

Data management: databases and standards

Motivation to submit your data into microarray database

Why would you publish your gene expression data in a standard form?

- Peer-review: If the data is going to be interpreted by independent parties, the information about the sample has to be in the database
- Building database that can be searched: Controlled vocabularies and ontologies (species, cell types, compound nomenclature, treatments, etc) are needed for unambiguous sample description
- Sample annotation critical: Gene expression data have meaning only in the context of detailed description of the sample
- Required by some journals

Publications on Microarray Data Exchange Standards

Minimum information about a microarray experiment (MIAME)

Alvis Brazma¹, Pascal Hingamp², Chris Stoeckert⁶, John Aach⁷, Wil Terry Gaasterland¹⁰, Patrick Glen Markowitz¹³, John C. Matousek⁴, H Schulze-Kremer¹⁴, Jason Stewart

Microarray analysis has become a key genomic scale. Although many significant advances have been made, a major barrier to the establishment of databases and tools with respect to MIAME, we information rather than the technical

Introduction

After genome sequencing, DNA microarray and the most widely used source of genome-scale data. Microarray expression studies are primary quantities of gene expression and other functional data, which promise to provide key insights into interactions within and across metabolic pathways. Unlike genome sequence data, however, which formats for presentation and widely used tools to analyze the microarray data generated so far remain limited.

Several factors contribute to the barrier to wide adoption of microarray data. The field is young and has not yet approached the maturity needed to identify and standardize the data. In addition, gene expression data are often presented in a way that they are meaningful only in the context of a detailed description of the conditions under which they were generated, including the particular state of the biological system and the perturbations to which it has been

BIOINFORMATICS

Editorial

AN OPEN LETTER TO THE EDITORS OF THE JOURNALS

One of the underlying principles of scientific publication in peer-reviewed journals has been the requirement that the authors make available the data and materials necessary for a reader to reproduce the experiment and to determine whether the data support the authors' conclusions. In many instances, such as DNA sequence or protein structure data, this has been a challenge, because the data generated in each experiment and the complexity of the ancillary information needed to interpret the results are unlike anything that has yet faced the research community. Databases to hold the data and the tools to annotate them properly are under development. As an interim solution, this guide defines types of data that are necessary to reproduce a microarray experiment. It should be noted that this information is only of value as long as it is available, and that every effort should be made to provide access to published data until such time as it is available from a public database.

The members of the Microarray Gene Expression Data (MGED) society (<http://www.mged.org>), working over the past few years to develop standards for the publication of microarray data. The authors of this guide represent a large cross-section of the scientific community that have worked with microarray data and are convinced of the importance of the data and strongly urge journals to use these standards when deciding whether to publish a paper on microarray data. In December 2001, we published in *Nature Genetics* in which we described the Minimal Information About a Microarray Experiment (MIAME) standard.

Science's

COMPASS

LETTERS SCIENCE & SOCIETY POLICY FORUM BOOKS ET AL. PERSPECTIVES REVIEWS



Standards for Microarray Data

ONE OF THE UNDERLYING PRINCIPLES OF scientific publication in peer-reviewed journals has been the requirement that the authors make available the data and materials necessary for a reader to reproduce the experiment or analysis and to determine whether the data support the authors' conclusions. In many instances, such as DNA sequence or protein structure data, this has been a challenge, because the data generated in each experiment and the typical complexity of the ancillary information needed to interpret the results are unlike anything that has yet faced the biological research community. Databases to hold microarray data and the tools to annotate them properly are under development. As an interim solution, we have described the types of data that are necessary to reproduce and interpret a microarray experiment. It should go without stating that this information is only of value as long as it is available, so every effort should be made to provide stable access to published data until such time as it is available from a public database.



In December 2001, we published a commentary in which we described MIAME—the Minimal Information About a Microarray Experiment (MIAME). MIAME is presented as a proposed standard for representation of array data that would be sufficient to allow readers of published reports to replicate the analysis presented and to facilitate the development of novel methods of data analysis by providing access to necessary primary data.

Community response to MIAME was favorable, and many instrument manufacturers, software developers, and international databases moved to adapt their systems to capture and manage MIAME-compliant data. However, by far the most common request from the community has been for a brief set of guidelines that could be used by authors, editors, and referees to try to meet the MIAME data standards.

These requirements can easily be met by adequately describing the experiment, the materials and methods used, and either (i) a relatively simple supplementary Web site or (ii) submission of this information to one of the public repositories [ArrayExpress (www.ebi.ac.uk/arrayexpress) or GEO (www.ncbi.nlm.nih.gov/geo/)]. Reviewers and editors should strive to help authors meet these requirements and should ensure that, if a publication cannot meet them, there are sound reasons. This document in no way attempts to eliminate the need for editors or reviewers to use their judgment on both the appropriateness of the

MIAME standard. In December 2001, we published in *Nature Genetics* in which we described the Minimal Information About a Microarray Experiment (MIAME) standard.

CATHERINE BROOKSBANK,⁷ HELEN C. CAUSTON,⁷ DUCCIO CAVALERI,⁴ TERRY GAASTERLAND,⁵ PASCAL HINGAMP,² FRANK HOEHLSTEIG,⁶ MARTIN RINGWALD,³ PAUL SPELLMAN,⁹ CHRISTIAN J. STOECKERT JR.,¹⁰ JASON E. STEWART,¹¹ RONALD TAYLOR,¹² ALVIS BRAZMA,^{2*} JOHN QUACKENBUSH¹³

¹Department of Genetics, Stanford University, Stanford, CA, ²EMBL-European Bioinformatics Institute, Cambridge, UK, ³Clinical Sciences Centre, Imperial College, London, ⁴Bauer Center for Genomic Research, Harvard University, Cambridge, MA, ⁵The Rockefeller University, New York, NY, ⁶Université D'Aix-Marseille II, Marseille, France, ⁷University Medical Center, Utrecht, Netherlands, ⁸The Jackson Laboratory, Bar Harbor, ME, ⁹University of California at Berkeley, ¹⁰University of Pennsylvania, Philadelphia, PA, ¹¹Open Informatics, Albuquerque, NM, ¹²Center for Computational Pharmacology, University of Colorado School of Medicine, Denver, CO, ¹³The Institute for Genome Research, Rockville, MD.

*To whom correspondence should be addressed. EMBL Outstation-Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. E-mail: brazma@ebi.ac.uk

References

1. A. Brazma et al., *Nature Genet.* **29**, 365 (2001)

Editor's Note: Science supports the evolving standardization of microarray data, one view of which is presented in this letter. We urge our authors to follow the criteria set forth here, although it is not a requirement for publication, and to let us know your experiences and reactions. Please send comments through the dEBates feature of Science Online (www.sciencemag.org), as they will help us in continuing to participate in this process and establish Science policy.

