BioMedBridges

Deliverable D3.1

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:	ESFRI BMS Online Dictionary of common molecular identifiers (eCMI)
WP No.	3
Lead Beneficiary:	1: EMBL
WP Title	3.1 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):	1: EMBL, 10: TMF, 11: HMGU, 4: STFC, 5: UDUS, 9: ErasmusMC

Authors: Jon Ison, Julie McMurry, Helen Parkinson, Nathalie Conte, Philipp Gormanns, Murat Sariyar, Gergely Sipos, Søren Brunak, Kristoffer Rapacki

Contents

Tables

1 Executive Summary

The provision and use of common and unambiguous identifiers 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 have determined an interoperable 'Dictionary' of identifier types (Appendix 1) used in this project, and within clinical/translational research more broadly. At the request of the scientific advisory board, we have also expanded our WP3.1 activities to include the development of best practices documentation for identifiers (Appendix 2-1) which was based on our identifiers landscape analysis (Appendix 2-2). Part of the expanded work on identifiers includes a shortlist of the most relevant Identifier Resolution and Conversion Tools (Appendix 2-3); these have also been registered with the BioMedBridges Tools and Data Services Registry.

Further documentation was developed to guide the selection of ontologies (Appendix 2-4) to support cross-domain data integration. Where no authoritative identifier standard exists, we have worked with the respective community to determine one that would support the activities of WP4 and BioMedBridges use cases. Relevant identifiers include those for samples (Task 2), small molecules, macromolecular assemblies, genes, proteins, drugs, diseases and phenotypes[1]. The summary of recommendations from the identifiers best practice document is shown in Box 1.

[1] Paper accepted: Deans et al, PlosOne, *Finding our way through Phenotypes*

Box 1 Excerpts from Identifiers Best Practice (Appendix 2)

Document purpose

This document describes experience and recommendations for best practice for the design and use of identifiers in the BioMedBridges (BioMedBridges) project. It covers identifier assignment, resolution, mapping, and provenance, based on community scenarios from the BMS RIs on the ESFRI roadmap partnered in BioMedBridges, and related projects and is limited in scope to the BioMedical field. It is produced as part of BioMedBridges deliverable 3.1 and is linked to a set of biological entities and their recommended identifiers type supplied by the BioMedBridges project partners as part of an analysis of biological entities and identifier usage within the BioMedBridges domains. The recommendations will be refined in light of discussions with the Research Data Alliance[2] (RDA), the European Data Infrastructure (EUDAT) and other domain specific organisations concerned with persistent identifiers e.g. Global Alliance for Global Health. This document is not intended as a developer specification, rather it provides pointers and context for existing specifications.

General recommendations

- Use any currently available identifier scheme that is "machine actionable, globally unique, widely (and currently) used by a community, and that has a long-term commitment to persistence (for example, see the W3C persistence policy[3]). Best practice is to choose a scheme that is cross discipline."[4]
- The primary identifier of an entity must be unique and unambiguous: i.e. a 1:1 relationship of identifier:entity, and designed so that it never has to be changed, retired, or reassigned.

- We recognize the need for formal specifications of identifier formats, and/or alignment between existing specifications. Key considerations for identifier format:
  - An identifier may be used in more than one format (e.g. a database accession number and URI), but it must be possible to transform one format to the other.
  - Identifiers should adhere to an unambiguous format, ideally one definable by a regular pattern[5] and whose prefix is unique with respect to other identifier schemes[6].
  - Consider the format http://{domain}/{dataset}/{identifier} for URL-based identifiers where {domain} is a stable domain name (e.g.www.uniprot.org), {dataset} is a descriptive tag for the type of

[2] https://rd-alliance.org/working-groups/pid-information-types-wg.html
[3] http://www.w3.org/Consortium/Persistence.html
[4] http://www.scc.lancs.ac.uk/research/projects/researchobject/mediawiki-1.22.6/index.php/Identifier_best_practices
[5] i.e. a regular expression
[6] Unique prefixing also facilitates creation of Compact URIs (CURIES) http://www.w3.org/TR/curie/; for an example of implemented CURIEs, see http://wiki.geneontology.org/index.php/Identifiers

entity for which the URL will return data, and {identifier} is the primary entity identifier (typically a database accession number)[7] For instance: http://www.informatics.jax.org/allele/MGI:3845668.

