Digital Preservation – Financial Sustainability of Biological Data and Material Resources

Christina Chandras, Thomas Weaver, Michael Zouberakis, John M. Hancock, Paul N. Schofield, and Vassilis Aidinis

***Abstract*—Following the technological advances that have enabled genome-wide analysis in most model organisms over the last ten years, there has been unprecedented growth in genome and post-genomic sciences with concomitant generation of an exponentially increasing volume of data. As a result numerous resources have been created to store and archive the data and biological materials produced, which are of substantial value to the whole community. Sustained access facilitating re-use of this primary data is vital, not only for validation, but for re-analysis, testing of new hypotheses and developing new technologies/platforms. A common challenge for most data resources and biological repositories today is finding financial support for maintenance and development so as to best serve the scientific community. In this manuscript we examine the problems that currently confront the data and resource infrastructure underlying the biomedical sciences. We discuss the financial sustainability issues and potential business models that could be adopted by biological resources and consider long term data preservation issues within the context of mouse functional genomics efforts in Europe.**

I. Introduction

In our attempt to better understand the biology of human disease we are generating increasingly diverse and specialized data sets, many of which are extremely large and complex, with the result that where primary data is put in the public domain it is scattered through an increasing number of databases and bio-repositories. These databases contain genomic (including sequencing, expression and microarray), proteomic (structure and function) and metabolomic data as well as information about function, structure, localization and clinical effects of mutations. Increased attention has recently been paid to mouse mutants that serve as models for human disease facilitating understanding of disease processes and supporting the development of novel therapeutic strategies. Biological databases have consequently become an important tool in assisting scientists to understand and explain biological molecules and processes, in addition to their interactions.

Since biological knowledge is distributed worldwide and therefore among many differently specialized databases, it is difficult and frequently impossible to ensure preservation and accessibility of information as well as data quality. Currently, much of the collected data are stored in a way that does not always guarantee future data retrieval by other researchers [1]. Assured growth, persistence and accessibility of databases are therefore imperative to encourage and support data deposition. Additionally, standardization of data representation and transfer is required for enabling the integration of existing and new databases. At present, biological databases cross-reference other databases with accession numbers or IDs as one way of linking their related knowledge together. Much current European effort is being expended in developing modes of data integration and database interoperability, either as a "one-stop-shop", through federation, or more recently in the development of "smart" clients which integrate data from multiple sources or run tailor-made workflows.

A major problem for most databases is securing financial support for the bioinformaticians and curators who create and maintain them [2]. Even popular databases commonly lack secure funding and frequently face loss of their original support after a few years in development. Hence, long-term sustainability of databases requires adequate and reliable sources of funding. In this paper we will give an overview of the current financial support situation, potential business models that could be adopted by databases for their long-term financial support, and the attempts that have been made so far.

CASIMIR (http://www.casimir.org.uk/), a Coordination Action funded by the European Commission, focuses on the dissemination and integration of databases containing biological collections, relevant to the mouse as a model organism for human disease. The overall aim of the project is to establish a framework for interoperable databases with concomitant added value to the scientific community, which should additionally become self-sustained in terms of data deposition, usage, development and financial support. In the context of this Coordination Action, CASIMIR aims to make recommendations to the European Commission on the problems that databases containing biological collections

This work was supported in part by the Sixth Framework Programme CASIMIR under Grant FP6-037811 (European Union) and in part by the Sixth Framework Programme MUGEN under Grant LSHG-CT-2005-005203 (European Union).

C. Chandras is with the B.S.R.C. Alexander Fleming, Vari, Greece (e-mail: chandras@fleming.gr).

T. Weaver is with Geneservice Limited, Cambridge, UK (email: tweaver@geneservice.co.uk)

M. Zouberakis is with the B.S.R.C. Alexander Fleming, Vari, Greece (e-mail: zouberakis@fleming.gr).

J. M. Hancock is with the Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire OX11 0RD, U.K. (e-mail: j.hancock@har.mrc.ac.uk).

P. N. Schofield is with the Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY U.K. (e-mail: PS@mole.bio.cam.ac.uk).

V. Aidinis is with the B.S.R.C. Alexander Fleming, Vari, Greece (corresponding author phone: +30-210-9654382; fax: +30-210-9654210 e-mail: v.aidinis@fleming.gr.

encounter and on potential business models that could be adopted by biological resources for their financial sustainability and data preservation.

