ELIXIR WP 7 - Data Integration & Interoperability

Status
Final report, 8th July 2009

Authors
Amos Bairoch¹, Michael Ashburner², Lydie Bougueleret¹, Vincent Breton³, Susanna-Assunta Sansone⁴. Authors can be contacted at WP7 mailing list.

Abstract
This document outlines the work of the ELIXIR Working Package (WP) 7 Committee, extending the interim report [InterimElixirWP7], and presents the final recommendations to achieve data integration and interoperability. This report is also the results of a close interaction with (i) other ELIXIR WPs, (ii) related activities in the other ESFRI, European and international projects and (iii) in the light of the results from the ELIXIR surveys.

Section 1 summarizes the scope of this WP, meetings, sponsored activities and links established to date; sections 2 to 5 present - for each of the key themes – a number of existing projects and relevant activities, along with their current status and recommendations; section 6 introduces the plans for adoption.

Table of Contents
1. WP7 Scope and Activities ....................................................................................................................................... 2
1.1. Meetings and Sponsored Workshops .................................................................................................................. 2
1.2. Synergies with Other ESFRI and relevant EU Projects ....................................................................................... 2
2. Programmatic Access – Theme A .......................................................................................................................... 4
2.1. Status and Opportunities ....................................................................................................................................... 4
2.2. Recommendations ................................................................................................................................................... 5
3. Nomenclatures – Theme B ......................................................................................................................................... 7
3.1. Status and Opportunities ....................................................................................................................................... 7
3.2. Recommendations ................................................................................................................................................... 7
4. Controlled Vocabularies and Ontologies – Theme C ............................................................................................... 8
4.1. Status and Opportunities ....................................................................................................................................... 9
4.2. Recommendations ................................................................................................................................................... 10
5. Reporting Requirements – Theme D ....................................................................................................................... 12
5.1. Status and Opportunities ....................................................................................................................................... 12
5.2. Recommendations ................................................................................................................................................... 15
6. Plan for Adoption– Themes B, C and D ....................................................................................................................... 17
7. References ................................................................................................................................................................... 19
8. Appendix 1 ............................................................................................................................................................... 24
9. Appendix 2 ............................................................................................................................................................... 28

¹ Swiss Institute of Bioinformatics, Geneva, Switzerland.
² Department of Genetics, University of Cambridge, UK.
³ LPC, CNRS-IN2P3, France, breton@clermont.in2p3.fr
⁴ EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, UK.
1. WP7 Scope and Activities

This WP has two broad objectives, defined as:

- Documentation of ELIXIR technical strategy on data integration and interoperability;
- Outline the work needed to ensure greater integration of public databases.

The objectives are achieved by addressing the following key themes:

a. Programmatic access: standardization of the technology to be used to build connections to databases and tools;

b. Nomenclatures: harmonization of names and symbols of biological objects;

c. Controlled vocabularies and ontologies: harmonization of the terminologies used to describe database content;

d. Reporting requirements: standardization of the minimal information content to be reported for a given domain and the format used for submission to databases and journals, to guide deposition and facilitate exchange of the information.

The first step has been the documentation of existing interoperability standardization efforts of the community databases and other ESFRI projects, but also of relevant European and international projects, such as the Cancer Biomedical Informatics GRID [CaBIG] and the Human Microbiome Project [HMP] Data Analysis and Coordination Center. Candidate technologies for theme “a” are those emerging from grid and web developments; themes “b”, “c” and “d” exploit the existing well-developed minimal information requirements, exchange formats and terminologies. The resulting picture has provided an insight into the current situation, highlighting areas requiring further development and assisted in the creation of a set of recommendations and a plan for the adoption of the agreed standards.

The second step is the development of the strategies required to fill the gaps and overcome current problems, in line with existing standards efforts.

The last step for this WP is the creation of a consensus set of recommendations and a plan for the adoption of the agreed standards.

1.1. Meetings and Sponsored Workshops

A number of meetings have been coordinated to tie in with existing workshops in the areas of standards and ontologies, co-sponsored by the EBI Industry Programme [DO meeting], UK’s NERC and BBSRC funds to Susanna-Assunta Sansone [WODS meetings]. Data interoperability issues were also addressed at CASIMIR-ENFIN workshop on Databases [CASIMIR-ENFIN meeting].

1.2. Synergies with Other ESFRI and relevant EU Projects

The WP7 Committee has established an initial contact with the leaders of WPs dealing with standardization and interoperability in the BBMRI [BBMRI], EATRIS [EATRIS] and InfraFrontier [INFRAFRONTIER] projects, where activities have substantial overlap with ours. To strengthen the communication and explore joint efforts the ELIXIR WP7 Committee has shared its preliminary report [InterimElixirWP7] with the leaders of the other ESFRI projects. Similarly, representatives of relevant EATRIS and BBMRI WPs have
organized two cross projects discussions, with the aim of bringing together the leaders of relevant WPs dealing with standardization and interoperability. However, despite those initiatives and the efforts of few motivated individuals, no common agreement has yet been reached on these issues across the ESFRI projects. It seems that these are at different stage of maturing and their internal reporting timelines are quite different, making problematic to tune the individual discussion. This final report, however, includes also specific comments and suggestions received from members of CASIMIR [CASIMIR], GEN2PHEN [GEN2PHEN] and EATRIS projects.
2. Programmatic Access – Theme A

One of the keys to ELIXIR success is to integrate distributed data resources throughout Europe. These data resources are mainly molecular biology databases which should be made available to the life sciences community but also to all the potential customers including the other European research infrastructures. Around a potential central ELIXIR node at EBI, a distributed physical infrastructure is required to enable the storage of databases but also their replication, curation, indexing and query.

Data integration can be achieved by providing common interfaces and rules to make all those resources interoperable. For this purpose, description of data should comply with common definitions using concepts from the scientific domain and interfaces should provide a common way to handle these data, which means in turn, to define a common protocol to make these services communicate with each other.

2.1. Status and Opportunities

ELIXIR is expected to investigate a service-oriented architecture making extensive use of the Web Services (WSs) technology. The main advantages of WSs are the following:

- They offer great interoperability (mainly because of standardised specifications).
- They enable communication between processes and data transfers independently of the programming language used in the underlying applications. Therefore, by extension, virtually almost any piece of software can be exposed as a WS.
- They can be considered as firewall-friendly, because they are based on standard internet protocols.

In terms of adoption, WSs are already widely used both in the bioinformatics and in the grid communities for the integration of heterogeneous tools and data repositories. In terms of sustainability, WSs have been largely promoted by the computing industry. Although specifications are still evolving significantly, many studies have shown that there is a large business in the years to come for internet-based services for which this technology is exactly fitted.