- Regardless of how the entity record will be accessed, it should be comprised solely of web-friendly and printable ASCII characters without whitespace.[8]
- For database accessions:
  - Consider a fixed alphabetical prefix that intuitively conveys the identifier type and authority, and a numerical suffix that confers uniqueness, whilst keeping overall length as short as is practicable.[9]
    - The alphabetical characters should be (non-accented) English letters, preferably not mixed case.[10]
  - Consider omitting a delimiter, or using an underscore if a delimiter is needed. (See delimiters section)
- Consider using check digits or similar scheme to guard against typos. Check digits is rarely implemented in bioinformatics because doing so is harder and lengthens the identifier.

- For ontology accessions, consider following established best practice[11]
- For identifier creation:
  - Where an entity is already well identified, re-use the existing canonical identifier. If multiple identifiers already exist for an entity, and none has broader adoption, consider using the identifier that has the best-maintained mappings to the others. Otherwise, it is acceptable to create a new identifier, and maintain and publish the mappings to the others.
  - Work with established authorities, e.g. major databases, on assignment of new identifiers, especially where they are expected to eventually host your dataset
  - Where practicable, work with dedicated and operationally independent services, e.g. identifiers.org[12], on issuance of new URL-based identifiers for database accessions
  - Management policy of identifiers must be well defined and

[7] http://{domain}/{entity}#{identifier} is also acceptable, where the range of identifiers for the entity is limited, e.g. this is common for classes defined in an RDF schema

[8] Avoid characters that require special encoding, e.g. Superscripts, subscripts, accented characters, whitespace, non-ASCII characters; all of these can pose problems when used in common exchange formats e.g. in URLs and XML. For details see https://support.google.com/dfp_premium/answer/1111200?hl=en#Safe1

[9] Separating the alpha and numeric portions (rather than interweaving them) avoids misinterpreting letters for numbers and vice versa. According to a Bell Labs study, the symbols l and 1, O and 0, Z and 2, and 1 and 7 accounted for more than 50% of the errors caused by symbol misidentification. Nierenberg GI. *Do it right the first time*. New York: John Wiley and Sons 1996:154-162

[10] Mixing upper and lowercase letters can make identifiers easier for humans to read, but if mixed case is used, it should always be used in the same (canonical) way. For example, 'THING_Abc' should never be represented as 'THING_ABC'. Furthermore, new entities should always be issued identifiers that are unique, regardless of case.

[11] http://www.obofoundry.org/id-policy.shtml

[12] http://identifiers.org

documented. Documentation should be publicly available and describe how ids are assigned and maintained.[13]
    - Versioning policy must be documented. I.e. what kind of changes in data triggers creation of a new version number and how to obtain the current version. See versioning section for details.
- For registration of identifier types, tooling, and related implementations
    - Where practicable, register new identifier authorities in Identifiers.org
    - Where practicable, register new identifier types in Identifiers.org
    - Register identifier-related services (e.g. a resolver or mapping service) in the BioMedBridges/ELIXIR Tools and Data Services Registry[14]
    - Register in the BioSharing registry[15,16,17] any public systems (such as databases and content standards) that make use of public identifiers.
- These registers are in the process of being connected (cross-referencing records) under the ELIXIR umbrella.
- When referencing identifiers from durable authorities and where practicable, reference the native URL rather than a resolver service.
- When referencing identifiers from durable authorities and where practicable, reference the native URL rather than a resolver service.

