

Concepts for biocomputing in a multiuser environment

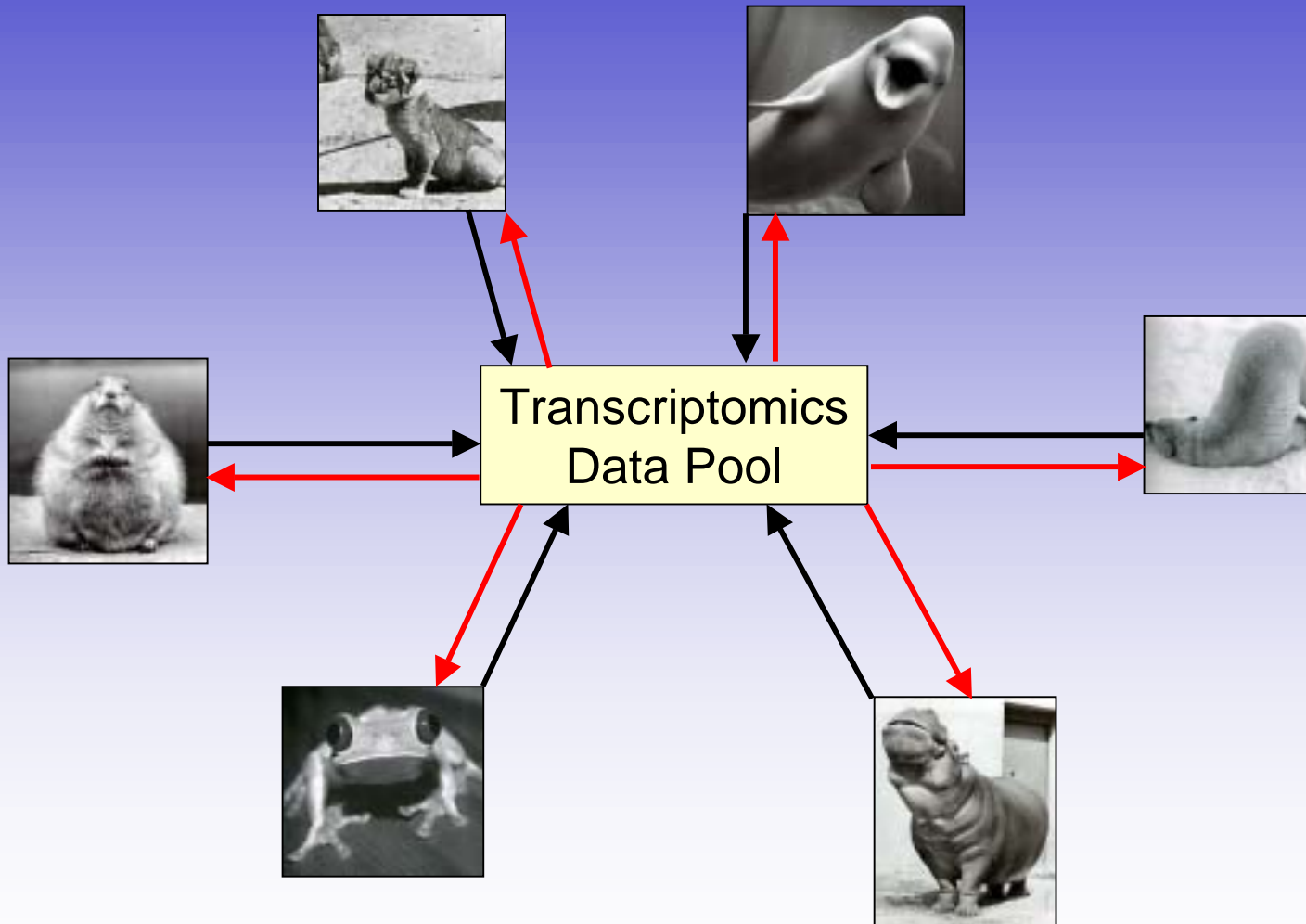
Ulrich Wagner

The Functional Genomics Center Zürich (FGCZ)



- „Joint venture“ of the ETH and the University of Zürich
- Inauguration in February 2002
- Braintrust and network for research and education
- Technology platform for transcriptomics, proteomics and bioinformatics (user lab)

The multiuser environment



Need for microarray data standards

- Large datasets
- Different platform types - nylon, glass
- Different technologies - oligos, cDNAs
- Gene expression data only make sense in the context of a detailed experiment description
 - Array annotation
 - Sample annotation
- Referencing to external db not stable
- **Data sharing** needs standardized way to annotate and record the information