II. Data and Biological Resources

Publication of experimental results and sharing of the related research materials have long been key elements of the life sciences. Indeed scientific progress depends on the ability of researchers to access and exploit data and materials reported in publications so that they can subsequently build on these findings. Publications also serve as a means of receiving intellectual credit and recognition which subsequently enhance a researcher's career prospects and potential for research support. It is however no longer adequate to share data through traditional modes of publication, and particularly with high throughput ("-omics") technologies sharing of datasets requires submission to public databases as has long been the case with nucleic acid and protein sequence data. This presents new challenges in extending the traditional publication model to the New Biology.

The traditional *quid pro quo* arrangement, where authors receive credit and acknowledgements in exchange for disclosure of their scientific findings, has been re-evaluated by a US National Academies committee. The responsibility of authors to share data and materials referenced in their publications, the role of journals to impose requirements for data and material sharing and whether a common set of requirements for sharing should exist has been closely examined and the concept of the "uniform principle for sharing integral data and materials expeditiously" (UPSIDE) [3], [4] has been established.

Biological Resource Centers (BRCs) are centralized repositories that specialize in storing and distributing data and information. Both repository and service functions contribute to the needs of national and international consortia, as well as individual laboratories and research institutes in support of academic research programs. A central role for the BRCs is to champion the principles set out by UPSIDE and embrace the open access policy, quality of material, data integration and sustainability. It is crucial that the scientific community, public funding bodies and governments acknowledge these issues as being of primary importance.

In accordance with the aforementioned responsibilities of authors, journals and BRCs came the recently published guidelines by the Organization for Economic Co-operation and Development (OECD) asserting that in order to comply with the data sharing imperative, adequate and reliable sources of funding are required to facilitate the sharing infrastructure and, as part of that, the long term stability of BRCs [5]. The notion that BRCs need to improve their management systems in order to control the quality of biological materials and related data is supported. If, for financial reasons, BRCs are unable to perform their tasks under conditions that meet the requirements of scientific research and the demands of industry, scientists will either see valuable information lost or being transferred into a strictly commercial environment with at least two consequences: (a) blockade of access to this information and/or high costs and (b) loss of data and potential for technology transfer for the foreseeable future. In either case the effect on both the scientific and broader community will be detrimental.

On the other hand, as the generation of certain data types (e.g., imaging, microarray, phenotypic etc) can include costly processes, requiring expensive consumables as well as specialized equipment and personnel for their generation, it can potentially be difficult to fulfil scientific duty and make resources available, unless these can be exchanged for public funding and recognition by peers.

Under ideal conditions, BRCs should conform to the web-data concept. The concept of creating a "hyper-mart", where everything is in one "data warehouse", has been more recently invoked. In principle this is a very practical and helpful notion as users will be able to find all the information in one place without interoperability issues. However, managing so much information in one place is very difficult and more importantly there are technical disadvantages limiting the applicability of this model. For this reason, decentralization is recommended given the existing technological infrastructures available, and as a result a more recent proposal suggests the formation of a "one-stop-shop" rather than a traditional "data warehouse", which will bring together data from multiple resources in a single web-interface, enabling collective data querying across different data sets and linking to biological material.

However, in order to achieve such a multi-resource portal, there are several hindrances to overcome in conjunction with some requirements that need to be met. All contributing BRCs should firstly be validated for their data/information quality according to accepted standards, and should be continuously updated, both at the level of material/data as well as incorporation of novel biological resources. Self-evidently, to achieve this constantly developing infrastructure, support from both biologists/curators and bioinformaticians is essential, which is a hindrance to the maintenance of a number of these databases. Furthermore, BRCs should all embrace open access policies upon publication of the related material, or the existence of simple material transfer agreements (MTA) and standards so that portals can integrate and become easily interoperable. Such restrictions should be eliminated as much as possible, especially for academic applications, to promote data sharing.

