BioMedBridges

Deliverable D3.2

Project Title:	Building data bridges between biological and medical infrastructures in Europe
Project Acronym:	BioMedBridges
Grant agreement no.:	284209
	Research Infrastructures, FP7 Capacities Specific Programme; [INFRA-2011-2.3.2.] “Implementation of common solutions for a cluster of ESFRI infrastructures in the field of "Life sciences"
Deliverable title:	Mapping and registry of ESFRI BMS standards (eSTR)
WP No.	3
Lead Beneficiary:	19: UMCG
WP Title	ESFRI BMS Standards Description and Harmonization
Contractual delivery date:	31 December 2014
Actual delivery date:	19 December 2014
WP leader:	Helen Parkinson (EMBL) and Morris Swertz (UMCG)
Contributing partner(s):	19: UMCG, 16: UH, 10: TMF, 11: HMGU, 20: CIRMMP, 3: KI, 4: STFC, 5: UDUS, 9: ErasmusMC

Authors and contributors: Julie McMurry, Helen Parkinson, Philip Gormanns, Juha Muilu, Murat Sariyar, Morris Swertz, Dennis Hendriksen, Fleur Kelpin, Jonathan Jetten, Chao Pang

Contents

1 EXECUTIVE SUMMARY 3
2 PROJECT OBJECTIVES 5
3 DETAILED REPORT ON THE DELIVERABLE 5
3.1 Metadata Model and Mappings Registry overview 5
3.2 Registry meta-model 7
3.4 Upload formats 8
3.5 Standards registry content 10
3.5.1 Overview of registration process 10
3.5.2 Overview of content 10
3.6 Sustainability 11
3.6.1 Content upkeep strategy 11
3.6.2 Software upkeep strategy 12
3.6.3 Hosting and integration strategy 13
3.7 User interfaces 13
3.7.1 Search 13
3.7.2 Search results 14
3.7.3 Model details 15
3.7.4 Visualisation 16
3.7.5 Upload 16
3.8 Future work 18
3.8.1 Enable end-users to contribute models 18
3.8.2 Expansion of contents 19
3.8.3 Mapping tool 19
3.8.4 Evaluation and Metrics 20
3.8.5 Integration with other registries 21
4 DELIVERY AND SCHEDULE 21
5 ADJUSTMENTS MADE 21
6 BACKGROUND INFORMATION 22
APPENDIX 1: OUTREACH AND DOCUMENTATION 27

Figures

Figure 1 Example of the meta-model usages the distributed annotation system, DAS . 7

Figure 2 Search interface for the Metadata Model and mapping registry 14

Figure 3 Example search results with records for the VCF and Chado standards 15

Figure 4 Model details showing record for the magetab gene expression standard ... 15

Figure 5 Example UML representation of a standard 16

Figure 6 Screenshot demonstrating the upload features using an XLS file 17

Figure 7 Intuitive feedback and validation information on completion of upload 18

Figure 8 Mapping results of measurement types vs. several available studies 20

Figure 9 UML diagram of the meta-model..... 28

Tables

Table 1 Current contents of the Metadata Models and Mapping Registry..... 10

1 Executive Summary

The development of a prototype data model registry is the objective of BMB Deliverable 3.2, with contributions from BMB partners and in collaboration with BBMRI. The overall aim is to promote FAIR principles for data (Find, Access, Integrate and Reuse)1, therefore the Meta Models and Mappings Registry is designed to make it easier for researchers, data stewards and tools producers to find, compare, and choose existing data models, formats, and guidelines, and in particular to promote the use of (de facto) standards. In contrast to currently fragmented resources where users typically need to manually explore very technical documentation, the Meta Models and Mappings Registry we have delivered, referred to throughout this document as the MMMR or simply as 'the registry', can quickly shortlist suitable data elements, entities and models using a simple Google-like search. The registry currently catalogues a representative collection of meta-data artefacts (models, formats, minimal information guidelines, biobank data dictionaries, detailed clinical/research models) of use to RIs/communities within BioMedBridges. Next to the description of model entities and attributes, the registry includes provenance details, links to relevant publications and key contact information. To achieve the objectives, it was necessary to develop a minimum information model to describe meta-models, building on existing open source standards and software as well as (meta) data cataloguing efforts in BBMRI. The registry content and software are open access and open source in order to further facilitate reuse and community participation. In the following report we summarise technical progress and outcomes of our work.

[1] http://www.datafairport.org

2 Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following:

No.	Objective	Yes	No
1	Provision and use of the ESFRI BMS common molecular identifiers (eCMI)		x
2	Identification, harmonization and integration of ESFRI BMS partner standards	x	
3	Provision of standards and harmonized elements in an accessible standards registry (eSTR)	x	
4	Provision and population of the ESFRI BMS Service Registry (eSR)	x	

3 Detailed report on the deliverable

3.1 Metadata Model and Mappings Registry overview

We built five key components in the Metadata Model and Mappings Registry:

1. **Meta-data model** – framework to capture detailed descriptions of the structural elements of standards models, formats, templates, data dictionaries, and guidelines, i.e., details on all model elements (entities, attributes and relationships including ontology annotations to promote data integration) while linking to more general information from existing registries such as biosharing.org and identifiers.org when available.
2. **Content** – i.e. actual meta-data on standard and individual models, guidelines and formats (with structure following the model above), with emphasis on broadly used resources. To demonstrate the system we have loaded 10 representative models having 379 entities and 3356 attributes. These were selected in consultation with the project partners

and represent their requirements and provide broad coverage of the standards domain.

3. **Registration mechanisms** – systems to enter meta-data via user interface, via batch upload using Excel, tab-delimited formats, or programmatically via REST. Upload of UML diagrams (using XMI format) and semantic web models (OWL files) are under development.
4. **Query interfaces** – alternative views to enable human users to search across all collected meta-data, drill down to the details of individual elements; and print the models. To facilitate integration we provide programmatic interfaces via REST/JSON.
5. **Mappings** – a system to view and curate entity/attribute mappings, including a computer aided tool to create new mappings using lexical/semantic matching (in-kind contribution BBMRI/BioSHaRE) is in beta. This can be used to map between related models, but also to find a suitable standard for an ad-hoc uploaded data sheet.
6. **Interoperability** – to ensure future integration into existing registries we organized a workshop[2]. To enable future integration we use the BioSharing[3] IDs to unambiguously identify model entries, and are in the processes of sharing ontologies (such as EDAM) between the registries, cross-link with the tools and service registry and are planning to annotate identifiers with links to Identifiers.org[4].

The registry content is open access and can freely be repurposed; accordingly, members of the community may create their own interfaces tailored to their specific needs, use the Software as a Service to host consortium standardization efforts or clone the system for internal use. To promote future maintenance, the software is open source, integrated in a long-existing open source software project (MOLGENIS[5]) which is related to BBMRI data cataloguing efforts and comes with federation REST interfaces to facilitate syndication with related efforts.

---

[2] Registries integration workshop documentation can be found at http://www.biomedbridges.eu/trainings/knowledge-exchange-workshop-resource-integration
[3] http://biosharing.org
[4] http://identifiers.org
[5] Swertz et al (2010)

3.2 Registry meta-model

A meta-model (or schema) for the data-model registry *itself* was refined based on the content from the previously fragmented structured (e.g. XMI, ODM) and unstructured (e.g. PDF files) sources, such as can be found when following links in for example biosharing.org. Importantly, the meta-model was designed such most cases can be easily covered (in contrast to more bespoke meta-models that are too complicated for non-technical users to fully understand). The meta-model was developed building on the open source MOLGENIS software platform to enable rapid prototyping of the registry.

