with LSBDBs. As specific healthcare roles are taken by different actors, these have been included in the process diagram to highlight data flow between different levels of the process hierarchy.


## 5.2. Figure 4 - Diagnostic lab data model



The diagnostic lab model has an emphasis on patients, the relationship between patients' phenotypes and genetic variants relationship to the disease which is suspected. It should be noted that only an ethically approved subset of the data could ever be exchanged and the model does not reflect which data could be submitted in this context, as this is subject to local ethical approval. This model is distinct from other GEN2PHEN modelling work to date and represents a special case.

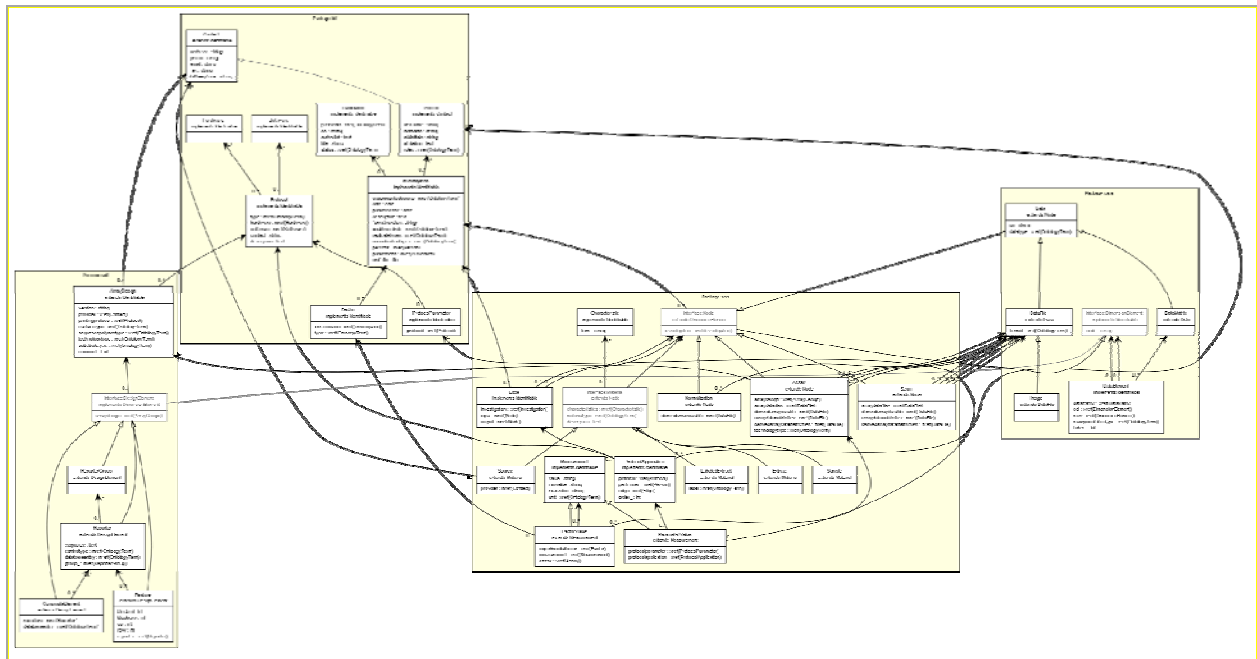
## 6. High throughput data model development

In order to comprehend the requirements for representing high throughput studies a specific use case relating to the representation of a GWAS based on genotyping chips was discussed. The aim of this discussion was to produce a detailed example to use in validation of available formats in use by GEN2PHEN Partners and external projects. The most commonly used format to date for such dataset is MAGE-TAB (MicroArray Gene Expression Tabular) [1] and in addition to the

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>	
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>
	<b>Authors:</b> Tomasz Adamusiak, Juha Muiilu, Helen Parkinson	<b>Version:</b> v 6.0 Final dfrac 13/16

specific model for GWAS described here, a test implementation for GWAS exists, and a generic model has been described. We expect to extend this model in future WP3 activities.

### 6.1. Figure 5 - Generic HTP data model




The MAGE-TAB object model was designed to contain the contents of files parsed in MAGE-TAB v1.0 and v1.1 format. This data exchange format is currently in use by ArrayExpress (<http://www.ebi.ac.uk/microarray-as/ae>), caBIG (cancer Biomedical Informatics Grid <http://cabig.cancer.gov>), TGCA (The Cancer Genome Atlas <http://cancergenome.nih.gov>), EGA (European Genotype Archive <http://www.ebi.ac.uk/ega>), and ENGAGE (European Network of Genomic and Genetic Epidemiology <http://www.euengage.org>). Both EGA and ENGAGE are actively involved in evaluating and extending this format.