If standardization of interfaces with WSs can drastically increase the interoperability between bioinformatics resources, it is however, only half of the effort required to build truly interoperable resources. By operating on standardized data formats, bioinformatics resources can be integrated almost readily in complete bioinformatics pipelines without having to restructure data between each service. It also opens the path to the sharing of data between initiatives such as the other European research infrastructures in the area of life sciences. The definition and adoption of common data formats requires the definition of agreed ontologies and syntaxes in close interaction with the user communities (see sections 3, 4 and 5).

WP7 builds on the work done in the EMBRACE Network of Excellence (NoE) [EMBRACE], funded by the European Commission. To achieve its objectives, EMBRACE adopted a set of recommendations, which can be summarized in three major points:
All resources (data bases, tools) should be exposed as web services obeying precise specifications (presently WSDL and WSs*) and exchanging information through the Simple Object Access Protocol (SOAP) protocol.

These WSs should have standard data types and be annotated according to community specific ontologies.

The WSs should be registered in a central registry of services.

After identifying the proper technical specifications, the EMBRACE project is now:
a) developing WSs interfaces to a number of databases and tools, including the most prestigious molecular databases (EnsEMBL, Hogenom, ProDom, UniProt) and bioinformatics algorithms (BLAST, CLustalW, EMBOSS) to facilitate their integration into biological analysis workflows; b) collecting web services produced in other European projects, such as BioSapiens [BioSapiens] and ENFIN [ENFIN]. The EMBRACE project is also deploying a number of biological use cases to validate the approach and identify potential bottlenecks.

The EMBRACE registry of services has been set up [EMBRACE Registry] where more than 800 services are now available. This registry will soon be replaced by BioCatalogue [BioCatalogue], a curated catalogue offering support to search, register, annotate and monitor life sciences web services.

The EMBRACE recommendations can be seen as a starting corpus for the design of the distributed data infrastructure that would be needed for a European project such as ELIXIR:

- A central node (tier-0) at EBI hosting core biomolecular resources including the ELIXIR registry of services
- Tier-1 nodes hosting additional core biomolecular resources such as SIB
- Tier-2 nodes hosting specialist biomolecular resources (BRENDA, IMGT, etc) and Model Organism resources (MGD, Flybase, etc)

The proposed approach for data integration relies also on the capacity to access computing resources needed for updating, indexing and curating the databases at tier-1 and tier-2 nodes. Such computing resources could be provided partly by the National Grid Initiatives (NGIs) coordinated within the European Grid Initiative (EGI) federation [EGI]. EGI is an ESFRI project under design in parallel with ELIXIR which aims at federating the National Grid Initiatives in Europe.

Progress with grid technology is such that security requirements for the handling of biological and medical data are now addressed by the middlewares deployed on NGIs in Europe [HealthGrid Conference proceedings].

2.2. Recommendations

Leverage on existing projects. ELIXIR should build a distributed data infrastructure on the foundations laid down by EMBRACE, namely a Service Oriented Architecture using WSs technology. CASIMIR, like EMBRACE, has identified the desirability of using a web-services-based mechanism to interlink databases [CASIMIR recommendations]. CASIMIR
and members of ENFIN are in the process of publishing a Database Description Framework of particular interest to the ELIXIR activities.

Coordination within ELIXIR nodes. The choice of WSs technology for interfacing data has a direct impact on the strategy for the integration of tool, databases. WSs technology is also foreseen as the mechanism to access computing and storage resources through grid services, and that impacts the integration of physical resources. A common policy for service provision needs to be developed and agreed by the participating ELIXIR nodes.

Coordination with other ESFRI projects. ELIXIR resources and services are needed by the ESFRI infrastructures in the field of biomedical sciences: contacts already established with BBMRI, EATRIS and InfraFrontier will be pursued actively in order to reach a consensus on programmatic access to the data. Moreover, other infrastructures such as Lifewatch dealing with biodiversity will benefit from the ELIXIR project. The corresponding ESFRI design studies should be involved as early as possible in the definition of interfaces in order to ensure that ELIXIR service oriented architecture fits their needs.

Coordination with the EGI. ELIXIR nodes will be involved in their respective NGI, which are setting up standard access to computing and storage resources. In this context, it is critical that ELIXIR works closely with the EGI to ensure that its resources are integrated into the local NGI and in turn the NGI provides storage and computing resources to manage ELIXIR data resources.
3. Nomenclatures – Theme B

The need for classification and naming of biological entities is as old as life sciences themselves. Historically, naming has often been seen as a reward to a ‘discoverer’ and nomenclatures were not really aimed at being useful to the whole community but were rather designed to fit a very specific field. Times have changed, data appear more and more interwoven, and the community as a whole is now looking forwards for ways of naming entities that could be informative for all.

3.1. Status and Opportunities

One of the areas in the life sciences that require a definite nomenclature effort is that of gene symbols and protein names. While quite a number of model organism databases or organizations have established nomenclature committees, guidelines and repositories, there is yet no pan-organism effort to establish consistency between these guidelines and the gene symbols that are being attributed. There are cases of “local” collaborations like that existing between the HUGO Gene Nomenclature Committee (HGNC) [HUGO GN] and the mouse genome informatics database (MGI) that tries to ensure the use of the same symbols in human and mouse in when genes are clearly orthologous. Some of these organism-centric resources also provide nomenclature resources concerning other biological objects such as alleles, mutations, chromosome aberrations, transposons and strain names.

In the field of proteins, an effort is lead by the EBI and the SIB, to establish, a protein nomenclature resource in the framework of the UniProtKB resource. As a first step, a compendium of guidelines on how to best name proteins has been produced [NameProt].

Taxonomy nomenclature is somewhat satisfactorily dealt with by the NCBI taxonomy database that contains the names of all organisms represented in the genetic databases by at least one nucleotide or protein sequence. The shortcoming of such a resource is that it does not address species that are not subject to any sequencing effort. More global resources are being put in place for example at ITIS (Integrated Taxonomic Information System) or in the framework of the Encyclopedia of Life (EoL) [EoL], which according to its mission statement is a “project to organize and make available via the Internet virtually all information about life present on Earth”. There are also many problems relevant to the precise definition of what really defines a species. In the cases of bacteria, archaea and viruses, the most prevalent organisms on earth, all of the classical definitions have been made obsolete by the advances brought along by classical sequencing as well as by environmental metagenomics efforts.

3.2. Recommendations

**Coordination on gene symbols.** A pan-organism effort should be initiated jointly by ELIXIR and HGNC, to harmonize the different guidelines and the gene symbols that are being attributed.

