# Integrating information from EU-funded mouse functional genomics projects: a questionnaire-based analysis

John M. Hancock[1], Christina Chandras[2], Michalis Zouberakis[2], Vassilis Aidinis[2], Paul N. Schofield[3]

[1] Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire OX11 0RD, U.K. [2] BSRC Fleming, 34 Fleming Street, 16672, Vari, Athens, Greece [3] Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3DY U.K.

[John M. Hancock j.hancock@har.mrc.ac.uk, Christina Chandras chandras@fleming.gr Michalis Zouberakis zouberakis@fleming.gr Vassilis Aidinis aidinis@fleming.gr Paul N. Schofield ps@mole.bio.cam.ac.uk]

**Abstract.** In recent years the European Commission has funded an increasing number of functional genomics projects aimed at using the mouse as a model of human disease. Many of these projects are producing large data volumes. A recently funded programme, CASIMIR (Coordination and Sustainability of International Mouse Informatics Resources) aims to make recommendations on the most efficient way to integrate these datasets. In Summer 2007 CASIMIR carried out a questionnaire survey of relevant EC-funded projects to determine their current use of data integration technologies and standards. This report describes the results of the survey and initial conclusions deriving from it.

**Keywords:** Database integration, EU-funded databases, CASIMIR

## 1 Introduction

The need for integration of data sets is well established in the computer science, bioinformatics and high throughput biology communities but is less well-established amongst bench biologists whose primary interest is hypothesis-driven experimental science and do not have experience of propagating large data sets to the wider community with a view to integrated analysis.

Over the last few years, the European Commission has supported an increasing number of functional genomics projects focusing on the use of the laboratory mouse as a model of human disease. The mouse has numerous advantages as a disease model including mammalian physiology and anatomy, short generation time and a well-developed genetic toolkit allowing, amongst other manipulations, knocking out and

knocking in of genes, production of tissue specific knockouts, and production of point mutations [1]. Mouse projects funded by the European Commission encompass methods for mouse phenotyping (EUMORPHIA: http://www.eumorphia.org/), archiving and distributing mutant mouse lines (EMMA: http://www.emmanet.org), large scale phenotyping of mouse lines (EUMODIC: http://www.eumodic.org/), systematic generation of knockouts of a significant proportion of all mouse genes (EUCOMM: http://www.eucomm.org), mapping of gene expression domains in mouse embryos (EurExpress: http://www.eurexpress.org), development of mouse models to investigate human immunological disease (MUGEN: http://www.mugen-noe.org), a database of images of mouse pathology (PATHBASE: http://www.pathbase.net) and numerous others (see http://www.prime-eu.org/euromouseiiprojects.htm for a fuller listing of current and recent projects). The diversity of these projects is so great that the Commission has also funded several Coordination Actions both to provide an overview of the activities of the various projects and to provide means for wider dissemination of the results. These have included PRIME (Priorities For Mouse Functional Genomics Research Across Europe: http://www.prime-eu.org/), which is a priority-setting organisation, and CASIMIR (Coordination and Sustainability of International Mouse Informatics Resources: http://www.casimir.org.uk) which is aimed at recommending standards to allow data sharing and integration between the different projects.

CASIMIR is an important initiative because bioinformatics is often not given serious thought when projects of this kind are planned. As a consequence, there is a risk that data will not be stored or preserved in a form amenable to future use or integration into other data sets. This would result in a massive waste of resources. CASIMIR spans a number of areas: data representation (in particular the use of shared ontologies), non-semantic, technical issues concerning database compatibility and interoperability, data acquisition, curation and ownership, integration of biological collections and material resources into the data network, and user interactions.

As a first step towards developing its recommendations, CASIMIR carried out a survey in summer 2007 to ascertain the sorts of database activities carried out by currently-active EC-funded mouse functional genomics projects and whether they are currently making use of community standards such as ontologies and minimum information standards for reporting experimental data. In this paper we report the results of the survey and discuss their consequences in the context of integration of these large projects into the wider data network.

## 2 Methodology

The questions included in the questionnaire are shown in Table 1. The questionnaire was circulated to a panel of recipients. These included bioinformatics representatives of projects funded by Europe-wide institutions (the European Commission and European Molecular Biology Laboratory) as well as contacts in other databases, many in the USA, which act as a control group and give the results a broader perspective. The list of EC-funded projects targeted is given in Table 2.