The meta-model is organized in two core groups. The first group of meta-data relates to the structural information about the meta-model, i.e., its sub-components, entities, attributes and their relations:

1. **(Sub)packages,** e.g., name, short description and tags that classify the model/format/guideline
2. **Entities**, e.g., name, short description and tags of the main and sub-components of the standards, e.g., classes, tables, file types, etc.
3. **Attributes**, i.e., name, short description, technical type, tags, and technical attributes (required, optional)

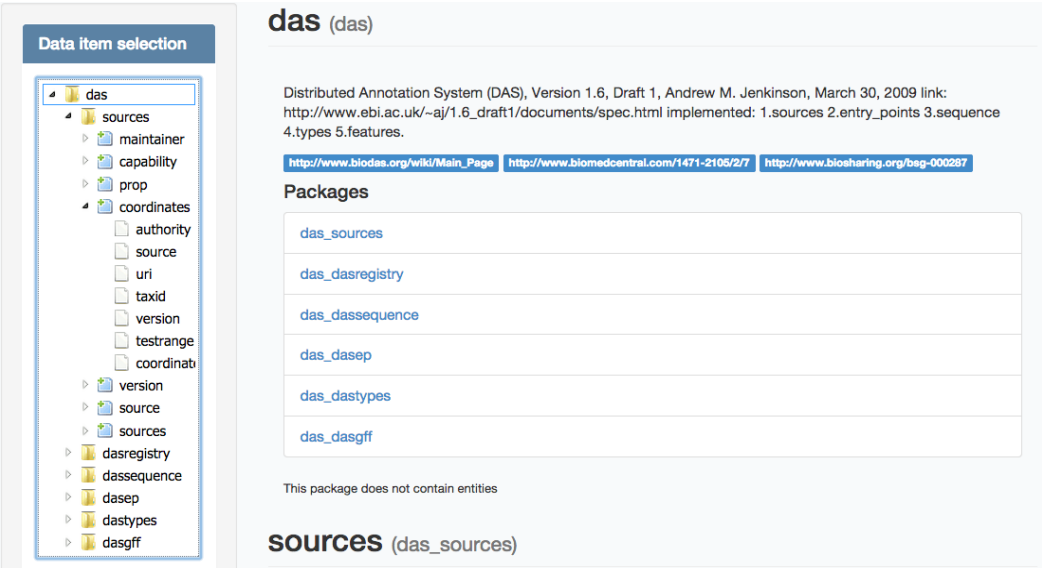


Figure 1 Example of the meta-model usages the distributed annotation system, DAS

In addition several extensions to this core meta-model were added to provide contextual information, which can be customized by the standard information providers using a flexible tagging system. Such information may include but is not limited to:

4. registration e.g. original model files uploaded, registrant name, stewardship, last update
5. authority, e.g., who is recommending the standard, what is the usage context
6. documentation, e.g. link to REST API documentation or WSDL file
7. support e.g. helpdesk, contact person, experts
8. restrictions e.g. license, terms of use
9. credits, e.g. developer, grants
10. literature, e.g. primary citation, relevant publications
11. see also e.g. URL of source registry or parent collection when this data was federated from another original source (e.g. biosharing.org)

Where possible the contextual information elements were aligned with the service registry (D3.3) to prepare for future integration. We expect that an entry in the MMMR will be linked to one or more related services in the service registry, for example pointing at services that implement, or validate a standard.

3.4 Upload formats

Interaction with different segments of the user community showed the wish to support model contributions from distinct user groups. Our current spreadsheet approach (Excel or CSV) is based on the system from the core MOLGENIS project, and has had good success with contributors such as bioinformaticians or biologists without a strong software engineering background. We called this spreadsheet format ‘EMX’ (entity model extensible) to describe all structural and contextual metadata. More technical users typically prefer use of a structured text file format; for this reason, we created JSON structures to easily script the generation of meta-data records in combination with a JavaScript/REST based programmatic interface. See

appendix B.2 for documentation on the EMX spreadsheet format and appendix B.3 for documentation on the REST API.

To engage the large community of (industrial) software engineers we also implemented support for several widely used meta-model standard formats. In many cases these groups use professional model engineering tools such as Enterprise architect based UML diagrams or Protégé semantic modelling tools. For these users we provide XMI (UML export) and OWL semantic web data converters[6].

It is inevitable that many (standard) model developers, providers, integrators, and cataloguers will continue to use their preferred models, methods and formats for software descriptions. Therefore, our schema is dynamically extensible with new meta-data entities so additional custom information can be added to the registry (e.g. new entities for example for 'contact details' or 'quality indicators', etc). Moreover, the visual presentation of the registry can be easily configured by the administrator using the 'menu-manager' were the admin can change the organisation of the screens, configure settings and enable/disable functionality. These for example include hyperlinks to project homepages, ontology annotations, example datasets, etc. Moreover, the original files can be attached as additional annotation.

Importantly, to keep the meta-data standard simple whilst retaining necessary utility we only extract a subset of information out of these standard files. For full provenance we keep the originally uploaded files so users can still drill down to the original details if desired. Currently data entry is enabled for authorised users via batch processing of spreadsheet templates. In the future we plan to extend data entry to the general public, including a simple graphical user interface for item-by-item data entry.

[6] These are currently available at https://github.com/jmuilu/molgenis-xmi

3.5 Standards registry content

3.5.1 Overview of registration process

Valuable repositories exist that catalogue parts of the meta-data landscape. For example, Biosharing.org contains summary level descriptions of standard models, formats and guidelines and associated databases, including hyperlinks to associated documentation. However, these resources usually don't define the data models in full detail and instead refer to the original documentation, making it cumbersome to quickly find and compare models in detail. We have collaborated with the BioSharing project in determining scope of the BioMedBridges activities to ensure that our deliverable complements the BioSharing project.

The BioMedBridges Metadata model and mapping registry aims to bridge this gap by exposing the full data models and providing deep mapping between models. To populate the prototype MMMR we used biosharing.org documentation hyperlinks as a primary gateway to existing PDF-based metamodels and manually added complete and representative models to the registry to enrich the content and to support data exchange use cases we have in BioMedBridges. See also the following section on data upkeep. Examples of the models are available on GitHub[7].

3.5.2 Overview of content

The system currently contains the following elements:

Table 1 Current contents of the Metadata Models and Mapping Registry

Model	Packages	Entities	Attributes
biosample	10	131	6739
Chado	1	133	673

[7] https://github.com/molgenis/molgenis/tree/master/molgenis-model-registry/src/test/resources

Model	Packages	Entities	Attributes
DAS	7	26	98
DICOM	1	47	379
ISA-TAB	1	8	190
MAGE-TAB	1	5	95
MIABIS	1	14	91
MIAME-ENV	1	6	22
SAM	1	3	19
VCF	1	3	45
system	1	3	8
Total	26	379	8359

3.6 Sustainability

3.6.1 Content upkeep strategy

For the coming years BBMRI-NL and ELIXIR-NL have adopted this system as the basis for their work on interoperability, in particular between the biobanks and in health related multi-center research within the Dutch academic hospital. BBMRI-NL has been recently funded with 9.8Meuro to further develop the national biobanking infrastructure. As part of their remit, they collaborate with ELIXIR-NL which has the responsibility for interoperability services within ELIXIR (in the context of Dutch Techcenter for Life Sciences, DTLS) to further harmonize their data models and content standards also working closely with the DataFAIRport initiative. In particular, between them these national ESFRI hubs will work on cataloguing and promoting data standards for molecular and phenotype data (such as currently collected in the Metadata Model and Mapping Registry) as well as cataloguing data dictionaries for all >200 Dutch biobanks which has already been piloted to use the same system as part of the EU-BioSHaRE project.

