

“MGED Standard implementation: establishing an infrastructure for sharing microarray data ”

Philippe Rocca-Serra

Scientific Data Manager

Microarray Informatics Team @ EMBL-EBI

European Bioinformatics Institute, Cambridge, UK

NETTAB meeting, Bologna, 27-28th November 2003



Talk structure

- Standardization efforts
 - Microarray standards by the MGED Society
- Infrastructure at the EBI
 - ArrayExpress Functional implementation of MGED standards
 - ExpressionProfiler
- Future and Development
 - Infrastructure
 - Standards



European Molecular Biology Laboratory

International network of 5 research institutes dedicated to **research** and **service** in **molecular biology**

Heidelberg

Services

Building, maintaining and making available **databases**

Grenoble

Hamburg

Research in

bioinformatics and computational molecular biology

Monterotondo

Industry

Promote **standards**

Hinxton
UK

EBI

Service

Research

Training

Industry

Microarray Informatics Team ?

Standards are important.....

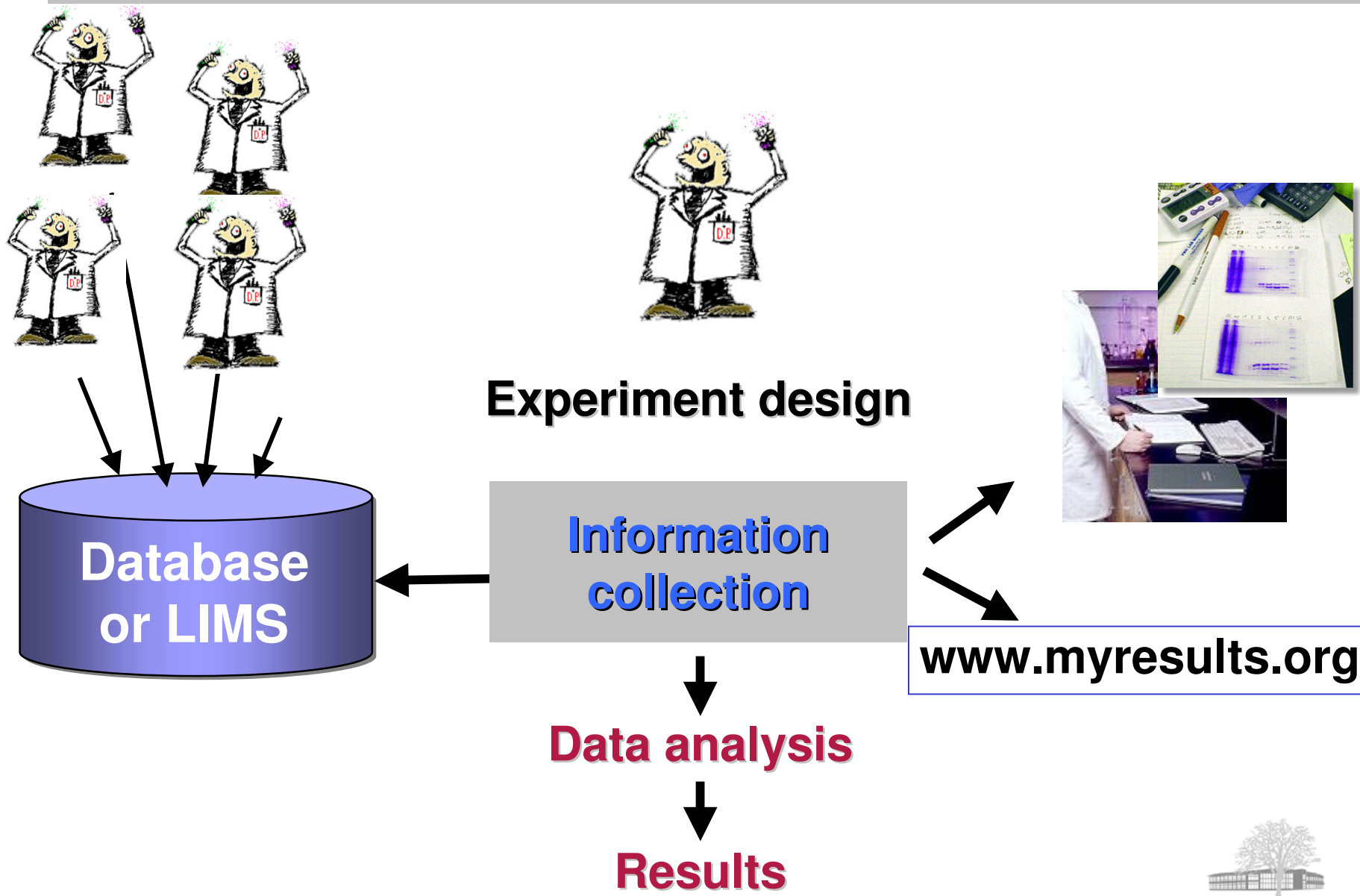
This is not a recent issue.....



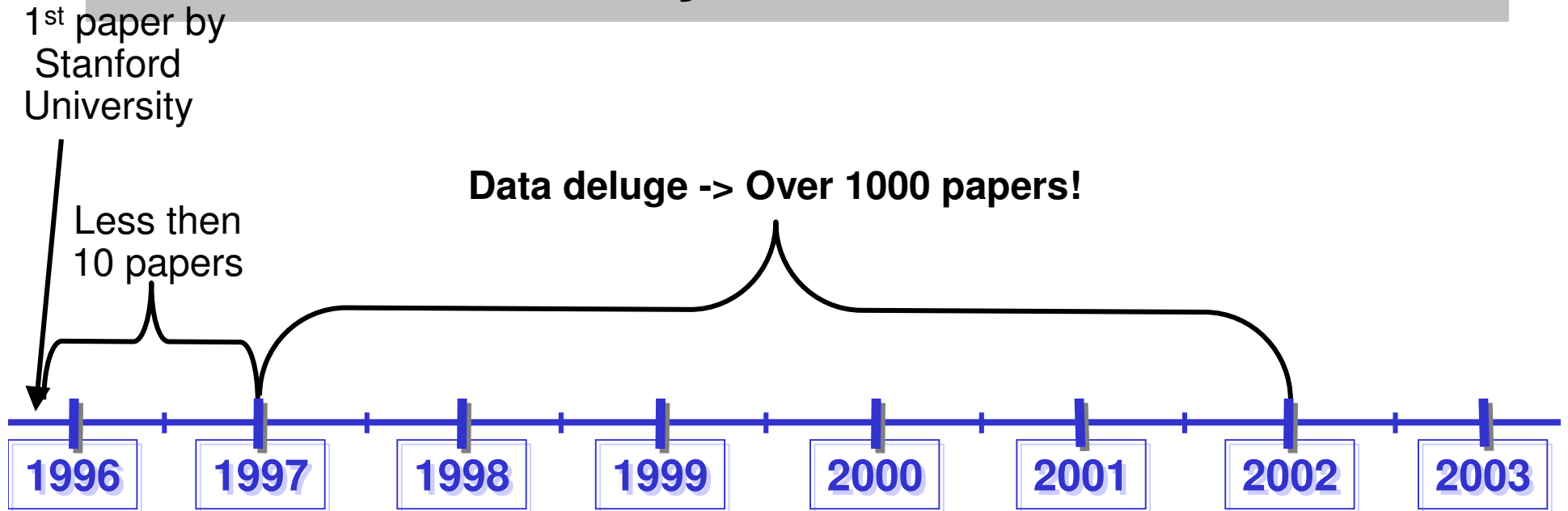
Mensa ponderaria (Pompei)



Why are standards so important ?