Results were gathered using a custom web form accessible via the CASIMIR web site (http://www.casimir.org.uk).

**Table 1.** Questionnaire.

| Question No. | Question |
|---|---|
| 1 | Are you using a relational database, object database or flat files? |
| 2 | If relational, what is your chosen RDBMS (Relational Database Management System)? |
| 3 | Is your database providing external links to other on-line resources; possibly via URL/HTTP (if yes please name them)? |
| 4 | Supported/Installed Web Services (if yes please name them)? |
| 5 | Please list the sorts of data entities you store (e.g. protein sequence data, mouse strain information etc...) |
| 6 | Can you provide a brief 'explanatory' description/schema of your data/data structure? |
| 7 | Are you willing to provide a entity relationship diagram and would you be willing to provide it under an open source license? |
| 8 | Are you currently using or do you intend to use any ontologies or controlled vocabularies to describe your data? |
| 9 | Do you plan to expand your use of ontologies in future? |
| 10 | Do you use OBO ontologies? |
| 11 | Do you perceive the need for additional ontologies to serve your domain of knowledge? |
| 12 | Do you make use of Minimum Information standards (such as MIAME for microarray experiments) to describe any data? If so, which ones? If you do not make use of these standards, are you likely to do so in future? |
| 13 | Do you have any comments/thoughts on standards for data representation that need to be developed or that you might like discussed in CASIMIR? |
| 14 | What do you perceive as the main limiting factor in data representation/interoperability etc. in European bioinformatics databases? |

**Table 2.** Targeted EC-funded projects. * Indicates did not respond.

| Project | URL |
|---|---|
| AnEUploidy | http://wp5.aneuploidy.eu |
| EMAGE | http://genex.hgu.mrc.ac.uk/Emage/database |
| EMMA (European Mouse Mutant Archive) | http://www.emmanet.org |
| EUCLOCK | http://www.euclock.org |
| EUCOMM | http://www.sanger.ac.uk/htgt |
| EUMODIC | http://www.europhenome.org |
| EURExpress 2 | http://www.eurexpress.org/ |
| EVI-GENORET* | http://www.evi-genoret.org |

| | |
|---|---|
| MAIN* | http://www.main-noe.org |
| MUGEN | http://www.mugen-noe.org/database/ |
| Pathbase | http://www.pathbase.net |
| QRISP | http://lymphomics.wall.gu.se |
| | |
| Related (EMBL-funded) | |
| Ensembl | http://www.ensembl.org/ |
| ArrayExpress | http://www.ebi.ac.uk/arrayexpress |

## 3 Results

28 responses were received, of which 11 were from the 13 targeted EC-funded projects (85% response rate). In the analysis the responses from the EC-funded projects were combined with responses from databases funded by the other pan-European funding agency, the EMBL, to give a broad picture of the state of European-funded databases. These results are presented under the heading "EC-funded" in Table 3. Results from non-EC or EMBL funded projects are presented separately in the category "Other". Results are summarized in Table 3 with additional detail below and in Table 4.

**Table 3.** Results of questionnaire.

| Question | Responses | N (%) | |
|---|---|---|---|
| | | EC-funded | Other |
| 1 | Relational | 13 (81) | 16 (89) |
| | Object | 1 (6) | 0 (0) |
| | Flat File | 2 (12) | 2 (11) |
| 2 | MySQL | 10 (67) | 6 (35) |
| | Postgresql | 2 (13) | 7 (41) |
| | Oracle | 2 (13) | 3 (18) |
| | DB2 | 1 (7) | 1 (6) |
| 3 | Yes | 13 (81) | 13 (87) |
| (See also Table 4) | No | 3 (19) | 2 (13) |
| 4 | Yes | 7 (44) | 4 (25) |
| | No | 4 (25) | 8 (50) |
| | Pending | 5 (31) | 4 (25) |
| 5 | See Main Text | | |
| 6 | Yes | 15 (94) | 11 (73) |
| | No | 0 (0) | 1 (7) |
| | No response | 1 (6) | 3 (20) |
| 7 | Yes | 14 (88) | 10 (67) |
| | No | 0 (0) | 4 (27) |

