# Promoting Coherent Minimum Reporting Requirements for Biological and Biomedical Investigations: The MIBBI Project

*Chris F Taylor[1,2], Dawn Field[2,3], Susanna-Assunta Sansone[1], Jan Aerts[4], Rolf Apweiler[1], Michael Ashburner[5], Catherine A Ball[6], Pierre-Alain Binz[7,8], Molly Bogue[9], Tim Booth[2], Alvis Brazma[1], Ryan R Brinkman[10], Adam Michael Clark[11], Eric W Deutsch[12], Oliver Fiehn[13], Jennifer Fostel[14], Peter Ghazal[15], Frank Gibson[16], Tanya Gray[2,3], Graeme Grimes[15], Nigel W Hardy[17], Henning Hermjakob[1], Randall K Julian, Jr.[18], Matthew Kane[19], Carsten Kettner[20], Christopher Kinsinger[21], Eugene Kolker[22,23], Martin Kuiper[24a,b], Nicolas Le Novère[1], Jim Leebens-Mack[25], Suzanna E Lewis[26], Phillip Lord[16], Ann-Marie Mallon[27], Nishanth Marthandan[28], Hiroshi Masuya[29], Ruth McNally[30], Alexander Mehrle[31], Norman Morrison[2,32], John Quackenbush[33], James M Reecy[34], Donald G Robertson[35], Philippe Rocca-Serra[1,36], Henry Rodriguez[21], Heiko Rosenfelder[31], Javier Santoyo-Lopez[15], Richard H Scheuermann[28], Daniel Schober[1], Barry Smith[37], Jason Snape[38], Keith Tipton[39], Peter Sterk[1], Stefan Wiemann[31]*

*[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK. [2] NERC Environmental Bioinformatics Centre, Mansfield Road, Oxford, OX1 3SR, UK. [3] Molecular Evolution and Bioinformatics Section, Oxford Centre for Ecology and Hydrology, Mansfield Road, Oxford, OX1 3SR, UK. [4] Roslin Institute, Roslin, Midlothian EH25 9PS, UK. [5] Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK. [6] Stanford Microarray Database, Department of Biochemistry, Stanford University School of Medicine, Stanford, CA 94305-5307, USA. [7] Swiss Institute of Bioinformatics, Geneva, Switzerland. [8] GeneBio SA, Geneva, Switzerland. [9] Jax Mouse PHenome Project, The Jackson Laboratory, 600 Main Street, Bar Harbor, Maine 04609, USA. [10] Terry Fox Laboratory, BC Cancer Research Centre, Vancouver, V5Z 1L3, BC, Canada. [11] The Lance Armstrong Foundation, PO Box 161150, Austin, TX 78716-1150, USA. [12] Institute for Systems Biology, 1441 N 34th Street, Seattle, WA 98103, USA. [13] University of California Davis, Genome Center, 451 E. Health Sci Dr., Davis, CA 95616, USA. [14] NIEHS/LMIT, Research Triangle Park, NC 27709-2233, USA. [15] Division of Pathway Medicine (DPM), University of Edinburgh Medical School, The Chancellor's Building, Little France Crescent, Edinburgh, EH16 4SB, UK. [16] School of Computing Science, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK. [17] Department of Computer Science, Aberystwyth University, SY23 3DB, UK. [18] Indigo BioSystems, Inc., 111 Congressional Blvd., Suite 160, Carmel, IN 46032, USA. [19] Division of Molecular and Cellular Biosciences, National Science Foundation, 4201 Wilson Boulevard, Arlington, VA 22230, USA. [20] Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Trakehner Strasse 7-9, D-60487 Frankfurt am Main, Germany. [21] Office of Technology and Industrial Relations, Office of the Director, National Cancer Institute, Bldg 31A, Rm 10A52, Bethesda, MD 20892, USA. [22] The BIATECH Institute, Suite 115, 19310 North Creek Parkway, Bothell, WA 98011, USA. [23] Seattle Children's Hospital and Regional Medical Center, 4800 Sand Point Way NE, Seattle, WA 98105, USA. [24a] Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium. [24b] Molecular Genetics, UGent, B-9052 Ghent, Belgium. [25] Department of Plant Biology, University of Georgia, Athens, GA 30602-7271, USA. [26] Department of Molecular and Cellular Biology, Life Sciences Addition, University of California, Berkeley, CA 94729-3200, USA. [27] Bioinformatics Group, MRC Mammalian Genetics Unit, Harwell, Oxfordshire OX11 0RD, UK. [28] Department of Pathology, University of Texas Southwestern Medical Center, Dallas, Texas 75390 USA. [29] RIKEN Genomic Sciences Center, 3-1-1 Koyadai, Tsukuba, Ibaraki, Japan. [30] ESRC Centre for Economic and Social Aspects of Genomics (Cesagen), Lancaster University, IAS County South, Lancaster, LA1 4YD, UK. [31] Division of Molecular Genome Analysis, German Cancer Research Center, Im Neuenheimer Feld 580, 69120 Heidelberg, Germany. [32] School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, UK. [33] Department of Biostatistics and Computational Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA. [34] Department of Animal Science, Center for Integrated Animal Genomics, Iowa State University, 2255 Kildee Hall, Ames, IA 50011-3150, USA. [35] Bristol-Myers Squibb, Route 206 & Province Line Road, Princeton, NJ 08543-4000, USA. [36] NuGO, The European Nutrigenomics Organisation. [37] Department of Philosophy and Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, NY 14260, USA. [38] AstraZeneca UK Ltd., Brixham, Devon, TQ5 8BA, UK. [39] School of Biochemistry and Immunology, Trinity College Dublin, College Green, Dublin 2, Ireland.*

## Abstract

**Throughout the biological and biomedical sciences, minimum information (MI) checklists are beginning to find favor with scientists, publishers and funders alike. Such checklists ensure that descriptions of methods, data and analyses support the unambiguous interpretation, corroboration and reuse of data. Hitherto, representatives of particular disciplines have developed MI checklists independently. Consequently, the full range of checklists can be difficult to establish without intensive searching, and tracking their evolution is non-trivial. Furthermore, overlaps in scope and arbitrary decisions on wording and substructuring inhibit their use in combination. These issues present difficulties for checklist users, especially those in systems biology who routinely combine information from several disciplines. To address the above, we present MIBBI (Minimum Information for Biological and Biomedical Investigations), a web-based communal resource designed to act as a one-stop shop for those exploring the range of extant checklist projects and to foster collaborative, integrative development of checklists.**

## Introduction

To fully understand the context, methods, data and conclusions that pertain to an experiment it is crucial to have access to a range of background information. However, the current diversity of experimental designs, analytical techniques and chemometric/bioinformatic approaches can greatly complicate the discovery, evaluation and review of experimental data; and the rate of production of that experimental data only serves to compound the problem. Community opinion increasingly favors that a regularized set of the available metadata ('data about the data') pertaining to an experiment[1,2] be associated with the statement of its results (*i.e.,* what is sometimes called primary data, together with conclusions), making explicit both the biological and methodological context. Many journals now require that authors reporting microarray-based transcriptomics experiments make available the metadata described by the Minimum Information about a Microarray Experiment (MIAME) checklist[3] as a prerequisite for publication[4,5,6,7]. Minimum information (MI) checklists such as MIAME promote transparency in experimental reporting, enhance accessibility to data and support effective quality assessment, thereby increasing the general value of a body of work (and by extension the competitiveness of the originators).