Microarray data.....inflation



1st paper by
Stanford
University

Less than
10 papers

Data deluge -> Over 1000 papers!

1996

1997

1998

1999

2000

2001

2002

2003

Alvis Brazma @ EBI
Industry Programme
'Promoter prediction
from co-expressed
genes'

- Large datasets
- Different platforms, surface types
- Different data formats
- Different level of details
- No infrastructure, no public database!

Storing DNA-Microarray generated information

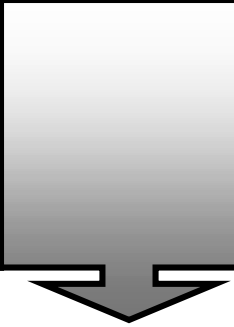
Key questions....

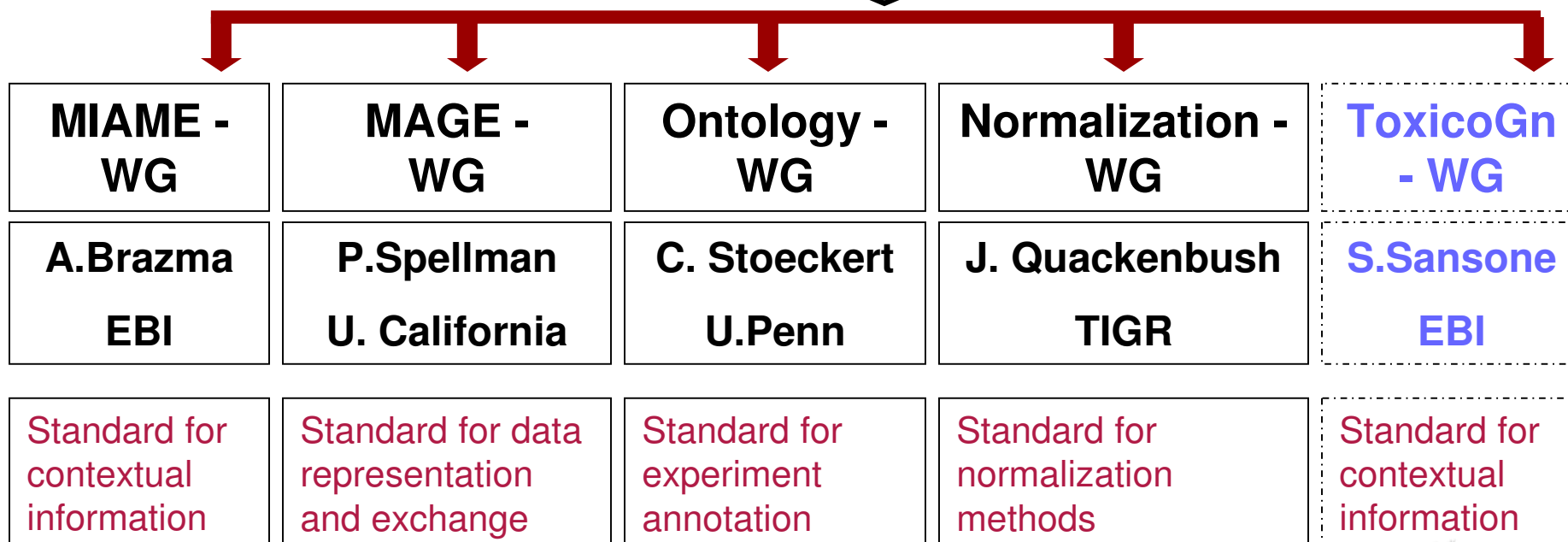
- **What should be the information to be recorded?**
- **How to make sets of information comparable?**
- **How to record the information ?**
- **Where to store the information ?**
- **How to retrieve and display information?**



To address these issues....The MGED response

MGED = Microarray Gene Expression Data Society

<u>EBI + Academics, e.g.</u>			<u>Companies e.g.</u>
TIGR	Stanford University.		Affymetrix
NCBI	Sanger Centre		Agilent / Rosetta
U Penn.	University of California		Iobion



Standard 1: What to store ?

The MIAME Requirements

Defining the Minimum Information About a Microarray Experiment

- **First major achievement of MGED** (*Nat.Genet. 2001, Dec;29(4):365-71*):
- **Defining the critical domains of a microarray experiment requiring sufficient annotation to be provided**
- **Set of guidelines issued for the microarray community**
- **Nailing down the rationales for accurate recording**
- **Providing a framework to start from for establishing public repositories**
- **Insisting on a need for infrastructure for data sharing**



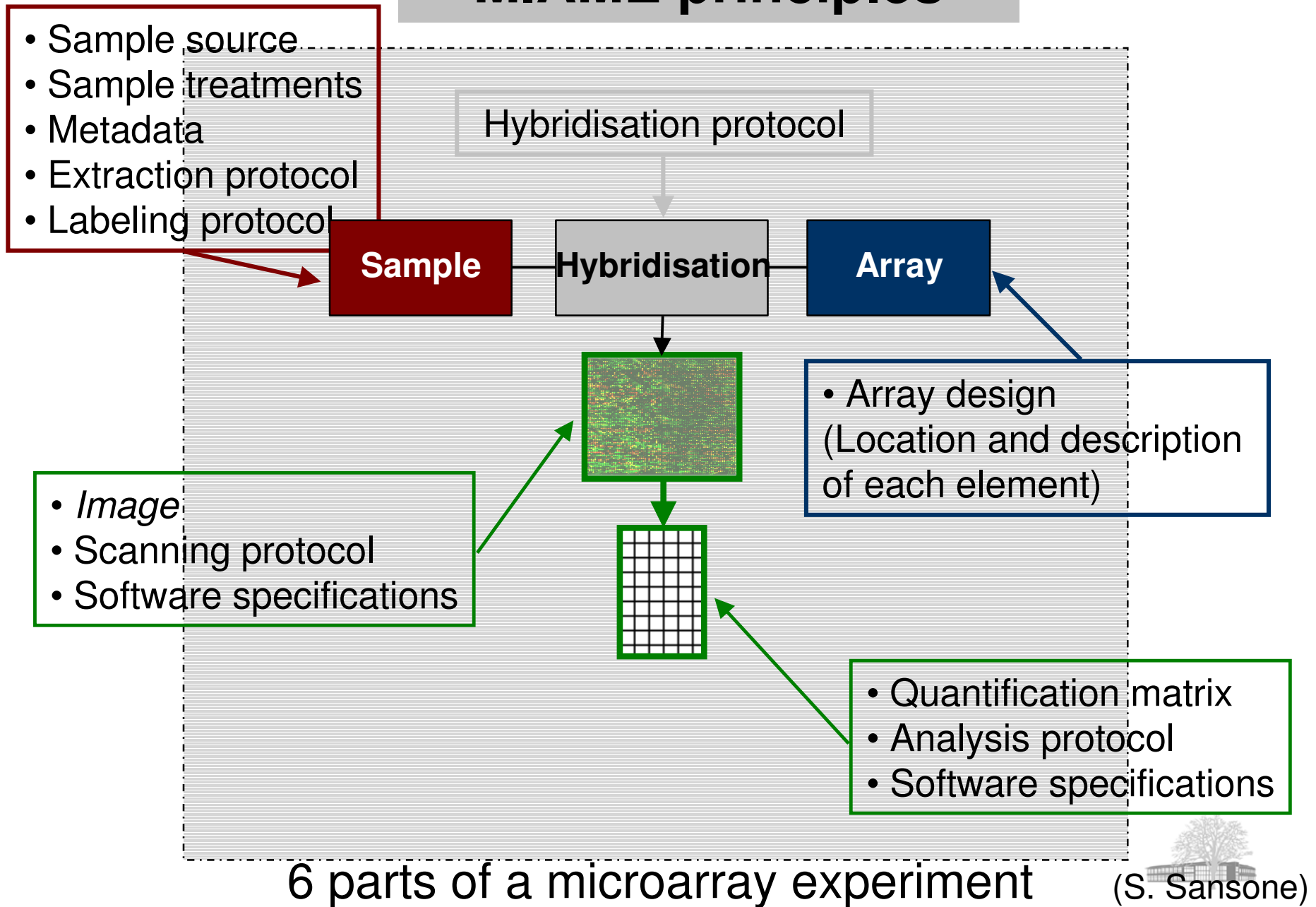
What does **MIAME** say ?