| | | | |
|---|---|---|---|
| | N/A | 2 (12) | 0 (0) |
| | No response | 0 (0) | 1 (7) |
| 8 | Yes | 9 (56) | 10 (67) |
| | CV | 0 (0) | 3 (20) |
| | No | 7 (44) | 2 (13) |
| 9 | Yes | 4 (44) / 4 (57) | 8 (80) / 3 (60) |
| (L: if using ontologies/ R: if not) | Possible | 2 (22) / 2 (28) | 0 (0) / 0 (0) 2 (20) / 2 (40) |
| | No | 3 (33) / 1 (14) | |
| 10 | Yes | 6 (67) | 6 (60) |
| | No | 3 (33) | 4 (40) |
| 11 | Yes | 8 (67) | 7 (50) |
| (See also Table 5) | Perhaps | 2 (17) | 2 (14) |
| | No | 2 (17) | 5 (36) |
| 12 | Yes | 2 (14) | 5 (33) |
| | Possible/Planned | 7 (50) | 3 (20) |
| | No | 5 (36) | 7 (47) |
| 13 | See Main Text | | |
| 14 | See Main Text | | |

**Table 4.** External links identified by databases. Asterisks indicate number of mentions for each external link.

| Link | EC | Other |
|---|---|---|
| ABG | | * |
| Allen Brain Atlas | * | |
| Arrayexpress | * | |
| ChEBI | * | |
| CisRed | * | |
| Compound | * | |
| DbSNP | * | |
| Dictybase | * | |
| EMBL(/GenBank/DDBJ) | ** | *** |
| EMMA | * | |
| EMPReSS | * | |
| Ensembl | ****** | **** |
| Entrez | * | |
| Entrez Gene | *** | ** |
| Enzyme Commission | | * |
| EUCOMM | | * |
| Eurexpress II | * | |
| Flybase | * | * |
| GenePaint | | ** |
| GeneTests | | * |

| | | |
|---|---|---|
| GO | ** | * |
| HapMap | * | |
| IMGT/LIGM | | ** |
| IMSR | | * |
| Interpro | | * |
| IPI | * | |
| JAX | * | |
| Kabat Database | | * |
| KEGG Gene | * | |
| KOMP | | * |
| Medline/Pubmed | **** | ***** |
| MGI | ****** | **** |
| MGI Vocabulary | | * |
| Mouse Tumor Biology Database | | * |
| MPD | * | |
| NCBI | * | ** |
| NCBI Ig germ-line genes | | * |
| NCBI Taxonomy | * | |
| NURSA | | * |
| OMIM | * | ** |
| PubChem | * | |
| PubGene | | * |
| Rat Genome Database | | * |
| Refseq | ** | |
| RIKEN Animal Search System | | * |
| RIKEN Omic Browse | | * |
| Sanger Team 109 | | * |
| Science Direct | * | |
| SwissProt | | * |
| Tbase | * | |
| UCSC | * | |
| Unigene | * | |
| Uniprot | *** | ** |
| ZFIN | | ** |

Question 5 returned a wide variety of terms. In summary these indicate a wide spread of data types, from genomic and proteomic (DNA sequence, Protein sequence, Gene name, Gene structure, Protein feature, Gene/protein function, Transcript sequence, Gene regulation, other genome features); gene expression data (from gene expression arrays and in situ hybridization); systems biology information at the level of pathways, DNA-protein interaction and systems models; cell lines and chemical interventions applied to them; information on individual mice and mouse lines and strains, including breeding history, genetic manipulations applied to them genotype, phenotype and pathology data and information concerning the welfare regulatory regime under which they were kept; more complex data types such as images and

their metadata and full descriptions and comparisons of ontologies; and information on researchers, publications and user requests.

Responses to Question 11 ("Do you perceive the need for additional ontologies to serve your domain of knowledge?") returned the following areas for future ontology development:

- Modified RNA/DNA molecules ("An ontology for RNA and DNA molecules with various characteristics, e.g. capped, methylated mRNA, nascent mRNA, DNA with 3` overhanging end, etc.")
- Phenotype ontology
- Experimental setups/standard operating procedures
- General (cross-species) anatomy
- Gene products

Questions 13 ("Do you have any comments/thoughts on standards for data representation that need to be developed or that you might like discussed in CASIMIR?") and 14 ("What do you perceive as the main limiting factor in data representation/interoperability etc. in European bioinformatics databases?") returned free-form responses which overlapped to a significant extent. The responses identified the following main themes as areas that need to be addressed:

- Databases built without regard to interoperability
- Lack of a "clear federating architecture"
- Lack of standards for data representation
- Lack of standards for many kinds of experimental data
- Lack of use of web services
- Duplication of resources
- Determining equivalence of entries. Mapping of equivalent IDs between different databases.
- Need for a central registry, or database of databases
- Data privacy (i.e. inaccessibility)
- Time and money
- Lack of expertise in database design within biological research institutes, especially outside the U.K.