MIAME Standards:

Nature family, Cell family, EMBO reports, Bioinformatics, Genome Research, Genome Biology, Science, The Lancet, NEJM, and others....

Slide by John Quackenbush, 2005

Where would you need other people's published data?

- Experiment design: See what other people have achieved, so that you can plan ahead your own experimental setup
- Getting more out of your own experiment (with less money)
- Model construction
- Key thing: comparison of larger amt of datasets

Comparing results: what you need to know about the datasets?

- sample annotations. Was the strain the same as yours?
- selection of genes on the chip(s)
- which platform was used
- study conditions (growth medium, temp, treatments)
- experimental conditions (labeling, hybridization, scanning)
- how the data was pre-treated and analyzed - should you re-analyze or can you compare normalized values?

Describing your own data well helps others

- These are the details you should describe when you publish your data
- The better you do this, the more time and efforts you use when describing and submitting well annotated data, the more useful your data will be for the research community.

Microarray data standards

Three parts of a gene expression database

- Gene annotation
- Sample annotation
- Gene expression matrix

Descriptive data and measured data

Metadata; Data to be stored **before** the experiment

- Description of the array and the sample.
- Direct access to all the cDNA and gene sequences, annotations, and physical DNA resources.

Gene Expression Matrix; Data to be stored **after** the experiment

- Raw Data - scanned images, processed data, and/or interpreted data, together with how the data were processed or analyzed.

Raw, intermediate and final data

Raw data

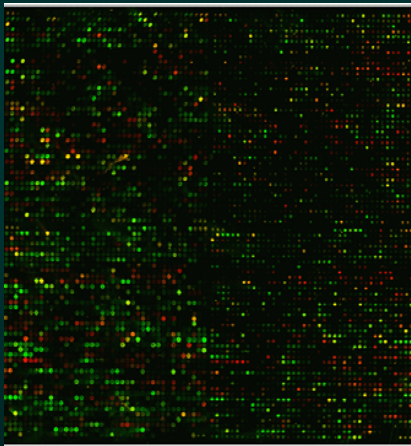
Intermediate data

Final data

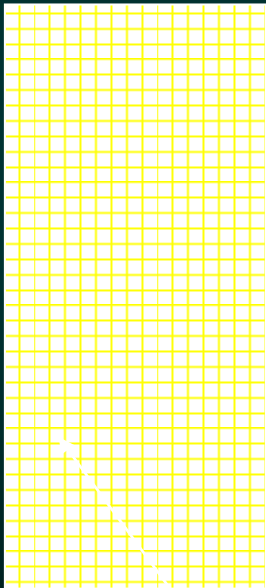
Array scans

Images

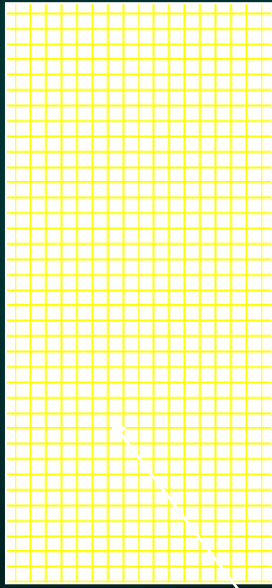
Samples



Spots



Genes



Spot/Image
quantiations

Gene
expression
levels

Sample annotation

- In general, avoid free text descriptions but some uses are unavoidable
- Controlled vocabularies and ontologies should be used wherever available
- Externally defined controlled vocabularies and ontologies should be used whenever they exist

Controlled vocabularies and Bio-ontologies

- Standardized language for describing an experiment, data analysis, etc.
- An ontology is a formal way of representing knowledge in which concepts are described both by their meaning and their relationship to each other.
- Unique identifiers that are associated with each concept in biological ontologies (bio-ontologies) can be used for linking to and querying molecular databases.
- GO Gene ontology
- MGED Ontology
- See OBO (Open Biomedical Ontologies) pages at <http://obo.sourceforge.net/>

MIAME = Minimum Information About a Microarray Experiment

What is MIAME?

- A document, the goal of which is to specify the minimum information that must be reported about a microarray experiment in order to ensure its interpretability, as well as potential verification of the results
- Underlying motivation –
 - to enable the establishment of public repositories for microarray data
 - to serve as a basis for designing a microarray data exchange format

How to think about MIAME

What minimum information about a microarray experiment should be recorded in a database for the database entries to be usable on stand-alone basis:

- the users may not know any background information that is not recorded
- the database should be usable for automated data analysis and mining, i.e. not only on record-by-record basis
- the data may be coming from different laboratories and different technology platforms

MIAME

- MIAME is not designed as a 'questionnaire' that can be filled in, but only as an informal specification based on which such a questionnaire, in fact, an annotation tool, can be based
- MIAME serves as a good document to help you think about what to save.

MIAME checklist

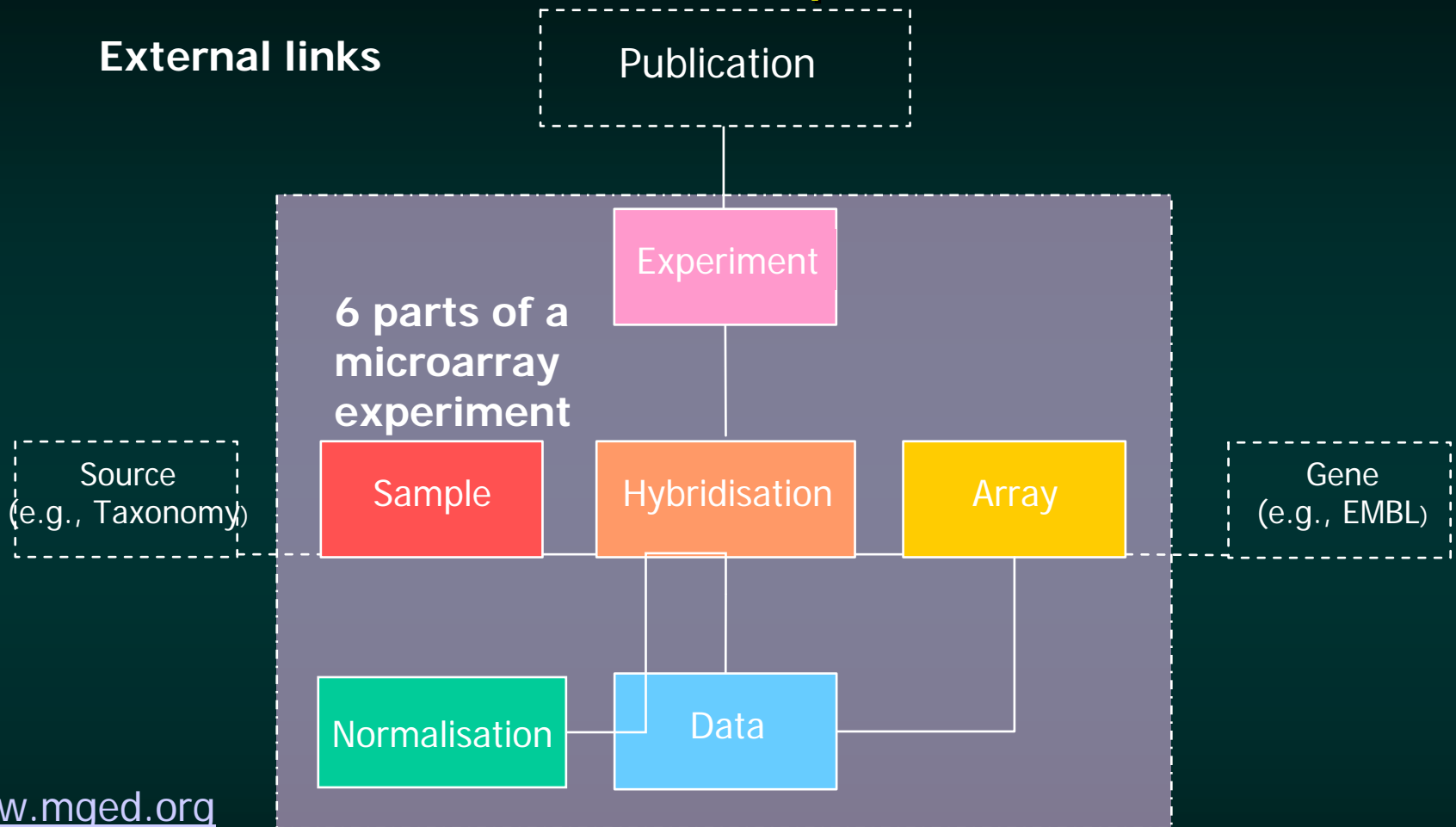
- MGED Guide to authors, editors and reviewers of microarray gene expression papers