- **Sufficient information must be recorded to :**
 - Correctly interpret and verify the results
 - Replicate the experiments

- **Structured information must be recorded to:**
 - Query and correctly retrieve the data
 - Analyse the data



MIAME principles



MIAME principles

- Type
- Factor
- Quality control
- Publication

Experiment

Sample

Hybridisation

Array

- Strategy
- Algorithm
- Array control elements

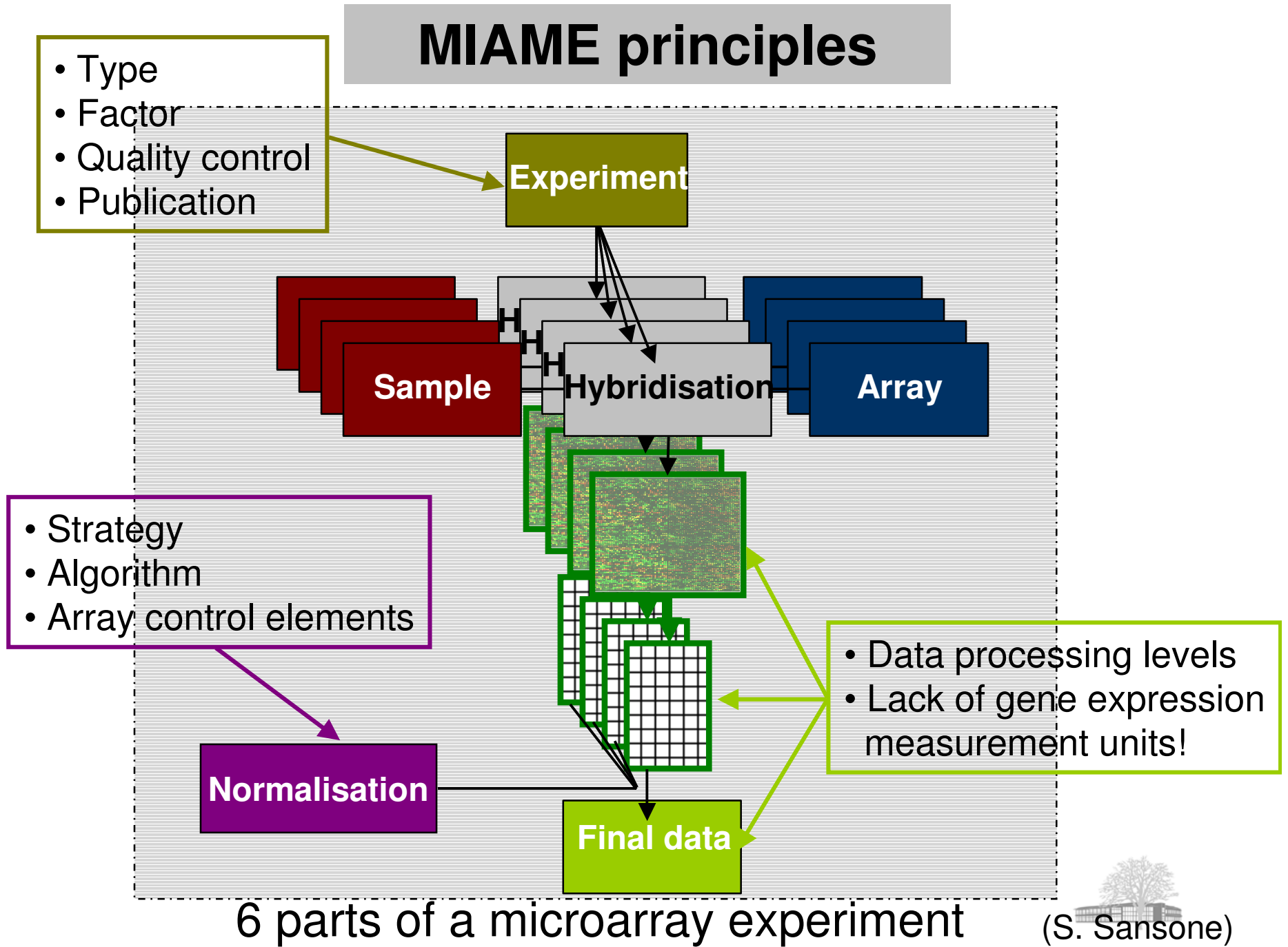
Normalisation

- Data processing levels
- Lack of gene expression measurement units!

Final data

6 parts of a microarray experiment

(S. Sansone)



Standard 2: How to efficiently annotate data ?

=>The MGED ontology (MO)

- **MIAME publication evokes the need for efficient annotation**
 - *Reducing use of synonyms or ambiguous terms*
 - *Avoiding free text description*
 - *Use of annotation good practice and use of structured controlled vocabularies and available ontologies*
- **Definition: An Ontology is a ...**
 - *Domain specific dictionary capturing semantic relationship between terms*



Standard 2: How to efficiently annotate data ?

=> **The MGED ontology (MO)**

- **Effort coordinated by C.Stoeckert & H. Parkinson**

- *Publication:Nature.Genet.2002, 32,pp 469-473*
- *Placed under the GOBO umbrella, a stable version has now been released*
- *MO is registered as GO Xref and at MOBY*
- *Opensource, modular and interconnected to existing ontologies*
 - ***Avoiding competing or redundant work***
 - ***Join and participate basis***
- *A set of rules on how to use it have been presented during MGED6 meeting*

MGED ontology available from: <http://mged.sourceforge.net/>



Standard 3: How to store microarray data ?