**Enhancement of taxonomy nomenclature.** Initiate a new activity jointly with the existing initiatives, such as ITIS and EoL, the environmental and metagenomics communities, for example the Genomics Standards Consortium [GSC].
4. **Controlled Vocabularies and Ontologies – Theme C**

The major biomedical databases understood from their inception the importance of controlled vocabularies for the annotation of data. This is evidenced, for example, by the development of the "Feature Table" for the Nucleotide Sequence Data Library [FT] and the development of a comprehensive keyword list for the Swiss-Prot Protein Database [SWKey]. The importance of these controlled vocabularies is that they provide a mechanism for the rigorous retrieval of data from databases. However, in the last decade or so, there has been the important development from controlled vocabularies to structured vocabularies (often known as ontologies). Structured vocabularies differ from controlled vocabularies in that, in the former, terms in the vocabulary bear a particular logically defined relationship to each other. The advantage of this is that it allows computational reasoning on data annotated with a structured vocabulary. A second development of the last decade has been the common adoption by diverse databases of the same ontology within a given biomedical domain. The advantage of this is that it brings around a degree of interoperability of these databases. This can be illustrated by the Gene Ontology (GO) project [GO]. The GO has developed a large structured vocabulary for the annotation of the functions of gene products. The GO is used to annotate gene products by over 20 model organism genomic databases and by large pan-organism databases such as UniProt. As a consequence, all of these databases can be queried – either by a browser or by computer program – from a single source.

It is increasingly apparent that the common use of ontologies by databases and other resources will be a very important factor to ensure the interoperability of biomedical data. For this reason we have seen, in recent years, the development of ontologies in several different biomedical domains. It is of great importance that these efforts are not made in isolation, but as the result of a close coordination and collaboration between the relevant groups. There are several reasons why this is so:

1. Ontology development should be a community effort, seeking broad community input and acceptance.
2. Ontologies should conform to a commonly accepted set of standards and be available in a small number of widely used formats.
3. Within any particular biomedical domain the community needs a single accepted ontology rather than several ontologies that compete (since that would defeat the purpose).
4. Ontologies themselves need to interoperate. For example GO needs to include terms from the ontology of chemicals (CheBI) [ChEBI].

For this reason an informal umbrella for biomedical ontologies has been formed jointly by groups in Europe and USA: OBO, Open Biomedical Ontology [OBO] portal. OBO has established a number of criteria for biomedical ontology development and those ontologies that have agreed to work towards satisfying these criteria are listed as candidate members of the OBO Foundry [Smith, 2007]. Three OBO Foundry meeting have been co-funded by ELIXIR and BBRSC funds to Susanna-Assunta Sansone (see section 1.1) [WODS meetings].

8 of 29
At of September 2008, the OBO portal includes 73 different ontologies (53 are candidate members of the OBO Foundry). Of these 33 are the sole or joint products of European groups. A list of these, with those with major European input marked, is attached as Appendix 1.

4.1. Status and Opportunities

Funding. Many, if not most, of the ontologies included within OBO are unfunded. There are exceptions, for example GO is funded by the NIH (USA) and PATO, the ontology of qualities, by the BBSRC (UK). We clearly need, at the European level, a modality by means of which new ontologies can be developed (see below) and existing ontologies maintained.

Tools. Ontologies require specialised software tools for several different purposes:

- Ontology building. The two major tools now in use for the construction and maintenance of ontologies are OBO-EDIT [OBOEDIT], a product of the Gene Ontology Consortium, and Protégé 4, the product of groups at Stanford and Manchester [Protégé]. In addition the Edinburgh group have developed the COBRA tool for ontology management [COBRA].

- Annotation. Annotation of data within databases or other data sources requires tools. In the past these have normally been home grown by the different databases, but now the Berkeley group have a more generic annotation tool under development, Phenote [Phenote], and this is being tested by several of the model organism database groups. Similarly, tools like Proteome Harvest [ProteomeHarvest] and ISAcreator [ISAcreator] assist users in reporting and describing the experimental metadata using ontologies, served via the EBI Ontology Lookup Service [OLS] (see also section 5).

Annotation. There is an increasing realisation that the bulk and complexity of biomedical data requires new approaches to the problem of annotation. At least four classes of initiative are now taking place:

1. The first of these is to develop mechanisms for community annotation [Mons, 2008], usually using wikimedia technologies. Examples of this are the Wikipedia pages for community annotation of RNA families used by the Rfam database [Rfam] and the community annotation pages for the E. coli genome [coli]. Here, the soon to be established International Society for Biocurators [BioCurators] can play a major role.

2. The second is the development of software plugins for general purpose utilities such as the Creative Commons and Microsoft ontology add-in for Word 2007 [Word07Ontology] which allows authors, at the time they actually are writing a scientific manuscript, to enrich their text by semantic markup with terms from ontologies or controlled vocabularies of, for example, gene and protein names [Fink and Bourne, 2007].

3. The third is the development of metadata standards and for the annotation of large corpora of experimental data, see section 5.

4. Finally methods for the extraction of structured data from published text and abstracts by natural language processing are being used both retrospectively and prospectively by the journal publishers and others (see [RSC] for an example).
One lack is the ability of the community using ontologies to communicate globally with the OBO community to request new terms, or suggest corrections, to an existing ontology. There has been discussion, but little progress, with the development of an Ontology Request Broker (ORB) that would allow users to easily interact with the developers of the individual ontologies through a single interface.

**Browsing and data analysis.** Annotated data can be either browsed or queried computationally. There has been considerable development of web-based ontology browsers (e.g. AMIGO [AMIGO]) and of open source or commercial software to analyse data (see [GOtools]).

4.2. Recommendations

**Coordination and funding of ontology development.** Above all action is needed to ensure funding instruments and the coordination of ontology development in the biomedical domain via the OBO Foundry (that at the present is unfunded). Coordination is of vital importance to avoid the duplication of effort and the effective interoperability of different ontologies. Despite recent progress there are major domains, which require concerted community effort to develop ontologies. Three of these can be listed:

1. Disease. There is a long tradition in medicine for the development of structured controlled vocabularies for the description of disease. These include such artefacts as the ICD codes [ICD] and SNOMED [SNOMED]. For the purpose of biomedical research these are far from adequate, having been primarily designed for epidemiology and hospital billing, respectively. There is an OBO disease ontology [DO] in development, but this is not yet funded. In addition there is a Human Phenotype Ontology from Berlin [HPO], but this does not yet meet OBO standards. Related to disease ontologies, and to many others, is PATO. Any future ontology in this field must take account of both the ICD codes (ICD-9 and ICCD-10) and SNOMED-CT, since it will be very important that legacy data be accommodated within any future structure. It is important to emphasise the role of domain experts in disease ontologies, such as human and model organism clinical and pathology communities