Six parts

1. Biological Samples
2. Hybridizations
3. Data Normalization and Transformation
4. Experimental Design and Factors
5. Array Design
6. Measurements

http://www.mged.org/Workgroups/MIAME/miame_checklist.html

MIAME six parts



www.mged.org

Samples used, extract preparation and labeling:

Sample

- The origin of the biological sample
- Manipulation of biological samples and protocols used
- Protocol for preparing the hybridization extract
- Labeling protocol(s)
- External controls (spikes)

Hybridization procedures and parameters:

Hybridisation

- Laboratory protocol including
 - The solution (e.g. concentration of solutes)
 - Blocking agent
 - Wash procedure
 - Quantity of labelled target used
 - Time, concentration, volume, temperature
 - Description of the hybridisation instruments
 - Optional additional 'qualifier, value, source' list

MIAME: Experimental Design

- The number of hybridizations performed in the experiment
- The type of reference used for the hybridizations, if any
- Quality control steps taken: for example, replicates or dye swaps
- URL of any supplemental websites

Array Design:

Array

- For each feature (spot) on the array, its location on the array and the ID of its respective reporter (molecule present on each spot) should be given.
- For each reporter, its type (e.g., cDNA or oligonucleotide) should be given, along with information that characterizes the reporter molecule unambiguously, in the form of appropriate database reference(s) and sequence (if available)
- Submitters are encouraged to submit seqs to public databases

Array Design:

- For commercial arrays: manufacturer, catalog number and references to the manufacturer's website
- For non-commercial arrays, the following details should be provided:
 - The source of the reporters (the cDNA or oligo collection used) with references
 - The method of reporter preparation
 - The spotting protocols used (array substrate, spotting buffer, and post-printing processing, including cross-linking)
 - Any additional treatment performed prior to hybridization

Measurement data and specifications:

Data

- Quantitations from all arrays upon which the authors base their conclusions.
- Scanning hardware and software used
- Image analysis software used
- A description of the measurements produced by the image analysis software and a description of which measurements were used in the analysis
- The complete output of the image analysis *before* data selection and transformation

Data Normalization and Transformation :

Normalisation

- Data selection and transformation procedures
- Final gene expression data table(s) used to by authors to make their conclusions *after* data selection and transformation

MIAME glossary available

- MIAME Home
- MIAME 1.1
- MIAME MAGE-OM**
- MIAME Checklist
- MIAME Software
- MIAME Archive
- MIAME Glossary
- MIAME to MAGE-OM
- MAGE-OM to Ontology

Home : Workgroups : MIAME : MIAME MAGE-OM : MIAME Glossary

MIAME concepts are listed in alphabetical order and definitions are provided.

Age	The time period elapsed since an identifiable point in the life cycle of an organism. (If a developmental stage is specified, the identifiable point would be the beginning of that stage. Otherwise the identifiable point must be specified such as planting) [MGED Ontology Definition]
Amount of nucleic acid labeled	The amount of nucleic acid labeled
Amplification method	The method used to amplify the nucleic acid extracted
Array design	The layout or conceptual description of array that can be implemented as one or more physical arrays. The array design specification consists of the description of the common features of the array as the whole, and the description of each array design elements (e.g., each spot). MIAME distinguishes between three levels of array design elements: feature (the location on the array), reporter (the nucleotide sequence present in a particular location on the array), and composite sequence (a set of reporters used collectively to measure an expression of a particular gene)
Array design name	Given name for the array design, that helps to identify a design between others (e.g. EMBL yeast 12K ver1.1)

http://www.mged.org/Workgroups/MIAME/miame_glossary.html



MGED and MAGE standards

MGED Society

MGED Home

[Home](#) [Meetings](#) [Workgroups](#) [Mission](#) [MGED Board](#) [Site Map](#)

Microarray Gene Expression Data Society - MGED Society

The Microarray Gene Expression Data (MGED) Society is an international organisation of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments.

The current focus is on establishing standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards, and promoting the sharing of high quality, well annotated data within the life sciences community. A long-term goal for the future is to extend the mission to other functional genomics and proteomics high throughput technologies.

Read more about:

[MGED Sponsors](#) • [Defined MGED Standards](#) • [Historical highlights](#) • [MGED Meetings](#) • [MGED Supported Meetings](#) • [Programming Jamborees](#) • [MGED Goals](#) • [MGED Workgroups](#) • [Relevant Publications](#) • [MGED Board](#)

Latest News:

New RSBI WGs	13/09/2004
Following plans of the MGED Society to extend its mission to other functional genomics technologies the Toxicogenomics WG is likewise enlarging objectives to include other communities. More about this new RSBI WGs .	
MIAME Open Letter	26/07/2004
An open letter about MIAME compliant data as a condition for publishing has been sent to scientific journals [RTF 14kb] [HTML 12kb].	
MGED 7 Registration	26/07/2004
MGED 7 meeting , September 8-10, Toronto, ON, Canada. Register now from here .	
Updated proposal on MIAME in MAGE-ML	13/05/2004
An updated version of the recommendations on how to encode MIAME required information in MAGE-ML is available from here .	