Standard for microarray data – MGED Group

- **Microarray Gene Expression Data Group:**
World's largest microarray labs and companies
(EBI, Sanger, Stanford, TIGR, Universite D'Aix-Marseille II,
Affymetrics, Agilent, NCBI, DDBJ, etc.)
- **MGED Group aims to**
 - **Facilitate adoption of standards for:**
 - Experiment annotation
 - Data representation
 - **Introduce standard for:**
 - Experimental controls
 - Data normalization methods
 - **Ultimately facilitate efficient data sharing**

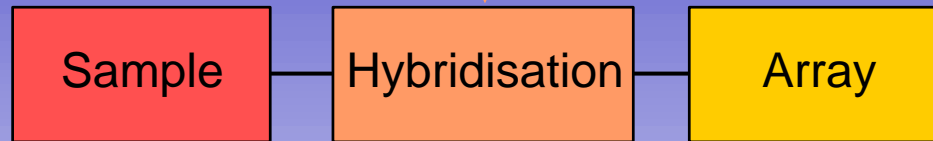
General MIAME principles

- **Minimum Information About a Microarray Experiment (MIAME)**
- **NOT a formal specification BUT a set of guidelines**
- **Sufficient information needed to:**
 - Correct interpretation and verification of results
 - Reproduction the experiments
- **Structured information allows:**
 - Query of data and correct retrieval
 - (Automated) data analysis/ data mining
- **Lit.: Brazma et al., *Nat. Gen.* (2001) 29: 365-371**

MIAME

- Sample source
- Sample treatments
- Extraction protocol
- Labelling protocol

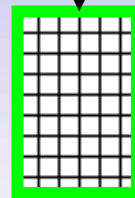
Hybridization protocol



Measurement

- Array design information
- Location of each element
- Properties of each element

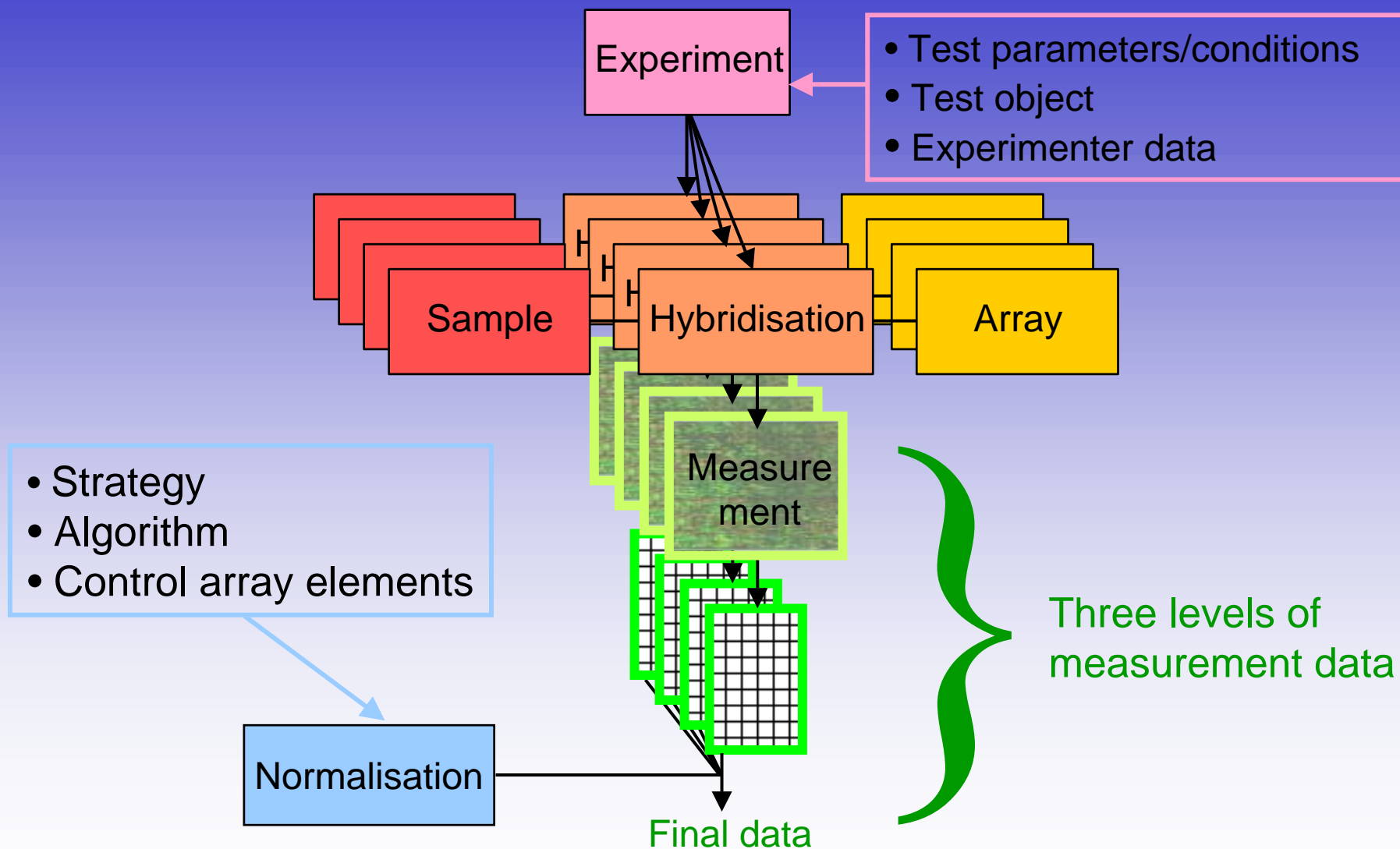
- Image (raw data)
- Scanning protocol
- Software specifications