2. Anatomies. Fundamental to the annotation of much biological data are anatomical structures. There are ontologies within the domains of human anatomy [FMA], mouse anatomy [JAX] and the anatomies of several model organisms (e.g. Drosophila [FBanatomy], zebrafish [ZFIN], medaka [medaka] and plants [PO]), but much effort is required to integrate these. In addition, work is needed to develop an ontology of anatomical homologies, so that information about, for example, mouse and human phenotypes can be seamlessly mined. Related to anatomical ontologies are those to be used for the description of cells and tissues. Although efforts have been made in these fields (see [CELL]) there is considerable work that needs to be done. Another aspect of anatomical ontologies and similar artefacts is important, that is the need to integrate with efforts to visualize anatomies. These are being developed for the human (e.g. The NLM’s Visible Human Project [NLM’s VHP]) as well as several efforts for the mouse, e.g. The Visible Mouse Project [VMP] and the EMAGE project [EMAGE]).
3. Organismal taxonomies. In the biomedical database field many use the Taxman product from the NCBI [Taxman], and the related NEWT product of the SIB [NEWT], to annotate organisms by their accepted names and taxonomic position. At present these efforts are independent of major work in the biodiversity field, such as the Encyclopedia of Life [EoL], the Tree of Life [ToL] and GBIF [GBIF], let alone more specialist databases such as the Index Fungorum [INDEX] and the International Plant Names Index [IPNI], to catalog the Earth’s biodiversity. These resources need to be integrated. Related to this are efforts to develop ontologies for the description of organismal environments [EnVO] and similar artefacts for geographical data [GAZ].

Software development. Most urgent is the development of further open source tools for annotation of data using ontologies and the development of open source tools which will allow data providers, including individual scientists that author papers, to richly mark-up their data and texts with terms extracted from relevant ontologies. Several tools are emerging, for example Terminizer [Terminizer] by the UK’s NERC Bioinformatic Center (NEBC) [NERC-NEBC] and Ontology Annotation sysTem [OATH] developed at Harwell.

Coordination with other ESFRI projects. The work described and proposed here is not only relevant to ELIXIR. Coordination efforts are also needed at the level of ontology users, to ensure the correct and efficient use of ontologies for data annotation. Other projects within ESFRI, in both the biomedical and environmental domains, will (or should) greatly benefit from the development of community ontologies under the auspices of ELIXIR. For example, INFRAFRONTIER will need murine anatomical ontologies and phenotype ontologies, and OBI, the Ontology of Biomedical Investigations [OBI]; BBMRI, ECRIN and EATRIS will need anatomical, tissue and disease ontologies and, probably, the structured geographical gazetteer (GAZ); LIFEWATCH will need the ontology of environments and the gazetteer.
5. Reporting Requirements – Theme D

In the area of life science, the cycle of data generation and processing is being vastly accelerated by the development of high-throughput experimental methods associated with genomic and post-genomic technologies (e.g., genomics, transcriptomics, proteomics, and metabolomics, hereafter referred as ‘omics’). Biological and biomedical studies commonly range from simple one assay-based to complex multi-assay studies. For the latter type, for example, consider the reporting of a complex multi-assay study looking at the effect on a number of subjects of a compound by characterizing the metabolic profile of their urine (i.e. by mass spectroscopy), measuring protein and gene expression in the liver (i.e. by mass spectrometry and DNA microarrays, respectively), and conducting conventional analysis (i.e. histology). Similar examples can be found in the environmental and other domains of the life science. Such studies are information intensive and to record their complex structure it is necessary to define and capture the experimental metadata, including experimental design, sample source(s) and treatment(s), the preparation of the sample for the analytical assay, the processes and instruments used throughout, and the final data. It is widely recognized that capturing experimental metadata on this level of granularity is necessary for enabling efficient data sharing and meaningful data mining.

The old adage “garbage in, garbage out” is constantly reiterated in the world of database development. Databases for ‘omics-based’ data are not immune to the pitfalls of the poorly guarded data storage system and may easily contain data with insufficient metadata to describe the studies. As the size and complexity of the datasets and the corresponding information stores grow, standards for collecting, describing, formatting, submitting and exchanging information are playing an increasingly active role. Consistent reporting of the experimental metadata and associated data has a positive and long-lasting impact on the value of collective scientific outputs. This has also been recognized by funding agencies that are therefore increasingly engaging in strategic planning for reporting and management of the datasets, often through the development of highly publicized data policies [ESRA], [NERC], [NERC-NEBC], [NSF], [NIH], [GBMF], [Genome Canada], [BBSRC], [MRC], [WT]. Many journals also require compliance with reporting requirements, contingent on their favourable reception by the scientific community and the availability of appropriate software tools and public repositories [Editorial Nature Biotechnology, 2006], [Editorial Nature Biotechnology, 2007].

5.1. Status and Opportunities

Reporting standards initiatives. To coordinate the description and the reporting of such heterogeneous studies, new approaches for communicating the complex metadata are required to correctly interpret the final results that they contextualize. Many groups have risen to this challenge and several standards initiatives occupy strategic positions in the international scenario, largely falling into two groups identifiable by the needs of their respective user communities.

One group of initiatives is driven by regulatory frameworks, and often supported by accredited (de jure) Standards Developing Organizations (SDOs). Most significantly, these
efforts focus on the Voluntary eXploratory Data Submissions (VXDS) and electronic data submission programs of the US Food and Drug Administration (FDA) [US HHS/FDA Guidance for Industry: Pharmacogenomic data submissions, 2005], [Frueh, 2006], [Tong, 2007] and the US Environmental Protection Agency (EPA) [US EPA Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA, 2004]. These initiatives also include long-standing efforts in the clinical and non-clinical domains [CDISC, SEND, CRIX, JANUS, FIREBIRD] alongside more recent activities in the pharmacogenomics area that add complex omics technologies to biomedical studies [Shabo, 2006].

A second group of initiatives that address particular technologies (i.e. genomics, microarray, proteomics and metabol/nomics) or defined domains of application (i.e. system biology, pathways, bio-banks) have emerged from the research community, in many cases with the support of commercial organizations such as instrument vendors and service providers. Such initiatives are focused on facilitating data management, supporting tool interoperability and enabling data exchange among public and proprietary systems, often by developing common three kinds of (de facto) reporting standards: minimal information checklists, ontologies and file formats (i.e., [Le Novère, 2006], [Ball and Brazma, 2006], [Orchard, 2007], [Sansone, 2007], [Field, 2008 a], [Field, 2008 b], [P3G], [PaGE-OM]). Minimal information checklists are easy-to-read, structured documents that reflect the consensus view of the essential pieces of information that should be reported; ontologies provide the semantics needed to describe the minimal information requirements and file formats the syntax to transmit and exchange these. Combining these three kinds of reporting standards a submission tool, for example, should guide researchers through the process of meeting the reporting requirements made by a given minimal information specification, enable straightforward practical use of ontology terms and export the collected information in a standard format to a given database (i.e. [Edgar R and Barrett, 2006], [Brazma and Parkinson, 2006], [Jones, 2006], [Hermjakob, 2006]).