MGED Sponsors:



International organization
Comprised of biologists,
computer scientists, and data
analysts

Aims to facilitate the sharing
and evaluation of microarray
data

- Establish standards for microarray data annotation
- Create microarray databases
- Promote sharing of high quality, well-annotated data
- Generalize to data generated by functional genomics and proteomics experiments

MAGE standard

- MAGE (MicroArray and Gene Expression)

is a standard for the representation of microarray expression data and it is able to capture information specified by MIAME.

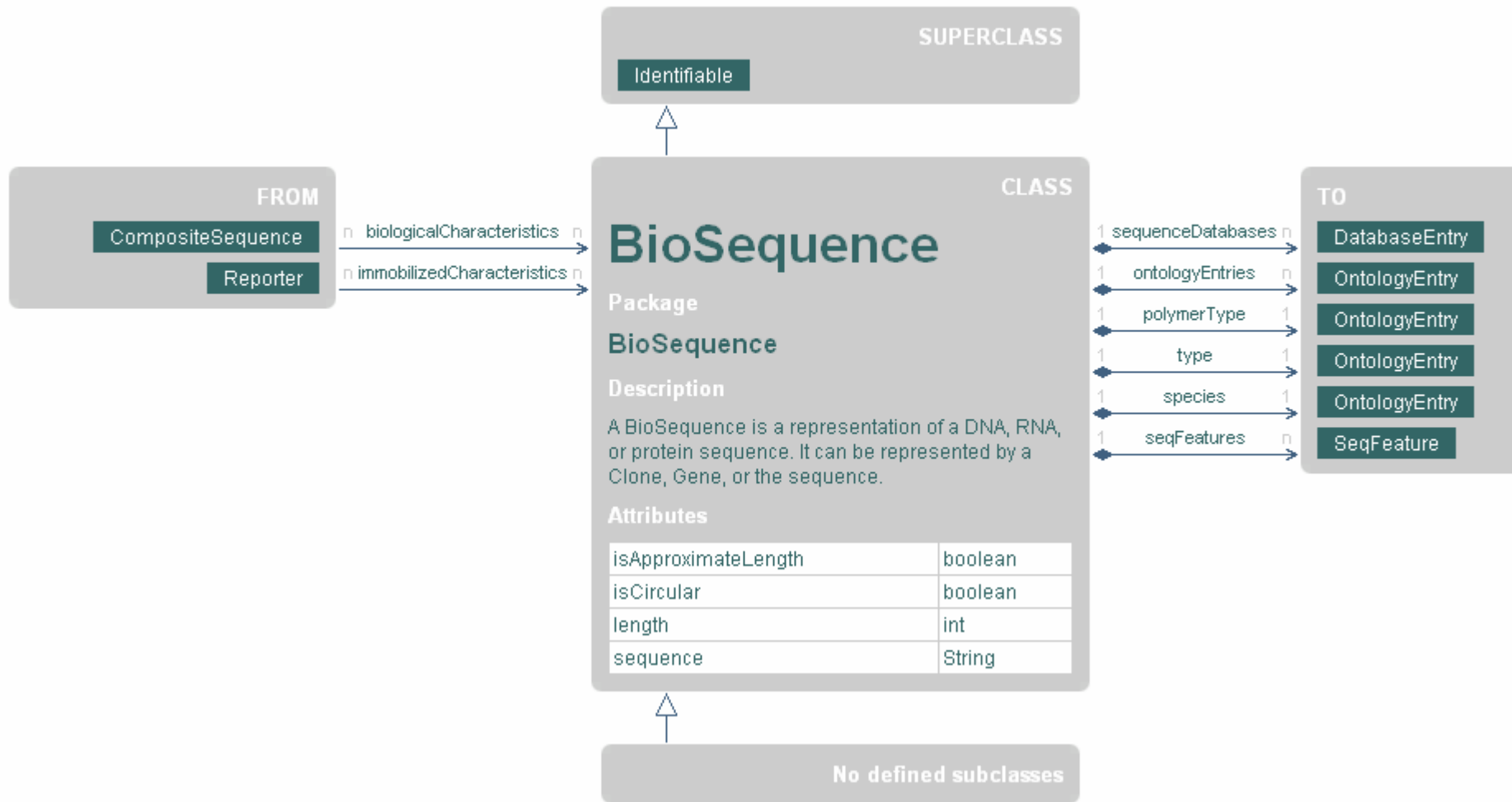
- MAGE consists of three parts:
 - An object model (MAGE-OM),
 - a document exchange format (MAGE-ML), which is derived directly from the object model, and
 - software toolkits (MAGE-STK), which help users to create MAGE-ML.
- Goal is to create computational tools that implement MIAME

MAGE-OM (Microarray Gene Expression Object Model)

- MAGE-OM is an object model for the representation of microarray expression data that facilitates the exchange of microarray information between different data systems and organizations.
- A standard underlying representation of expression data
- A database schema
- Database created using MAGE-OM is able to store experimental data from different types of DNA technologies such as cDNA, oligonucleotides or Affymetrix.
- It is also capable of storing experiment working processes, protocols, array designs and analyzing results.

MAGE-OM

- Key components
 - **Experiment** package: experiment goals and design
 - **BioMaterial** package: biological materials used and description of their creation
 - **ArrayDesign, BioSequence** packages: array design and purpose
 - **Array** package: array manufacture
 - **BioAssay** package: hybridization, wash and scan information
 - **BioAssayData** package: gene-expression data



Why good object models are important

- Good models
 - reduce the amount of data storage that is required
 - capture all (most) of the required cases (prevent “lying” to the database)
 - create a mechanism for storing controlled vocabularies (e.g. a list of words that are “allowed” as descriptors of common data types
 - man, human, homo sapiens, H. sapiens ---> homo sapiens - makes querying possible

MAGE-ML (MicroArray Gene Expression Mark-up Language)

- MAGE-ML is an XML-based file format able to capture all MIAME required information, and it is derived directly from MAGE-OM.
- In other words, MAGE-ML is an XML representation of the MAGE-OM
- A standard for data communication
- A MAGE-ML document contains up to 16 different packages, their order and content being specified by the MAGE-ML.dtd, a document type definition for this type of file.

MAGE-ML

- It is important to remember that valid MAGE-ML file does not necessarily mean that it contains all the information required by MIAME, *e.g.* it can be used to transfer some manufacturer specific data without any expression information.
- MAGE-ML files can be downloaded from ArrayExpress (<http://www.ebi.ac.uk/arrayexpress>).

MIAME is what to store, MAGE-OM is how to store it and MAGE-ML is how to communicate it.