Collaborative MI checklist development projects for a wide range of biologically- and technologically-delineated subject areas are ongoing. A special issue of the journal OMICS[8] included invited pieces from eight communities supporting MI checklist development projects, including the Microarray and Gene Expression Data (MGED) Society[9], from which sprang the well-established MIAME checklist mentioned above. However, until recently there were no established mechanisms through which such projects might coordinate their development. Here we explore some of the issues arising from the development of checklists in relative isolation, discuss the potential benefits accruing to greater coordination and describe the mechanisms we have put in place to facilitate such coordination. In summary, we present the MIBBI (Minimum Information for Biological and Biomedical Investigations) project, which maintains a web-based freely-accessible resource for checklist projects providing straightforward access to extant checklists (and to complementary data formats, controlled vocabularies, tools and databases), thereby enhancing both transparency and accessibility, as discussed above. MIBBI enables more efficient checklist development, both by increasing connectivity between MI checklist development projects, and by disseminating best practice both in relation to process (*e.g.,* open mechanisms to receive and respond to public comment) and presentation (*e.g.,* use of shared language, documentation style and structure, production of user-friendly summaries).

MIBBI is managed by representatives of the various participant communities and is fully open to comment from any interested party. Our goal is to facilitate the development of an integrated checklist resource site for bioscientists, clinicians, bioinformaticians and others. An example of a potential consumer of MIBBI products is the US National Cancer Institute, which recently launched the Clinical Proteomic Technologies for Cancer initiative (http://proteomics.cancer.gov/) to build the foundation of technologies, data, reagents, reference materials, and analysis systems needed to systematically advance understanding of protein biology in cancer, thereby accelerating discovery research and the development of clinical applications. This large scale project will require detailed annotation to successfully share, compare and analyze experimental and clinical data.

## On the need to harmonize minimum information checklists

The current proliferation of documents specifying the minimum information to provide when reporting particular kinds of experimental data has in large part been driven by the advent of a range of so-called 'omics' (and allied) technologies, many of which operate in a high-throughput mode, thereby generating large volumes of data. These documents have been developed independently for the most part, and as a result feature numerous arbitrary differences in both wording and structure. This greatly complicates the integration of data sets that comply with different MI checklists. Increasing appreciation of the potential value accruing to 'secondary use' of data is also a significant factor[10], reflecting the general increase in frequency of data-driven (as opposed to hypothesis-driven) investigations in recent years. These trends have together made the need for coordination and harmonization between groups developing data format and reporting standards a critical issue[11]. (*N.B.* Throughout this document, the words 'standard' and 'standardization' are used to refer only to the regularization of data capture, representation, annotation or reporting, as opposed to best practices for experimental procedures, often referred to as Standard Operating Procedures or SOPs.)

While it is clear that checklists should be developed through close consultation with their sponsoring practitioner communities, such checklists should also, we believe, be designed to anticipate 'cross-domain' integrative activities. It is unhelpful to confine checklists for the use of particular technologies to a limited set of biologically-delineated communities, or to conceive of any such community as being restricted to a particular set of technologies. Consider mass spectrometry, which is employed in the study of proteins, metabolites and even to sequence genes; or consider toxicology, which may employ any or all of the available 'omics' technologies in pursuit of the greater understanding of the mode of action of a particular compound. Clearly the vistas from any two locations can overlap significantly, so who can claim sole ownership of any part of the scientific landscape? Initiatives such as that to harmonize the description of 'sample' (the biological source material for a study)[12] or to develop (separable) community-level extensions to shared core standards such as MIAME to better describe domain-specific studies (for example, in environmental biology[13]) are clearly the order of the day. This throws into relief an important division between analytical approaches and the various subdivisions of the biosciences. Checklists that do not span that division will always achieve greater utility, because they can be reused more straightforwardly to construct new, bespoke checklists for a wider range of workflows.

Any reporting structure (a term potentially comprehending data formats, controlled vocabularies or ontologies, minimum information checklists, software tools and databases) can be defined as a protocol that is approved by a community as a specification of the (required) procedure for disseminating the results of a particular experiment for a specific purpose, such as publishing work in a journal. The formulation of an MI checklist as the first step in developing such a reporting structure has now been widely accepted, based largely on the perceived success of the MIAME checklist in driving the development of appropriate tools, controlled vocabulary, formats and databases. Ideally, any such checklist should reflect a consensus view of the essential data and metadata to be reported in a particular context. As such, these checklists have general utility. While their primary purpose is to guide researchers in reporting their experiments, checklists also constitute realistic test scenarios for software and database developers (whose products should be able to handle the specified data appropriately). This is especially true for instrument vendors, with respect to checklist-compliant data set export from their instrument management software. It is also likely that journals and funders will adopt some checklists wholesale, incorporating them into their guidance for authors/applicants. Some communities are anticipating this situation. The STRENDA initiative for protein function data is developing a system to support direct electronic submission to a public database prior to publication (http://strenda.bioinfo.nat.tu-bs.de/strenda2/). This has been shown to be the gold standard for comprehensive data acquisition in macromolecular sequencing and structural biology[14].

The management of information from experiments (both data and metadata) requires the adoption of standards that ensure transparency and interoperability and that facilitate the integration and exchange of data from different sources. Standards (whether checklists, data formats, controlled vocabularies or ontologies) that are integrable may facilitate the execution of more powerful queries against repositories of experimental data (for example, a query such as "find me all the studies that used technology *x*" relies on all analytical techniques being flagged as such (either in the data format, or the ontology), regardless of origin. This will be possible because core information will be regularized and extended information will be supplied in a well-characterized manner. This long-term vision will require significant effort and buy-in from a range of scientific communities spread across many nations, but development of some of the kinds of components required to establish such infrastructure is well underway; examples are given in Table 1: Functional Genomics Experiment (FuGE) is an object-oriented data model (with an associated XML-based syntactic format) capable of capturing a wide range of (meta)data in a consistent manner; Ontology for Biomedical Investigations (OBI) is a broad-scope ontology providing a self-compatible set of terms with which to describe a wide range of biological and medical studies; Reporting Structure for Biological Investigation (RSBI) provides a foundational *lingua franca* for standards projects (described further below); the Open Biomedical Ontologies (OBO) Foundry coordinates the activities of a set of ontologies (including OBI) to ensure orthogonality and promote stylistic consistency; INFOBIOMED seeks to integrate tools and resources from the disparate worlds of medical informatics and bioinformatics. For all the projects listed in Table 1, and other similar projects, there is a valuable role (as discussed above) for MI checklists as key 'use cases' in that they represent the distilled opinion of a particular community as to the information that should normally be captured to effectively describe an experiment. They therefore provide a realistic scenario with which to test any resource's suitability for use by a community.

## A resource for minimum information checklists: MIBBI

The lack of cross-domain coordinating structures in the biological and biomedical standards communities requires that individuals participate in several initiatives, both to publicize existing products and to coordinate new development. The activities of standardization groups often go unpublished and may not be accessible at all, practically speaking, hindering attempts at bridge building. The establishment of a common resource for minimum information checklists, coordinated by a group of community representatives from ongoing standardization activities, will help to forge a sense of a unified mission within the standardization community, and among experimentalists and clinicians generating various kinds of data. It will assist in recruiting participants to ongoing activities and it will help to maintain transparency of process by providing access to project-related information (*e.g.,* status, key players and plans). It will also ease the establishment of new initiatives by providing answers to questions such as: "How do we get started?" And importantly, "How do we make sure we don't reinvent the wheel?" Such an effort will improve communication, knowledge transfer and integration between checklist development projects hailing from different scientific communities, and further, between different *kinds* of data standards projects (*i.e.,* data formats, controlled vocabularies, ontologies, tools and databases), ultimately resulting in simplified access to a broad range of richly-annotated data for the end user. Thus, we have established the MIBBI project; a web-based communal resource designed to act as a 'one-stop shop' for those exploring the range of extant checklist projects and to foster collaborative, integrative development of checklists (see Box 1). In common with many other public standardization projects, the MIBBI website is hosted by SourceForge (http://sourceforge.net/), who provide a rich set of facilities for no charge; including web space, discussion lists and fora, and a version-controlled file repository.