III. Problems Encountered

As previously mentioned, one of the biggest concerns that BRCs encounter is their financial sustainability beyond their creation and after the original funding has ended [2]. Typically, BRCs may obtain an initial funding for a project relatively easily where a community need is clear. As a result many biological resources and databases have been designed in various research institutes and are commonly created without meeting validated quality standards. Furthermore, they are developed with varying formats and

quality, and occasionally exhibit limited international access. Consequently integration of these BRCs into the international data network is often not possible, an action that would enable completeness in data acquisition for the scientific community, a link between databases and biological material and results in addition to a simultaneous avoidance of BRC redundancy. For prolonged data archiving and curation, long term financial support is required which is frequently a stumbling block for BRCs today. Lack of secure funding may frequently result in database or biological resource decommissioning as well as loss of valuable and irreplaceable data. An obvious question that arises is to examine who would provide the required financial support for the archiving of these valuable data and the distribution of biological material, as well as the customer service/user support. How does one support a useful BRC to ensure appropriate data/information archival and curation?

IV. Models Examined

Whereas BRCs are expected to embrace an open access policy and be accessible to the broad scientific community, pharmaceutical and biotechnology companies do not share data generously. Some companies like Incyte (http://www.incyte.com/), a provider of integrated platforms of genomic technologies, apply a subscription fee, or pay-per-view policy. Other companies, such as Exelixis (http://www.exelixis.com/), employ marketing and public relations policies to help them sell their products or demonstrate their product and technology utility. Finally, some corporations like Wyeth (http://www.wyeth.com/), engage in research collaborations for research they are unable to perform in-house, an effort which indirectly promotes knowledge and information.

Many BRCs currently charge fees to those who want to obtain biological materials and gain access to associated databases. Varying fee structures can be applied for access depending on the nature of the biological material, the status and constraints of the institution holding the resources and its relationship with the public and private sectors, national policies and relevant international frameworks.

There are four major models that have been examined and are currently in use by different BRCs.

A. Cost Recovery

The "cost recovery" model entails establishment of an annual, equitable rate structure for the standing expenses towards the major utilities of the respective BRC. A project that adopted this policy was the Human Genome Project (HGM; http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml) which was completed after a 13-year-long effort. The model of financial support that was employed included a list of world-wide funders comprising government programs, nonprofit organizations, pharmaceutical firms, and dedicated genomics firms. The Australian Genome Research Facility (AGRF; http://www.agrf.org.au/) is a project established in 1997, and is nowadays maintained in a similar fashion, through the National Collaborative Research Infrastructure Strategy (NCRIS) of the Australian government for the provision of specialist services that support bio-molecular research.

There are a few projects which utilize the "cost recovery" model in conjunction with other methods of financial support. These include: the Drosophila Genomics Resource Center (https://dgrc.cgb.indiana.edu/), the Bloomington Drosophila Stock Center at Indiana University (http://fly.bio.indiana.edu/) and the John Innes Centre Genome Laboratory (JGL; http://jicgenomelab.co.uk/). Finally, the Jackson Laboratory (JAX; http://jaxmice.jax.org) receives operating revenue from several sources, including the public sector, in the form of federal grants, the private sector in the form of private foundation grants and philanthropic contributions, and resource revenue in the form of cost and fees collected for JAX Mice and Services.

B. Fee-for-service

'Fee-for-service' is a standard business model where services are unbundled and paid for separately. BRCs that adopt this model acquire a pay per view interface, which is of course in opposition to the principles reported by UPSIDE since it entails a restricted access policy. Some of the BRCs that have partially adopted this business model are: Incyte and JAX Mice as previously mentioned.

C. Institutional Funding

Another common model for the financial sustainability of a resource is through allocated funds obtained from a particular institution towards the respective BRC. An example of a database that operates through this model is the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) which receives funds from both the National Library of Medicine (NLM; http://www.nlm.nih.gov/) and the National Institutes of Health (NIH; http://www.nih.gov/).

D. Top Slicing publicly funded grants

The NERC Environmental Bioinformatics Centre (NEBC; http://nebc.nox.ac.uk/) operates on yet a different model. This is based on the 'top-slicing' of NERC project grants within a given programme, which in this case is the NERC Post-Genomics and Proteomics Programme (http://www.nerc.ac.uk/research/programmes/proteomics). This project is funded by the Natural Environment Research Council (NERC; http://www.nerc.ac.uk/) where it is envisaged that approximately 10% of the allocated funds will be set aside to ensure the effective development and implementation of a data management plan to ensure the long-term accessibility of the related project data.