Expanding into the European infrastructures, BBMRI-NL plays an active role in the BBMRI-ERIC common services for IT that is due to be launched next year and we expect the Metadata Model and Mapping Registry to be a major contribution. Moreover, we are active in various other EU consortia, such as EU-BioSHaRE and RD-connect with several H2020 proposals submitted, building on data catalogues where also the Model registry is a core component. We expect this use of the model registry and extensive interaction with user groups will enable content upkeep in the wider European ESFRI context for the near future and promote usage of the resource.

3.6.2 Software upkeep strategy

The Metadata Model and Mapping Registry is built as part of the MOLGENIS open source project which supports >25 active projects (100 servers) covering applications from biobank catalogues and multi-center omics research databases up to rare disease patient registries and molecular diagnostics tools. The choice of Molgenis allowed us to qucikly generate the data model and supporting applications and will allow us to evolved these if needed. Next to the leading installation that has been developed in BioMedBridges, we have decided to make local copies of the Metadata Model and Mapping Registry available in all the Molgenis installations to ensure the dissemination of the project as well as to promote maintenance and future development of the Metadata Model and Mapping Registry software. Interested consortia can even choose to download and use the Metadata Model and Mapping Registry software as a local tool to maintain their meta models and their mappings, e.g. data dictionaries used by biobanks. Moreover, the system is designed such that it can be used in a federated way (using JSON/REST) or that models can be easily downloaded and uploaded to ease interoperability between the systems, a feature we plan to further develop to ease the sharing of models across instances of the Metadata Model and Mapping Registry. We expect that this larger ecosystem of stakeholders will increase the quality of software upkeep beyond the end of BioMedBridges.

3.6.3 Hosting and integration strategy

Multiple resources are currently being developed and maintained to facilitate and promote information on existing standards, policies, models, identifiers, databases, services, and ontologies. This may be sub-optimal and therefore we organized a workshop on 1 October 2014 to (i) identify common goals shared between these resources, (ii) Identify areas of duplication and gaps and (iii) Define a common integration and development strategy. As a first step the participating resources (biosharing.org, identifiers.org, tools & data services registry[8] and Metadata Model and Mapping Registry[9]) committed to consolidate the identification and ontology schema's used and to enable interoperation using APIs, first steps from the Metadata Model and Mapping Registry (MMMR) are reported here. For the long run, we envision that the content of the MMMR can be databased as part of the biosharing.org service, details of implementation are still under discussion. Practically, we expect users of biosharing.org to have the ability to view MMMR records directly from within the biosharing.org website with the ability to drill down and explore the details and/or to compare with other models.

3.7 User interfaces

3.7.1 Search

Users can search across all collected standards using a simple search box or choose to view all meta-data models.

[8] http://wwwdev.ebi.ac.uk/fgpt/toolsui/
[9] https://molgenis08.target.rug.nl/

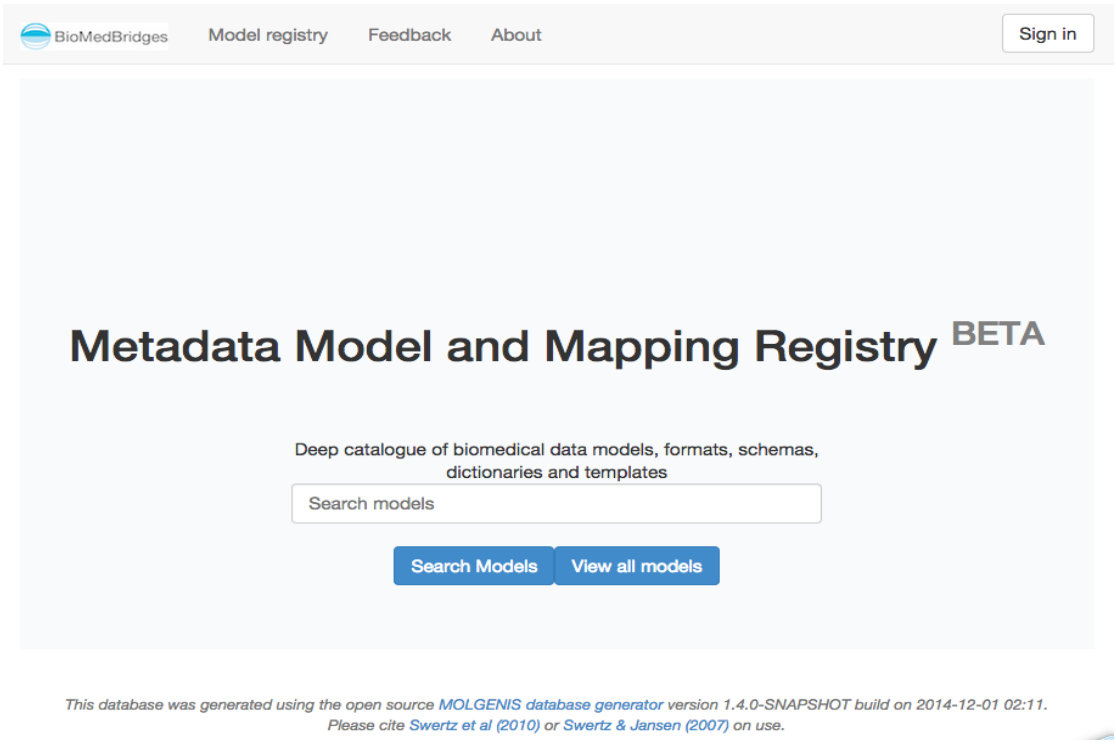


Figure 2 Search interface for the Metadata Model and mapping registry

3.7.2 Search results

Search results are made based on matches at a variety of levels: package, entity and/or attribute descriptions, as well as their associated tags. Users can quickly review what kind of match it was by scanning the “matched:” field (grey below) and can drill down to view the ‘model details’.

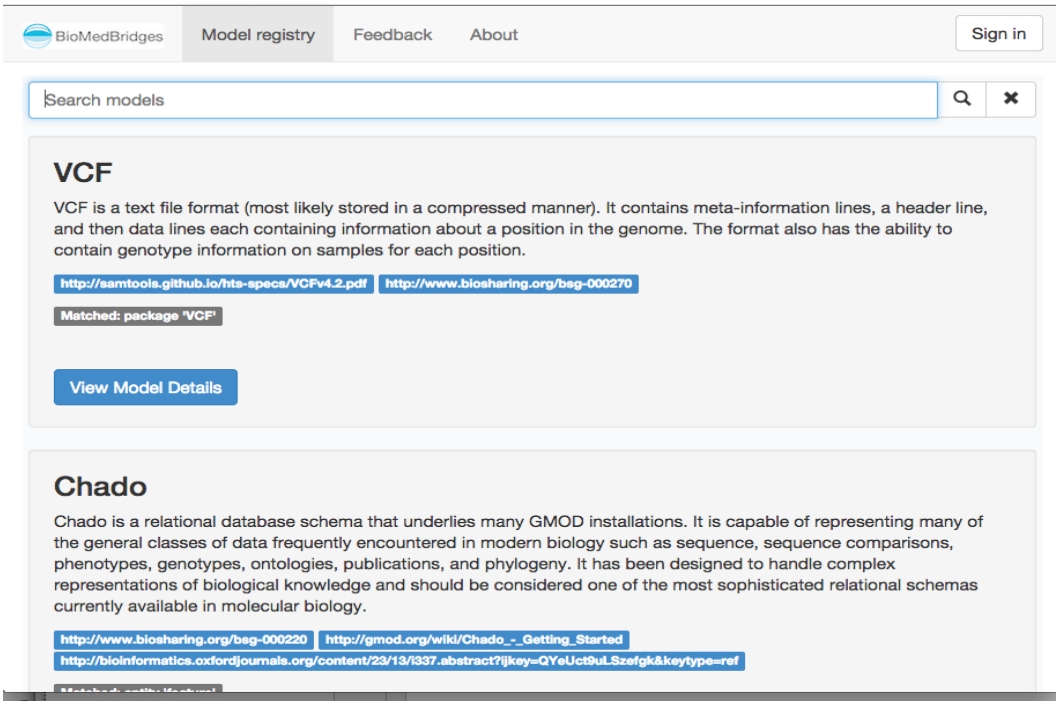


Figure 3 Example search results with records for the VCF and Chado standards

3.7.3 Model details

Users can drill down on the details of each model. A tree view is shown on the left to easily navigate all data model (sub)packages, entities and attributes. Full documentation of the model is shown on the right, including the option to print the results.