=> The MAGE-OM

- **Second Milestone delivered by the MGED society**
- **Effort coordinated by Paul Spellman**
- **Model developed by Ugis Sarkans at the EBI**
- **Joined submission by MGED / Rosetta (Michael Miller)**
- **Officially approved by the OMG in October 2002**
- **Model now « frozen » for 2 years**



MAGE-OM: an overview

- **MAGE-OM a formalized representation of the world of microarray independent of**

- *Experimental platform*
- *Image analysis method*
- *Normalization method*

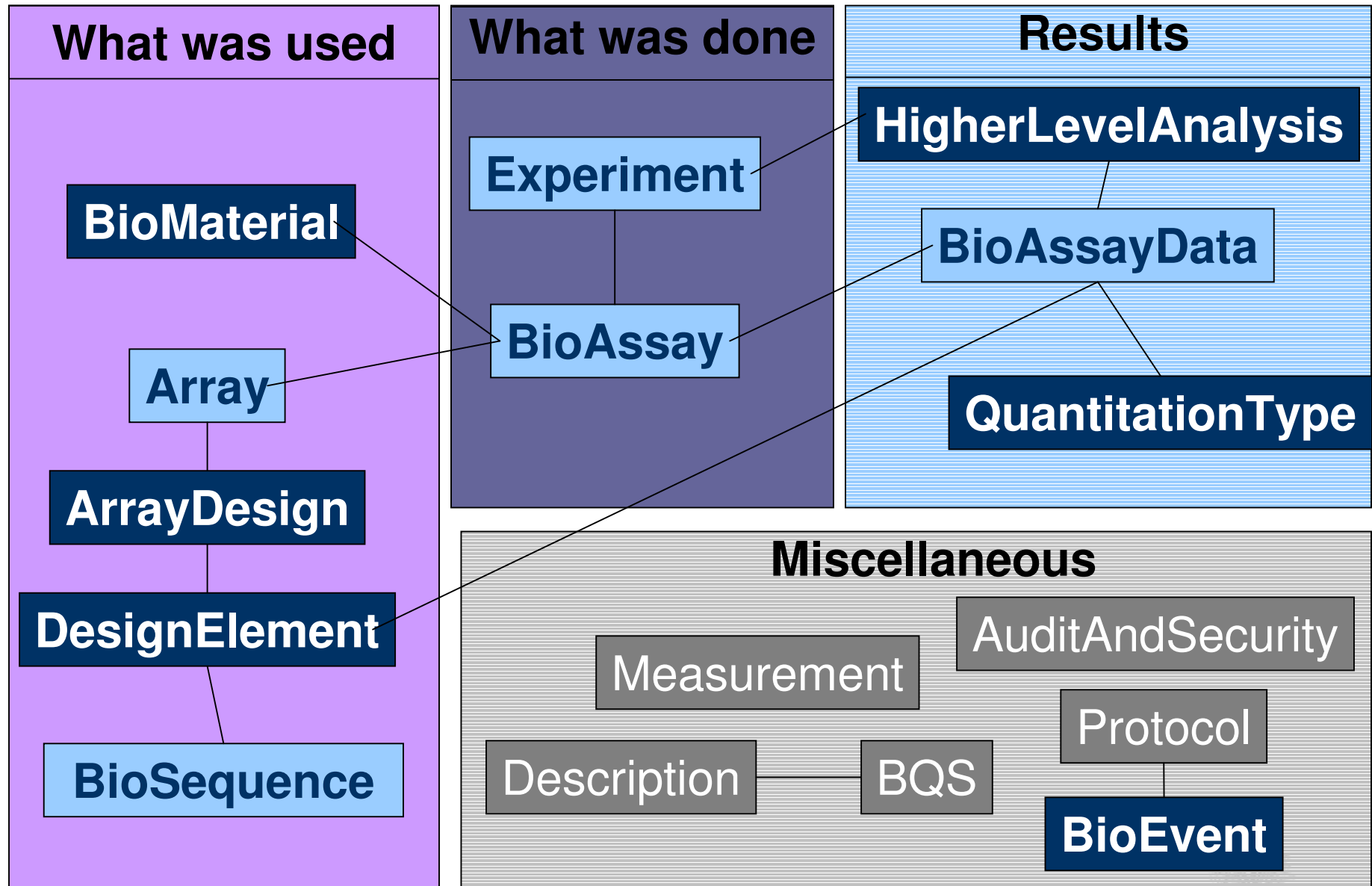
- **MAGE-OM contains 16 main groups (or packages)**

- **In the UML formalization:**

- *Objects are described by Classes*
 - *Related classes are grouped together into packages*



MAGE Object Model made simple



(U. Sarkans)

MAGE Object Model: zooming in...



Standard 4:How to exchange microarray data?

=>The MAGE-ML format

■ What is MAGE-ML ? A byproduct of the MAGE-OM

- Publication: Genome Biol. 2002 Aug 23;3(9):RESEARCH0046.
- *Basically, it's an XML file:*
- *The corresponding DTD has been automatically generated from the MAGE Object Model*
- *it is (almost) human readable*

■ Why MAGE-ML ?

- *Technically, a broad range of tools already available to handle, parse XML*
- *Ease of use in object oriented programming software development environment*



Standard 4:How to exchange microarray data?

=>**The MAGE-ML format**

- Oct 2002: Press release by Rosetta/Agilent:
 - Resolver 6.0 now accepts MAGE-ML format
- NCI's Director Challenge adopts MAGE-ML as internal data exchange format (Ken Buetow, SOFG meeting 2002)
- BASE Lund University LIMS works on MAGE-ML export
- NHIES & NCT: adoption of MAGE-ML for toxicogenomics data



Standard 5: How to make data comparable ?

=> **Normalization and data transformation**

- Preliminary studies showed limitation in data comparison
- Need for an independant MGED Working Group
- Effort coordinated by J. Quakenbush

(Review in Nature.Genet.2002, 32,pp 496-501)

- *Taking into account within experiment hybridization variability to allow for quantitation level comparison*
- *Defining the comparison procedures and algorithm*
- *Defining the necessary control element types, amount of replications*
- *Issuing recommandations on data comparison, data transformation and potentially on experimental design*



MGED standards -Summary

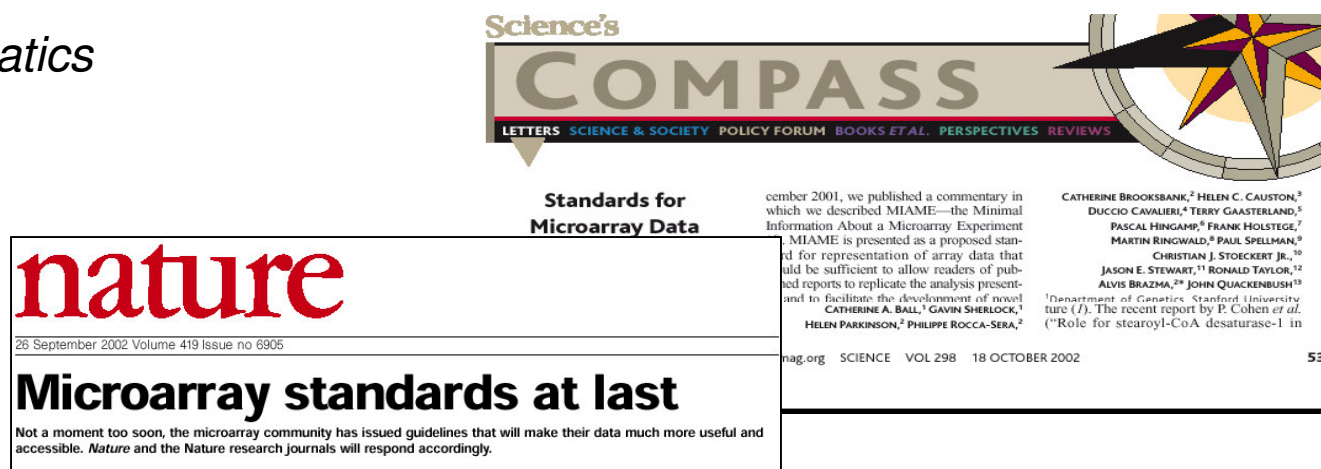
- Standard for quantitative information
 - **MIAME**
- Standard for qualitative annotation
 - **MGED Ontology**
- Standard for data representation/exchange
 - **MAGE-OM** & derived **MAGE-ML language**
- Standard for recording controls, normalization methods