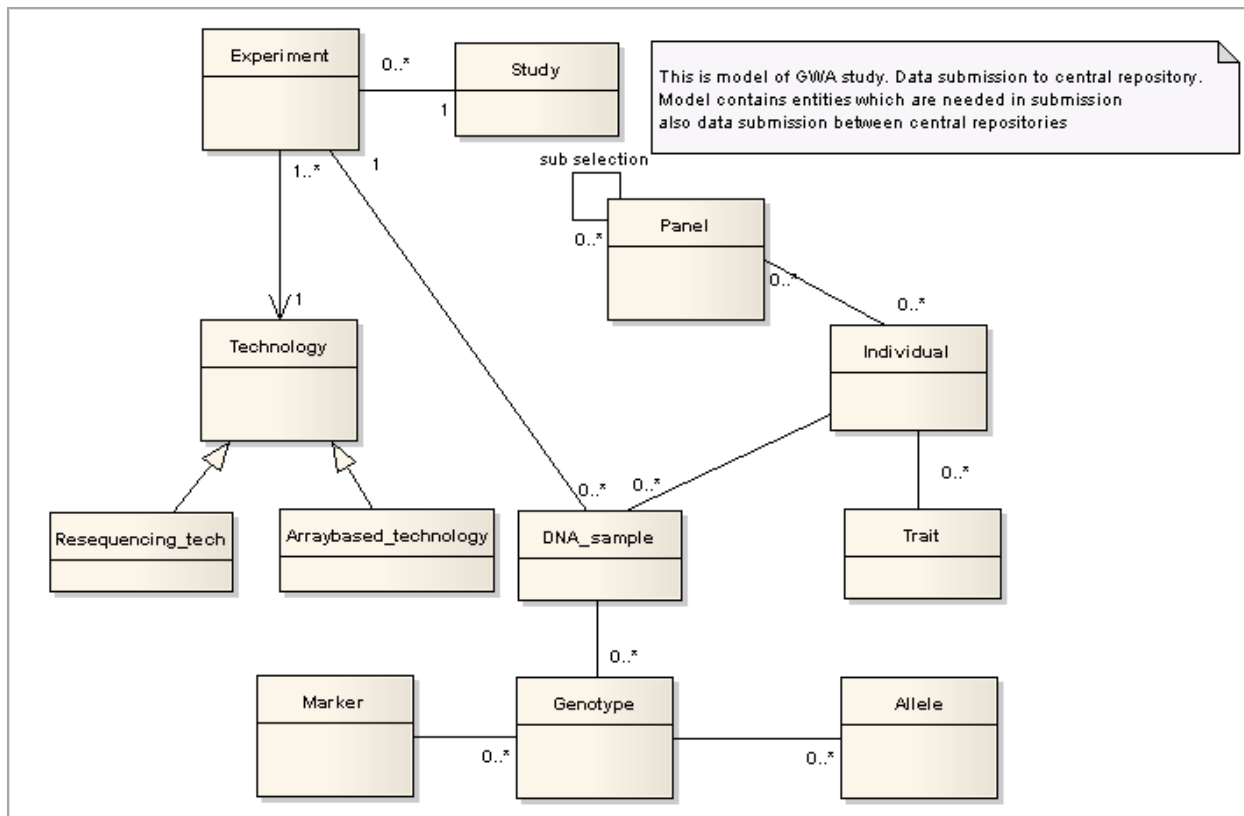
The motivation for developing the test implementation was primarily to integrate MAGE-TAB format data sets from existing resources into one simple system, an open source MAGE-TAB database with import/export tools capable of reading in and outputting MAGE-TAB formatted data. This proof of principle was created using the MOLGENIS (Molecular Genetics Information System) platform [2]. MAGE-TAB object model could be seen as an implementation model in contrast to PAGE-OM, which is considered a reference model by the GEN2PHEN Consortium.

Full documentation of the model is available from <http://magetab-om.sourceforge.net> and also in the Appendix 3.