V. The Role of Industry versus The Role of Government

Both the role of industry and that of government have proven to comprise the majority of funding bodies supporting BRCs. Some BRCs have attained use of a dual

support system, where the research councils provide grants for specific projects and programs, whereas governments funding councils, usually supported by different ministries/departments, provide block grant funding to support the research infrastructure and enable the institutions to undertake ground-breaking research of their choosing. Such funding also provides the capacity to undertake research commissioned by the private sector, government departments, charities, the European Union and other international bodies. The European Bioinformatics Institute (EBI; http://www.ebi.ac.uk/) is a good example of this dual funding support practice, as it is funded by the governments of EMBL's member states, as well as other major funders such as the European Commission, Wellcome Trust, US National Institutes of Health, UK research councils and some industry partners.

Furthermore, there are specific 'projects' (e.g., biobanks; i.e. collection of cell, tissue, blood, DNA samples) that may have a two-fold character, as collections of both samples and data. These may be operated under the auspices of either the public sector institutions (i.e. university departments) or of individual or private bodies (e.g., pharmaceutical companies). Irrespective of the responsible institution, they may be funded from public or private resources. One could expect that some funding from these projects may be dedicated towards resource integration and dispersal.

Development of the business model aforementioned, as a supplementary activity towards cost recovery, is not as effective for underpinning the infrastructure, as it does not cover Full Economic Costing (FEC; real costs of running an infrastructure, including all costs above and beyond consumables and direct staff costs. These involve rent for space, overheads, staff salary/benefits, staff training and any business development support) or opportunity costs. It is not easy for BRCs to adopt such a model as they would become liable for the service charges imposed towards these costs rather than the direct service provided.

With regard to industry investments, although extremely essential since they provide invaluable support towards further developing the resource assets (e.g., validated assays, new applications), implementing the notion of advertising may give the impression of being advantageous; however the benefits will not be enough to cover the infrastructure and the business development overhead will outweigh any benefit from original attempt.

VI. A Model with potential: academic – Commercial Partnership Focus on Core Competencies

Another model that has been examined and appears to have great potential in being successful towards the prolonged financial sustainability of BRCs is an 'academic-commercial partnership'. Academic laboratories, mostly sustained by institutional funding, or grants, develop new applications and tools as well as analysis systems, whereas concurrently they support the identification of communal needs and define quality standards all of which prove to be beneficial to the research community. Commercial organizations on the other hand, which are financially supported by their own commercial activities, function in a collaborative way between research and licensing (Pharmaceutical and Biotechnology companies) and operate as service providers, offering standard technologies and quality systems, sales and marketing distributors.

In the context of CASIMIR, and in the course of examining the potential financial models that resource centres could adopt for their maintenance, the MUGEN Mouse database (MMdb; http://www.mugen-noe.org/database/), a virtual mutant mouse repository, created in the context of the MUGEN Network of Excellence (http://www.mugen-noe.org/) to provide on-line information on murine models for immunological disease [6], serves as a use-case example. For demonstration purposes, MMdb, taking advantage of its simplicity and useful size, currently provides direct trial links, under the gene information, to Invitrogen (http://www.invitrogen.com/) and Geneservice (http://www.geneservice.co.uk/) through the gene IDs (Fig. 1). The user may therefore be directly transferred to the respective product page, where all the gene-related products (e.g., antibodies, RNAi, primers, cDNA clones, proteins, assay kits etc) are presented. Ensuing the overall discussions regarding financial sustainability of databases, and following a successful connection, MMdb has approached Invitrogen as well as other potential companies, asking them to link their individual products with the respective mouse model and also examine the possibility that such big vendor corporations would be interested in linking with MMdb and explore their willingness towards marketing/advertisement service charges which could help maintain the databases. Indeed Invitrogen responded very positively towards this effort, and has pledged to undertake a survey with regard to the company's perspectives and willingness to financially support this effort. Unfortunately, the overall response was not as expected, since only one out of the six companies approached responded to the request, demonstrating some enthusiasm and feedback in this attempt. The suggested approach, although in principle appearing to have great potential, in practice it is somewhat harder to achieve, as companies are not that willing to sponsor academic institutions. This may of course be a matter of time and should big vendor corporations be appropriately primed this arrangement may indeed prove to be beneficial towards prolonged sustainability.

VII. The Role of Consortia

The European Commission in support of the fifth and sixth Framework Programmes has over the last seven years sponsored a number of projects generating biological experimental data, including sequences, and material resources such as biological collections. Some of these consortia (e.g., Eumorphia, Eucomm, Eumodic, Eurexpress, Emma, Mugen etc) also serve as liaisons towards the European Commission, giving advice with respect to specific areas of interest and their respective needs for further development and also suggesting potential future directions that the European Commission should pursue.