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

[13] This is crucial for platforms like identifiers.org to be able to assign regex patterns.
[14] http://bioregistry.cbs.dtu.dk
[15] http://www.biosharing.org
[16] BioSharing is also operating as a working group under the Research Data Alliance (RDA) https://rd-alliance.org/group/biosharing-registry-connecting-data-policies-standards-databases-life-sciences/case-statement
[17] BioSharing is also part of the NIH BD2K CEDAR centre: http://metadatacenter.org

3 Detailed report on the deliverable

3.1 Background

We performed a landscape analysis for the identifier types used and required by the BioMedBridges partners. This analysis directly informed the development of the dictionary of molecular identifiers (Appendix 1) and best practices documentation (Appendix 2-1) reported here. We believe these tools will improve the design and use of identifiers by clarifying identifier concepts, illustrating identifier usage with real-world examples and offering recommendations on best practice, including which types of identifiers should be used for which entities across the BioMedBridges project.

3.2 Identifier Landscape Analysis

The identifier landscape analysis survey was sent out to all individuals in the BioMedBridges project; 30 responses were received representing all the BMS-ESFRIs in BioMedBridges. The full details of the survey are provided in Appendix 2 (2). The survey informed us about identifier usage and challenges. The survey covered 20 different types of identifiers and ontologies relevant to clinical and translational research. Most of the identifier types were relevant in a majority of respondents; those of highest relevance are listed at the top[18].

Table 1 Identifier types relevant to clinical and translational research

Identifier type	Example entity
Phenotypes / symptoms	albinism
Biosamples derived from an organism or group of organisms	Blood from human subject with leukemia
Individual genes	human p53 gene
Diseases	Acute Myeloid Leukemia
Cell types	T cell

[18] Because the number of respondents varies from question to question, it is difficult to state definitively a rank order

Identifier type	Example entity
Proteins	Cellular tumor antigen p53
Individual subject	John Doe
Cell lines	Hela cervical cancer cell line
Whole genomes	human genome assembly
Gene/Transcript/Protein variants	gene variant
Protocols	Illumina sequencing
Chemicals incl drugs, metabolites	aspirin
Species	Mouse (Mus musculus)
Experiments	Transcription profiling of mammalian male germ cells
Transcripts	human myosin VI
Etiologic agents/isolates	Mycobacterium tuberculosis strain H37Rv
Researcher	Helen Parkinson
Individual images	IMG_0123456
Antibodies	anti-CD52

The survey responses, together with follow up discussions in an identifiers workshop held in Amsterdam in June 2014, reflect that although identifier needs within the BioMedBridges community are heterogeneous, there are common concerns and challenges. For example, the overwhelming majority of respondents indicated that identifiers for phenotypes, genes, diseases and biosamples were either highly relevant or somewhat relevant; below is a summary of the main challenges associated with these important identifiers. A list of all of the pain points noted by respondents is included in Appendix 2 (2).

Table 2 Primary identifier-related challenges to clinical and translational research

Identifier type	Main challenges	Proposed action
Phenotypes	Lack of coverage of terms amongst existing ontologies; ambiguous and restrictive licences hamper data sharing and mining; difficult to identify phenotypes across species bridges	Coordination between developers of existing ontologies, WP7 activities to extend this space
Diseases	There are many disease ontologies with different scope; lack of coverage of terms e.g. rare genetic disease; partial overlap between existing ontologies, paid subscription services hamper open data sharing, mapping, and	Further develop open-access disease ontologies, improve cross references between disease ontologies.

Identifier type	Main challenges	Proposed action
	mining.	
Biological samples	Biosamples DB19 provides the tooling necessary to link a given biosample across projects, however more multi-omic studies should deposit information in the biosamples database.	Publicise the BiosamplesDB more widely; target the multi-omic research community. Encourage users to generate Biosamples DB identifiers to generate for samples at the time they are collected e.g. by pre-registration or via connectivity to aaLIMS systems; this way they are maintained throughout the data lifecycle.
Images	No established standard exists. Even the practice of identifying images within a research group is not widely practiced.	Coordinate imaging communities to develop standards and identifier generation/resolution platforms. This issue will be raised at an upcoming workshop in February 2015 sponsored by Systems Microscopy and endorsed by BioMedBridges.