Enforcing standards:

▪ December 2002: A New Publication from the MGED society

- *Nature and related Journals of the Nature Publishing Group*
- *Bioinformatics*
- *Lancet*
- *Science*

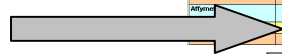
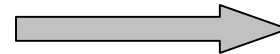
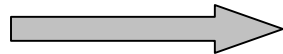


▪ Guide to authors, reviewers and editors of microarray gene expression papers.

▪ More journals ask for accession number as prerequisite for submission !



In the (MGED) standards we trust...

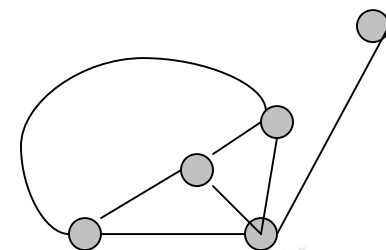


MAGE classes	Standard Quantitation Type	Measured Signal	Derived Signal	Specialized Quantitation Type	Standard Quantitation Type	Measured Signal	Derived Signal	Specialized Quantitation Type	Derived Signal
ArrayExpress Terms	SIGNAL_AREA	SIGNAL_RAW	SIGNAL_MEAN	SIGNAL_ID	SIGNAL_AREA	SIGNAL_RAW	SIGNAL_MEAN	SIGNAL_ID	SIGNAL_MEAN
GO Definition	SIGNAL_AREA	SIGNAL_RAW	SIGNAL_MEAN	SIGNAL_ID	SIGNAL_AREA	SIGNAL_RAW	SIGNAL_MEAN	SIGNAL_ID	SIGNAL_MEAN
Allen GenePhi	Yes	F1 Total Intensity	F1 Total Intensity	F1 Total Intensity	Yes	F1 Total Intensity	F1 Total Intensity	F1 Total Intensity	F1 Total Intensity
Reduction Images	Area	Total_sq	Mean_sq	Std_sq	Background Area	Background Area	Mean_sq	Std_sq	Std_sq
Agilent	ghumPa rHumPa	ghumPa rHumPa	ghumPa rHumPa	ghumPa rHumPa	ghumPa rHumPa	ghumPa rHumPa	ghumPa rHumPa	ghumPa rHumPa	ghumPa rHumPa
Imaging Research ArrayVision	Area	Density	Median Density	Std. Deviation	Area	Density	Median Density	Std. Deviation	Std. Deviation
Affymetrix	Intensity	CELIntensity	CELIntensity	CELIntensity	Intensity	CELIntensity	CELIntensity	CELIntensity	CELIntensity
	SGR_1_Std	SGR_1_Mean	SGR_1_Std	SGR_1_Std	SGR_1_Std	SGR_1_Mean	SGR_1_Std	SGR_1_Std	SGR_1_Std

MGED standards

- 1-MIAME
- 2-MAGE-OM
- 3-MAGE-ML
- 4-MO

Tumor classifier



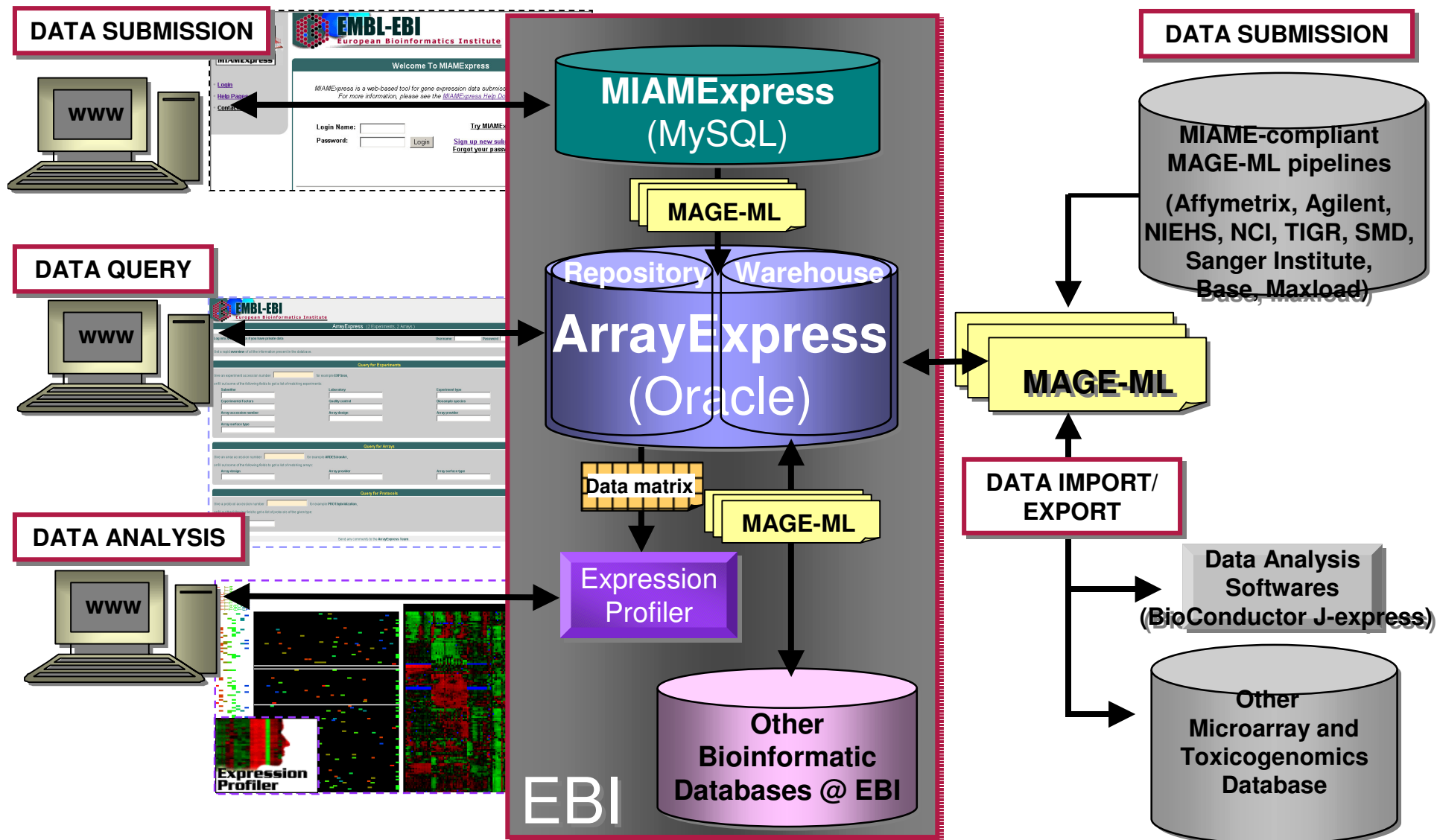
Regulatory networks

Talk structure

- Standardization efforts
 - Microarray standards by the MGED Society
- Infrastructure at the EBI
 - ArrayExpress Functional implementation of MGED standards
 - ExpressionProfiler
- Future and Development
 - Infrastructure
 - standards