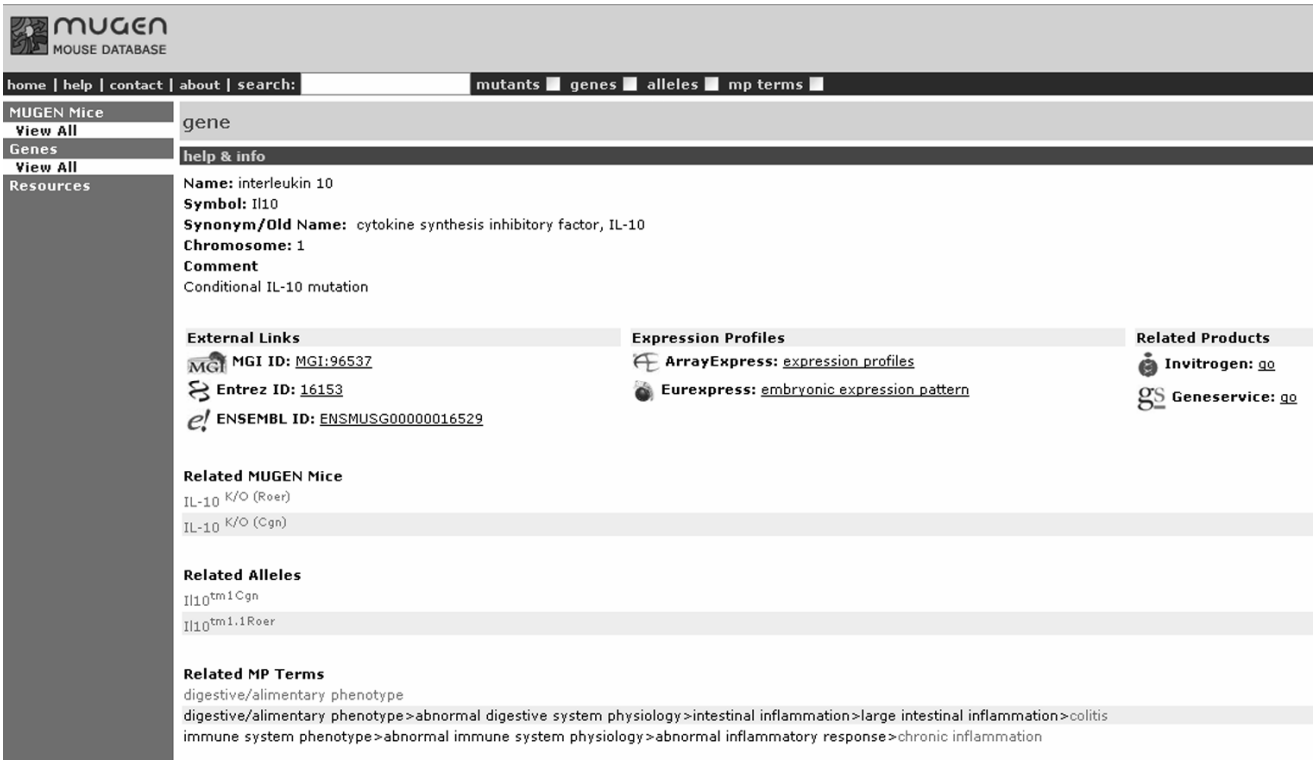

Fig. 1. Sample screen shot of MMdb "IL-10" gene with the direct trial links, under the gene information, to Invitrogen and Geneservice through the gene ID.

Furthermore, the European Commission has also supported some co-ordination actions (e.g., PRIME, CASIMIR) especially to organize and bring together the individual European efforts as well as survey the scientific community needs. These consortia also play an intermediary role between the scientific community and the European Commission, making recommendations to the latter with respect to the needs that the scientific community has, thus aiming to improve scientific development. This interactive relationship allows networks to lobby both national and international funders, for example to improve application practices and for funders to approach and consult with the network with regard to issues and priorities.