## 4 Discussion

An increasing number of large projects, generating high volumes of functional genomics data, are being established in Europe to exploit the mouse as a model of human disease [1]. It is crucial that the best use is made of these large data sets. To do this, it is essential that any large project of this kind establishes a database which can be integrated into the wider mouse data network. This survey was designed to investigate the current state of the art in European-funded projects, to identify

strengths and weaknesses, and to drive further discussions under the auspices of CASIMIR, leading to a set of recommendations on how to facilitate the data integration process. Any such process should be compatible with developments world-wide, where both in the US (through projects such as caBIG [2]), Japan (through a new initiative to integrate all RIKEN's biological databases [3]) and Australia (http://www.ncris.dest.gov.au/capabilities/integrated_biological_systems.htm) major data integration initiatives are being established.

The results of the CASIMIR questionnaire suggest that in general European projects are well-placed to respond to the challenges of integration but that some issues need to be addressed. The range of data being stored in the databases we involved in the questionnaire is wide and covers most of the areas that are important in modern biology (Question 5). Relatively few projects are relying on flat-file formats for storing data - most are using relational or object technology (Questions 1&2). In this they are consistent with practice on the non-European-funded projects that responded to the questionnaire. We asked if databases were willing to make their relational schemas publicly available (Questions 7&8). Most were willing to do so but some were not. The main argument from those databases not willing to make their schemas public was that they did not wish to do so before publishing a journal article on their database, after which most were willing to publish their schemas. We therefore conclude that most databases operate in a spirit of openness. Most databases in the survey provide external links to data in other databases, linking them into the wider data network at the level of the user of the web interface.

An increasingly important route for making data accessible to external "power" users is the implementation of web services. Less than half of the European-funded databases we involved currently had web services available (Question 4) although the proportion (44%) was higher than for the non-European Commission or EMBL-funded databases (25%). A significant proportion declared an intention to implement web services (31% for EC+EMBL-funded databases, 25% for the others) but a large group also declared no intention to do so (25% of EC+EMBL-funded projects and 50% of others). This may reflect an opinion that web services are of no obvious value to the users of a given database. This might change over time as more and more useful implementations making use of web services are demonstrated, for example the new generation of workflow clients such as Taverna [4]. One of the aims of Work Packages 5 ("Technical issues concerning database compatibility and interoperability") and 8 ("User Interactions") of CASIMIR is to implement such a demonstration with the hope that this will stimulate other databases to make web services available.

An essential element for developing the potential for applications that mine data across multiple databases is consistent nomenclature. In the biological sciences the development of domain-specific ontologies, particularly the Gene Ontology [5] has played an important role in widening the acceptance and use of consistent nomenclature in biological databases. Consistent nomenclature across databases demands use of the same core set of broadly accepted ontologies by all databases. The OBO foundry family of ontologies, which developed from the original GO concept, is intended to act as a set of consistent, broadly orthogonal ontologies for the biological

sciences [6]. We therefore asked about the use of ontologies in our database set and whether they favoured OBO foundry ontologies. A majority of databases currently use ontologies to represent their data but a significant minority do not. Some (exclusively in this sample amongst the non-EC-funded databases) use in-house controlled vocabularies (CVs) rather than ontologies. When asked if they intended to expand their use of ontologies, the majority said yes but a few again said no indicating that there is a core of resistance to the use of ontologies. This may be because they are not seen to be necessary, or because some developers find them difficult to implement. In Question 10 we asked if databases made use of OBO ontologies. A slim majority did so, but a proportion did not and either developed their own or used some nomenclatures not part of the OBO "family", such as NCBI Taxonomy. At least one respondee was unaware whether the ontologies they used were OBO ontologies. It would seem that a valuable way forward in this area would be the development of a forum involving OBO and other ontology providers that could work towards a self-consistent set of usable ontologies. Increased involvement with the user community (defined here as the database managers and programmers who might be expected to implement ontologies) may also be worthwhile.