The Genomic Standards Consortium (GSC) [GCS] stands out among others - in this second group - for an innovative initiative. Recently the GSC moved beyond the development of standards to improve data capture at the level of the scientific publication, by launching a new electronic journal with highly structured "Genome Notes" that must be standards compliant [Garrity, 2008]. This initiative marks the growing shift away from a traditional dichotomy between "database entries" and "journal articles" and an increasing adoption of hybrid models of collecting and disseminating scientific information.

**Integrative cross-domain initiatives.** Domain-specific initiatives are regarded as important because they address ‘real world’ data reporting requirements; either for the particular technologies being used or the particular biologically- or biomedical-delineated community concern. Being focused, however, leads to duplication of effort, and more seriously, the development of (largely arbitrarily) different and fragmented reporting standards, severely hindering their application. Nowadays it is becoming increasingly more frequent to get a complete picture of the fundamental biological processes under study. Researchers are able to perform multi-assay studies where the same sample is run through the full range of ‘omics and conventional technologies, in combination. For
example, reiterating the scenario illustrated above (in section 5), when metabolic profile, protein and gene expression are measured in subjects treated with a compound. In this specific case, it is critical that the standards developed for metabol/nomics, proteomics and microarray are designed to be interoperable and fit neatly into a jigsaw, with users being able to take the pieces that are relevant to report their study.

Fortunately, amongst the academic community several synergistic activities are fostering the harmonization and consolidation of the three kinds of (de facto) standards being developed. In addition to the OBO Foundry effort, described in section 4, other synergistic activities include:

- **Content**: Twenty-two groups now participate in the Minimum Information for Biomedical or Biological Investigations (MIBBI) project, which offers a one-stop shop for those exploring the range of extant ‘minimum information’ checklists [Taylor, 2008], [MIBBI] (see Appendix 2). MIBBI significantly fosters collaborative, integrative development of minimum information and has strong links with the Equator Network, an activity leading the standardization of the reporting of clinical trials in the medical literature [EQUATOR].

- **Format**: Several groups participate in the Functional Genomics (FuGE) project to develop a single generic data model that will underpin a variety of file formats -based on the extensible markup language (XML) - based by providing a single common framework [Jones, 2007], [FuGE]. In parallel, another complementary initiative has sprung up from a growing number of communities that work collaboratively on a general purpose, common tabular framework with which to collect and communicate complex metadata [ISA], [Sansone et al., 2008], (i) as a user-friendly presentation layer for XML-based formats (via an XSL transformation), and (ii) to complements existing biomedical formats such as the CDISC Study Data Tabulation Model [SDTM]. A list of reporting formats is being created and will be available in the next version of this report.

Adherence to reporting standards goes a long way to ensuring that databases are stocked with useful information ultimately [Editorial Nature Cell Biology, 2008]. It is critical, however, to maintain a compromise between detail and practicality in reporting, so that compliance with the standards is not so onerous as to inhibit their adoption. Minimum requirements, format and ontology are not a panacea for all potential problems. These are proposed as a method to effectively describe a certain experiment rather than a prescriptive for how one should do experiments or analyze the data they generate [Quackenbush, 2006].

*Implementation of synergistic standards*. The fragmentation severely hinders the interoperability of databases and tools, implementing such reporting standards: this scenario is illustrated by the ArrayExpress [ArrayExpress] ENA-Reads [ENAreads] and PRIDE [PRIDE] - EBI production systems for microarray, sequencing and proteomics data respectively. These systems implement (non-interoperable) standards applicable only for their ‘omics’ technology. Consequently users have to deal with different submission formats and tools, diverse representations of the metadata and terminologies when depositing their datasets in these systems, and similarly when downloading other
datasets. Such fragmentation has a strong impact on the user community, particularly by hampering deposition of complex multi-assay studies.

An example, of how such fragmentation can be solved is illustrated by the ISA infrastructure [ISA], developed to manage metadata from for biological and biomedical studies, which commonly range from simple one assay-based to complex multi-assay studies. The ISA infrastructure’s software components leverage on MIBBI, OBO ontologies and the ISA-Tab format, and can work independently, or as unified system for local use:

- The ISAconfigurator enables a power users (e.g. a curator) to regulate the minimal requirement fields, according to the relevant MIBBI checklist(s), and set their allowed values, for example ontology terms.
- The ISAcreator drives users to report the metadata following the configured requirements, search and select terms from OBO Foundry ontologies, particularly OBI, for relevant metadata fields, using web services provided by the OLS.
- The ISAconverter transforms ISA-Tab formatted metadata into several other related tabular and XML-based formats for submission to ArrayExpress, ENA-Reads and PRIDE.
- The BioInvestigation Index database enables storing and querying functionalities. An instance of the BioInvestigation Index database has been installed as prototype at EBI [BioInvIndex].

Funded to manage complex multi-assay studies from toxicogenomics and nutrigenomics European projects, the ISA infrastructure is being used by several communities in different domain, including NERC NEBC, to ensure the information is collected using common ontologies and reporting requirements.

### 5.2. Recommendations

**Pan-domain coordination and funding.** Most urgent are the funds to manage the process of consensus-building from start to finish. This takes time and expertise; but the time invested in these efforts to build commonalities and synergies among projects is difficult without central grants or with limited funds [WODS]. Above all action a ‘top-down’ coordination is needed to help bringing these standardization efforts closer, addressing the fragmentation and making reporting standards interoperable, as initiated by MIBBI, OBO Foundry and the ISA-Tab communities. Although, regulatory- or biomedical-driven initiatives have far stricter guidelines than academia, much could be learned from exchange of ideas and practices of these sectors. To achieve interoperability from a technical perspective, these ‘meta’ standardization projects need to (i) resolve overlaps between domain-specific reporting standards and (ii) fill gaps where they exist. It is anticipated, however, that some reporting standards will be more mature – ‘ready’ to be integrated – than others, particularly because development takes time and 'buy-in' both from potential users and those that govern them (journals, funders, regulators). These are technically complex, but demonstrably tractable tasks. By contrast, the sociological barriers facing these kinds of large-scale collaborations can be far more challenging, mandating extensive liaison between communities.
Software development. Both software interoperability and the data integration remain challenging as things stand due to the fragmentation of the reporting standards. One of the many benefits accruing to the development of interoperable reporting standards is the increased ease of the development of standards-compliant products by academic and commercial software developers, instrument vendors and others. They do so by limiting the range and variability of standards for such parties to consider, thereby reducing development time (cost). The job of harmonising reporting standards is still very much a work in progress. However, tools such as the ISA components demonstrate that something can be achieved to enable consisted reporting of the experimental information despite fragmented standards scenario.