MIBBI has two key parts: Firstly, the 'Portal', which exists simply to raise awareness of, and afford more straightforward access to a wide range of checklists by providing researchers, journal editors, reviewers, funders and the wider community of checklist developers with a quick and simple way to discover (whether there is) a checklist addressing a particular area, and to establish the scope and progress of the underlying project. The Portal provides summary information for each of the MIBBI-affiliated projects; specifically, the primary contact(s) and web site (where available), an overview of the project's scope and developmental status, and links to publications and other documents (including, where possible, a link to the most recent version of that project's checklist). Box 2 offers brief textual descriptions of the twenty projects currently registered with MIBBI; Table 2 provides a tabular representation of the concepts (akin to 'natural parts') that comprise each project's scope, along with their checklist's developmental status and, where applicable, an indication that a checklist is composed of separate modules.

By signing up to the MIBBI Portal and thereby attracting more intensive peer oversight, communities will come under pressure to maintain their checklist in light of scientific advances, to provide open access to their processes and to respond to comments. We hope that one of the primary benefits of the Portal will be to raise awareness in the biological and medical communities of the importance of standardization, thereby increasing willingness amongst researchers to become involved in guiding and shaping the evolution of these activities. We hope it will help push the community to strive for compliance in their own publication and data dissemination practices by facilitating access to relevant information about these efforts. We also see this as an excellent artifact with which to promote collaboration within and between communities: The principle we endorse is that if a broadly relevant effort already exists (for example, describing the use of a particular technology), individuals with an interest should seek to join that effort rather than compete with it. However, it is absolutely crucial that MIBBI should never preclude revisions or innovations; the hoped-for kudos and enhanced coordination accruing to membership should not translate to a possible dominion.

The second key part of MIBBI is the 'Foundry'; communities can, if motivated, sign up to the foundry to jointly examine ways to refactor the checklists over which they have control and then to develop a suite of self-consistent, clearly bounded, orthogonal, integrable checklist modules. These modules will then be made available to the community; a tool (MICheckout), which will assist users in compiling the correct list of modules and downloading them in a form that they can use, is in the early stages of development. It is important to state that registering a project with MIBBI implies no commitment by a project to participate in the Foundry activity. It is also important to recognize that attempts to integrate checklists through the foundry should be managed through a community-driven mechanism that relies primarily on openness and transparency to encourage (voluntary) uptake. The products of the foundry will fulfill our stated aim of supporting cross-domain activities such as those driven by systems biology, or the development of personalized medicine (theranostics). The Foundry is modeled on the Open Biomedical Ontologies (OBO) Foundry[15] (http://obofoundry.org/), a newly established initiative in the field of ontology development. The goal of the OBO Foundry is to develop a set of 'gold standard reference ontologies', which can be used in combination because they are based on common principles and, importantly, because procedures have been established to ensure resolution of the conflicts which may arise where ontologies overlap. Communities working together through MIBBI will similarly produce orthogonal (*i.e.,* non-overlapping) MI modules.

High-level abstractions of the components of experimental workflows offer a useful framework to support the integration of checklists. An example of a group attempting to produce such abstractions is the MGED Society's RSBI working group[16], which interacts with a number of other initiatives[17,18,19] in working towards an integrated view of functional genomics investigations. In their characterization, an *Investigation* is a self-contained unit of scientific enquiry, with a holistic hypothesis or objective and a design that is defined by the relationships between one or more *Studies* and *Assays*. A *Study* represents the part of an experiment containing information about the biological material, and an *Assay* is the part employing particular technologies that produce data. The RSBI's proposed framework of well defined high-level abstractions (such as the three just described) was developed because the above concepts are duplicated, but differently named, across different checklists, confounding the uniform description of the diverse events that may occur within a *Study* (*sensu* RSBI). Additionally, the current checklists are almost uniformly designed around one technology (or type of *Assay*); for example, microarrays in MIAME. According to the scheme proposed by the RSBI, an *Assay* could involve any of a series of distinct analytical processes such as mass spectrometric analysis, the use of a microarray, histopathology, the gathering of a set of biometrics or *in situ* hybridization.

When considering this long-term aim of fostering the harmonization of minimum information checklists to provide a non-redundant set of such documents, we can look to the proteomics community, which provides a good example of a group of practitioners unified by an overarching concern (the study of proteins) but divided into several distinct groups (delineated in that instance by their focusing on particular technologies). The Proteomics Standards Initiative (PSI) has moved forward by first defining general guidelines for the development of their (modular, integrable) checklists, the Minimum Information About a Proteomics Experiment (MIAPE) document[10]; this has prefaced the development of a group of intercompatible, non-overlapping checklists for the various relevant technologies. The Metabolomics Standards Initiative (MSI) has created a similar series of working groups, but covering a far broader range of topics; sample description, analytical techniques and statistical analysis are all addressed from the perspective of that community (http://msi-workgroups.sourceforge.net/).

All the projects listed in Table 2 represent significant investments of 'person hours'. For example, the genome-wide RNAi global initiative is developing its 'Minimum Information About an RNAi Experiment' (MIARE) checklist through an iterative and evaluative strategy of collaborative experimentation between a range of stakeholders. Given these levels of investment, it is clear that the community that generated the checklists would be in the best position to make adjustments. Plainly; all Foundry activities must be driven by the member communities (acting through their representatives). In preparation for the Foundry activity we will therefore simply establish discussion forums in the first instance, facilitating communication between communities to encourage discussion of the overlaps between two or more checklists and possible ways of working. Exploratory studies will then be performed, initially based on coarse comparison tables that highlight areas addressed by one or more projects (as is presented in Table 2), then by using 'wiki' software to draft new jointly-developed modules for those shared areas. Throughout this gradually intensifying activity, we will hold regular face-to-face meetings, to act both as development workshops and as a means of establishing good working relationships between project representatives.

There are extended benefits accruing to this federated approach to checklist development, such as; generation of consensus guidelines for the development of new checklists; promotion of the re-use of components of existing reporting structures (both syntax and semantics); highlighting 'gaps in the market' (*i.e.,* areas not addressed by any existing project); encouraging new checklist development projects to be initiated; or exploration of general issues, such as the appropriate level of detail for an average publication versus that for a notional repository. The utility of tiered checklists requesting additional detail in particular contexts is likely to be high. For example, on the 'depth' of description of the origin of a sample being studied, contrast the minimal needs of genomics with the extensive requirements of metabolomics for information about potentially confounding factors related to the history of the source material or organism.

## Foundational analysis of MIBBI-registered projects

To better understand the scope and depth of the various MIBBI-registered MI checklists, a comparative analysis was performed. Table 2 presents a projection of the checklists onto a coarse-grained list of *ad hoc* concepts constructed specifically for the purpose (see Materials and Methods). It will be clear to the reader that some of these concepts are almost universal, such as the *organism* under study, while others may relate to one group alone. It is also clear that, as discussed above, the depth of description required in relation to particular concepts varies widely across projects suggesting a 'tiered' approach (*i.e.,* some of the checklist modules generated by the MIBBI Foundry should, in some cases, require a different depth of description contingent on the particular experimental context). Row and column totals (summing presence/absence only) are provided in Table 2, roughly approximating the breadth of scope of individual projects (column totals) and the level of interest in particular concepts (row totals). These totals have been used to rank-order projects and concepts. Figure 1 (A) lists the twenty most common *ad hoc* concepts, rank-ordered by the number of projects whose scope they fall within; concepts such as the general description of an organism, a literature reference, and research personnel figure highly. Figure 1 (B) rank-orders all twenty registered projects by the number of *ad hoc* concepts they comprise; note though that these concepts are extremely diverse with respect to their content (contrast the description of a literature citation with the description of the design of an entire study) and as such this table should not be taken to reflect the significance of any one project.