- Quantification matrix
- Analysis protocol
- Software specifications

MIAME: Six parts of a microarray experiment

MIAME



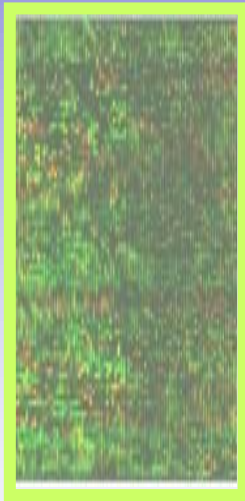
MIAME: Six parts of a microarray experiment

MIAME

Three processing levels of measurement data:

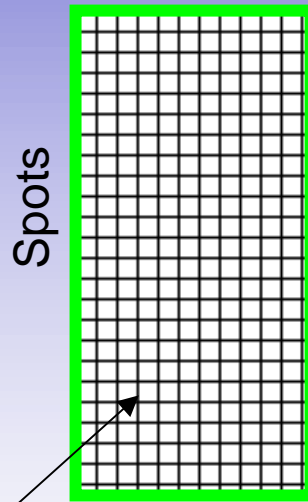
Raw data

Array scans



Intermediate data

Quantifications

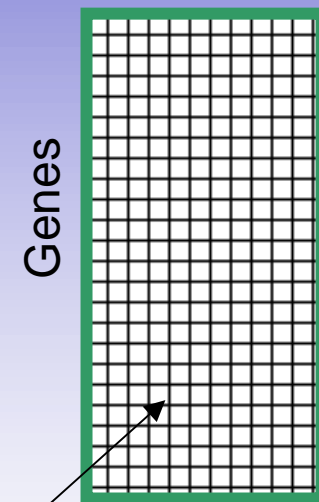


Spot quantification



Final data

Conditions



Gene expression levels

•Lack of gene expression measurement units !

MIAME – Annotation challenge

- **Annotation implementations are required**
 - Avoid/reduce free text descriptions
 - Use controlled vocabulary (CV)
 - Define and add sources for each term
 - Remove synonyms or use of synonym mappings
 - Data curation at source (LIMS)
 - Integration of controlled terms in query interfaces
- **Facilitate data queries-analysis**

Annotation – implementation issues

Bioinformatics has a communication problem :

- **Many researcher/ databases use their own labels and categories for storing data objects**
- **Some use identical labels and categories but with a different meaning.**
- **Conversely, one concept is often found under different names.**
- **Exemples : concepts of "gene" and "protein sequence" which are used with different semantics by major international genomic and protein databases thereby making database integration difficult and strenuous.**

Solution: Make use of controlled vocabulary or ontologies and avoid free text wherever possible ...

What is an ontology?