In Question 11 we asked whether additional ontologies were needed to improve databases' data representation. Some of the areas mentioned here are already the subject of ontology development - specifically phenotype, general anatomy and gene products (although the exact meaning of the latter response is unclear). It is possible that the responses reflect dissatisfaction over lack of clarity or over-complexity in these areas. Phenotype can now be represented either by unitary ontologies (such as MP for mouse) [7] or using the EQ formalism and PATO [8]. The latter is developing rapidly and offers the opportunity to represent detailed experimental data, but examples of its use are still limited; an example of an experimental use can be found in the MUGEN database [9] (http://www.mugen-noe.org/database/). CASIMIR aims to hold one or two workshops in this area over the next year or two, which it is hoped will help to clarify some of the issues in this area. The area of anatomy ontologies is also a complex one as anatomy changes during development and mapping between homologous structures in different species is not trivial [10]. As a consequence there are competing approaches to the representation of anatomy using ontologies. Again, it is to be hoped that this area will become clearer over time.

The last area investigated by the questionnaire was the use of Minimum Information (MI) standards. MI standards define the information that needs to be collected to adequately describe specific types of high throughput, functional genomics experiment. The original example was MIAME for microarray-based gene expression experiments [11], but numerous standards are now under development by various communities, many under the auspices of the MIBBI (Minimum Information for Biological and Biomedical Investigations; mibbi.sourceforge.net/ [12]) consortium. Relatively few of the responding databases currently implemented MI standards - in nearly all cases this was MIAME although one implements MISFISHIE (Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments) [13]. It is likely that the uptake of MI standards protocols will increase as they become available for more areas. As with all such computational tools, it will be important that these are easy to use as well as powerful.

Finally we asked two open questions, the aim being to elicit opinions on the most important areas in which development was needed to further database interoperability. Many of the areas mentioned in these responses also emerge in the discussion above. However a theme that clearly emerges is the need for overarching advisory bodies that can help individual database managers and programmers design their databases optimally for data integration, recommend on standards, and so on. Some technical needs were also raised, specifically a resource providing mappings between equivalent IDs that would enable mapping of data from different databases. Another technical suggestion was the establishment of a "database of databases" that could be automatically queried to provide information on issues such as accessibility of web services or usage of ontologies in a specific database. Work package 7 of CASIMIR has established a draft database of databases (available at http://www.fleming.gr/mrb/) and further work is being carried out to develop a more mature resource. Establishment of these kinds of resource on a large scale would require some financial investment and it is not clear at present where that might come from. It might, though, form part of a distributed mouse functional genomics data network which has been discussed a possible offshoot of CASIMIR.

Several funding agencies have developed data sharing or data management policies in response to their agreed responsibility for the preservation and dissemination of data from publicly funded science. The obligations of data users, originators and funding agencies towards maintenance, distribution and use of data are clearly articulated in several key documents [14, 15], which form the core of the data sharing policies of many of those funding agencies which have explicitly adopted such a policy. In order to function effectively, however, such data sharing requires the adoption of shared reporting standards, such as the minimal standards discussed above, ontologies and syntax. Without advice, support and the incentive to comply with these requirements the responses we have obtained suggest that broad implementation of these policies will be difficult if not impossible. Both sanctions and rewards for data sharing may be required to help with the generation of a climate of trust and co-operation necessary to ensure compliance with policy.

The approach of two funding bodies in the UK exemplifies the current range of data policies. The Medical Research Council (MRC) expects data arising from MRC-funded research to be made available to the scientific community with as few restrictions as possible and whilst accepting that there may be reasons for delay in public release of data (e.g. Protection of IPR), expect that data should be released in a "timely and responsible manner". The emphasis of the MRC policy is that the originators should be responsible for appropriate curation, retention and dissemination [16]. The UK Natural Environment Research Council (NERC) on the other hand operates a policy of data management whereby a central, domain oriented infrastructure has been created to support data management and archiving with guidance and support in using the appropriate standards for data structure description and metadata. Such a policy goes a significant way to address many of the issues raised in the questionnaire responses, but is of course expensive to implement [17, 18]. The NERC approach provides a potential template for a training network to enable the dissemination of the necessary database skills to laboratories primarily involved in biological research. This could address the perception that there was a

widespread shortage of expertise in database design, especially outside the UK (which hosts the European Bioinformatics Institute and many of the other databases involved in this survey).