Infrastructure @ EBI



ArrayExpress: Public Repository Infrastructure

- MAGE-OM Implementation in Oracle 9i
- 132 SQL tables: oracle/SQL scripts available at:

www.ebi.ac.uk/microarray/ArrayExpress/Implementation/implementation.html

- DB tools : (available from the same URL)
 - Validator: checking datafiles against our implementation of the OM
 - Loader/Unloader: input is a MAGE-ML file
 - Query interface: data access and retrieval
(Java servlets on Tomcat server)



Submission to ArrayExpress (part 1)

■ Direct submission as MAGE-ML documents:

Creation of pipelines from local /private databases/LIMS:

• **Commercial resources:**

- Affymetrix GDAC-Exporter SDK.
- Oracle 9i XDK

• **Open source resources:**

- MAGE-Stk Api (Perl and Java)

**EBI MAGE Validator/Loader
distributed**

Valid MAGE-ML file



Submission to ArrayExpress (part 2)

■ Web based submission using MIAMExpress:

- URL: www.ebi.ac.uk/miamexpress/
- *Based on MIAME questionnaire*
- *Built-in Controlled Vocabulary*
- *Opensource Perl-CGI/mysql/Apache (GNU license)*
- *Usable as a Lab Notebook for daily data storage*
- *Automatically generates valid MAGE-ML documents*



MIAMExpress Submission (part 1)

• MIAMExpress official release 15/12/2002:

- *Already 26 successful experiment submissions released*

Technology Type - Microsoft Internet Explorer

Technology Type

A drop down list of method is provided. Select one or more of these to describe your method used for generating the probes and spotted onto the array. If your type is not on the list, please specify a description in the text box provided. We may contact you for further information. For further information on the technology types, see this [MGED Ontology link](#).

Close Help Window

design submissions

Array design

EBI [Homo sapiens] 151

version 1.0

spotted ds DNA features

unknown

in situ oligo features

spotted antibody features

spotted colony features

spotted ds DNA features

spotted protein features

spotted ss oligo features

other

If other, please specify:

If other, please specify:

If other, please specify:

If other, please specify:

class TechnologyType

namespace: <http://mged.sourceforge.net/ontologies/MGEDOntology.daml/>

documentation: [The technology type or platform of the reporters on the array.](#)

type: primitive

superclasses: [ArrayDesignPackage](#)

used in classes: [FeatureGroup](#)

used in individuals: [in situ oligo features](#), [spotted antibody features](#), [spotted colony features](#), [spotted ds DNA features](#), [spotted protein features](#), [spotted ss oligo features](#)

individual spotted_ds_DNA_features

namespace: <http://mged.sourceforge.net/ontologies/MGEDOntology.daml/>

documentation: [A descriptor for the TechnologyType for a group of features where double stranded DNA is spotted on the array e.g. a PCR of a cDNA clone.](#)

instance of: [TechnologyType](#)

Stored files for this array:

ADF-Human.gal

Strandedness type:

Array description:

Array manufacturing protocol:

Equipment:

1. Lucidea microarray spotter (Amersham Biosciences UK).
2. 384 microtiter plates (Amersham Biosciences UK).

Array description file (ADF):

Browse... Upload Remove

Date for public release:


20 May 2003

Submit



MIAMExpress Submission (part 2)

• User assistance and annotation standardization: EnsMART + ADF converter

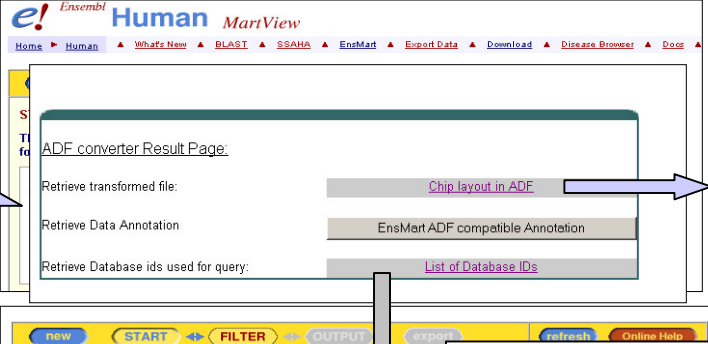


Welcome to ADF converter

Select species:

Upload file: (max: 2.5 Mb)

Select identifier type and rank in file:



ADF converter Result Page:

Retrieve transformed file:

Retrieve Data Annotation:

Retrieve Database ids used for query:

FILTER

Further refine your search or click 'next':

Converted file

	A	B	C	D	E
1	MetaColumn	Metarow	column	row	Genbank
2	1	1	1	1	1 null
3	1	1	1	1	2 U44378
4	1	1	1	1	3 U43746
5	1	1	1	1	4 X76132
6	1	1	1	1	5 M92424
7	1	1	1	1	6 L11353
8	1	1	1	1	7 X74594
9	1	1	1	2	1 M14694
10	1	1	1	2	2 M15400
11	1	1	1	2	3 X51630
12	1	1	1	2	4 U92436

Genomic Annotation file

	A	B	C	D	E	F	G	H	I	J	K	L	M
Reporter Name	Reporter BioSequence DatabaseEntry [ens_gene_id]	Reporter BioSequence DatabaseEntry [ens_trscpt_id]	Reporter BioSequence DatabaseEntry [embl]	Reporter BioSequence DatabaseEntry [refseq]	Reporter BioSequence DatabaseEntry [locus]	Reporter BioSequence DatabaseEntry [swall]	Reporter BioSequence DatabaseEntry [swissprot]	Reporter BioSequence DatabaseEntry [pdb]	Reporter BioSequence DatabaseEntry [interpro]	Reporter BioSequence DatabaseEntry [ens_fam_id]	Reporter BioSequence DatabaseEntry [go]	Reporter BioSequence DatabaseEntry [omim]	


Nucleic Acid Component					Protein Component					Functional Component			
Ensembl Gene ids					Swall ids					GO ids			
Ensembl transcript ids					Swissprot ids					GO descriptors + EvC			
Embl-GenBank DDBJ ids					PDB ids					Phenotype Component			
NCBI RefSeq ids					Interpro ids					OMIM id			
Locus Link ids					Ensembl Protein Family								