(In the computer scientific not in the philosophical sense)

- more than just nomenclature, more than classification, more than taxonomy, more than (controlled) vocabulary (CV), more than dictionary
- a collection of common terms, the meaning of the terms and the formal relations between the terms agreed by a group so that entites (people, computers, DB, ...) can communicate.
- removal of ambiguity, providing semantics and constraints
- allowing for computational inferences and reliable comparisons

MIAME- motivated ontology

- Under construction by MGED ontologists
- Incorporation of existing ontologies (e.g. GeneOntology, NCBI taxonomy database, TAIR etc.)
- Example: MGED BioMaterial Ontology



















MGED BioMaterial Ontology

External References

Instances

©-BioMaterialDescription

©-Biosource Property

- ©-Organism   **NCBI Taxonomy**  ***Mus musculus musculus* id: 39442**
- ©-Age   7 weeks after birth
- ©-DevelopmentStage   **Mouse Anatomical Dictionary**  **Stage 28**
- ©-Sex   Female
- ©-StrainOrLine   **International Committee on Standardized Genetic Nomenclature for Mice**  **C57BL/6**
- ©-BiosourceProvider   Charles River, Japan
- ©-OrganismPart   **Mouse Anatomical Dictionary**  **Liver**

©-BioMaterialManipulation






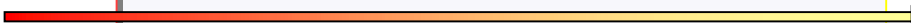

©-EnvironmentalHistory

©-CultureCondition

- ©-Temperature   22 ± 2°C
- ©-Humidity   55 ± 5%
- ©-Light   12 hours light/dark cycle
- ©-PathogenTests   Specified pathogen free conditions
- ©-Water   *ad libitum*
- ©-Nutrients   MF, Oriental Yeast, Tokyo, Japan

©-Treatment

©-CompoundBasedTreatment

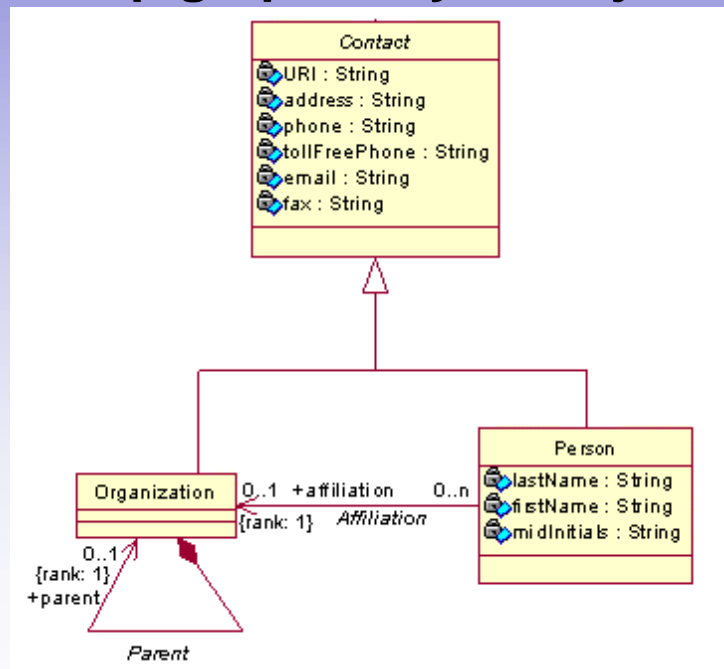
- (Compound)   **ChemIDplus**  **Fenofibrate, CAS 49562-28-9**
- (Treatment_application)   *in vivo*, oral gavage
- (Measurement)   100mg/kg body weight

Making use of the MIAME concept

- Development of MIAME-compliant databases or LIMS (e.g. ArrayExpress)
- Creation of submission/annotation tools for generating MIAME-compliant information (e.g. *MIAMExpress*)