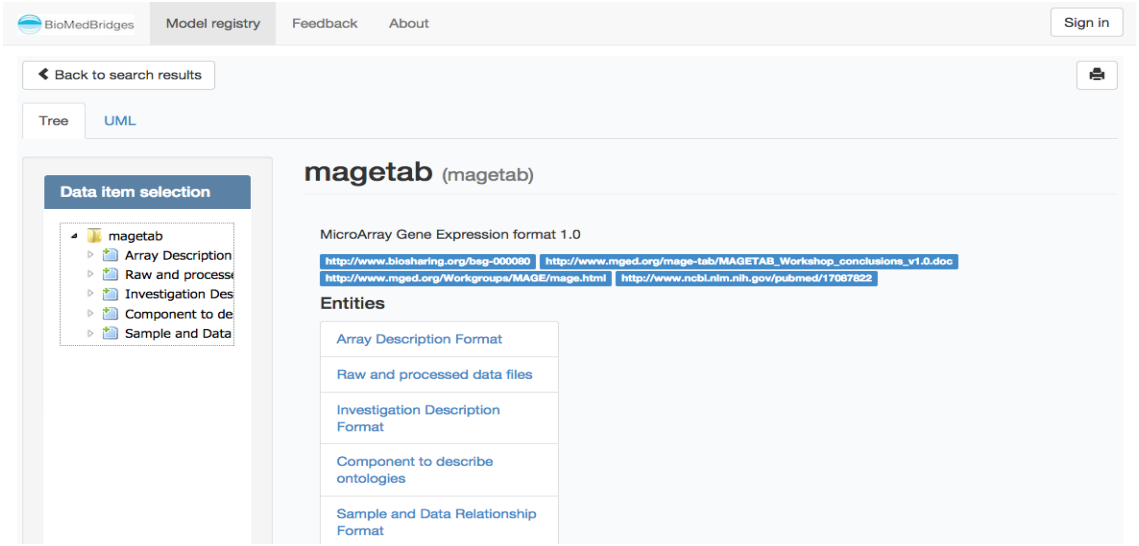


Figure 4 Model details showing record for the magetab gene expression standard

3.7.4 Visualisation

Alternatively, users can view the model in the ‘UML’ view which is shown on top of the details view.

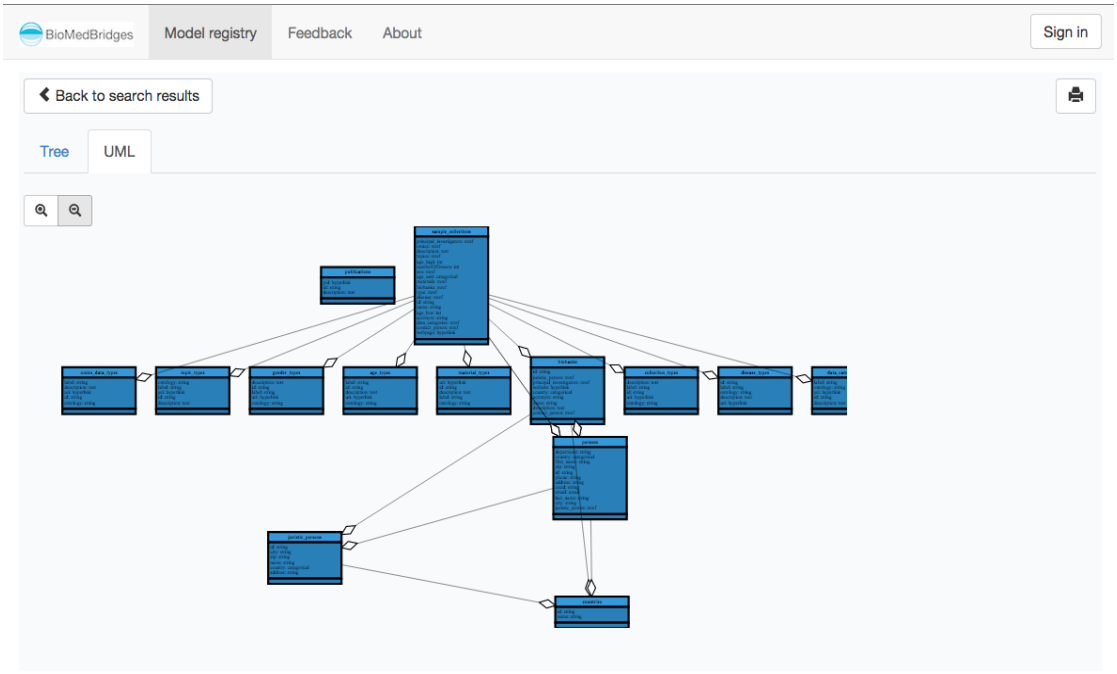


Figure 5 Example UML representation of a standard

3.7.5 Upload

Authorized users can use a simple upload box to upload models, entities and their elements in Excel or zipped TSV format.