Coordination with other ESFRI projects. Other projects within ESFRI will greatly benefit from the development and use of common reporting standards. Effective data deposition, management and sharing are essential across all scientific disciplines.
6. Plan for Adoption– Themes B, C and D

Resolving inconsistencies and conflicts between nomenclature, ontological, or reporting resources is of course best addressed by discussions between the different actors. It is obvious that any European-wide funding mechanism should include provisions to make sure that overlaps are avoided and that key players collaborate in the framework of such integrative efforts. This results in more appropriate resources for the biomedical and scientific community, which means that the job of capturing, annotating, integrating, sharing and exploiting (meta)data is simplified, increasing the return on the (largely public) investment of funds that supported their generation.

The massively-collaborative nature of this undertaking requires frequent face-to-face workshops to create the necessary conditions for the building of consensus; unfortunately - for the initiatives emerged from the academic community – this is chronically limited due to lack of financial resources. However, the lack of standardization is quite simply an unacceptable state of affairs, for the researchers, repeatedly proving to be a significant bottleneck in the collection, sharing, and integration of data, for funders and for society at large. This has motivated both developers, and the potential users with whom they consult in the relevant communities, to participate on an almost exclusively voluntary basis.

A few stakeholders also have pivotal roles to play as enablers, to maximize the adoption of common nomenclature and ontological resources, or reporting standards. Many publishers will require compliance with reporting standards, contingent on their favourable reception by the scientific community and the availability of appropriate standards-compliant software tools and public repositories. BioMed Central's journals - with clinical content and BMC Bioinformatics - now endorse the MIBBI portal in the instructions for authors [BMCauthors] [BMCreview] and encourage data deposition, where appropriate. In that respect, the emergence of databases and repositories as media for publishing experimental results will also be crucial. This is already the case in the field of microarrays and in proteomics where repositories are somehow constraining their submitters to abide to existing standards. This trend is expected to be a feature of all life sciences resources in the near future, along with hybrid models of collecting and disseminating scientific information, such as the e-journal initiative launched by the GSC towards standards-compliant genomic and metagenomic publication record [Garrity, 2008].

Another way to ensure compliance with existing standards is to enforce them when researchers are writing papers. This can be best achieved by using semantic tagging methodologies, as described in section 4.2. In such an approach, the researcher has to indicate, while she/he is writing her paper not only exactly what biological objects they are referencing to, but also, if appropriate the relationships between those objects and the methodologies used in the experiments.

Funding agencies also play an active role in the strategic stewardship of omics data, often through the development of data policies encouraging the use of (existing) nomenclatures,
ontologies, and reporting standards and public standards-compliant repositories for data collection and management (e.g. BBSRC, NERC-NEBC) [BBSRC], [NERC-NEBC]. Exceptionally, funders may also track, and commit to contribute additional investment and organizational support (e.g., dedicated staff, meetings *inter alia*) to the evolution of such standards [NERC-NEBC].
7. References

[AMIGO] http://amigo.geneontology.org/cgi-bin/amigo/go.cgi
[ArrayExpress] http://www.ebi.ac.uk/arrayexpress
[BioInvIndex] http://www.ebi.ac.uk/bioinvindex
[CaBIG] https://cabig.nci.nih.gov/
[CELL] http://www.obofoundry.org/cgi-bin/detail.cgi?id=cell
[ChEBI] http://www.ebi.ac.uk/chebi
[Ccobra] http://www.ai.aed.ac.uk/project/cobra-ct/
[DO meeting] http://www.ebi.ac.uk/industry/Workshops/workshops.html


[MRC] http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/PolicyonDataSharingandPreservation/MRC002551
[NameProt] http://www.uniprot.org/docs/nameprot
[NERC] http://www.nerc.ac.uk/research/sites/data/policy.asp;
[NERC-NEBC] http://nebc.nox.ac.uk,
http://nebc.nox.ac.uk/datapolicy/NEBCDataPolicy.pdf
[NEWT] http://www.ebi.ac.uk/newt/display
[OATH] http://www.har.mrc.ac.uk/news/?id=15
[OBI] http://www.obofoundry.org/cgi-bin/detail.cgi?id=obi
[OLS] http://www.ebi.ac.uk/ontology-lookup

[PRIDE] http://www.ebi.ac.uk/pride
[ProteomeHarvest] http://www.ebi.ac.uk/pride/proteomeharvest
[Rfam] http://www.sanger.ac.uk/Software/Rfam/
[SWkey] http://expasy.org/cgi-bin/keywlist.pl
[Tong, 2007]
[WODS meetings] http://www.ebi.ac.uk/net-project/projects.html#workshop
8. Appendix 1
Ontologies available from OBO (26th September, 2008). Those in italics have exclusive or substantial European input.