3.3 Dictionary of common molecular identifiers

Identifiers.org is a platform providing resolvable persistent URIs used to identify data for the scientific community, with a current focus on the Life Sciences domain. The provision of a resolvable identifiers (URLs) fits well with the Semantic Web vision, and the Linked Data initiative. The EDAM bioinformatics ontology contains types of identifiers, as well as data types and data formats. As part of this deliverable, EDAM ontology identifiers branch[20] was substantially extended and further mapped to entries in Identifiers.org. We expanded the scope of the identified entities beyond molecules (e.g. DNA) to any entity of biological or clinical interest (e.g. human subjects and specimens). The EDAM and Identifiers.org efforts are complementary: In addition to providing persistent resolvable URIs, Identifiers.org provides various ways to browse and programmatically access the information about identifier metadata. EDAM provides the stronger typing of identified entities.

[19] http://nar.oxfordjournals.org/content/42/D1/D50
[20] http://bioportal.bioontology.org/ontologies/EDAM?p=classes&conceptid=data_0842

Details about this work are included in Appendix 1. 43 new terms have been modeled for EDAM; this will bring the total number of terms in the branch to 566. Many of these new terms would also be new additions to Identifiers.org, due to the recency of the identifier authority (e.g. RNA central). Mapping EDAM terms with Identifiers.org data provided 2,000 additional annotations including those for regular expression and identifier authority. EDAM topics and tags for each identifier type were added so that they could be retrieved according to an area of interest (e.g. Metabolomics, or Mouse).

Table 3 summarises the common entities of clinical and translational interest with the corresponding recommended identification authority in the scope of the project.

Table 3 Summary of identifier types and recommended authorities for BioMedBridges

Identifier type	Recommended identification authority	Usage	Example entity	Example ID
Phenotypes and disease				
Phenotypes, symptoms, and diseases	Human Phenotype Ontology[21]	Translational research	albinism	HP:0001022
	Mammalian Phenotype Ontology[22]	Translational research	absent coat pigmentation	MP:0005171
	MeSH[23]	Indexing medical literature	albinism	mesh:68000417
	ORDO[24]	Rare disease	Oculocutaneous albinism type 1B	Orphanet:79434
	Experimental Factor Ontology[25]	Translational research	Oculocutaneous albinism type 1B	Import from Orphanet

[21] http://www.human-phenotype-ontology.org/
[22] http://www.informatics.jax.org/searches/MP_form.shtml
[23] http://www.nlm.nih.gov/mesh/meshhome.html
[24] http://www.orphadata.org/cgi-bin/inc/ordo_orphanet.inc.php
[25] http://www.ebi.ac.uk/efo/

Identifier type	Recommended identification authority	Usage	Example entity	Example ID
	ICD-926	Medical billing, eHR, USA	Other disturbances of aromatic amino-acid metabolism	ICD-9-CM 270.2
	ICD-1027	Medical billing, eHR, Europe	Albinism	ICD-10-CM E70.30
	UMLS28	licensed, merges of lots of terminologies	(available via license only)	(available via license only)
	Disease Ontology29	Generated computationally, human curated, highly xref'd	oculocutaneous albinism	DOID:0050632
Cellular phenotypes	CMPO30	Translational research	apoptotic cell shape phenotype	CMPO:0000048
Organisms, species and samples				
Biosamples derived from an organism or group of organisms	BioSamples Database31	Translational research spanning multi-omic studies	Blood from human subject with leukemia	SRS346051
Cell lines	BioSamples Database		Hela cervical cancer cell line	SAMN01728936
Individuals, or samples from individual organisms	Locally minted	If privacy concerns apply (see also identifier formats)	Individual human subject Individual mouse	Patient ID 1234 Mouse ID 4567
Species	NCBI Taxonomy32	Translational research	Mouse (Mus musculus)	NCBITaxon_10090

[26] http://www.cdc.gov/nchs/icd/icd9.htm ICD-11 is anticipated in 2017
[27] http://www.cdc.gov/nchs/icd/icd10cm.htm ICD-11 is anticipated in 2017
[28] http://www.nlm.nih.gov/research/umls/
[29] http://disease-ontology.org/
[30] http://www.ebi.ac.uk/cmpo/
[31] http://www.ebi.ac.uk/biosamples/
[32] http://www.ncbi.nlm.nih.gov/taxonomy