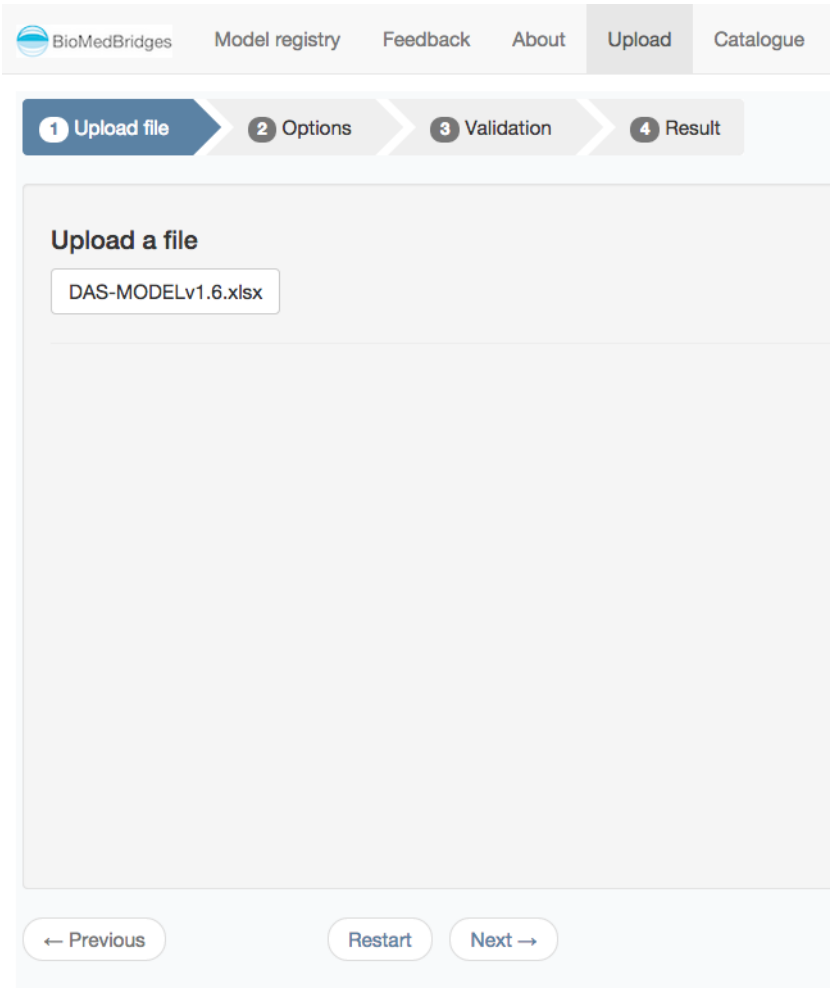


Figure 6 Screenshot demonstrating the upload features using an XLS file

An validation step reports if the uploaded metadata conforms to the EMX data model by marking acceptable elements in green and problematic elements in yellow (for issues that can be ignored) or red (for showstoppers). Information can be uploaded both to an existing model entry (automatically updating the values) or by creating a new entry.

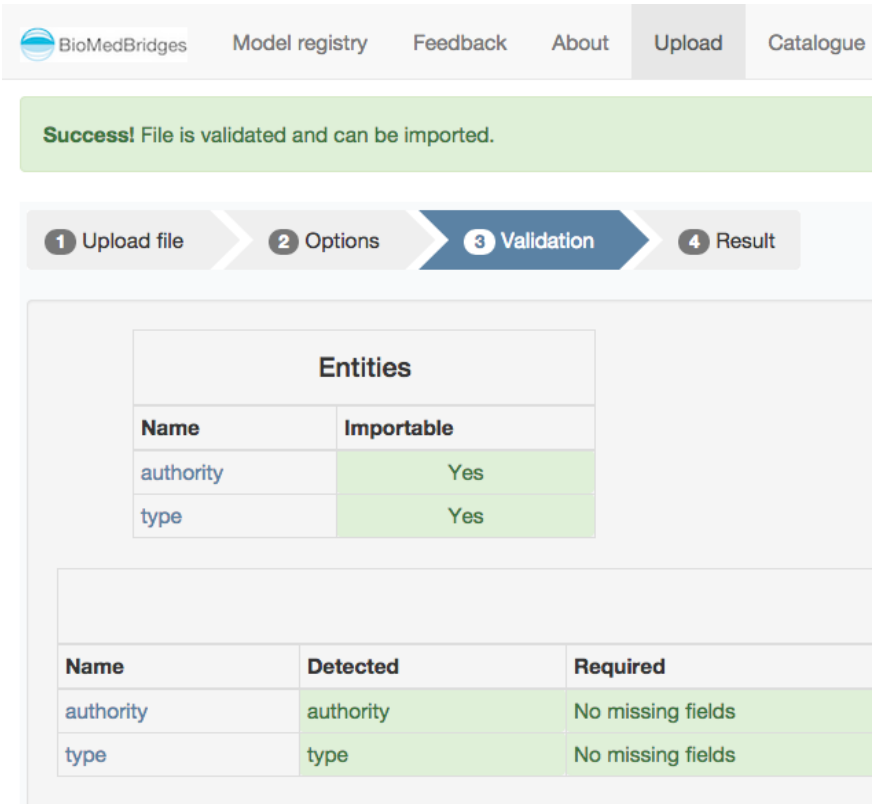


Figure 7 Intuitive feedback and validation information on completion of upload

3.8 Future work

Mission of this project is to facilitate FAIR data (findable, accessible, interoperable, reusable) data. Therefore we envision three major new components in the near future:

3.8.1 Enable end-users to contribute models

Currently only system administrators can upload models into the Metadata Model and Mapping Registry. In the coming year we will enhance the registry such that modelers/consortia can upload their own models and/or can edit the models by hand. Moreover, consortia can first create the models in a private environment before they choose to 'publish' the model for public use. We base the design on this editing component on the extensive experience we developed within BioMedBridges whilst coordinating the MIABIS (minimum information about a biobank information system) working group. Here we learned the dynamic involved in drafting the meta model elements, the

processes needed to agree on its entities and attributes, and finally the release of first and subsequent versions. First iteration of the editor is available, however the fine-grained permissions and versioning systems needed are still under development.

3.8.2 Expansion of contents

While 10 representative models based on survey needs of the project partners (representing theire respective community of users) have been loaded, many more models are currently being prepared. On one hand we will further expand the details of leading standards as already catalogued on a general level in biosharing.org. On the other hand there are many non-standard meta-models that are currently being used in software tools (such as reported in the tools registry) and in databases (such as the data dictionaries of biobanks or the schema's model organism databases). Driven by the needs of the BMS ESFRI partners we will plan to load a few dozen more models (9 from the biobanking domain are already in draft), and train other users how to also catalogue models. In addition, we want to increase the quality control such that also modeling quality is more consistent and of higher level. This will support real use cases and enable data exchange for these models.

3.8.3 Mapping tool

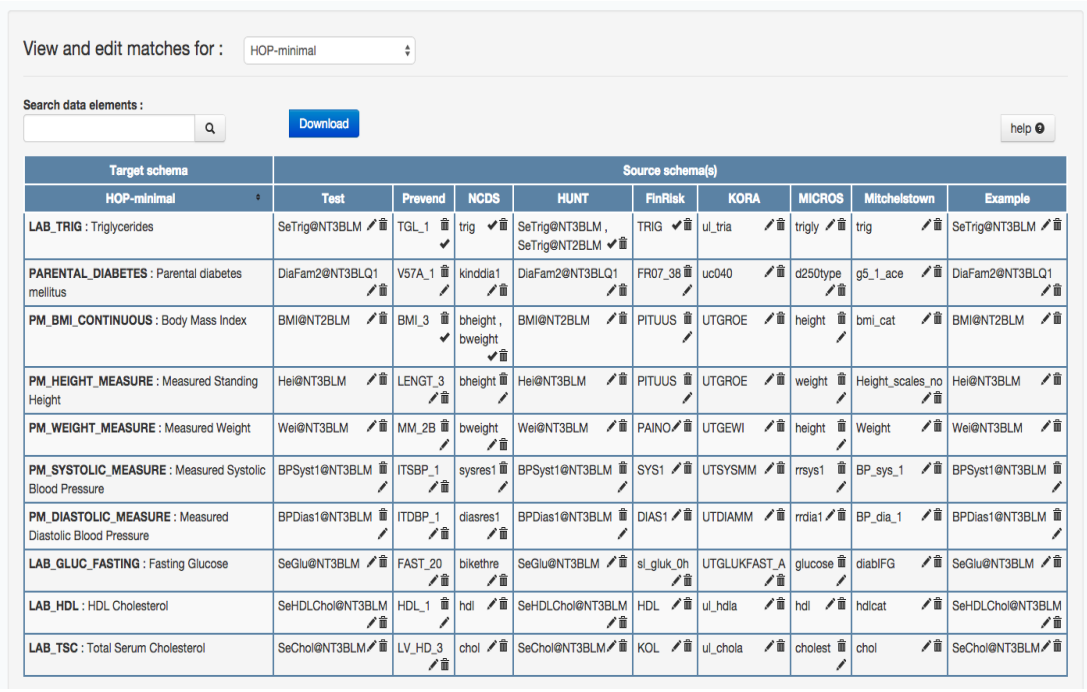
We have piloted a tool where users can generate mappings across data models to establish harmonization rules that would enable data integration. We plan to bring this tool into public production the coming months:

The figure below shows an example: There is a standard data model for biobanking called 'HOP-minimal' which describes minimal parameters for studying 'healthy obese' which are individuals that are obese but still surprisingly healthy. For these studies parameters like 'triglycerides', 'parental diabetes' and 'bmi' are needed. The mapping view shows how other data models map onto these desired parameters. The figure shows 9 meta models that were loaded from the biobanks onto the model registry and then each cell shows candidate mappings that are auto-generated using lexical and

ontological matching. This is implemented in collaboration with EU-BioSHaRE project using the BiobankConnect software (Pang et al. 2014)[10].