ArrayExpress – Access to Data

- **ArrayExpress Query interface:**

URL: www.ebi.ac.uk/arrayexpress/query/entry

- **Providing private and public access to data : close cooperation with journals and editors**



The screenshot shows the EMBL-EBI ArrayExpress homepage. The header includes the EMBL-EBI logo and navigation links: EBI Home, About EBI, Research, Services, Toolbox, Databases, Downloads, and Submissions. The main content area is titled "ArrayExpress at the EBI" and describes it as a public repository for microarray data. It includes a "Current Content Overview" table with the following data:

Category	Count	Action
Experiments	56	View
Arrays	81	View
Protocols	393	View
Hybridizations	1341	

There are also links for "Browse Database", "Query Database", "Login To Database", "Submissions", "Help & Documentation", "Microarray Standards", "Schema", "Implementation", and "EBI Microarray Home". A "Latest News" section mentions the "New MIAMExpress Release 1.5" and "Mapping the MAGE-OM to data within the Stanford Microarray Database". A "Today's release" section mentions "E-MEXP-24, A-MEXP-26 by C. Whitfield, A.M. Cziko and G. Robinson, Entomology Department, University of Illinois, Urbana-Champaign".



The screenshot shows the ArrayExpress query interface. The header includes the EMBL-EBI logo and navigation links. The main content area is titled "ArrayExpress" and shows a search results page. The search criteria are "You are logged in as guest" and "with species = Homo sapiens". The results show 8 matches. The first result is for Experiment E-MANP-1, submitted by Muckenthaler, with a description of HeLa cells grown to subconfluent density and CaCo-2 cells grown to high density. The second result is for Experiment E-MEXP-1, submitted by Zenke, with a description of Dendritic Cell differentiation. The interface includes links for "See details", "Providers", "Retrieve data", "Biosources used", and "Experimental protocols".

ArrayExpress Query interface: Access to Data

ArrayExpress

Help

ArrayExpress

Help

1 / 1

Experiment : E-SNGR-3

Author : Mata

Lab : The Wellcome Trust Sanger Institute

Type : time course

Fission yeast cells undergo meiosis and sporulation under conditions of nutritional stress, most frequently nitrogen starvation. This is a complex developmental process, which results in the formation of four spores that are

ArrayExpress

Export data »

BioAssays

☒ DerivedBioAssay[170142] jm_11h_131-15 DBA:jm_11h_131-15
☒ DerivedBioAssay[170143] jm_12h_135-12 DBA:jm_12h_135-12
☒ DerivedBioAssay[170130] jm_0h_131-2 DBA:jm_0h_131-2
☒ DerivedBioAssay[170140] jm_9h_131-13 DBA:jm_9h_131-13

BioAssayData

☒ DerivedBioAssayData[170184] DATA:JM_0h_131-2
☒ DerivedBioAssayData[170186] DATA:JM_0h_135-24
☒ DerivedBioAssayData[170194] DATA:JM_1h_131-3
☒ DerivedBioAssayData[170196] DATA:JM_2h_131-4

QuantitationTypes

☐ F635 Median »
☐ F635 Mean »
☐ F635 SD »
☐ B635 Median »
☐ B635 Mean »
☐ B635 SD2 »
☐ % > B635+1SD2 »
☐ % > B635+2SD2 »
☐ F635 % Sat. »
☐ F532 Median »
☐ F532 Mean »
☐ F532 SD »
☐ B532 Median »
☐ B532 Mean »
☐ B532 SD2 »
☐ % > B532+1SD2 »
☐ % > B532+2SD2 »
☐ F532 % Sat. »
☒ Ratio of Medians »
☒ Ratio of Means »
☒ Median of Ratios »
☒ Mean of Ratios »


ArrayExpress

1.55 114.93 67.32 24.3
2.48 89.71 38.45 25.87
0.89 98.52 36.21 12.58
0.02 215.3 45.02 78.21
0.36 51.43 85.32 25.8

See data matrix »

Quick data analysis in Expression Profiler

Complete data upload to Expression Profiler



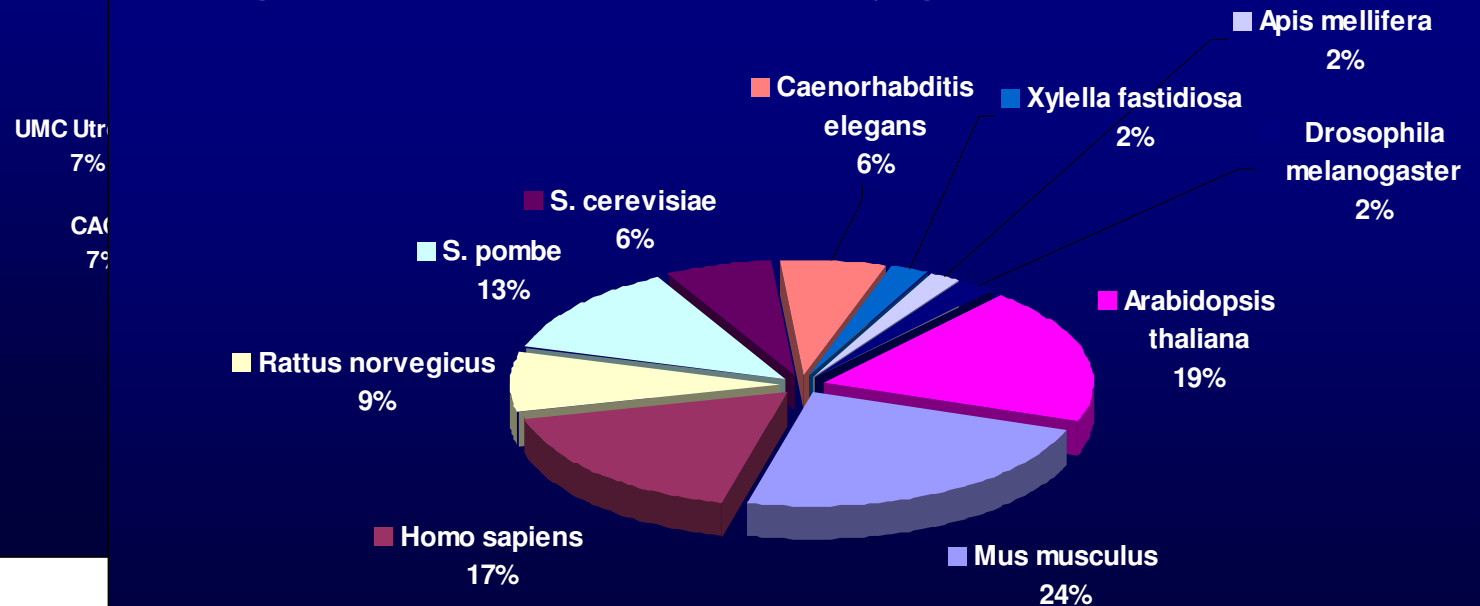
ArrayExpress DataSets

- In the Production Database: > 2700 hybridizations

Array Manufacturers

Experiment Submissions broken down by Pipelines

Experiment Submissions Broken Down by Species



Talk structure

- Standardization efforts @ EBI
 - Microarray standards by MGED Society
- Standard-supporting microarray infrastructure@EBI
- Future and Development
 - Infrastructure
 - standards