Identifier type	Recommended identification authority	Usage	Example entity	Example ID
Cell types	Cell Ontology33	Translational research	T cell	CL_0000084

Identifier type	Recommended identification authority	Usage	Example entity	Example ID
Genomics and proteomics				
HGNC	Human Gene names34	Clinical, translational	breast cancer 1, early onset	BRC1
Individual gene identifiers	Ensembl35	Clinical, translational	human p53 gene	ENSG00000141510
Whole genomes	Genome Reference Consortium36	Clinical, translational	human genome assembly	GRCh37.p13
Gene/Transcript/ Protein variants	HGVS conventions37 Locus Reference Genome38	Clinically-relevant loci	gene variant	NG_007400.1:g.9595G>A
	dbSNP39	Any loci	gene variant	
	Mouse Genome Informatics40	Translational research	Mouse allelic variant	
Transcripts	Ensembl	Clinical, translational	human myosin VI	ENST00000369985

[33] https://code.google.com/p/cell-ontology/
[34] http://www.genenames.org/
[35] http://www.ensembl.org/index.html
[36] http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
[37] http://www.hgvs.org/
[38] Locus Reference Genomic: reference sequences for the reporting of clinically relevant sequence variants, MacArthur JA et al., Nucleic Acids Res. 2014 Jan doi: 10.1093/nar/gkt1198
[39] http://www.ncbi.nlm.nih.gov/SNP/
[40] http://www.informatics.jax.org/

Proteins	UniProt41	Clinical, translational	Cellular tumor antigen p53	P04637
Macromolecular assemblies	Enzyme Commision42	Classes of enzymes	Alcohol dehydrogenase	EC 1.1.1.1
Macromolecular assemblies	Antibody Registry43	Antibodies	anti-CD52	AB_10763735
Chemistry				
Chemicals including drugs, metabolites	ChEMBL44	Drug discovery, dev't	aspirin	CHEMBL25
Experiments and other entities				
Gene expression experiments	Varies by experiment type; many experiments covered by Array Express[45]	Translational research	Transcription profiling of mammalian male germ cells	E-MEXP-31
Bioassays	Bio Assay Ontology[46]	Translational research	protein-protein interaction assay	BAO_0002990
Experimental protocols and metadata	Experimental Factor Ontology	Translational research	Illumina HiSeq 1000 standard manufacturer's protocol	EFO_0005085
Person (Researcher)	ORCID[47]	Any scientific researchers	Helen Parkinson	http://orcid.org/0000-0003-3035-4195

3.3 Conclusion and future work

Preliminary work (Appendix 1) has been done to enumerate the identifier types and the authorities that issue them. This information needs to be imported

[41] http://www.uniprot.org/
[42] http://en.wikipedia.org/wiki/Enzyme_Commission_number
[43] http://antibodyregistry.org/
[44] https://www.ebi.ac.uk/chembl/
[45] https://www.ebi.ac.uk/arrayexpress/
[46] http://bioassayontology.org/
[47] http://orcid.org/

from the collaboratively curated spreadsheet into a new release of EDAM, and into Identifiers.org. One of the main challenges moving forward will be to develop ontologies in order to harmonise concepts of biological entity type (protein, gene etc.). EDAM and Identifiers.org, like other biological databases, organise information according to the type of biological entity using terms from several ontologies. However, there is no broad consensus for an ontology of biological entities to use across different resources; harmonisation would therefore be beneficial. Strongly typed biological entities would also be beneficial for catalogues of identifier types such as Identifiers.org and the EDAM ontology identifiers branch. Identifier types also need to be better incorporated into the Tools and Data Services Registry[48] and Metadata Model and Mapping Registry[49]; this would facilitate the discovery and integration of relevant tools and databases.

4 Delivery and schedule

The delivery is delayed: ☐ Yes ☑ No

5 Adjustments made

No adjustments were made to the deliverable.