We foresee several major use cases for this mapping view:

1. existing standards can use it to create mapping between them to ease data flow between BMS domains
2. new standard proposals can use the mapping view to discover 'model modules' that they can reuse instead of creating standards from scratch
3. end-users can upload their local data (e.g. from a locally executed study) and get assistance in converting their data to ease data upload to public repositories (e.g. to aid upload to the European Genotye and Phenotype archive).



Target schema	Source schema(s)								
HOP-minimal	Test	Prevend	NCDS	HUNT	FinRisk	KORA	MICROS	Mitchelstown	Example
LAB_TRIG : Triglycerides	SeTrig@NT3BLM	TGL_1	trig	SeTrig@NT3BLM , SeTrig@NT2BLM	TRIG	ul_tria	trigly	trig	SeTrig@NT3BLM
PARENTAL_DIABETES : Parental diabetes mellitus	DiaFam2@NT3BLQ1	V57A_1	kinddia1	DiaFam2@NT3BLQ1	FR07_38	uc040	d250type	g5_1_ace	DiaFam2@NT3BLQ1
PM_BMI_CONTINUOUS : Body Mass Index	BMI@NT2BLM	BMI_3	bheight , bweight	BMI@NT2BLM	PITUUS	UTGROE	height	bmi_cat	BMI@NT2BLM
PM_HEIGHT_MEASURE : Measured Standing Height	Hei@NT3BLM	LENGT_3	bheight	Hei@NT3BLM	PITUUS	UTGROE	weight	Height_scales_no	Hei@NT3BLM
PM_WEIGHT_MEASURE : Measured Weight	Wei@NT3BLM	MM_2B	bweight	Wei@NT3BLM	PAINO	UTGEWI	height	Weight	Wei@NT3BLM
PM_SYSTOLIC_MEASURE : Measured Systolic Blood Pressure	BPSyst1@NT3BLM	ITSBP_1	sysres1	BPSyst1@NT3BLM	SYS1	UTSYSMM	rrsys1	BP_sys_1	BPSyst1@NT3BLM
PM_DIASTOLIC_MEASURE : Measured Diastolic Blood Pressure	BPDias1@NT3BLM	ITDBP_1	diasres1	BPDias1@NT3BLM	DIAS1	UTDIAMM	rrdia1	BP_dia_1	BPDias1@NT3BLM
LAB_GLUC_FASTING : Fasting Glucose	SeGlu@NT3BLM	FAST_20	bikethre	SeGlu@NT3BLM	sl_gluk_0h	UTGLUKFAST_A	glucose	diablFG	SeGlu@NT3BLM
LAB_HDL : HDL Cholesterol	SeHDLChol@NT3BLM	HDL_1	hdl	SeHDLChol@NT3BLM	HDL	ul_hdla	hdl	hdlcat	SeHDLChol@NT3BLM
LAB_TSC : Total Serum Cholesterol	SeChol@NT3BLM	LV_HD_3	chol	SeChol@NT3BLM	KOL	ul_chola	cholest	chol	SeChol@NT3BLM

Figure 8 Mapping results of measurement types vs. several available studies

3.8.4 Evaluation and Metrics

To focus future development we plan to implement extensive user evaluation as well as automated metrics into the system. We will co-organize a series of

[10] http://www.ncbi.nlm.nih.gov/pubmed/25361575

'Bring your own datamodel' workshops (BYOD) in collaboration with DTL Netherlands and BBMRI-NL (in January), the BioMedBridges annual meeting (in February) as well as participating in relevant meetings with other consortia (such as RD-connect, following up on a previous meeting in Rome, November 2014). These workshops will help to add to the contents as well as to pinpoint shortcomings in the user interface design. In addition we will implement automatic logs within the software itself so that we can collect statistics on usage of the different MMR features as well as the ability to highlight what searches and models are particularly popular. This enables us to focus the future collection of new metamodel information as well as learn what components are particularly popular and would be good to develop further. Finally, we have implemented a feedback component to enable users to easily report issues with the registry.

3.8.5 Integration with other registries

Finally, we aim to further the integration of the MMR with the other registries in the standards domain. As described in section 3.6, we will add bi-directional cross links with identifiers.org and the service registry, in particular to enable users to find how the models/formats map onto identifiable records in (public) databases and what models/formats are used within the various services and tools. Moreover, in collaboration with biosharing.org we want to add more references on the usage of the models by adding URIs refering to instances where the model is used.

4 Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

5 Adjustments made

No adjustments were made to the deliverable.

6 Background information

This deliverable relates to WP 3; background information on this WP as originally indicated in the description of work (DoW) is included below.

WP 3 Title: ESFRI BMS Standards Description and Harmonization

Lead: Helen Parkinson (EMBL-EBI, Morris Swertz (UMCG)

Participants: EMBL, KI, STFC, UDUS, TUM-MED, ErasmusMC, TMF, HMGU, VU-VUMC, UCPH, UH, UMCG, CIRMMP

Standardization is necessary to ensure infrastructures can work together (syntactic interoperability: data models, data formats, API's, services descriptions, registration and discovery of services), understand each other data (semantic interoperability: ontologies, vocabularies, coding systems, common identifiers), have analysis and supporting tools that complement each other and can be combined in a pipeline (process interoperability) and allow multiple data sets from different origins (including public resources) to be analysed together.

This work package (WP) requires close collaboration with domain experts, research infrastructures, WP4 which will provide implementation based on standardization deliverables described here, and WP5 which will address security issues and use case work packages 6-10. In order to work efficiently a nominated individual from each ESFRI BMS expert area will be responsible both for tasks in this WP, registration of standards, representation of, and correspondence with, relevant domain specific external standardization parties and to represent their community requirements in this WP. WP3 partners are also represented in the use case work packages and will ensure their requirements are supported here.

This WP involves the majority of partners, and exchange of information, registry of services and meta mapping activities will require a diverse set of personnel. The design of this WP therefore includes an allowance for exchange of personnel between this WP and others to facilitate the implementation of deliverables in other WPs and to support interaction with

external experts at meetings and workshops where necessary. This will ensure that relevant experts have the opportunity to actively solve problems by working closely with individuals from work packages to which they have not been assigned. We have also allowed developer time for the creation of training materials and delivery of training at BioMedBridges workshops, as described in WP12.

Work package number	WP3	**Start date or starting event:**	month 1
Work package title	ESFRI BMS Standards Description and Harmonization		
Activity Type	RTD		

Participant	1: EMBL	3: KI	4: STFC	5: UDUS	7: TUM-MED	9: ErasmusMC	10: TMF	11: HMGU	22: VU-VUMC	15: UCPH	16: UH	19: UMCG	20: CIRMMP
Person months	42	21	6	28	4	5	16	30	16	8	11	32	14

Objectives

Addition of scientific value and support for the integration of data between the ESFRI BMS domains by catalogue, review, modification, harmonization, registration and implementation of existing identifier, content, syntactic and semantic standards across the ESFRI BMS projects to support data exchange, integration and infrastructure development.