VIII. Recommendations for the Mouse Functional Genomics Community

Having reviewed extensively the substantial amount of information provided by BRCs and the importance of making the data freely available to the research community, it is clear that it is imperative to promote data preservation and dissemination, for secure storage and easy retrieval of information. Moreover, BRCs should not exist as data warehouses, but rather a cluster of activities supporting the community of academic and commercial researchers all aiming, through a unified effort, towards providing information for the progression of research. CASIMIR is indeed already taking action in the direction of promoting database integration and interoperability, and should authors conform to their responsibilities and share data as recommended by UPSIDE [3], [4] this would obviously greatly promote research advances.

Furthermore, following the close examination of setbacks that most of these BRCs today encounter and existing business models that they could potentially adopt in order to reinforce database sustainability, the conclusion that can be drawn is that long-term sustainability of databases requires adequate and reliable sources of funding so that data is preserved and disseminated properly.

With regard to the business models examined in this manuscript as potential patterns to be adopted by BRCs for their financial sustainability, the "full cost recovery" model which has already been tested by some resources has proved to not be viable. The "fee-for-service" model is already practiced, at least in part, by some BRCs, however, this opposes to UPSIDE recommendations, according to which data should be shared. The most promising models examined in this manuscript are the "Institutional Funding" and "Top Slicing of Public funding" both of which seem to provide a secure environment for the BRCs to develop and implement a secure data management plan and potentially ensure the long-term accessibility of the related project data. Indeed agencies around the world such as the National Institute of Health (NIH), are now turning their attention to working out how best to assist the growth of standardized and accessible databases. This should involve, at the least, development of policies for evaluating proposals on databases and associated analytic tools, for their sustained funding, and for ensuring that the data deposited remain accessible long after the project originators have moved on.

The aforementioned model of academic-commercial partnership may appear to have potential should vendor corporations become involved in this collaborative effort. In all cases, funders should be aware of the need to support viable career paths for the software engineers and bioinformaticians who create the knowledge environments and curate the data in them. In order to obtain value for money, it will be vital for funding agencies to carefully select the databases they choose to support and then to support them for the long term. They must encourage the sustained availability of these data and build incentives for the development of cross-querying capability.

IX. Conclusion

The last decade has seen a rapid growth in the genome sciences, through modern advances in biological sciences, molecular biology and genetics, which have enabled genome-wide analysis in most model organisms, and the generation of high-throughput of data. To facilitate the secure storage and easy retrieval of this substantial amount of information, numerous data and biological material resources have been created which are of significant value and should be openly accessible to all scientists for the purposes of result validation, testing new hypotheses and developing new technologies/platforms. An inevitable consequence that has arisen from this data and biological material resource boom is the significant challenge in the access and sustainability of these databases. Preservation of these centralized repositories that specialize in storing and distributing data is therefore imperative. In this manuscript, CASIMIR reviewed the potential business models that biological resources could adopt for their financial sustainability and prolonged data storage and aims to appropriately make recommendations to the European Commission.

References

[1] T. Weaver, J. Maurer, and Y. Hayashizaki, “Sharing genomes: an integrated approach to funding, managing and distributing genomic clone resources,” *Nat. Rev. Genet.*, vol. 5, pp. 861-866, November 2004

[2] Editorial, “The database revolution; Funding agencies face conflicting challenges in supporting the databases essential to modern biology,” *Nature*, vol. 445, pp. 229-230, January 2007

[3] T. R. Cech, S. R. Eddy, D. Eisenberg, K. Hersey, S.H. Holtzman, G. H. Poste, N. V. Raikhel, R. H. Scheller, D.B. Singer, M. C. Waltham, and National Academics Committee on Responsibilities of Authorship in the Biological Sciences, “Sharing publication-related data and materials responsibilities of authorship in the life sciences,” *Plant Physiol.*, vol. 132, pp. 19-24, May 2003

[4] N.R. Cozzarelli “UPSIDE: Uniform principle for sharing integral data and materials expeditiously,” *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 3721-3722, March 2004

[5] Organization for Economic Co-operation and Development (14 June 2007). “OECD best practice guidelines for biological resource centers” [Online]. Available: http://www.oecd.org/document/36/0,3343,en_2649_34537_38777060_1_1_1_1,00.html

[6] V. Aidinis, C. Chandras, M. Manoloukos, A. Thanassopoulou, K. Kranidioti, M. Armaka, E. Douni, D. L. Kontoyiannis, M. Zouberakis, G. Kollias and the Mugen consortium, “MUGEN mouse matabase; animal models of human immunological diseases,” *Nucleic Acids Res.*, vol. 36, pp. D1048-54, January 2008