To support greater understand of the relatedness of the different projects, and of the various *ad hoc* concepts, two pairwise comparisons have been conducted using the data presented in Table 2; *i.e.,* concepts 'shared' between pairs of projects, and pairs of concepts co-occurring within projects (counting presence/absence only). Figure 2 illustrates the interrelatedness of the twenty MIBBI-registered projects both as a tree and as an interaction graph. These two representations make clear that there is a subset of closely-related (*i.e.,* heavily-overlapping) projects; these are, broadly speaking, the 'technologically-delineated' projects such as MIAME and MIAPE. It is also clear that there are a large number of projects that are 'related' (according to the tree, if considered in isolation) only by their low degree of relatedness to any other project (as the interaction graph makes clear). These analyses make two things plain; firstly, that there are standout priority areas for the MIBBI Foundry (*e.g.,* the uniform description of an organism); and secondly, that there are many 'niche' areas where little or no collaborative activity is required (*e.g.,* the process of mouse phenotyping) — a simple endorsement by MIBBI of the products of a particular project being sufficient (as things stand).

Figure 3 presents an unrooted tree expressing the relatedness of individual concepts; again, branch lengths have been adjusted for clarity. This analysis is based on the various projects' scopes, rather than any sense of the similarity of the concepts themselves, nevertheless it produces some sensible-looking groupings. All the highly-ranked ('high-priority') concepts from Figure 1 cluster towards the bottom of the figure; it is not unexpected that this should be the case – they cluster together because the majority of the projects share an interest in many of them, so they are often found to co-occur in individual projects' scopes. Such an analysis can help in deciding how the *ad hoc* concept-based analysis presented herein should be developed into a bauplan for the various checklist modules that will ultimately be developed by participants in the MIBBI Foundry's activities (*i.e.,* can some concepts be combined, should others be further subdivided, and so on).

## Managing MIBBI

To be of use, standards must gain widespread acceptance both across the user community and among institutional stakeholders (funders, publishers and regulatory bodies). Managing the process of consensus-building and adoption takes time, resources and expertise; this has led to the formation of a range of standardization projects, each focused on a particular (if not completely discrete) domain. The committee that coordinates the MIBBI project comprises community representatives from many such standardization projects. An important task for this Coordination Committee is to ensure that MIBBI maintains a high level of visibility. General promotion of MIBBI on the web and in appropriate print media will ensure that level of visibility is achieved, ensuring that affiliated projects are publicized effectively, and that unaffiliated projects are made aware of MIBBI's existence. Funders and publishers represent an important special constituency for MIBBI to address; both could ultimately be seen as consumers of MI checklists, and of course funding is vital for the maintenance of project resources over the long term and for underwriting meetings between project participants to facilitate rapid progress.

The ontology community offers some useful models for this type of activity. The OBO Foundry (briefly described above and listed in Table 1) provides a precedent for integrative, non-redundant development. Additionally, OBI (an ontology development project, also listed in Table 1) provides an organizational model for a project that necessarily involves collaboration between diverse communities. The MIBBI project's charter, which describes our principles, structures and regulations in detail, is available from our website (http://mibbi.sourceforge.net/).

## Conclusions

By providing easy access to all checklist development projects and their products through MIBBI we will facilitate the discovery of checklists appropriate to the needs of practitioners from diverse parts of biological and biomedical science (the 'one-stop shop' principle). More widespread use of minimum information checklists will promote greater transparency in experimental reporting (more detail supports greater understanding), enhance accessibility to data and support more effective quality assessment, thereby increasing the general value of a body of work (and by extension the competitiveness of its originators).

MIBBI will increase connectivity between participants in MI checklist development projects and more widely. The resultant evolution of an interdisciplinary community of checklist developers will bring into focus the collective expertise residing in that group. It will accelerate the establishment of mutually-beneficial networks of expertise, and advance (through the MIBBI Foundry, building on the foundational analysis presented here) our jointly-held long-term vision of a fully-integrated, broad-coverage suite of minimum information checklists, in step with the general movement in the biological and medical sciences towards integrated multifaceted investigations of the puzzles that remain to be addressed in the post-genomic era.

## Materials and Methods

For the foundational analysis, a base data set (Table 2) was created by analyzing the content of the registrant projects' checklists and deriving the list of sixty-four concepts presented. These concepts were created for the purpose of this analysis and are not taken from any other source, although the meanings of *study* and *assay* where they appear are as set by the RSBI. The concepts have been designed to capture the content of a checklist in an intuitive but compact manner, which means that some concepts represent a large body of methods and technologies (*e.g., nucleic acid sequencing*). However, where a component of such a broad concept was found to have an analog in another project's checklist, that component was factored out to form a new standalone concept (*e.g., detection/tagging/staining*, is a concept common to workflows involving microarraying, gel electrophoresis and mass spectrometry) the better to highlight the commonalities between projects. Note also that some concepts are just 'naturally' narrow (such as *citations* et alia). The sixty-four *ad hoc* concepts thus derived have been used throughout the analyses presented here. In some cases, concepts in Table 2 are indented; this is to indicate that they represent a further specialization of the last less-indented concept above (*e.g.*, a *human* is an *animal*, which is a *generic organism*). However, the specialization of more general concept ***does not*** imply that those concepts' content overlaps as might be the case in an ontology (*i.e., human* cannot be taken to imply '*animal* plus additional information'), and having a specific requirement (*e.g., human*) does not imply that there is also generic guidance (*i.e.*, for any organism). The concepts have been represented thus simply to guide the eye while demonstrating that a project may address a concept in a generic, or specific manner, or may actually provide both kinds of requirement (six projects do this, to varying degrees).

Pairwise 'similarities' between projects were calculated by summing the 'total occurrences' (*i.e.*, the row totals from Table 2) of all the concepts addressed by both projects, then scaling that figure by the sum of all such totals (*i.e.*, the sum of all row totals). This has the effect of weighting a concept's contribution to the total pairwise similarity score between any two projects by its 'importance' (*i.e.*, its total occurrence in all checklists). A similar procedure was followed to gauge the pairwise similarities between concepts. Pairwise distances between projects and between concepts (used for the two trees) were calculated as one minus their similarity score (calculated as above). To produce a clearer final tree (both rooted and unrooted versions), all distances were rescaled (to the range 1,0), and as stated in the text, the trees themselves have been heavily manipulated for presentational purposes; therefore while the gross structure of the trees is correct, branch lengths do not reflect distance.

## Acknowledgements

Work on MIFlowCyt is supported by the NIH/NIBIB (EB005034-01). RRB is a Michael Smith Foundation for Health Research Scholar. DF, SAS and CT acknowledge funding from the NERC Environmental Bioinformatics Centre (NEBC) and BBSRC (BB/E025080/1) to DF and SAS for supporting CT and MIBBI activities.

## Disclaimer

Opinions, findings and conclusions or recommendations expressed in this paper are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

*Box One. The MIBBI Project: Parts and Purposes*

**The Portal** (*http://mibbi.sourceforge.net/portal.shtml*)

Lists all registered minimum information checklist development projects and provides:

— Description and links (to the various projects' homepages, publications, *etc*.).

— Raises visibility of projects both to each other and to the wider community.

— Simple registration procedure (spreadsheet-based form).

**The Foundry** (*http://mibbi.sourceforge.net/foundry.shtml*)

Web-based infrastructure to support the refactoring and extension of existing checklists:

— Wiki-based development environment

— Document repository with versioning system

— Supplemented by discussion fora, email, teleconferences and face-to-face meetings

**MIBBI Search** (*http://mibbi.sourceforge.net/search.shtml*)

Google™ Custom Search Engine covering; the entire MIBBI site; registered projects' sites (if available) and other germane web-based resources (*e.g.*, policy-forming bodies' statements).

**Relevant Resources** (*http://mibbi.sourceforge.net/resources.shtml*)

Miscellaneous links to; relevant software, data formats and ontologies; publications and policy statements; and other projects of note such as EQUATOR (*http://www.equator-network.org/*).