<table>
<thead>
<tr>
<th>Full name</th>
<th>Domain</th>
<th>Short name</th>
<th>File name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amphibian gross anatomy</td>
<td>anatomy</td>
<td>AAO</td>
<td>amphibian_anatomy.obo</td>
</tr>
<tr>
<td>Biological process</td>
<td>biological process</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
</tr>
<tr>
<td>C. elegans development</td>
<td>anatomy</td>
<td>WBl</td>
<td>worm_development.obo</td>
</tr>
<tr>
<td>C. elegans gross anatomy</td>
<td>anatomy</td>
<td>WBbt</td>
<td>WBbt.obo</td>
</tr>
<tr>
<td>C. elegans phenotype</td>
<td>phenotype</td>
<td>WBP</td>
<td>worm_phenotype.obo</td>
</tr>
<tr>
<td>Cell type</td>
<td>anatomy</td>
<td>CL</td>
<td>cell.obo</td>
</tr>
<tr>
<td>Cellular component</td>
<td>anatomy</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
</tr>
<tr>
<td>Cereal plant trait</td>
<td>phenotype</td>
<td>TO</td>
<td>plant_trait.obo</td>
</tr>
<tr>
<td>Chemical entities of biological interest</td>
<td>biochemistry</td>
<td>CHEBI</td>
<td>chebi.obo</td>
</tr>
<tr>
<td>Common Anatomy Reference Ontology</td>
<td>anatomy</td>
<td>CARO</td>
<td>caro.obo</td>
</tr>
<tr>
<td>Dictyostelium discoideum anatomy</td>
<td>anatomy</td>
<td>DDANAT</td>
<td>dictyostelium_anatomy.obo</td>
</tr>
<tr>
<td>Drosophila development</td>
<td>anatomy</td>
<td>FBdv</td>
<td>fly_development.obo</td>
</tr>
<tr>
<td>Drosophila gross anatomy</td>
<td>anatomy</td>
<td>FBbt</td>
<td>fly_anatomy.obo</td>
</tr>
<tr>
<td>Environment Ontology</td>
<td>environment</td>
<td>ENVO</td>
<td>envo.obo</td>
</tr>
<tr>
<td>Evidence codes</td>
<td>experiments</td>
<td>ECO</td>
<td>evidence_code.obo</td>
</tr>
<tr>
<td>Fly taxonomy</td>
<td>taxonomy</td>
<td>FBsp</td>
<td>fly_taxonomy.obo</td>
</tr>
<tr>
<td>Foundational Model of Anatomy (subset)</td>
<td>anatomy</td>
<td>FMA</td>
<td>fma.obo</td>
</tr>
<tr>
<td>Fungal gross anatomy</td>
<td>anatomy</td>
<td>FAO</td>
<td>fungal_anatomy.obo</td>
</tr>
<tr>
<td>Human developmental anatomy, abstract</td>
<td>anatomy</td>
<td>EHDAA</td>
<td>human-dev-anat-abstract.obo</td>
</tr>
<tr>
<td>Human developmental anatomy, timed version</td>
<td>anatomy</td>
<td>EHDA</td>
<td>human-dev-anat-staged.obo</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>---------</td>
<td>------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Human disease</td>
<td>health</td>
<td>DOID</td>
<td>human_disease.obo</td>
</tr>
<tr>
<td>Infectious disease</td>
<td>health</td>
<td>IDO</td>
<td>infectious_disease.obo</td>
</tr>
<tr>
<td>Mammalian phenotype</td>
<td>phenotype</td>
<td>MP</td>
<td>mammalian_phenotype.obo</td>
</tr>
<tr>
<td>Mass spectrometry</td>
<td>experiments</td>
<td>MS</td>
<td>psi-ms.obo</td>
</tr>
<tr>
<td>Medaka fish anatomy and development</td>
<td>anatomy</td>
<td>MFO</td>
<td>medaka_ontology.obo</td>
</tr>
<tr>
<td>Molecular function</td>
<td>biological function</td>
<td>GO</td>
<td>gene_ontology_edit.obo</td>
</tr>
<tr>
<td>Mosquito gross anatomy</td>
<td>anatomy</td>
<td>TGMA</td>
<td>mosquito_anatomy.obo</td>
</tr>
<tr>
<td>Mosquito insecticide resistance</td>
<td>environment</td>
<td>MIRO</td>
<td>mosquito_insecticide_resistance.obo</td>
</tr>
<tr>
<td>Mouse adult gross anatomy</td>
<td>anatomy</td>
<td>MA</td>
<td>adult_mouse_anatomy.obo</td>
</tr>
<tr>
<td>Mouse gross anatomy and development</td>
<td>anatomy</td>
<td>EMAP</td>
<td></td>
</tr>
<tr>
<td>Mouse pathology</td>
<td>health</td>
<td>MPATH</td>
<td>mouse_pathology.obo</td>
</tr>
<tr>
<td>NMR-instrument specific component of metabolomics investigations</td>
<td>experiments</td>
<td>NMR</td>
<td>NMR.owl</td>
</tr>
<tr>
<td>OBO relationship types</td>
<td>all</td>
<td>OBO_REL</td>
<td>ro.obo</td>
</tr>
<tr>
<td>Ontology for biomedical investigations</td>
<td>experiments</td>
<td>OBI</td>
<td>obi.owl</td>
</tr>
<tr>
<td>Pathogen transmission</td>
<td>health</td>
<td>TRANS</td>
<td>transmission_process.obo</td>
</tr>
<tr>
<td>Phenotypic quality</td>
<td>phenotype</td>
<td>PATO</td>
<td>quality.obo</td>
</tr>
<tr>
<td>Plant growth and developmental stage</td>
<td>anatomy</td>
<td>PO</td>
<td>po_temporal.obo</td>
</tr>
<tr>
<td>Plant structure</td>
<td>anatomy</td>
<td>PO</td>
<td>po_anatomy.obo</td>
</tr>
<tr>
<td>Protein modification</td>
<td>proteins</td>
<td>MOD</td>
<td>PSI-MOD.obo</td>
</tr>
<tr>
<td>protein ontology</td>
<td>proteins</td>
<td>PRO</td>
<td>pro.obo</td>
</tr>
<tr>
<td>Protein-protein interaction</td>
<td>experiments</td>
<td>MI</td>
<td>psi-mi.obo</td>
</tr>
<tr>
<td>Sequence types and features</td>
<td>biological sequence</td>
<td>SO</td>
<td>so.obo</td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------------</td>
<td>--------</td>
<td>--------------</td>
</tr>
<tr>
<td>Spatial Ontology</td>
<td>anatomy</td>
<td>BSPO</td>
<td>spatial.obo</td>
</tr>
<tr>
<td>Spider Ontology</td>
<td>anatomy</td>
<td>SPD</td>
<td>spider_comparative_biology.obo</td>
</tr>
<tr>
<td>Subcellular anatomy ontology</td>
<td>anatomy</td>
<td>SAO</td>
<td>SAO.owl</td>
</tr>
<tr>
<td>Suggested Ontology for Pharmacogenomics</td>
<td>health</td>
<td>SOPHARM</td>
<td>sopharm</td>
</tr>
<tr>
<td>Systems Biology</td>
<td>biochemistry</td>
<td>SBO</td>
<td>SBO_OBO.obo</td>
</tr>
<tr>
<td>Teleost anatomy and development</td>
<td>anatomy</td>
<td>TAO</td>
<td>teleost_anatomy.obo</td>
</tr>
<tr>
<td>Teleost taxonomy</td>
<td>taxonomy</td>
<td>TTO</td>
<td>teleost_taxonomy.obo</td>
</tr>
<tr>
<td>Tick gross anatomy</td>
<td>anatomy</td>
<td>TADS</td>
<td>tick_anatomy.obo</td>
</tr>
<tr>
<td>Units of measurement</td>
<td>phenotype</td>
<td>LIO</td>
<td>unit.obo</td>
</tr>
<tr>
<td>Xenopus anatomy and development</td>
<td>anatomy</td>
<td>XAO</td>
<td>xenopus_anatomy.obo</td>
</tr>
<tr>
<td>Yeast phenotypes</td>
<td>phenotype</td>
<td>YPO</td>
<td>yeast_phenotype.obo</td>
</tr>
<tr>
<td>Zebrafish anatomy and development</td>
<td>anatomy</td>
<td>ZFA</td>
<td>zebrafish_anatomy.obo</td>
</tr>
<tr>
<td>Bilateria anatomy</td>
<td>anatomy</td>
<td>BILA</td>
<td>bilateria_mrca.obo</td>
</tr>
<tr>
<td>Biological imaging methods</td>
<td>experiments</td>
<td>FBbi</td>
<td>image.obo</td>
</tr>
<tr>
<td>BRENDA tissue / enzyme source</td>
<td>anatomy</td>
<td>BTO</td>
<td>BrendaTissueOBO</td>
</tr>
<tr>
<td>Event (INOH pathway ontology)</td>
<td>biological process</td>
<td>IEV</td>
<td>Event Ontology Download</td>
</tr>
<tr>
<td>eVOC (Expressed Sequence Annotation for Humans)</td>
<td>experiments</td>
<td>EV</td>
<td>evoc_v2.7.obo.tar.gz</td>
</tr>
<tr>
<td>Gene Regulation Ontology</td>
<td>genomic</td>
<td>BOOTStrep</td>
<td>GRO_v0.3.owl</td>
</tr>
<tr>
<td>human phenotype ontology</td>
<td>phenotype</td>
<td>HP</td>
<td>human_phenotype.obo</td>
</tr>
<tr>
<td>Microarray experimental conditions</td>
<td>experiments</td>
<td>MO</td>
<td>MGEDOntology.owl</td>
</tr>
<tr>
<td>Molecule role (INOH Protein name/family name)</td>
<td>proteins</td>
<td>IMR</td>
<td>molecule_role.obo</td>
</tr>
<tr>
<td>ontology)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>NCBI organismal classification</td>
<td>taxonomy</td>
<td>NCBITaxon</td>
<td>taxonomy.dat</td>
</tr>
<tr>
<td>NCI Thesaurus</td>
<td>health</td>
<td>NCIt</td>
<td>EVS ftp site</td>
</tr>
<tr>
<td>Pathway ontology</td>
<td>biological process</td>
<td>PW</td>
<td>pathway.obo</td>
</tr>
<tr>
<td>Physico-chemical methods and properties</td>
<td></td>
<td>FIX</td>
<td>fix.obo</td>
</tr>
<tr>
<td>Physico-chemical process</td>
<td></td>
<td>REX</td>
<td>rex.obo</td>
</tr>
<tr>
<td>Plant environmental conditions</td>
<td>environment</td>
<td>EO</td>
<td>environment_ontology.obo</td>
</tr>
<tr>
<td><em>Plasmodium</em> life cycle</td>
<td>anatomy</td>
<td>PLO</td>
<td>PLO_ontology.txt</td>
</tr>
<tr>
<td>Protein covalent bond</td>
<td>proteins</td>
<td>RESID</td>
<td>RESIDUES.XML</td>
</tr>
<tr>
<td>Protein Domains</td>
<td>proteins</td>
<td>IPR</td>
<td>InterPro FTP directory</td>
</tr>
<tr>
<td>Proteomics data and process provenance</td>
<td>proteins</td>
<td>ProPreO</td>
<td>ProPreO-060506.owl</td>
</tr>
<tr>
<td>Uber anatomy ontology</td>
<td>anatomy</td>
<td>UBERON</td>
<td>uberon.obo</td>
</tr>
<tr>
<td>UniProt taxonomy</td>
<td>taxonomy</td>
<td>NEWT</td>
<td>Organism identification code list</td>
</tr>
</tbody>
</table>
9. Appendix 2