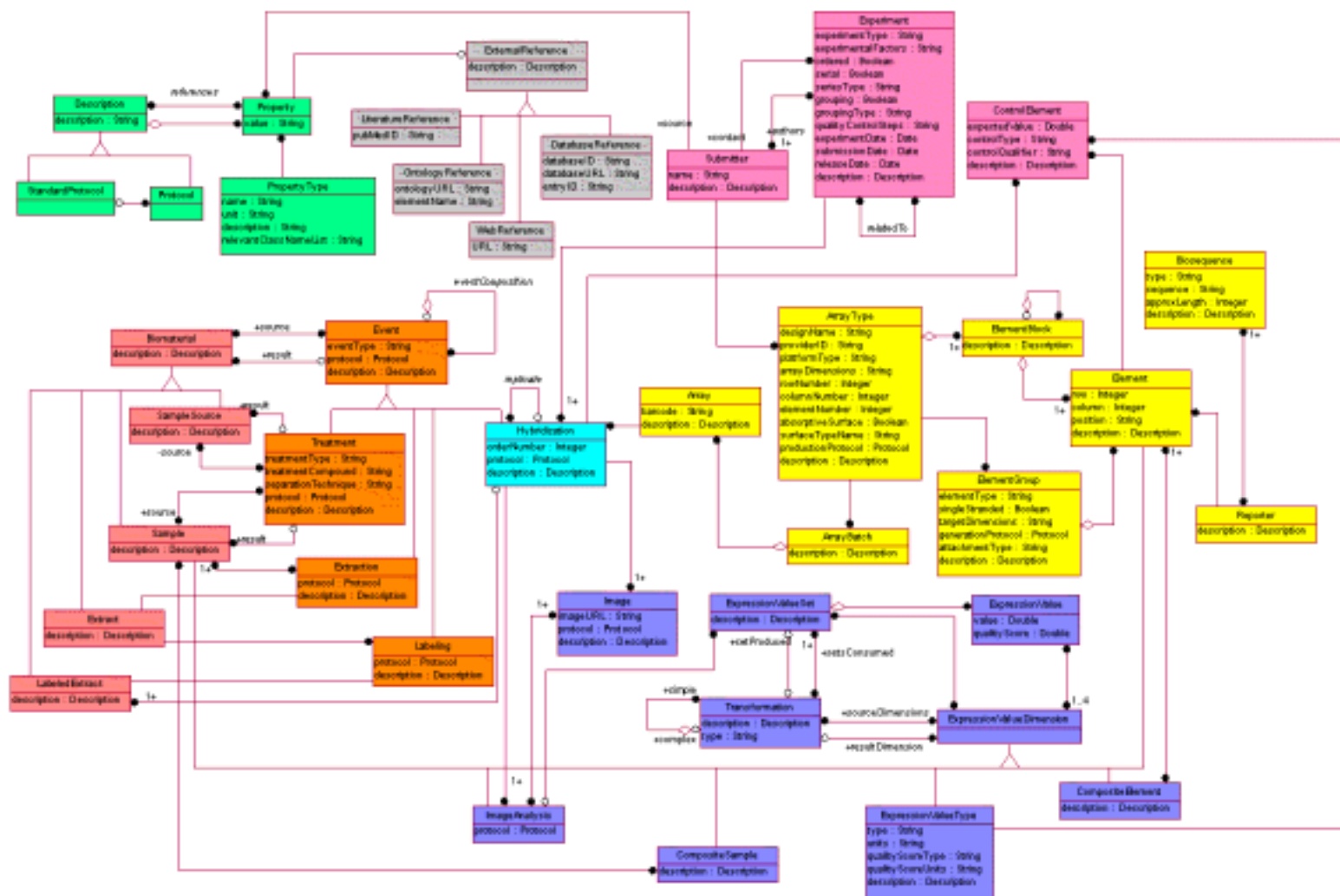
Making use of the MIAME concept

- MicroArray Gene Expression- Object Model (MAGE-OM) is a so-called Object Modelling Language
- MAGE-OM helps to represent MIAME-compliant the data and their relationship graphically as objects:



- Database scheme is derived from MAGE-OM

A complex example (conceptual model from *Arrayexpress*)



MAGE-ML as exchange data format

- Platform for moving data between data generators and shared databases
 - Support for all known types of microarray data
 - Flexibility
- Internal format to communicate data from databases to third part applications
- Development of a Microarra Gene ExpressionMarkup Language (MAGE-ML), which is derived from MAGE-OM and is close to XML