**About Us** (*http://mibbi.sourceforge.net/about.shtml*)

Provides some background information about the MIBBI Project and links to policy documents, discussion lists (email-based) and discussion fora (online only).

**News** (*http://mibbi.sourceforge.net/news.shtml*)

Date-ordered announcements of all significant project-related developments (new registrations, product developments, website updates, *etc*.).

***Box Two. Checklist development projects registered with MIBBI***

**CIMR** (*http://msi-workgroups.sourceforge.net/*)

The Metabolomics Standards Initiative's Core Information for Metabolomics Reporting (CIMR) comprises modules for particular aspects of metabolomics workflows; various biological disciplines (*e.g.*, microbiology, mammalian biology, plant biology); analytical techniques such as chromatography and NMR; and the use of various statistical tools.

**IMIAGE** (*http://www.immport.org/*)

Immport's Minimum Information About a Genotyping Experiment (IMIAGE) addresses genotyping based on single nucleotide and microsatellite repeat polymorphisms, and genetic association and linkage analysis in humans, with special reference to immunology.

**MIACA** (*http://miaca.sourceforge.net/*)

The Minimum Information About a Cellular Assay (MIACA) checklist relates to the perturbation of cells with various classes of molecule, such as small interfering RNA (siRNA) or small chemical compounds. They also provide guidance on environmental stressors such as temperature shift or starvation, and combinations thereof.

**MIAME** (*http://www.mged.org/Workgroups/MIAME/miame.html*)

The Microarray and Gene Expression Data Society's well-established Minimum Information About a Microarray Experiment (MIAME) checklist relates to the use of (micro)arrays to assay nucleotide abundance (most commonly, messenger RNA) and analysis of the data generated.

**MIAME/Nutr** (*http://www.mged.org/Workgroups/rsbi/rsbi.html*)
**MIAME/Tox** (*http://www.mged.org/Workgroups/rsbi/rsbi.html*)
**MIAME/Env** (*http://nebc.nox.ac.uk/miame/miame_env.html*)
**MIAME/Plant** (*http://www.ebi.ac.uk/at-miamexpress*)

The MIAME checklist has recently been extended to capture parameters appropriate to nutrigenomics (/Nutr), toxicogenomics (/Tox), environmental biology (/Env) and phytology (/Plant); in each case adding relevant information about the background to the experiment.

**MIAPA** (*http://mibbi.sourceforge.net/projects/MIAPA/*)

The Minimum Information About a Phylogenetic Analysis (MIAPA) checklist relates to the use of software to align biological sequences, and the subsequent use of algorithms to construct phylogenies/cladograms and to draw inferences from them.

**MIAPE** (*http://www.psidev.info/*)

The Minimum Information About a Proteomics Experiment (MIAPE) checklist comprises modules for reporting the use of various analytical techniques such as mass spectrometry, gel electrophoresis or liquid chromatography. Modules addressing the description of the biological material under study have not yet been produced.

**MIARE** (*http://www.miare.org/*)

The Minimum Information About an RNA interference Experiment (MIARE) checklist identifies minimal reporting parameters for aspects of high-throughput RNA interference (*e.g.*, siRNA, shRNA) screens, such as the use of cellular assays (*cf.* MIACA checklist discussed above) and flow cytometry (*cf.* MIFlowCyt,  discussed next).

**MIFlowCyt** (*http://flowcyt.sourceforge.net/*)

The Minimum Information for a Flow Cytometry Experiment (MIFlowCyt) checklist addresses the use of flow cytometry, especially to measure the phenotype and function of cells; information is required about the sample analyzed, the probe, fluorochrome and instrument used, and the analysis of the data collected.

**MIGS/MIMS** (*http://gensc.sourceforge.net/*)

The Minimum Information About a Genome Sequence (MIGS) specification is an extension of the metadata traditionally captured by the International Nucleotide Sequence Databases (DDBJ/EMBL/GenBank). It captures information relating to nucleic acids sequence, location, and sequencing method. The description of habitat is also being extended via the tightly integrated Minimum Information About a Metagenomic Sequence/Sample (MIMS) checklist.

**MIMIx** (*http://www.psidev.info/*)

The Minimum Information required for reporting a Molecular Interaction experiment (MIMIx) checklist addresses the reporting of a molecular interaction experiment; including the identity of molecules that participate in an interaction (with accession number), the methods by which both the interaction and the identity of the participants were established, and the role of these molecules in the context of the experiment (as distinct from their biological role).

**MIMPP** (*http://www.interphenome.org/*)

The Minimum Information for Mouse Phenotyping Procedures relate to the diverse protocols deployed to characterize the phenotype of a mouse. The checklist addresses both behavioral and physiological traits.

**MINI** (*http:// www.carmen.org.uk/*)

The Minimum Information about a Neuroscience Investigation (MINI) checklist identifies the minimum information required to report the use of electrophysiology in a neuroscience study.

**MIQAS** (*http://miqas.sourceforge.net/*)

The Minimum Information for QTLs and Association Studies (MIQAS) checklist relates to the mapping of quantitative trait loci (QTLs) and their association with genetic markers.

**MIRIAM** (*http://biomodels.net/index.php?s=MIRIAM*)

The Minimum Information Requested In the Annotation of biochemical Models (MIRIAM) checklist offers formal requirements for describing theoretical models of biochemical systems.

**MISFISHIE** (*http://mged.sourceforge.net/misfishie/*)

The Minimum Information Specification For *In Situ* Hybridization and Immunohistochemistry Experiments (MISFISHIE) checklist addresses those performing visual interpretation-based tissue gene expression localization experiments such as those using *in situ* hybridization or immunohistochemistry.

**STRENDA** (*http://www.strenda.org/*)

The Standards for Reporting Enzymology Data (STRENDA) initiative, along with participants in the biannual ESCEC (Experimental Standard Conditions of Enzyme Characterizations) symposia, maintain a series of checklists addressing the description of enzyme activity data and the experiments in which they were collected. These checklists are subject to permanent review by the community involved.

| ***Project* [*URL*]** | ***Products*** |
|---|---|
| FuGE [*fuge.sourceforge.net*] | Object model (and markup language) to support the description of diverse experiments and development of new formats |
| OBI [*obi.sourceforge.net*] | Ontology providing descriptors for a wide range of experimental and clinical research workflows, equipment and data types |
| RSBI [*www.mged.org/ Workgroups/rsbi*] | Cross-domain analysis of project structures; development of well-characterized generic concepts to facilitate integrative activities |
| OBO Foundry [*obofoundry.org*] | Collaborative management of orthogonal (*i.e.*, non-overlapping) ontologies covering diverse domains |
| INFOBIOMED [*www.infobiomed.org*] | Collaborative integration of biological and medical informatics resources; development of novel applications and technologies |

**Table 1.** Example resources of various kinds focused on supporting cross-domain activities.

**Table 2.** Matrix describing the composition of each of the twenty checklists visible through the MIBBI Portal. Concepts (row headings) were derived as described in the Materials and Methods section. Color coding of cells and bullets indicate granularity of coverage and developmental status respectively. *N.B.* Some bullets have been placed within the matrix itself to provide a finer-grained view of developmental status. Row and column totals (counting presence/absence only) are provided in the rightmost column and lowermost row. Analyses of these data are provided in Figures 1–3.