Minimum information guidelines for diverse bioscience domains available from MIBBI Portal (13th November, 2008).

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIMR</td>
<td>Core Information for Metabolomics Reporting</td>
</tr>
<tr>
<td>MIABE</td>
<td>Minimal Information About a Bioactive Entity</td>
</tr>
<tr>
<td>MIACA</td>
<td>Minimal Information About a Cellular Assay</td>
</tr>
<tr>
<td>MIAME</td>
<td>Minimum Information About a Microarray Experiment</td>
</tr>
<tr>
<td>MIAME/Env</td>
<td>MIAME / Environmental transcriptomic experiment</td>
</tr>
<tr>
<td>MIAME/Nutr</td>
<td>MIAME / Nutrigenomics</td>
</tr>
<tr>
<td>MIAME/Plant</td>
<td>MIAME / Plant transcriptomics</td>
</tr>
<tr>
<td>MIAME/Tox</td>
<td>MIAME / Toxicogenomics</td>
</tr>
<tr>
<td>MIAPA</td>
<td>Minimum Information About a Phylogenetic Analysis</td>
</tr>
<tr>
<td>MIAPAR</td>
<td>Minimum Information About a Protein Affinity Reagent</td>
</tr>
<tr>
<td>MIAPE</td>
<td>Minimum Information About a Proteomics Experiment</td>
</tr>
<tr>
<td>MIARE</td>
<td>Minimum Information About a RNAi Experiment</td>
</tr>
<tr>
<td>MIASE</td>
<td>Minimum Information About a Simulation Experiment</td>
</tr>
<tr>
<td>MIENS</td>
<td>Minimum Information about an ENvironmental Sequence</td>
</tr>
<tr>
<td>MIFlowCyt</td>
<td>Minimum Information for a Flow Cytometry Experiment</td>
</tr>
<tr>
<td>MIGen</td>
<td>Minimum Information about a Genotyping Experiment</td>
</tr>
<tr>
<td>MIGS</td>
<td>Minimum Information about a Genome Sequence</td>
</tr>
<tr>
<td>MIMIx</td>
<td>Minimum Information about a Molecular Interaction Experiment</td>
</tr>
<tr>
<td>MIMPP</td>
<td>Minimal Information for Mouse Phenotyping Procedures</td>
</tr>
<tr>
<td>MINI</td>
<td>Minimum Information about a Neuroscience Investigation</td>
</tr>
<tr>
<td>MINIMESS</td>
<td>Minimal Metagenome Sequence Analysis Standard</td>
</tr>
<tr>
<td>MINSEQE</td>
<td>Minimum Information about a high-throughput Sequencing Experiment</td>
</tr>
<tr>
<td>MIPFE</td>
<td>Minimal Information for Protein Functional Evaluation</td>
</tr>
<tr>
<td>MIQAS</td>
<td>Minimal Information for QTLs and Association Studies</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>MIqPCR</td>
<td>Minimum Information about a quantitative Polymerase Chain Reaction experiment</td>
</tr>
<tr>
<td>MIRIAM</td>
<td>Minimal Information Required In the Annotation of biochemical Models</td>
</tr>
<tr>
<td>MISFISHIE</td>
<td>Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments</td>
</tr>
<tr>
<td>STRENDÁ</td>
<td>Standards for Reporting Enzymology Data</td>
</tr>
</tbody>
</table>