Example for MAGE-ML file

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE MAGE-ML SYSTEM "../MAGE-ML.dtd">
<MAGE-ML identifier="Protocol_test">
  <AuditAndSecurity_package>
    <Contact_assnlist>
      <Person
        identifier="Person:John_Smith"
        name="John_Smith"
        lastName="Smith"
        firstName="John"
        phone="(800) 555 4325 x 561"
        email="jsmith@imakearrays.com"
        URI="http://www.imakearrays.com/~jsmith">
        <Affiliation_assnref>
          <Organization_ref
            identifier="Organization:imakearrays"/>
          </Affiliation_assnref>
        </Person> .....
```

Conclusion

- **MIAME is an evolving concept of standards integrated into a wider model of gene expression data sharing**
- **The developers are anxious to integrate user's wishes and needs into new releases**
-> **fixing the standards is an issue for the whole scientific community**
- **Support by companies (Affymetrix, Agilent etc.) makes MIAME very likely to establish as a global standard**



Happy user !

Literature

- **Brazma et al., (2001)Minimum information about a microarray experiment (MIAME)- towards standards for microarray data. *Nat. Gen.* 29: 365-371**
- **<http://www.mged.org/Workgroups/MIAME/miame.html>**
- **http://www.mged.org/Workgroups/MIAME/miame_mage-om.html**
- **http://www.cbil.upenn.edu/Ontology/MGED_ontology.html**
- **www.ebi.ac.uk/microarray**

Latest news...



nature

26 September 2002 Volume 419 Issue no 6905

Microarray standards at last

Not a moment too soon, the microarray community has issued guidelines that will make their data much more useful and accessible. *Nature* and the Nature research journals will respond accordingly.

You read a paper with a fascinating conclusion about the expression of several genes. You decide to use some of the same experiments on your system of choice. But when you wade through hundreds of pages of supplementary information

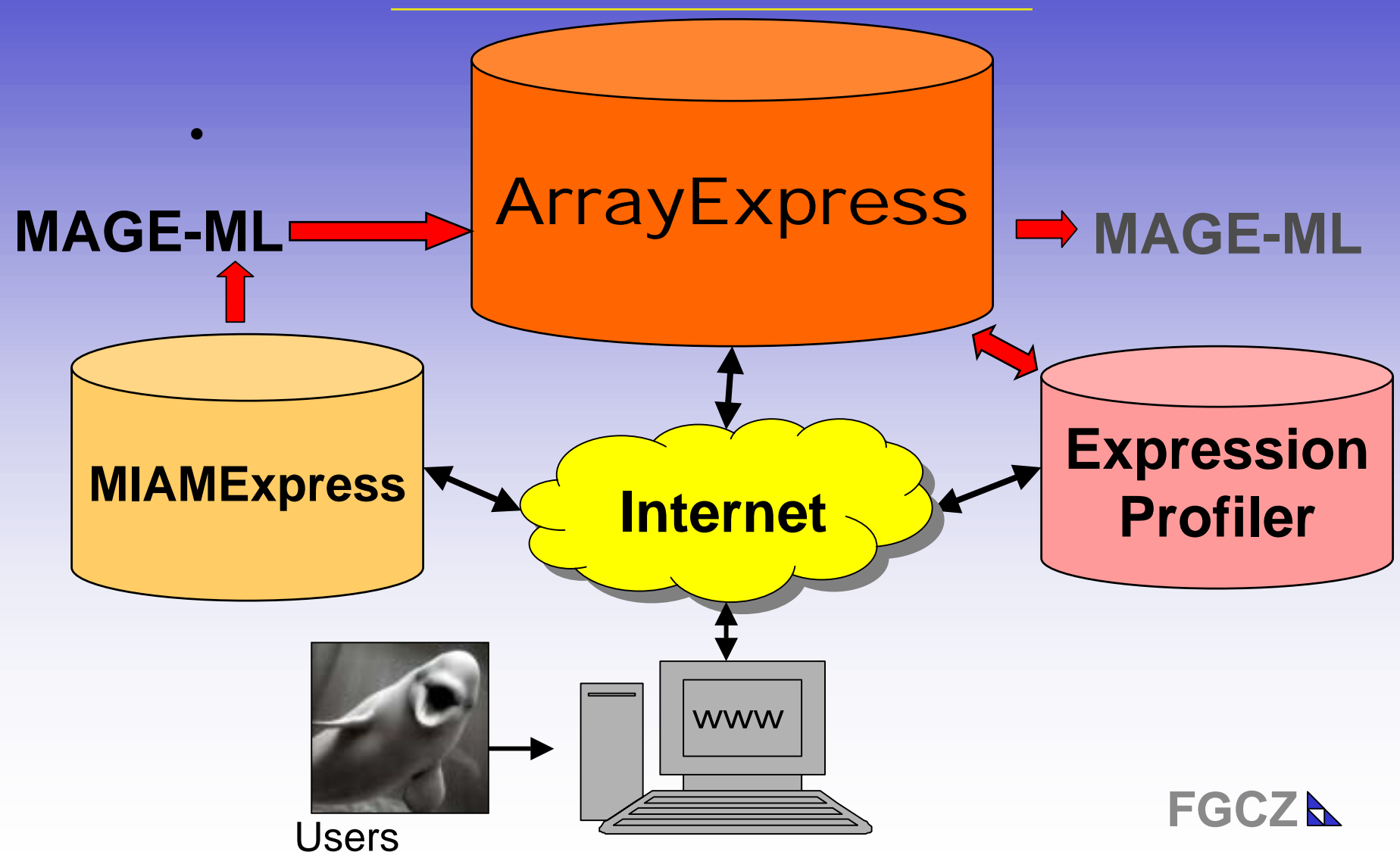
For authors, the proposal provides a checklist of variables that should be included in every microarray publication, at http://www.mged.org/Workgroups/MIAME/miame_checklist.html. This checklist, with all variables completed, would be supplied as supplementary

- From 1st of December on, new microarray experiments have to include MIAME- compliant data
- Data should be submitted to ArrayExpress or GEO database