[48] Deliverable 3.3 ESFRI BMS Meta Service Registry (eSR)
[49] Deliverable 3.2 Mapping and registry of ESFRI BMS standards (eSTR)

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: Dictionary of common molecular identifiers

Table A 1 Summary resources developed as part of the dictionary of common molecular identifiers, the scope of which has been expanded to include identifiers of clinical and translational interest

Resource	Location
The source file containing the manual curation of the EDAM identifiers branch enriched with manual annotations and imports through Identifiers.org cross references.	http://tinyurl.com/identifiersdictionarysource
User interface destination of above ontology development (future work).	www.identifiers.org
Location of resulting EDAM ontology version in bioportal (future work).	http://tinyurl.com/edamidentifiersbranch

Table A 2 Summary of BioMedBridges-sponsored curation within EDAM ontology identifiers branch

Type of curation	Explanation	Before	After	Total
MIRIAM cross references	There was a considerable amount of content overlap between Identifiers.org and the identifiers branch of the EDAM ontology. Cross-references to Identifiers.org have been added to the EDAM identifiers branch; the corresponding references within Identifiers.org	0	228	228
Term authority	For each identifier, a corresponding authority was determined. Where a cross-reference to Identifiers.org was possible, the authority was imported; otherwise it was manually added.	0	298	298
Last known update by ID Authority	Year of last update has been added for several identifier authorities whose corresponding website appeared to be, or was expressly stated to be, no longer maintained. REBASE enzyme number is issued by the REBASE DB which was last updated in 2010.	0	60	60
EDAM Topic tags	EDAM topics were used to tag each of the identifier types; See summary in Table 3 below. This makes it possible to filter terms according to their relevance to	0	708	708

Type of curation	Explanation	Before	After	Total
	area(s) of interest. Eg. IntAct accession number (http://edamontology.org/data_1130) is relevant to both Protein interactions (http://edamontology.org/topic_0128) as well as to Drugs and targets (http://edamontology.org/topic_0620)			
Identified entity	EDAM data types were used to specify the type of entity that an identifier referred to. Eg. IntAct accession number (http://edamontology.org/data_1130) identifies a Protein-ligand interaction (http://edamontology.org/data_1566)	81	482	563
Taxonomic tags	27 of the identifiers were limited in scope or relevance to a specific taxon. Human, Mammal, Vertebrates, C elegans, Algae, Protozoa, Arabidopsis, Yeast, Fungi, Rat, Plant, Mouse, Drosophila, Amphibia. These have now been annotated accordingly.	0	27	27
Total terms	48 terms were added	523	48	571

Table A 3 Frequency of EDAM Identifier term corresponding to EDAM Topic

Topic	Term frequency
Genetics	104
Proteomics	37
Molecular interactions, pathways and networks	35
Chemistry	24
Genotype and phenotype	18
Drugs and targets	17
Gene structure	15
Protein families	13
Gene expression	11
Organisms	9
Protein domains and folds	7
Genetic variation	6
Literature and reference	6
Protein interactions	6
Sequence clustering	6
Enzymes	5
Genomics	5

Topic	Term frequency
Proteins	5
Bioinformatics	4
Carbohydrates	4
Data identity and mapping	4
Microarrays	4
Structural biology	4
Transcriptomics	4
Biobanks	3
Clinical Trials	3
Disease	3
Drug discovery	3
Immunology	3
Immunoproteins, genes and antigens	3
Ontology and terminology	3
Phenomics	3
RNA	3
Sequence sites, features and motifs	3
Gene regulation	2
Laboratory experiments	2
Mouse biology	2
Neurobiology	2
Neurology	2
Pharmacogenomics	2
Phylogenetics	2
Protein structural motifs and surfaces	2
Transcription factors and regulatory sites	2
Anatomy	1
Biological models	1
Biological processes	1
Cell biology	1
Cell lines	1
Cheminformatics	1
Clone library	1

Appendix 2: Identifiers best practice and resources

1. Identifiers best practice: http://tinyurl.com/identifiersbestpractice
2. Identifiers Landscape Analysis: https://www.surveymonkey.net/results/SM-WXVRJ778/
3. Identifier Resolution and Conversion Tools: http://tinyurl.com/identifiertools
4. Ontology selection: http://tinyurl.com/rulesforontologyselection (in preparation)
5. Glossary: http://tinyurl.com/bmbstandardsglossary