In recent years the "bottom-up" approach to developing standards through community consensus has proved to be the most effective way of establishing usable data standards and resources, such as ontologies tailor-made to the needs of that community. Global adoption will only happen if standards are easy to apply and meet the current and projected requirements of the community. Projects such as CASIMIR and the Gene Ontology can act as forums for the generation of community consensus and represent an important social integration of the resources and expertise within the biological community. It is hopefully through initiatives like this we can move to a seamless data network in the life sciences with all the power that will bring.

## 5 Acknowledgements

The authors CASIMIR (funded by the European Commission under contract number LSHG-CT-2006-037811) for financial support.

## 6 References

1. Rosenthal, N., Brown, S.: The mouse ascending: perspectives for human-disease models. Nat Cell Biol **9** (2007) 993-999
2. Oster, S., Langella, S., Hastings, S., Ervin, D., Madduri, R., Phillips, J., Kurc, T., Siebenlist, F., Covitz, P., Shanbhag, K., Foster, I., Saltz, J.: caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research. J Am Med Inform Assoc (2007)
3. Toyoda, T., Wada, A.: Omic space: coordinate-based integration and analysis of genomic phenomic interactions. Bioinformatics **20** (2004) 1759-1765
4. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P., Oinn, T.: Taverna: a tool for building and running workflows of services. Nucleic Acids Res **34** (2006) W729-W732
5. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nat.Genet. **25** (2000) 25-29
6. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., OBI_Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol **25** (2007) 1251-1255

7. Smith, C.L., Goldsmith, C.A., Eppig, J.T.: The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biol **6** (2005) R7
8. Gkoutos, G.V., Green, E.C.J., Mallon, A.-M., Hancock, J.M., Davidson, D.: Using ontologies to describe mouse phenotypes. Genome Biol **6** (2005) R8
9. Aidinis, V., Chandras, C., Manoloukos, M., Thanassopoulou, A., Kranidioti, K., Armaka, M., Douni, E., Kontoyiannis, D.L., Zouberakis, M., Kollias, G., Mugen NoE Consortium: MUGEN mouse database; animal models of human immunological diseases. Nucleic Acids Res **36** (2008) D1048-1054
10. Burger, A., Davidson, D., Baldock, R. (eds.): Anatomy Ontologies for Bioinformatics Principles and Practice, Vol. 6. Springer (2008)
11. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C.P., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo.Jaak, Vingron, M.: Minimum information about a microarray experiment (MIAME) - towards standards for microarray data. Nat Genet **29(4)** (2001) 365-371
12. Taylor, C.F.: Standards for reporting bioscience data: a forward look. Drug Discov Today **12** (2007) 527-533
13. Deutsch, E.W., Ball, C.A., Bova, G.S., Brazma, A., Bumgarner, R.E., Campbell, D., Causton, H.C., Christiansen, J., Davidson, D., Eichner, L.J., Goo, Y.A., Grimmond, S., Henrich, T., Johnson, M.H., Korb, M., Mills, J.C., Oudes, A., Parkinson, H.E., Pascal, L.E., Quackenbush, J., Ramialison, M., Ringwald, M., Sansone, S.A., Sherlock, G., Stoeckert, C.J.J., Swedlow, J., Taylor, R.C., Walashek, L., Zhou, Y., Liu, A.Y., True, L.D.: Development of the Minimum Information Specification for In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE). OMICS **10** (2006) 205-208
14. Wellcome Trust: Sharing data from large-scale biological research projects: a system of tripartite responsibility. Wellcome Trust Meeting Report, January 14–15, 2003. (2003) http://www.wellcome.ac.uk/assets/wtd003207.pdf
15. OECD: Recommendation of the Council concerning Access to Research Data from Public Funding. (2006) http://www.oecd.org/dataoecd/9/61/38500813.pdf
16. Medical Research Council: Policy on data sharing and preservation. (2006) http://www.mrc.ac.uk/PolicyGuidance/EthicsAndGovernance/DataSharing/PolicyonDataSharingandPreservation/index.htm
17. NERC: NERC Data Policy. (2002) http://www.nerc.ac.uk/research/sites/data/policy.asp
18. Field, D., Tiwari, B., Snape, J.: Bioinformatics and data management support for environmental genomics. PLoS Biol **3** (2005) e297