| ***Table Key*** | | | | |
|---|---|---|---|---|
| ***Granularity*** | Coarse | Medium | Fine | |
| ***Maturity*** | ● Planned | ● Drafting | ● Release | ● Published |
| [†] Denotes that a specification is provided as a suite of related documents | | | | |

| CONCEPT | SPECIALISATION | CIMR [†] 🔴 | IMIAGE 🔵 | MIACA 🟢 | MIAME 🔴 | MIAME/Env 🟢 | MIAME/Nutr 🟢 | MIAME/Plant 🔴 | MIAME/Tox 🟢 | MIAPA ● | MIAPE [†] 🔴 | MIARE 🟢 | MIFlowCyt 🟢 | MIGS/MIMS 🔴 | MIMIx 🔴 | MIMPP 🔵 | MINI 🔵 | MIQAS 🔵 | MIRIAM 🔴 | MISFISHIE 🔴 | STRENDA 🟢 | Row totals |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| study inputs | study design | 🟨 | 🟨 | 🟨 | 🟩 | 🟨 | 🟨 | 🟨 | 🟩 | 🟨 | 🟨● | 🟨 | 🟩 |  |  | 🟩 | 🟩 |  |  | 🟨 | 🟨 | 16 |
|  | generic organism | 🟥 | 🟨 | 🟩 | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 | 🟩 | 🟨● | 🟨 | 🟨 | 🟨 | 🟩 | 🟨 | 🟩 | 🟨 |  | 🟨 | 🟩 | 19 |
|  | cells/microbes | 🟥 |  | 🟥 |  |  | 🟥 |  | 🟨 |  |  |  |  |  |  |  | 🟩 |  |  |  |  | 5 |
|  | plant | 🟥 |  |  |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  |  |  |  |  | 2 |
|  | animal | 🟥 |  |  |  |  | 🟥 |  | 🟨 |  |  |  |  |  |  |  | 🟩 |  |  |  |  | 4 |
|  | human | 🟨 | 🟨 |  |  |  | 🟥 |  | 🟨 |  |  |  |  |  |  |  | 🟩 |  |  |  |  | 5 |
|  | environment/habitat | 🟥 | 🟨 |  |  | 🟥 |  | 🟥 |  |  |  |  | 🟨 | 🟥 |  |  |  |  |  |  |  | 6 |
|  | *in silico* model |  |  |  |  |  |  |  |  | 🟨 |  |  |  |  |  |  |  |  | 🟨 |  |  | 2 |
|  | population | 🟥 | 🟥 |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  | 3 |
| study procedures | animal husbandry | 🟥 |  |  |  |  | 🟨 |  | 🟨 |  |  |  |  |  |  | 🟩 |  |  |  |  |  | 4 |
|  | cell/microbe culture | 🟥 |  | 🟥 |  |  | 🟨 |  | 🟨 |  |  |  |  |  |  |  |  |  |  |  |  | 4 |
|  | medical intervention |  | 🟨 |  |  |  | 🟥 |  | 🟨 |  |  |  |  |  |  |  |  |  |  |  |  | 3 |
|  | plant cultivation | 🟥 |  |  |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  |  |  |  |  | 2 |
|  | preconditioning/pretreatment | 🟥 |  |  |  | 🟥 |  | 🟥 |  |  | 🟨● | 🟨 |  |  |  |  | 🟩 |  |  |  | 🟩 | 7 |
| assay inputs | organism part | 🟥 | 🟨 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |  | 🟨● | 🟩 | 🟩 |  | 🟩 |  | 🟨 |  |  | 🟥 | 🟩 | 15 |
|  | organism state |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟨 |  |  |  |  | 1 |
|  | organism trait |  | 🟥 |  |  |  |  |  |  |  |  |  |  | 🟨 |  |  |  | 🟥 | 🟨 |  |  | 4 |
|  | synthetic analyte |  |  |  |  | 🟩 |  | 🟩 |  |  | 🟨● |  |  |  |  |  |  |  |  |  |  | 3 |
|  | enzyme |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 | 1 |
|  | silencing RNA reagent |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  | 1 |
|  | sample collection | 🟥 | 🟨 |  |  |  |  |  |  |  |  |  |  | 🟩 |  |  |  |  |  |  |  | 3 |
|  | sample processing | 🟥 |  |  | 🟨 | 🟥 | 🟨 | 🟨 | 🟨 |  |  | 🟩 | 🟩 | 🟨 |  |  |  |  |  | 🟨 | 🟩 | 11 |
|  | sample storage | 🟨 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟨 |  | 2 |
|  | sample transport | 🟩 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |
| assay procedures | detection/tagging/staining | 🟩 |  | 🟨 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |  | 🟨 | 🟨 | 🟥 |  |  |  |  |  |  | 🟥 | 🟩 | 12 |
|  | generic analysis |  |  | 🟩 |  |  |  |  |  |  | 🟩 | 🟨 |  |  |  |  |  |  |  |  |  | 3 |
|  | array-based assay |  | 🟨 | 🟨 | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |  | 🟥● | 🟨 |  |  |  |  |  |  |  | 🟥 |  | 10 |
|  | capillary electrophoresis | 🟥● | 🟩 |  |  |  |  |  |  |  | 🟥🔵 |  |  |  |  |  |  |  |  |  |  | 3 |
|  | cell phenotyping |  |  | 🟨 |  |  |  |  |  |  |  | 🟩● |  |  |  |  |  |  |  |  |  | 2 |
|  | clinical test/examination |  | 🟨 |  |  |  | 🟥 |  | 🟨 |  |  |  |  |  |  |  |  |  |  |  |  | 3 |
|  | column chromatography | 🟥 |  |  |  |  | 🟩 |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  |  | 3 |
|  | electrochemical detection | 🟥● |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |
|  | electrophysiology mensuration |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  |  | 1 |
|  | enzyme activity assay |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 | 1 |
|  | flow cytometry |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  |  |  |  |  |  | 1 |
|  | gel electrophoresis |  |  |  |  |  | 🟩 |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  |  | 2 |
|  | image acquisition | 🟩 |  | 🟨 |  | 🟩 | 🟩 |  |  |  | 🟥 | 🟩● |  |  |  |  | 🟩 |  |  | 🟥 |  | 8 |
|  | infrared spectroscopy | 🟥● |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |
|  | mass spectrometry | 🟥 |  |  |  |  | 🟩 |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  |  | 3 |
|  | molecular interaction detection |  |  |  |  |  |  |  |  |  | 🟨 |  | 🟥 |  | 🟩 |  |  |  |  |  |  | 3 |
|  | mouse phenotyping |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  |  |  | 1 |
|  | nmr spectroscopy | 🟥 |  |  |  |  | 🟩● |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 2 |
|  | nucleic acid sequencing |  | 🟥 |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  |  |  |  |  | 2 |
|  | toxicology assay |  |  |  |  |  | 🟥 |  | 🟨 |  |  |  |  |  |  |  |  |  |  |  |  | 2 |
| data | generic data description | 🟥 | 🟨 |  |  |  |  |  |  |  | 🟨● | 🟨 |  |  |  |  | 🟩 |  |  |  | 🟥 | 6 |
|  | confidence indicator | 🟥 |  |  |  |  |  |  |  |  |  |  |  |  | 🟩 |  |  | 🟨 |  |  | 🟥 | 4 |
|  | enzyme activity data |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 | 1 |
|  | generic data analysis | 🟨 | 🟨 | 🟨 | 🟨 |  |  |  |  |  | 🟨● | 🟨 |  |  |  |  |  |  |  |  |  | 5 |
|  | flow cytometry informatics |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  |  |  |  |  |  | 1 |
|  | gel electrophoresis informatics |  |  |  |  |  | 🟩 |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  |  | 2 |
|  | genetic linkage analysis |  | 🟥 |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  | 2 |
|  | image characterisation |  |  | 🟨 |  | 🟨 | 🟨 |  |  |  | 🟨 |  |  |  |  |  | 🟩 |  |  | 🟥 |  | 6 |
|  | mass spectrometry informatics | 🟥 |  |  |  |  | 🟩 |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  |  | 3 |
|  | microarray informatics |  | 🟨 |  | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |  |  |  |  |  |  |  |  |  |  |  |  | 6 |
|  | nmr spectroscopy informatics | 🟥 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |
|  | nucleic acid sequence assembly |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  |  |  |  |  | 1 |
|  | phylogenetic analysis |  |  |  |  |  |  |  |  | 🟥 |  |  |  |  |  |  |  |  |  |  |  | 1 |
|  | population genetic analysis |  | 🟥 |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  | 2 |
|  | QTL description & map |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 🟥 |  |  |  | 1 |
| data availability | raw data |  |  |  | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |  | 🟥 | 🟩 |  |  |  |  | 🟩 |  |  |  | 🟩 | 9 |
|  | processed data |  | 🟨 |  | 🟥 | 🟥 | 🟥 | 🟥 | 🟥 |  | 🟥 | 🟩 |  |  |  |  |  |  |  |  |  | 8 |
| administrative | citations *et alia* | 🟩 | 🟨 |  | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 |  | 🟨 | 🟩 | 🟨 |  | 🟨 | 🟨 | 🟨 | 🟨 | 🟥 | 🟨 | 🟨 | 17 |
|  | supporting data |  | 🟨 |  |  |  |  |  |  | 🟥 |  | 🟩 |  |  |  | 🟩 |  | 🟨 | 🟨 |  | 🟩 | 7 |
|  | personnel | 🟩 | 🟨 |  | 🟨 | 🟨 | 🟨 | 🟨 | 🟨 |  | 🟨 | 🟩 | 🟨 |  | 🟩 |  | 🟨 | 🟨 | 🟨 | 🟩 |  | 15 |
|  | Column totals | 32 | 22 | 12 | 11 | 16 | 27 | 16 | 19 | 5 | 23 | 18 | 11 | 7 | 6 | 6 | 15 | 10 | 5 | 11 | 14 |  |