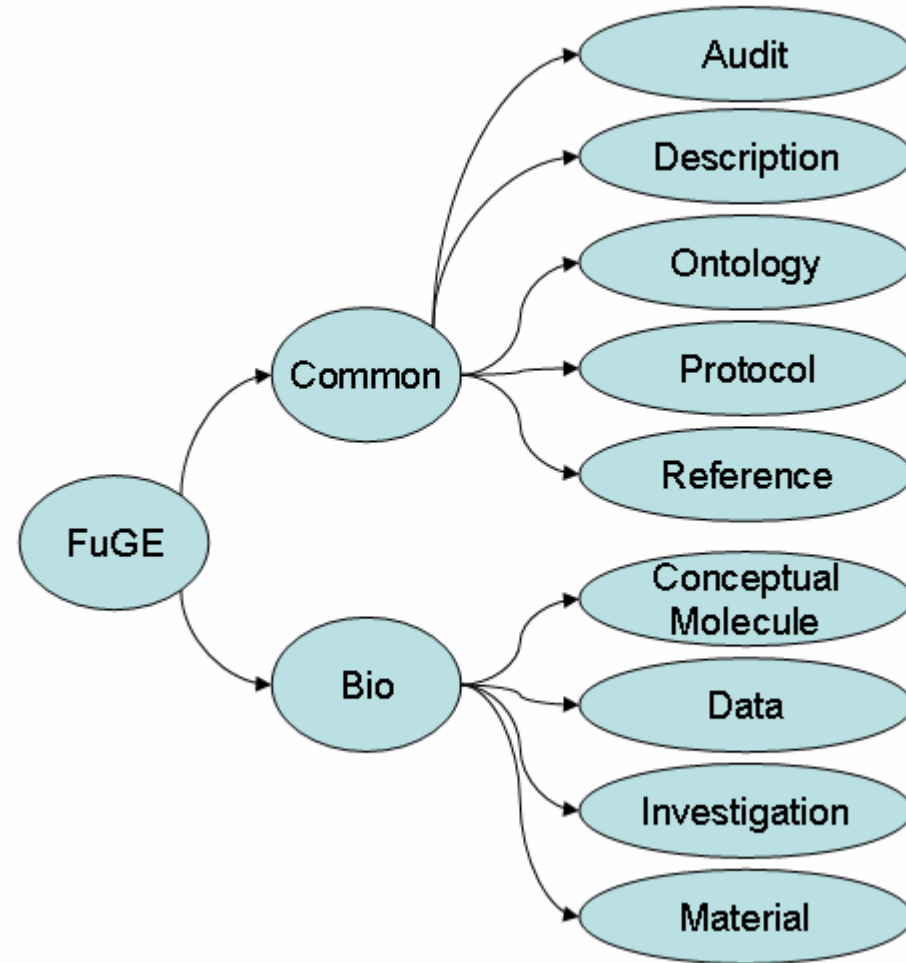
MAGE-STK (MicroArray Gene Expression Software Toolkit)

- Because of the complexity of the model it's not quite human readable (though it's XML), but a software is needed to support importing and exporting MAGE-ML.
- The MAGE Software Toolkit is a collection of Open Source packages that implement the MAGE Object Model in various programming languages.
- The suite currently supports three implementations: MAGE-Perl, MAGE-Java, and MAGEC++.
- The idea is to be able to have an intermediate object layer that can then be used to export data to MAGE-ML, to store data in a persistent data store such as a relational database, or as input to software-analysis tools.

MAGE-OM future

- New version of MAGE-OM2 coming within a year
- FuGE-OM (The Functional Genomics Experiment Object Model) = Completely new, more generic object model under development that will tie up OMs for *e.g.* microarrays and proteomics
- Divided into generic (Common) and technology-specific (Bio) areas

FuGE package structure:



Glossary

- MIAME is a standard
- MAGE-OM is an object model
- ArrayExpress is a database implementation which uses that model
- MAGE-ML is a mark-up language auto generated from MAGE-OM
- MIAMExpress is a tool for generating data in MAGE-ML format

slide by Helen Parkinson, EBI

Public data repositories

Three main public data repositories

- Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI)
- ArrayExpress at the European Bioinformatics Institute (EMBL-EBI)
- the Center for Information Biology Gene Expression (CIBEX) database at DNA Data Bank of Japan (DDBJ)
today, we'll talk about the first two

General, local and/or private databases:

- SMD, Stanford Microarray Database <http://genome-www.stanford.edu/microarray> (Oracle)
- BASE <http://base.thep.lu.se/> (MySQL)
- GeneX <http://www.ncgr.org/genex/> (PostgreSQL)

Stanford Microarray Database

<http://genome-www.stanford.edu/microarray>

- Largest academic database
- Uses Oracle as DBMS
- Stores data from Stanford investigators and their collaborators
- Contains 61013 experiments (BioAssays), of which 10514 are public, from 37 organisms
- Data is from 2-color DNA arrays
- Data tables more consistent with MAGE-OM

Example databases with public microarray data:

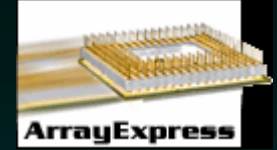
specific databases:

- GERMonline, germ cell expression in 11 model organisms
<http://www.germonline.org/>
- HugeIndex, Affy data for 19 normal human tissues
<http://www.hugeindex.org>

some yeast-specific databases:

- SGD (*Saccharomyces* Genome Database)
<http://www.yeastgenome.org/>
- yMGV (Yeast Microarray Global Viewer)
<http://www.transcriptome.ens.fr/ymgv/>
- YeastBASE (at CSC)

ArrayExpress by EBI

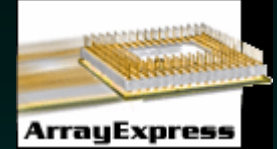


<http://www.ebi.ac.uk/arrayexpress/>

Data organized as:

- Experiments: 1373
- Arrays: 891
- Protocols: 6512
- Hybridizations: 39916

ArrayExpress by EBI

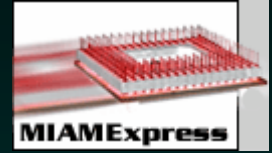


- Database implementation of MAGE-OM
- Oracle as DBMS
- Able to import MAGE-ML format
- Can deal with both raw and processed data
- Independence of:
 - experimental platforms
 - image analysis methods
 - data normalization methods

ArrayExpress - three ways to submit data:

1. Via web-based tools, MIAMExpress (MIAME compliant microarray data submission tool) or Tox-MIAMExpress (toxicogenomics specific version)
2. Tab2MAGE (spreadsheet)
3. As MAGE-ML formatted files

ArrayExpress - submitting data :



1. MIAMExpress

- Based on MIAME concepts and questionnaire
- Experiment, Array, Protocol submissions
- Uses Controlled Vocabulary/Ontology wherever possible

MIAMExpress - all the protocols are described first:

Protocols

- Array preparation
- Sample Growth Condition
- Sample Treatment
- Extraction
- Pooling
- Labeling
- Hybridization
- Scanning
- Normalization
- Transformation

MIAMExpress - Array design is next:

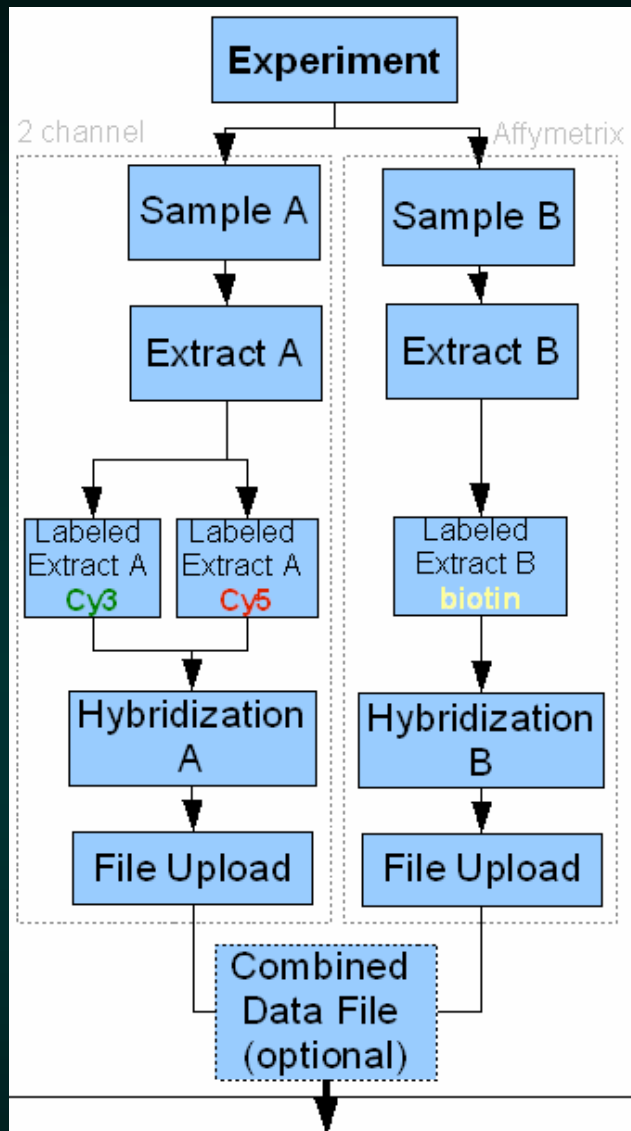
- For commercial chips, pick from the ones already in ArrayExpress (commercial chips), ask ArrayExpress staff to make one for you, or create a new array design
- From home-made chips, you have to generate Array Description File (ADF)

Protocols

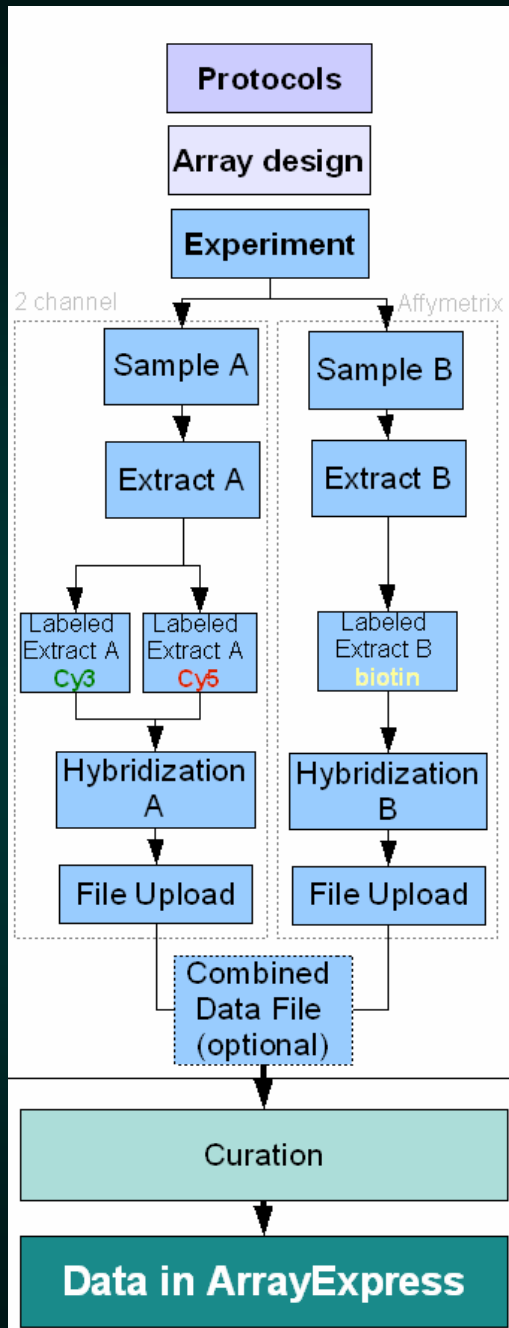
Array design

MIAMExpress - Experiment submission follows:

1. Experimental design
2. Publication
3. Samples
4. Extracts
5. Labeled extracts
6. Hybridizations
7. Complete submission



MIAMExpress – finally curation and publishing:



- Data is checked by ArrayExpress staff
- You get an accession number for your experiment
- The submission will remain private until the release date that you have specified



ArrayExpress - submitting data :



2. Tab2MAGE

- a spreadsheet parser that generates MAGE-ML and comes with some handy checking scripts
- tool for submitting large datasets (>50)
- three main sections: Experiment, Protocol and Hybridization in a defined spreadsheet (or tab-delimited text file) format
- two (Perl) scripts validate the data and generate MAGE-ML of it

ArrayExpress - submitting data :

3. MAGE-ML

- tool for submitting very large datasets (>500 per year)
- needs expertise & dedication; you need to learn the MAGE-OM
- you have to map your schema to MAGE-OM and use MAGE-STK locally (perl or java) to export MAGE-ML

GEO - Gene Expression Omnibus by NCBI

<http://www.ncbi.nlm.nih.gov/geo/>

Data organized into 3 entities:

- Platforms (GPL) 2208
- Samples (GSM) 79982
- Series (GSE) 3520

GEO - Gene Expression Omnibus by NCBI

Platforms (GPLxxx):

- Corresponds to the *Array* package in MAGE-OM
- Describes the list of elements on the array (e.g., cDNAs, oligonucleotide probesets, ORFs, antibodies) or the list of elements that may be detected and quantified in that experiment (e.g., SAGE tags, peptides).
- A platform may reference many samples that have been submitted by multiple submitters.
- For commercial chips, platform may already be in GEO

GEO - Gene Expression Omnibus by NCBI

Samples (GSMxxx):

- Corresponds to the *Experiment* package in MAGE-OM
- Describes the conditions under which an individual sample was handled, the manipulations it underwent, and the measurements recorded
- A sample entity must reference only one platform and may be included in multiple series.