ArrayExpress's Future: Input wise (1)

- **New pipeline implementation with large centres**
 - *(e.g. MAGE-ML adopted by NCI for file exchange)*
- **Development of Quality Control rating procedures**
 - Creation of quality metrics to rate Experiment submission
 - Creation of MIAME validator
- **Development of appropriate curation tools**
 - *Automated submission tracking*
 - *Better integration with Ontology browsers*
 - *Improvement of the Validator to enhance CV checking*



ArrayExpress's Future: Input wise (2)

- **Online submission and MIAMExpress Mk II:**
 - *Scaled up to match the full MAGE model*
 - *Integration of MGED ontology in the interface*
 - *Extension according to the user needs & feedback*
 - *Increase of usability, flexibility and scalability*



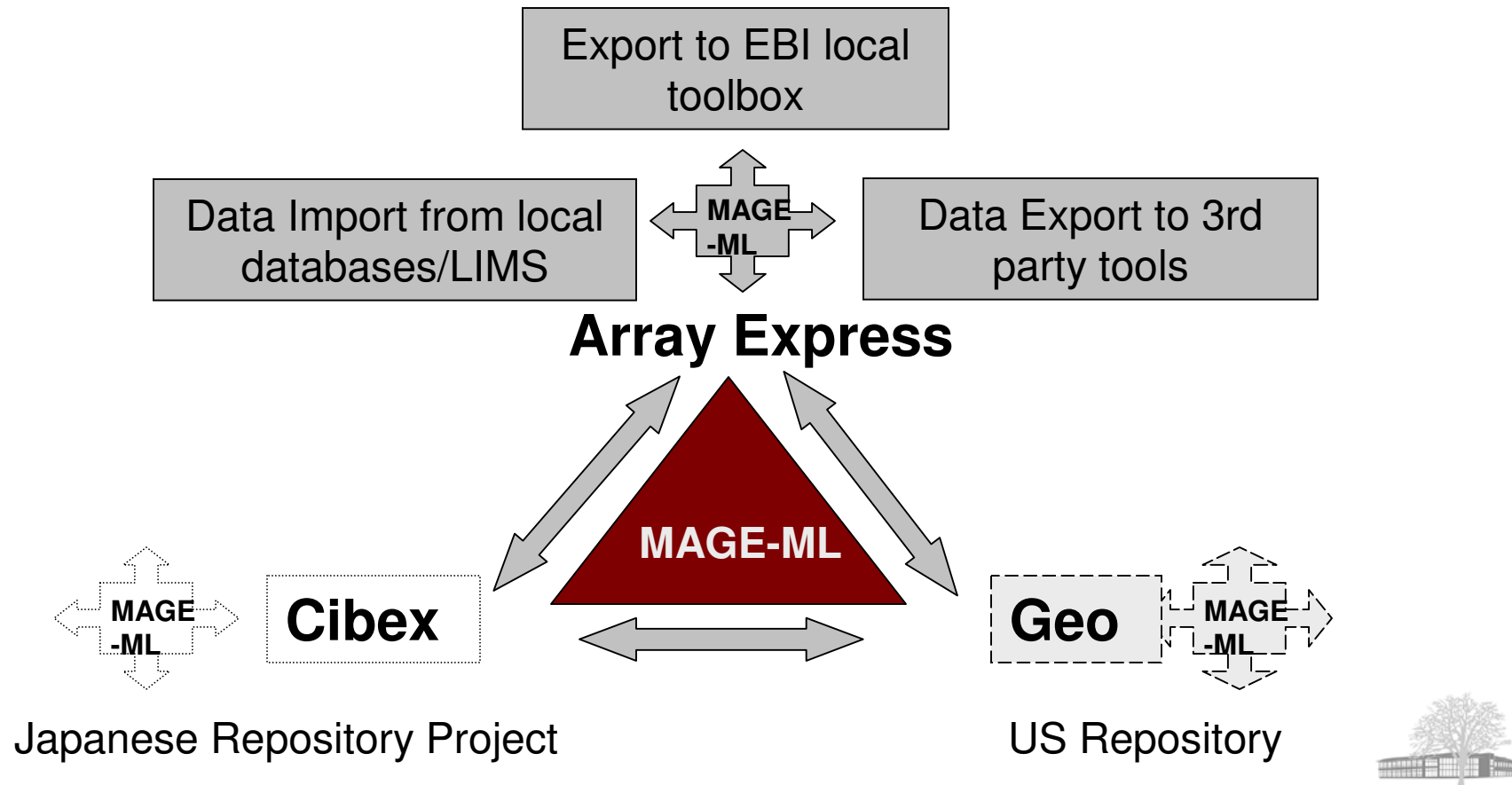
ArrayExpress's Future: Output wise

- **Development of graphical display interface**
 - *Creation the interface for displaying results in a meaningful way.*
- **Development of Gene-centric / Sequence-centric queries**
 - *Several hurdles: Generating a “consensual” Gene Index*
 - *Standardizing Array Annotation: re-annotating Arrays ?*
 - *=> unavoidable disclosure of sequences used in array*
 - *=> IP issues*
 - *Solution: follow Affymetrix example ! working along with Ensembl!*
- **Development of query storage/tracking tools**



ArrayExpress's Future: Output wise (2)

- a possible view of MAGE-ML based data exchange in action ...(if all repositories rely on standards)



ArrayExpress's Future:

- **Better integration with Expression Profiler**
- **Development of a Datawarehouse in a production scale** *(for the moment prototyping)*
- **New MAGE-OM** (U. Sarkans, P. Spellman)
 - *To ensure better description of Biological material used in experiments*
 - *Application to new emerging technologies in the field of high throughput biology (application to Proteomics)*
- **MIAME-Toxicogenomics guidelines:**
 - *Consequence of ILSI project and interest expressed by NHIES and FDA (sansone@ebi.ac.uk: Toxicogenomics coordinator @ EBI)*



Microarray Informatics team at EBI

Alvis Brazma - *group leader*

EU Temblor funding

ArrayExpress

- Gonzalo Garcia
- Ahmet Oezcimen
- Anjan Sharma
- Ugis Sarkans

MIAMExpress

- Mohammadreza Shojatalab
- Niran Abeygunawardena
- Sergio Contrino

Expression Profiler

- Misha Kapushesky
- Patrick Kemmeren
- Jaak Vilo

Toxicogenomics / Nutrigenomics effort

- Susanna Sansone
- Philippe Rocca-Serra
- Sergio Contrino

Research

- Thomas Schlitt
- Katja Kivinen
- Lev Soinov
- Anastasia Samsonova
- Aurora Torrente

Curation

- Helen Parkinson
- Philippe Rocca-Serra
- Ele Holloway
- Gaurab Mukherjee



Reminders

- **MIAME Checklist and standard enforcement policies**
 - Journals
 - Funding agencies
- **EBI:** www.ebi.ac.uk/microarray
 - Data storage resources
 - MAGE-OM and MAGE-ML dtd
 - ArrayExpress: SQL scripts + Validator + MAGE-ML files
 - MIAMExpress: sourceforge project ID49908
- **MGED:** www.mged.org
 - MGED Ontology GOBO project
 - Meetings:
 - MAGE jamboree meeting Hinxton 1-6 December 2003