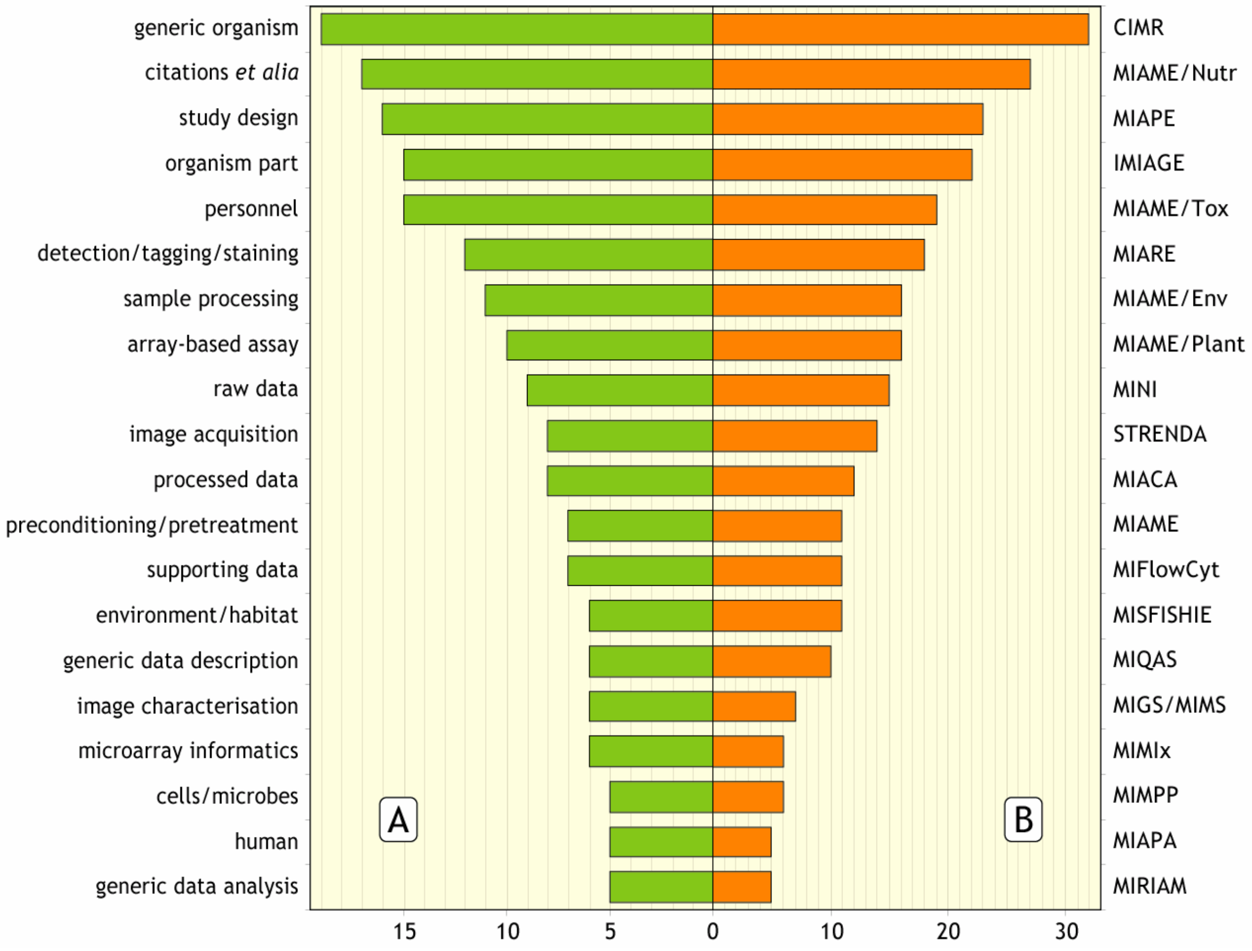

**Figure 1.**

(A) Concepts addressed by five or more projects, rank ordered. The highest-ranked concepts are naturally the highest priority areas for the MIBBI Foundry. (B) Projects rank-ordered according to the number of *ad hoc* concepts they comprise. Clearly, such a 'concept count' is no measure of worth, but it does illustrate the differences in breadth of scope to be found between projects.

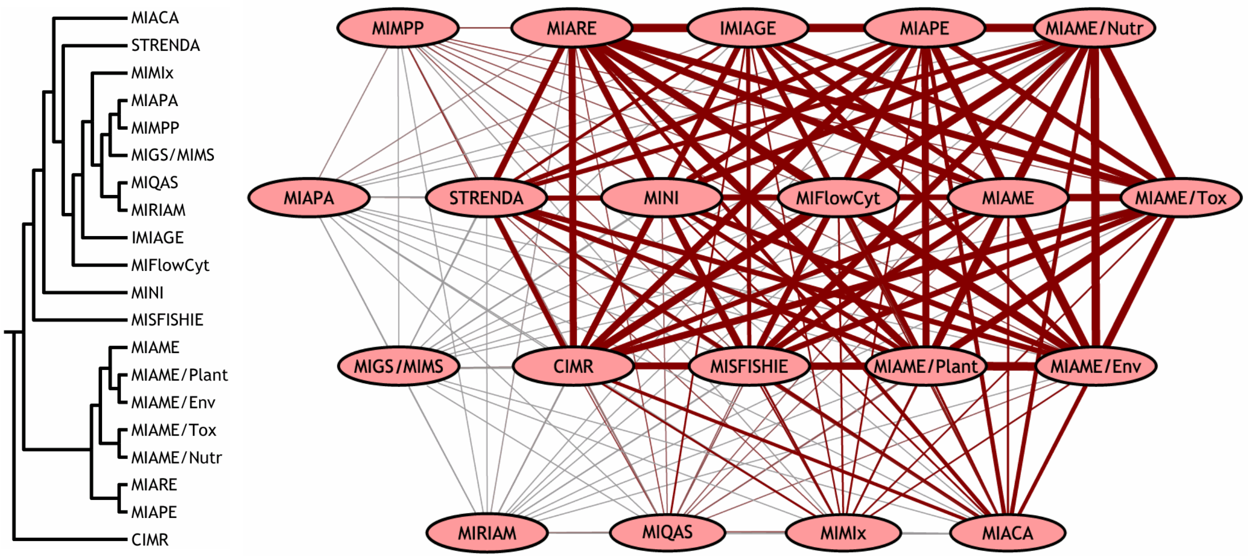

**Figure 2.**

(Left) A rooted tree based on the pairwise distances between projects (see 'Materials and Methods'). There are two large clusters visible (along with the outgroup, CIMR); however, while the lower cluster is apprehensible because most of the projects deal with overlapping sets of omics-related technologies, the nature of the upper cluster is less obvious.