GEO - Gene Expression Omnibus by NCBI

Series (GSExxx):

- Corresponds to the *BioAssay* package in MAGE-OM
- A series provides a focal point and description of the experiment as a whole.
- You can specify experimental subsets
- Series records may also contain tables describing extracted data, summary conclusions, or analyses (these are emailed directly to GEO staff)

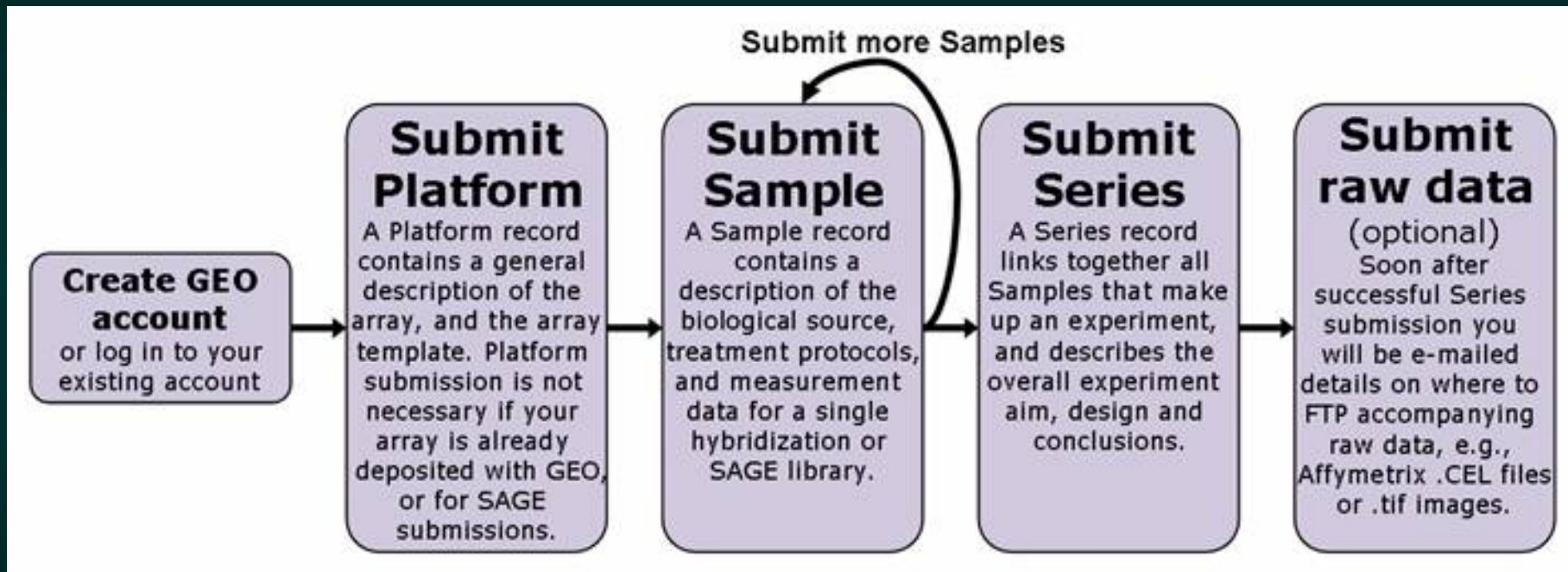
GEO - Gene Expression Omnibus by NCBI

GEO Datasets (GDSxxx):

- Curated sets of GEO Sample data
- Collection of biologically and statistically comparable GEO Samples
- Same Platform
- Value measurements for each Sample within a GDS are calculated in an equivalent manner
- Experimental design is explained through GDS subsets.

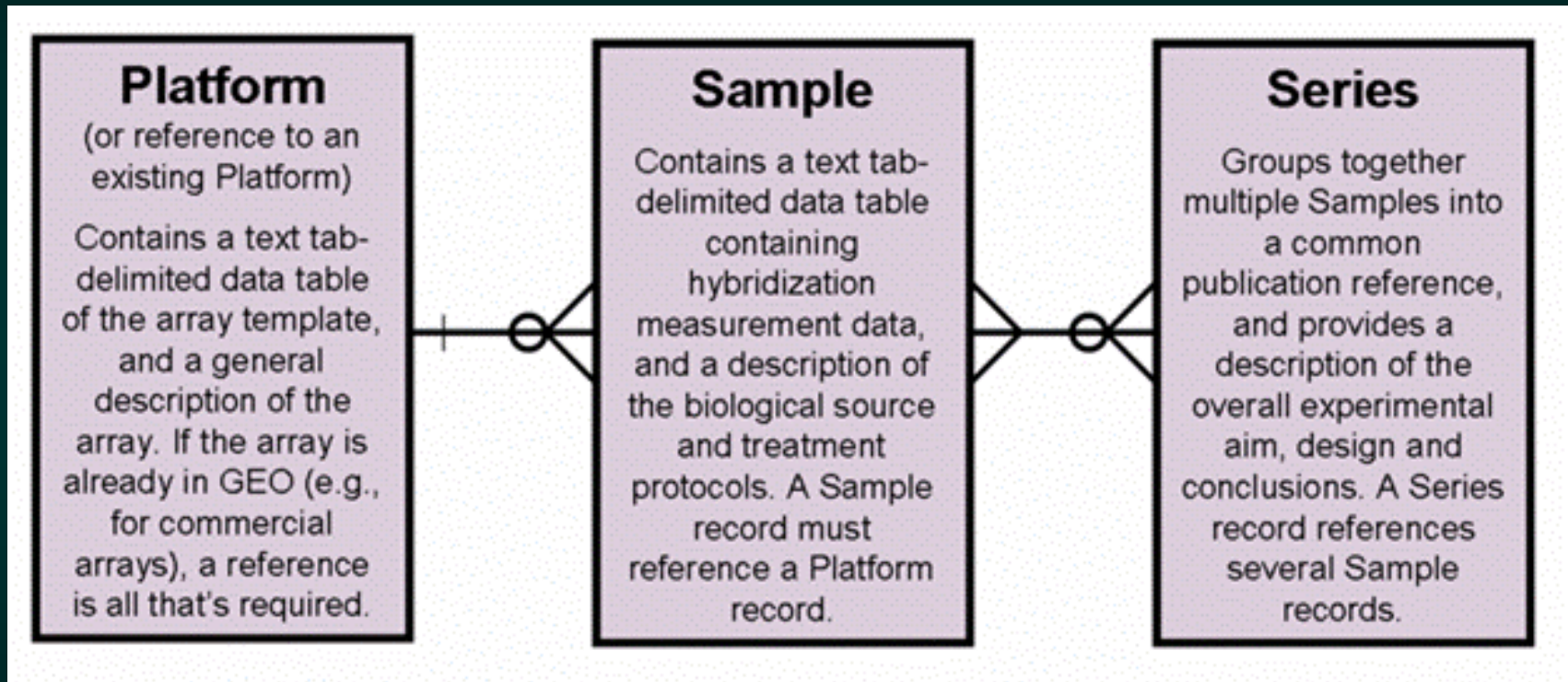
GEO - You can submit data in two main ways:

1. Via Internet using interactive web forms (for individual datasets)



GEO - You can submit data in two main ways:

2. Batch Submission (if your data are already in a database, or if you have many samples to submit)



GEO - submitting data

Batch Submission:

- Four formats accepted:
 1. SOFTtext file
 2. SOFTmatrix (spreadsheet or text, *e.g.* as an Affymetrix pivot spreadsheet for many samples)
 3. MINiML
 4. MAGE-ML
- SOFT format = Simple Omnibus Format in Text
can also be used to download data from GEO ftp
site <ftp://ftp.ncbi.nih.gov/pub/geo/data/geo/>

GEO - submitting data

SOFT template for platform:

```
^PLATFORM=[required]
!Platform_title = [required]
!Platform_technology = [required][VALUE=spotted DNA/cDNA, in situ oligonucleotide, spotted oligonucleotide, antibody, tissue, MS, MPSS]
!Platform_distribution = [required][VALUE=non-commercial, commercial, custom-commercial]
!Platform_organism = [required]
!Platform_manufacturer = [required]
!Platform_manufacture_protocol = [required]
!Platform_catalog_number = [optional]
!Platform_support = [optional]
!Platform_coating = [optional]
!Platform_description = [optional]
!Platform_web_link = [optional]
!Platform_contributor = [optional; 1 per author; use 'firstname,lastname' or 'firstname,middleinitial,lastname']
!Platform_pubmed_id = [optional]
#ID = [required; a unique id should be provided for each 'spot' on the array]
#HEADER_2 = [required; all elements of the array should be identified using one or more columns in addition to the ID column]
#HEADER_3 = [optional]
#HEADER_N = [optional; provide as many headers as needed to fully describe the elements of the array]
!Platform_table_begin
ID  HEADER_2  HEADER_3  HEADER_N
...insert data table here; columns may appear in any order after the ID column...
!Platform_table_end
```

GEO - submitting data

Attributes explained in GEO help pages:

Label	Number of allowed labels	Allowed values and constraints	Content guidelines
^PLATFORM entity			
!Platform_title	1	string of length 1-120 characters, must be unique within local file and over all previously submitted Platforms for that submitter	Provide a unique title that describes your Platform. We suggest that you use the system [institution/lab]-[species]-[number of features]-[version], e.g. "FHCRC Mouse 15K v1.0".
!Platform_distribution	1	commercial, non-commercial, custom-commercial, or virtual	Microarrays are 'commercial', 'non-commercial', or 'custom-commercial' in accordance with how the array was manufactured . Use 'virtual' only if creating a virtual definition for MS, MPSS, SARST, or RT-PCR data.
!Platform_technology	1	spotted DNA/cDNA, spotted oligonucleotide, in situ oligonucleotide, antibody, tissue, SARST, RT-PCR, MS, or MPSS	Select the category that best describes the Platform technology.
!Platform_organism	1 or more	use standard NCBI Taxonomy nomenclature	Identify the organism(s) from which the features on the Platform were designed or derived.
!Platform_manufacturer	1	any	Provide the name of the company, facility or laboratory where the array was manufactured or produced.
!Platform_manufacture_protocol	1	any	Describe or reference the array manufacture protocol. Include as much detail as possible, e.g., clone/primer set identification and preparation, strandedness/length, arrayer hardware/software, spotting protocols. It is strongly recommended that complete protocol descriptions are provided within your submission. References to published protocol descriptions are acceptable - please provide complete citation information. Links to web sites that provide protocol information are not recommended since Web addresses and content often change.
!Platform_catalog_number	0 or more	any	Provide the manufacturer catalog number for commercially-available arrays.
!Platform_web_link	0 or more	valid URL	Specify a Web link that directs users to supplementary information about the array. Please restrict to Web sites that you know are stable.
!Platform_support	0 or 1	any	Provide the surface type of the array, e.g., glass, nitrocellulose, nylon, silicon, unknown.

GEO - submitting data

MINiML and MAGE-ML :

- both are XML and follow MIAME, but MINiML is a standalone XML schema definition and MAGE-ML is a DTD (Document Type Definition) generated automatically from the object model (MAGE-OM).
- MINiML (MIAME Notation in Markup Language) only as beta version and subject to change
- MAGE-ML can structure data in a variety of ways and is mostly suitable when using the MAGE-OM as object model in an underlying database.
- GEO can accept MAGE-ML submissions, but cannot currently export MAGE-ML.

References

- O. Bard & Rhee. Ontologies In Biology: Design, applications and Future challenges. *Nature Reviews Genetics* 5, 213-222 (2004); doi:10.1038/nrg1295
1. 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., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, 29, 365-71.
 2. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Garcia, L. G., Oezcimen, A., Rocca-Serra, P., and Sansone, S-A. (2003) ArrayExpress-a public repository for microarray gene expression data at the EBI, *Nucleic Acids Research*, 31, 68-71.
 3. Edgar, R., Domrachev, M., and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Research*, 30, 207-10.
 4. Gardiner-Garden, M. and Littlejohn, T. G. (2001) A comparison of microarray databases, *Briefings in Bioinformatics* 2, 143-158.
 5. Gollub, J., Ball, C.A., Binkley, G., Demeter, J., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J. C., Schroeder, M., Brown, P. O., Botstein, P., and Sherlock, G. (2003) The Stanford Microarray Database: data access and quality assessment tools, *Nucleic Acids Research*, 31, 94-96.
<http://genomebiology.com/2002/3/9/research/0046.1>
 6. Penkett, C. J. and Bähler, J. (2004) Navigating public microarray databases. *Comp Funct Genom* 2004; 5: 471-479.
 7. Saal, L. H., Troein, C., Vallon-Christersson, J., Gruvberger, S., Borg, Å., and Peterson, C. (2002) BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data, *Genome Biology*, 3, software0003.1-0003.6.
 8. Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W. L., Goncalves, J., Markel, S., Jordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, D., Senger, M., Aronow, B. J., Robinson, A., Bassett, D., Stoeckert, C. J. Jr, and Brazma, A. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML), *Genome Biology*, 3, research0046.1-0046.9.
 9. Stoeckert, C. J., Causton, H. C., and Ball, C. A. (2002) Microarray databases: standards and ontologies, *Nature Genetics*, 32, 469-473.