1. Provision and use of the ESFRI BMS common molecular identifiers (eCMI)
2. Identification, harmonization and integration of ESFRI BMS partner standards
3. Provision of standards and harmonized elements in an accessible standards registry (eSTR)

4. Provision and population of the ESFRI BMS Service Registry (eSR)

Description of work and role of participants

The standardization task is large as ESFRI BMS projects have been active in this area evaluating intra-domain standards, bottlenecks and solutions and there are numerous external standards efforts corresponding to content, data format, semantic and identifier standardization in this domain in which many project partners are involved. Examples include the gene ontology (GO) as an example of a semantic standard, DICOM as an imaging format standard, MIMPP as a content standard from EUROPHENOME, the LCF/MTZ file format, and the CCPN data model for macromolecular NMR. WP will address the following tasks to provide focus:

1. Common identifiers (Task Lead ELIXIR)

The provision and use of common identifiers to determine unambiguous molecular identity for bio-molecules such as genes, proteins and bioactive compounds is key to supporting the information flow from basic science, model organism biology, bioinformatics and structural biology through to translational research and clinical care. The ESFRI BMS project partners will work together to determine a 'Molecular Dictionary' of identifier types and their attributes for use in this project which will constitute best practice for cross domain integration. Where no authoritative identifier standard exists, we will work with the respective community to determine one to support the activities of WP4 and use cases. Relevant identifiers include those for samples (Task 2), small molecules, macromolecular assemblies, genes, drugs and proteins especially where these relate to clinical scenarios.

2. Sample meta data standards (Task Leads BBMRI)

The ability to identify samples and describe their attributes, so data relating to them can be integrated and analysed is common to all ESFRI BMS domains. Content standards which determine exist for given experimental scenarios which data should be collected e.g. age, sex, phenotype, disease state, sampling time, processing state, etc. These are typically determined based on

requirements for analysis, data sharing needs and regulations within a research or technology based domain. For example, the MIAME standard determines which information should be stored about a gene expression experiment performed on a microarray. This is not necessarily consistent with core information about the same sample stored in a BioBank which may include sample processing state, disease and tissue, a sample used to determine a protein structure, or a live animal sampled from the ocean. Where processing states, provenance, storage conditions, or other experimental context are important for a domain e.g. INSTRUCT or for re-use of data relating to samples across domains, these will also be explored with respect to the use cases. The clinical data community have specific requirements relating to integration of Electronic Health Records (EHR), use of clinical terminologies such as SNOMED-CT, description of medical imaging procedures and provision of molecular data in clinical context with appropriate quality control data and translation across these domains is relevant to this task, Task 4 and WP10. Standards in use within the ESFRI BMS projects for data content and semantics will be documented in a public interactive matrix consisting of project, standard and individual elements of standards. Comparable elements across standards will be identified by a harmonization and mapping process across partners. For example BBMRI has produced a lexicon which defines important concepts for the bio-banking domain and EATRIS has analysed standards relating to inter and intra operability between organisations. Standards in use by partners relating to samples will be meta-mapped; common elements e.g. from BBMRI will be cross referenced to relevant concepts from ELIXIR, ECRIN and EATRIS. Where standards are in development e.g. from 2008 roadmap ESFRI BMS projects these will be added and harmonized once they are determined to be stable and valid within a domain, e.g. imaging standards are under development by EuroBioImaging. We do not expect all standards to be fully interoperable and the process of meta-mapping and presentation of these data in an interactive and updated form will inform partners and focus use cases. We will pay specific attention to widely adopted standards, and supporting integration rather than development of standards de novo.

3. Service registration and annotation (Task Lead ELIXIR)

The description of where data and services exist, and by what mechanism these are accessible is key to integrating and exchanging data and has been identified by ELIXIR, EATRIS and others as a blocker to integration especially across domains. Therefore we will develop the Meta-Services Registry comprising tools and terminology for annotation of services (eSR) to catalogue services across partners, domains allowing partners to self-register their own and others services. This will build on previous work in the Bioinformatics domain (EMBRACE, BioCatalogue) and will be extended this with the 2008 roadmap ESFRI BMS partners and throughout the grant as services appear and are used. This will promote the use of domain specific services across partners and also internationally.

4. Semantic standards – ontologies and annotation (Task Lead ELIXIR)

Content standards define what data about a sample in a context or domain. However the meaning of data can be made explicit only by the use of defined terminologies. The use, standardization and mapping of terminologies across domain and species will be explored in the context of use case Work Packages 7 and 10. WP7 explores the semantic integration between mouse models of disease, phenotype and WP10 explores integration of sample data of different types. In order to make these tasks feasible prioritized dataset(s) will be identified with WP7/10 by means of integration criteria which will be developed jointly with these work packages. For example – availability of data in the public domain and /or focus on a key disease type which is well represented in the terminologies to be integrated and available datasets.

Appendix 1: Outreach and documentation

1 Online documentation, feature requests, bug tracker, email list

- Online documentation can be found as part of the MOLGENIS open source project at http://github.com/molgenis/molgenis
- Documentation of the EMX spreadsheet format: https://github.com/molgenis/molgenis/wiki/EMX-upload-format
- Documention of the JavaScript/REST api: https://github.com/molgenis/molgenis/wiki/REST-API-v1
- Issue tracker to report bugs: https://github.com/molgenis/molgenis/issues?q=is%3Aissue+is%3Aopen+model-registry (use the tag 'model-registry')

2 Metadata model

For the capture of the metamodels we have adopted the Observation Entity Model Extensible (Observ-EMX or EMX for short) from the EU-BioSHaRE project. EMX is based on four simple concepts: *Entity, Attribute, Package and Tag,* which can be used to model most data structures used in life science experiments. Examples of Entities are 'protocols', 'experiments', 'mutation', and 'samples' – or essentially any structured collection of information. Examples of Attributes are 'name', 'genomic position', 'weight' and 'sample type' - these being the components described within an Entity. Examples of Packages are 'Biobank LifeLines study', 'Genome of the Netherlands study' and 'Genome wide association study experiment' - in that Packages are the larger containers that hold the Entities. Finally, Tags enable flexible annotations of any Entity, Attribute or Package, for example to add 'homepage' or to refer to 'standard code for cardiovascular disease'.

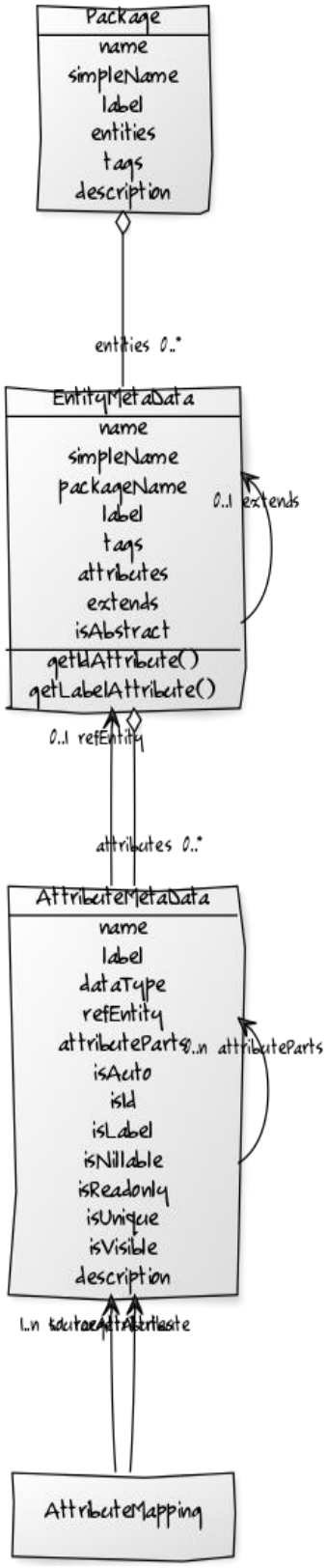


Figure 9 UML diagram of the meta-model

3 EMX upload format

EMX comes with a spreadsheet based upload format. The most recent documentation can be found on GitHub[11].