(Right) Pairwise similarity of projects (see 'Materials and Methods') represented as an interaction graph (constructed using Cytoscape – http://www.cytoscape.org/); both the color saturation and thickness of the lines joining projects indicate the similarity between them. This graph makes clear that the checklists in the upper cluster in the tree (discussed above) share only their unrelatedness to other checklists.

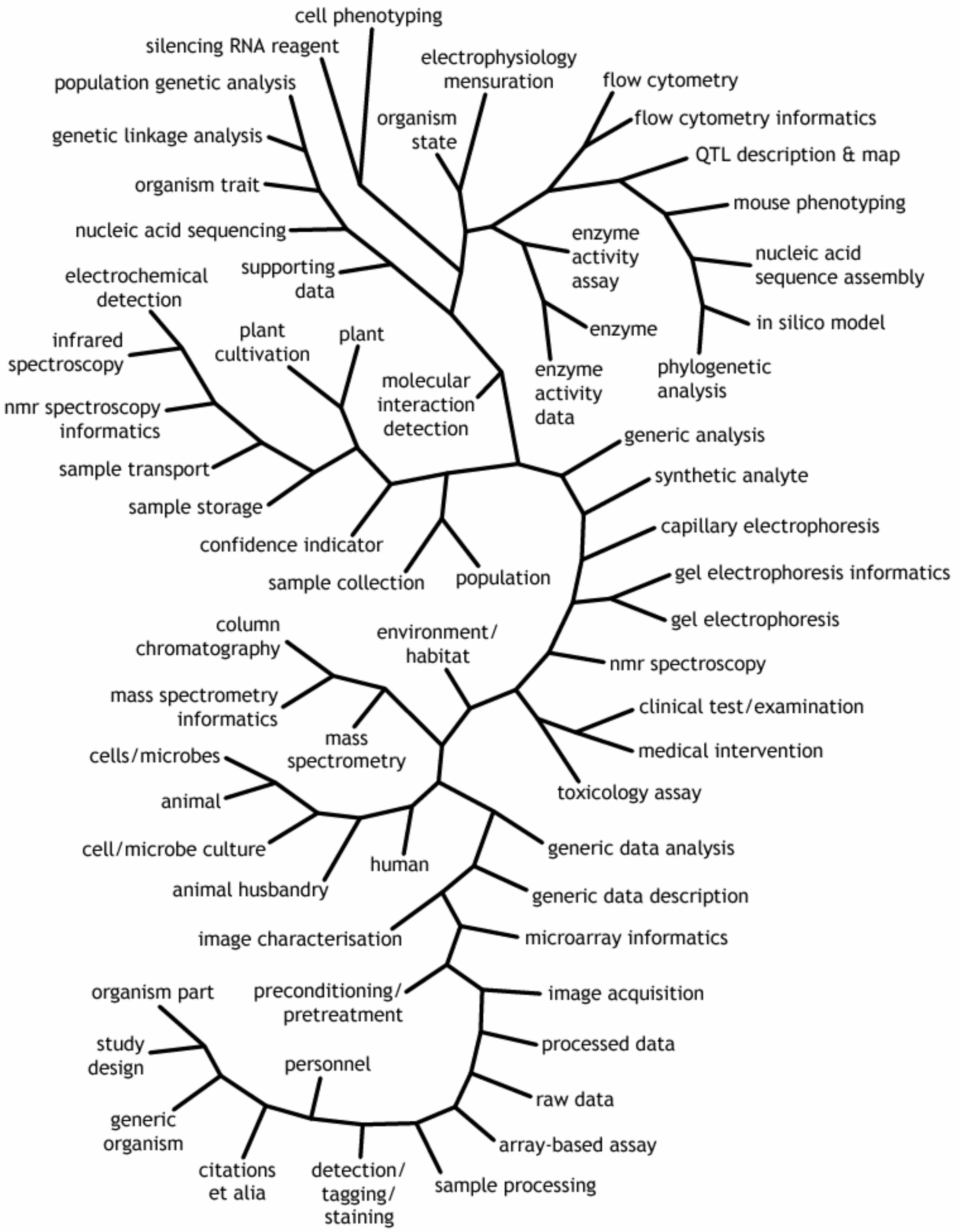

**Figure 3.**

An unrooted tree built on the pairwise distance between concepts (see ‘Materials and Methods’). Several interesting concept clusters emerge from this analysis; the vast majority of the highly-ranked concepts from Figure 1 cluster together towards the bottom of the figure; other appropriate clusters can be found throughout the tree. It is worth stressing that the algorithm used to generate this tree did not cluster using scores based of the similarity of these concepts, but simply on the frequency with which they co-occur in different projects’ scopes (*i.e.*, there is an orthogonal ‘psychosociological’ aspect to this data set).

## References

[1] Quackenbush, J. Standardizing the standards. *Mol. Syst. Biol.* **2**, 2006.0010 (2006).

[2] Anonymous. *Nat. Methods* **3**(6), 415 (2006).

[3] Brazma, A., *et al*. Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).

[4] Anonymous. Microarray standards at last. *Nature* **419**, 323 (2002).

[5] Ball, C.A. et al. A guide to microarray experiments—an open letter to the scientific journals. *Lancet* **360**, 1019 (2002).

[6] Ball, C.A., *et al*. Standards for microarray data. *Science* **298**(5593), 539 (2002).

[7] Ball, C.A., *et al*. The underlying principles of scientific publication. *Bioinformatics* **18**(11), 1409 (2002).

[8] Field, D., Sansone, S.-A. A Special issue on data standards. *OMICS* **10**(2), 84-93 (2006).

[9] Ball, C.A., Brazma, A. MGED Standards: Work in Progress. *OMICS* **10**(2), 138-144 (2006).

[10] Taylor, C.F., *et al*. The Minimum Information About a Proteomics Experiment (MIAPE). *Nature Biotechnol.* **25**(8):887-93 (2007).

[11] Sansone, S.-A., Morrison, N., Rocca-Serra, P., Fostel, J. Standardization initiatives in the (eco)toxicogenomics domain: a review. *Comp. Funct. Genom.* **5**, 633-641. (2004).

[12] Morrison, N., *et al*. Concept of sample in Omics Technology. *OMICS* **10**(2), 127-137 (2006).

[13] Morrison, N., *et al*. Standard Annotation of Environmental Omics Data: Application to the Transcriptomics Domain. *OMICS* **10**(2), 172-178 (2006).

[14] Apweiler, R., *et al*. (2005). The importance of uniformity in reporting protein-function data. *Trends Biochem. Sci.*, **30**(1), 11-12 (2005).

[15] Smith, B., *et al*. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnol.* **25**(11), 1251-1255 (2007).

[16] Sansone, S.-A., *et al.* A Strategy Capitalizing on Synergies: The Reporting Structure for Biological Investigation (RSBI) Working Group. *OMICS* **10**(2), 164-171 (2006).

[17] Taylor, C.F., *et al.* The Work of the Human Proteome Organisation's Proteomics Standards Initiative (HUPO PSI). *OMICS* **10**(2), 145-151 (2006).

[18] Fiehn, O., *et al*. Establishing Reporting Standards for Metabolomic and Metabonomic Studies: A Call for Participation. *OMICS* **10**(2), 158-163 (2006).

[19] Field, D. *et al.* Toward a richer description of our complete collection of genomes and metagenomes: the minimum information about a genome sequence (MIGS) specification. *Nature Biotechnol.* [*in press*] (2008).