The model was validated with available dbGaP (database of Genotypes and Phenotypes <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap>) data and the implementation is available from <http://magetab-om.sourceforge.net/molgenis4magetab.html>

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>		<b>Security: PU</b>
	<b>Authors:</b> Tomasz Adamusiak, Juha Muilu, Helen Parkinson		<b>Version:</b> v 6.0 Final dfrac


## 6.2. Figure 6 - GWAS sub domain model



Also available in Enterprise Architect format from

<http://askja.gene.le.ac.uk/drupal5/filemanager/active?fid=46>

This model contains specific elements required for GWAS only. The GWAS model has been used to validate the aforementioned generic MAGE-TAB model, which can be additionally extended. An important new feature in this model is that sample (DNA\_Sample) is specific to experiment (as compared to PAGE-OM for example). This option fits the implementation cases where there are multiple high throughput measurements using the same sample using same technology and experimental settings. However, this implementation has a normalization problem, because there can be multiple measurements using the same sample but using different technologies. This can be fixed by adding a separate 'source sample' class between the DNA\_sample and Individual. Similar thinking applies to the marker side of measurement as well. In case there are multiple high throughput measurements done from thousands of different samples using same marker assay settings, then a separate assay level is needed between the marker and genotype measurement (this option is available in PAGE-OM, where Genomic\_variation\_assay provides assay specific link between the marker (PAGE-OM

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>	
	<b>Authors:</b> Tomasz Adamusiak, Juha Muiilu, Helen Parkinson	<b>Version:</b> v 6.0 Final draft	15/16

Genomic\_variation) and genotype (Assayed\_genomic\_genotype). These issues will be discussed in the next round of modelling.

## 7. Model validation

It is difficult to assess object model usability without a working implementation. For this we have used MOLGENIS [2]. MOLGENIS is an open source software platform to efficiently design, implement, and autogenerate web applications from object models. Its power is in the use of models and generators so the best solutions are easily reused between applications. MOLGENIS in one simple step generates a database (MySQL or PostgreSQL), a web-based GUI, programmatic interfaces including Java API, SOAP web services usable in tools like Taverna (<http://taverna.sourceforge.net>) and by statistical scripts written in the R language (<http://www.r-project.org>), as well as a full documentation of the object model. Several Java plug-in mechanisms are also available to customize the generated software.

MOLGENIS has been successfully used by:

1. MAGE-TAB OM:  
<http://magetab-om.sourceforge.net>
2. LSDB object model developed in the course of the Second Modelling Workshop:  
[http://magetab-om.sourceforge.net/lsdb/1.0/object\\_model.html](http://magetab-om.sourceforge.net/lsdb/1.0/object_model.html)
3. An example LSDB - Findis, the Finnish National Mutation Database (NMDB):  
<http://www.schemalet.org/mediawiki/index.php/FINDIS:Database>


## 8. FUTURE PLANS

### 8.1. Scope and Range Requirements of Specialized Domain Models. (D3.4)

The second phase of domain modelling will focus on complex representations of Phenotype which have not yet been addressed in detail. This work will include the model organism community, detailed examination of relevant datasets and more use case gathering from GEN2PHEN partners on their requirements for representing phenotype. This work commenced in the second modelling workshop Jan 2009.

### 8.2. A High-Level Domain Model Version 2, with Sample/Phenotype Focus. (D3.5)

The current core use cases do not address requirements for Phenotype. A survey will recruit information from partners and a workshop in mid-2009 will identify existing models and tools within Gen2Phen and in the public domain. These use cases are likely to be complex, to support the needs of multiple different recording systems and to support longitudinal measurements. This work will be accomplished with interaction with external projects, specifically CASIMIR (mouse phenotyping) and BBRMI (phenotyping for biobanks).

 HEALTH-200754	<b>Development of High-Level Domain Model Version 1</b>		
	<b>WP3 – Standard data models and terminologies</b>	<b>Security: PU</b>	
	<b>Authors:</b> Tomasz Adamusiak, Juha Muiilu, Helen Parkinson	<b>Version:</b> v 6.0 Final draft	16/16

### 8.3. Derivation and Specification of Exchange Format (D3.7)

The priorities for data formats in GEN2PHEN are the data exchange between locus specific databases and central repositories and HTP data. The modelling work to date has separated these domains to support immediate needs for data exchange.

Validation of LSBD data model will commence in 2009 by working with the existing LSDBs inside and outside the GEN2PHEN consortium, most of who have existing data formats. The model reported here will be used to align formats and as the basis for the derived formats.

Validation of the MAGE-TAB OM is underway and progress is promising. We envisage that the Phenotypic descriptors, e.g. membership of a cohort through a shared phenotype, or trait will require an extension of MAGE-TAB, and the requirement to provide details of markers in context of HTP data will also require an extension.

## REFERENCES

1. Rayner TF, Rocca-Serra P et al., A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB, BMC Bioinformatics. 2006 Nov 6;7:489
2. Swertz MA et al., Molecular Genetics Information System (MOLGENIS): alternatives in developing local experimental genomics databases, Bioinformatics. 2004 Sep 1;20(13):2075-83. Epub 2004 Apr 1