Using the Metadata capabilities you can define your data structures. This section is optional, e.g. because somebody else has already uploaded the metadata.

The core section is the 'attributes' sheet that describes the smallest atomic data elements of the data structure. Per attribute you can define name, type and a series of additional constraints. Attributes are grouped into 'entities' which can be thought of as data tables. When defining your own metadata the 'attributes' section is required.

Optionally, additional meta-data can be provided. The section on 'entities' enables the description and tagging of the data tables, although in more advanced models they can also be (object oriented) data classes, categorical/ontological value sets, or even more abstract entities. The section on 'packages' enables grouping of entities into '(sub)packages' which depending on the context can relate to organisation (e.g. EBI), standardization body (e.g. MIABIS), or data organisation (e.g. investigations, studies, assays). And finally, the section on 'tags' enables definition of simple text tags, ontology references up to full RDF triples.

'Attributes' sheet

The attributes sheet is used to define the data elements per data entity. The example below defines a simple data structure with entities 'city', 'person' and 'user'. Note that 'user' had exactly the same attributes as 'person' so we will use 'object orientation' to say that 'user' is a special kind of 'person'.

[11] https://github.com/molgenis/molgenis/wiki/EMX-upload-format#example-meta-data

entity	attribute	dataType	nillable	refEntity	idAttribute	description
cities	cityName				TRUE	unique city name
persons	displayName				TRUE	unique name
person	firstName					first name
persons	lastName					family name
persons	birthdate	date	TRUE			day of birth
persons	children	mref	TRUE	person		parent-child relation
persons	birthplace	xref	TRUE	city		place of birth
users	username				TRUE	unique login name
users	active	bool	TRUE			whether user is active

'Attributes' options

Required columns:

- entity : name of the entity this attribute is part of
- attribute : name of attribute, unique per entity

Optional columns (can be omitted):

- dataType: defines the data type (default: string)
  - string : character string of <255 characters
  - text : character string of unlimited length (usually <2Gb)
  - int : natural numbers like -1, 0, 3. Optionally use rangeMin and rangeMax
  - long : non-decimal number of type long
  - decimal : decimal numbers like -1.3, 0.5, 3.75 (float precision)
  - bool : yes/no choice
  - date : date in yyyy-mm-dd format
  - datetime : date in yyyy-mm-dd hh:mm:ss
  - xref : cross reference to another entity; requires refEntity to be provided
  - mref : many-to-many relation to another entity; requires refEntity to be provided

    - compound : way to assemble complex entities from building blocks (will be shown as tree in user interface); requires refEntity to be provided

- refEntity : used in combination with xref, mref or compound. Should refer to an entity.
- nillable : whether the column may be left empty. Default: false
- idAttribute : whether this field is the unique key for the entity. Default: false
- description : free text documentation describing the attribute
- rangeMin : used to set range in case of int attributes
- rangeMax : used to set range in case of int attributes
- lookupAttribute : true/false to indicate that the attribute should appear in the xref/mref search dropdown in the dataexplorer
- label : optional human readable name of the attribute
- aggregateable : true/false to indicate if the user can use this atrribute in an aggregate query
- labelAttribute : true/false to indicate that the value of this attribute should be used as label for the entity (in the dataexplorer when used in xref/mref)
- readOnly true/false to indicate a readOnly attribute
- tags : ability to tag the data referring to the tags sections, described below

'Entities' sheet (optional)

In most cases the 'attributes' sheet is all you need. However, in some cases you may want to add more details on the 'entity' that the attributes are part of, or even use more advanced data modelling concepts such as 'abstract' (for interfaces) and 'extends' (for inheritance). Optionally, you can group your entities using 'packages'.

For example:

entity	package	extends	abstract	description
person	people		true	person defines general attributes like firstName, lastName
user	people	person		users extends persons, meaning it 'inherits' attribute definition
patient	people	person		patient extends person, adding patientNumber

'Entities' options

Required columns:

- entity : unique name of the entity. If packages are provided, name must be unique within a package.

Optional columns:

- extends : reference to another entity that is extended
- package : name of the group this entity is part of
- abstract : indicate if data can be provided for this entity (abstract entities are only used for data modeling purposes but cannot accept data)
- description : free text description of the entity
- tags : ability to tag the data referring to the tags sections, described below

'Packages' sheet

When data structures become larger, or when many data tables are loaded then the package mechanism enables to group your (meta)data. The packages sheet enables addition of meta-data describing the packages.

For example:

name	description	parent	tags
root	my main package		
people	sub package holding entities to describe all kinds of persons	root	homepage

'Packages' Options

Required columns:

- name : unique name of the package. If parent package is provided the name is unique within the parent.

Optional columns:

- description : free text description of the package
- parent : use when packages is a sub-package of another package
- tags : mechanism to add flexible metadata such as ontology references, hyperlinks

'Tags' sheet (BETA)

Optionally, additional information can be provided beyond the standard metadata described above. Therefore all meta-data elements can be tagged in simple or advanced ways (equivalent to using RDF triples). For example, above in the packages example there is a 'homepage' tag provided.

For example:

identifier	label	objectIRI	relationLabel	codeSystem	relationIRI
like	like				
homepage	http://www.molgenis.org	http://www.molgenis.org	homepage		
docs	http://some.url	http://www.molgenis.org	Documen-tation and Help	EDAM	http://edamontology.org/topic_3061

'Tags' options

Required columns:

- identifier : unique name of this tag, such that it can be referenced
- label: the human readable label of the tag (e.g. the 'like' tag as shown above).

Optional columns:

- objectIRI: url to the value object (will become an hyperlink in the user interface)
- relationLabel: human readible label of the relation, e.g. 'Documentation and Help'
- relationIRI: url to the relation definition, e.g. http://edamontology.org/topic_3061
- codeSystem: name of the code system used, e.g. EDAM

4 JavaScript/REST Programmatic interfaces

Documentation can be found on GitHub[12]. Assuming that you have entities 'datasets', 'protocol' and 'features' then you can retrieve the metadata as follows:

Endpoints

- http://www.example.org/api/v1/dataset/meta
- http://www.example.org/api/v1/protocol/meta
- http://www.example.org/api/v1/feature/meta

Retrieve resource metadata

Request

```
GET                    http://your.molgenis.url/api/v1/dataset/meta
```

Response

```
200                                                              OK
{
    "href":                                    "/api/v1/DataSet/meta",
    "name":                                                "DataSet",
    "label":                                                      "",
    "attributes":                                                 {
        "Identifier":                                            {
            "href":                "/api/v1/DataSet/meta/Identifier"
        },
        "Name":                                                   {
```

[12] https://github.com/molgenis/molgenis/wiki/REST-API-v1#resource-meta-data

```
            "href":                          "/api/v1/DataSet/meta/Name"
        },
        "description":                                                {
            "href":                   "/api/v1/DataSet/meta/description"
        },
        "ProtocolUsed":                                               {
            "href":                 "/api/v1/DataSet/meta/ProtocolUsed"
        },
        "startTime":                                                  {
            "href":                    "/api/v1/DataSet/meta/startTime"
        },
        "endTime":                                                    {
            "href":                      "/api/v1/DataSet/meta/endTime"
        }
    },
    "labelAttribute":                                      "Identifier"
}
```