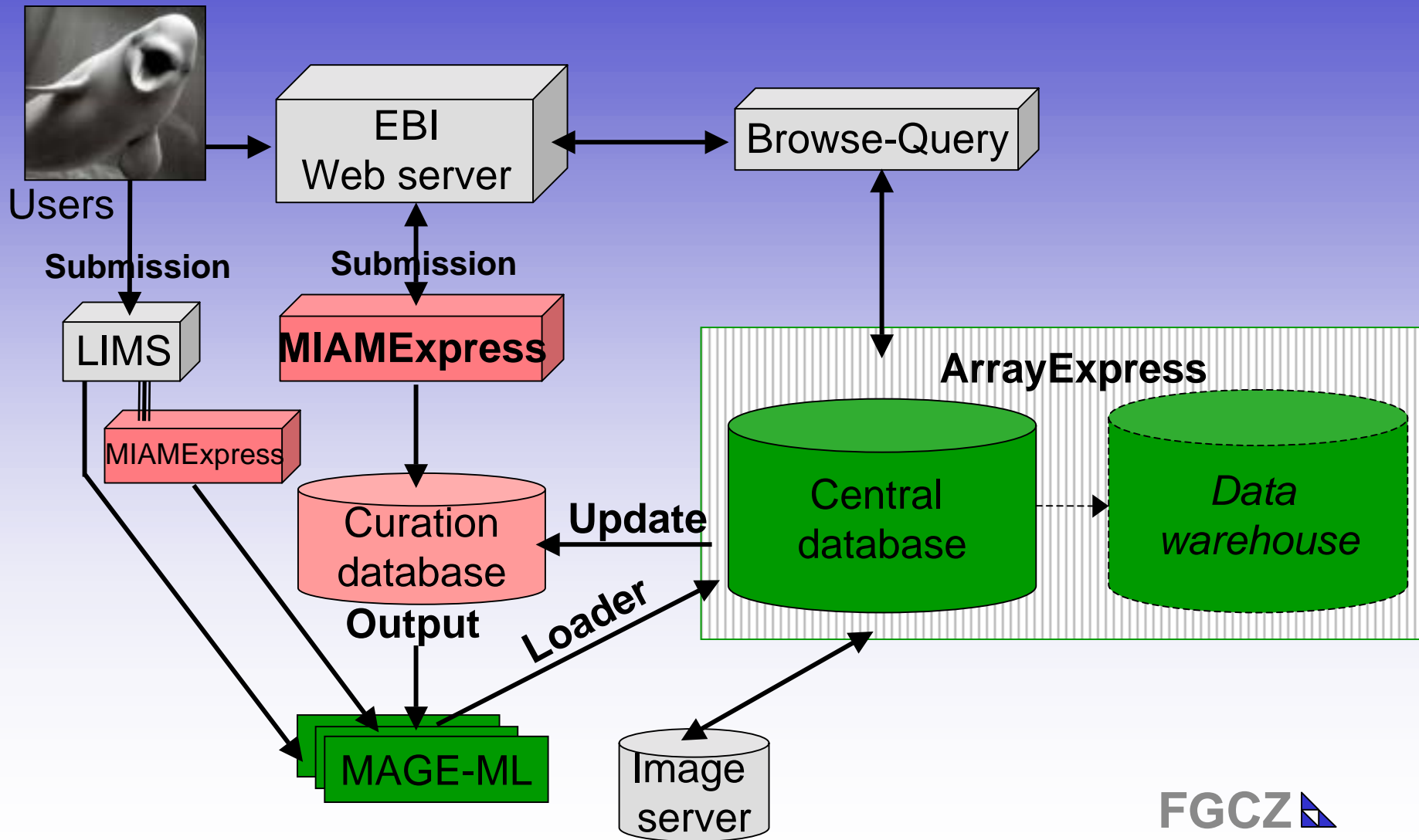




Data flow concept- *Arrayexpress*



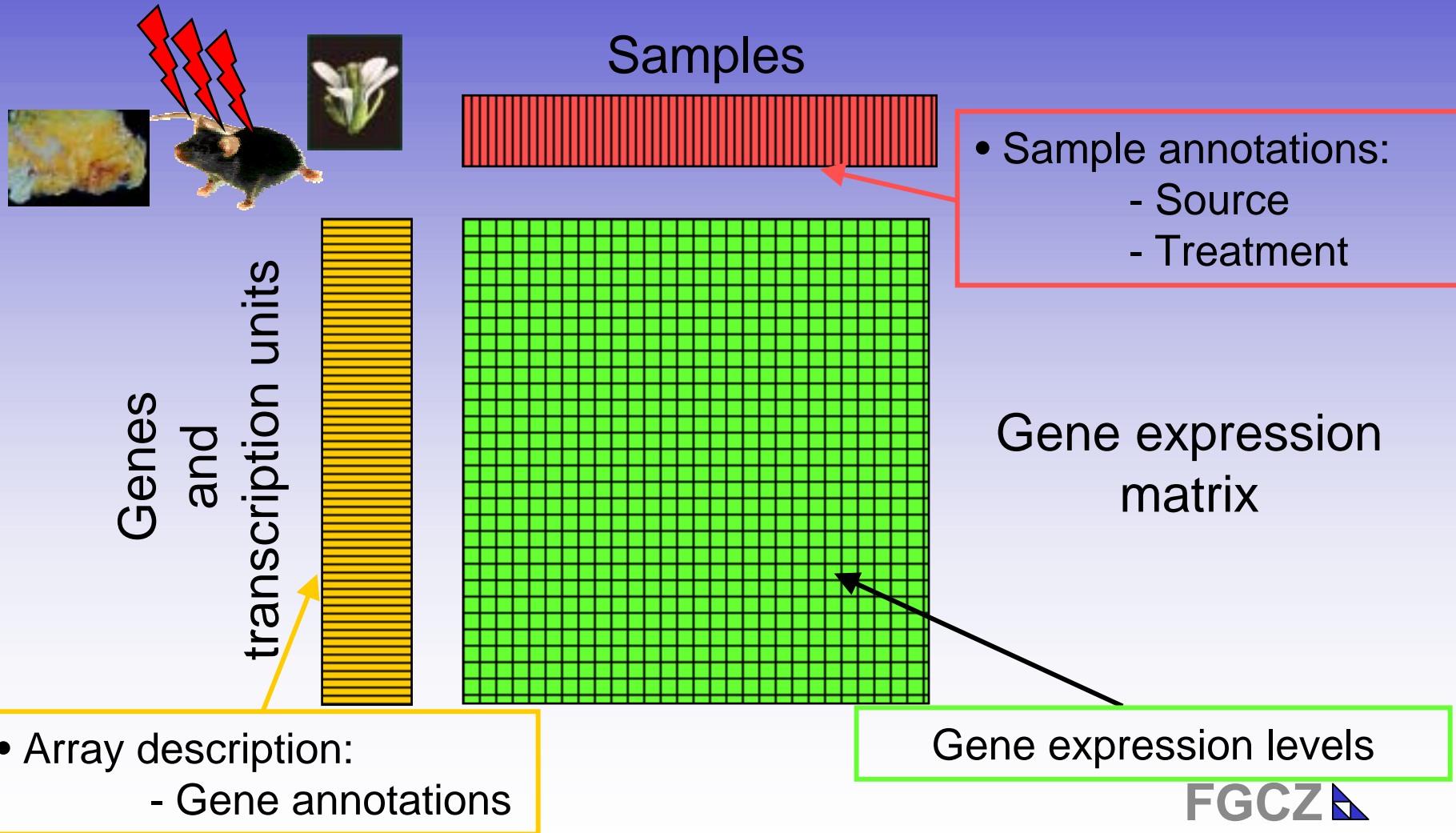
Data flow concept- *Arrayexpress*



History of MAGE-ML

- **Two independent efforts started to develop data storage formats for microarray data**
 - Rosetta and Collaborators (GEMML)
 - MGED (MAML)
- **Both groups submitted XML based proposals to the OMG in response to an RFP covering large scale gene expression**
- **Joint development process since January 2001**
 - Many meetings
 - One programming “Jamboree”, another planned

A gene expression database from the data analyst's point of view



MIAME - Gene annotation

- **Unambiguous identification**
- **Synonyms !**
 - Community approved names
 - Alternative to gene names
- **Usable external sources e.g.:**
 - EMBL-GenBank - sequence accession #
 - Jackson Lab - approved mouse gene names
 - HUGO - approved human gene names
 - GO categories - function, process, location

MIAME - Sample annotation

- **Gene expression data only have a meaning in the context of detailed sample descriptions !**
- **Usable external sources e.g.:**
 - NCBI Taxonomy - organisms
 - Jackson Lab - mouse strains names
 - Mouse Anatomical Dictionary – mouse anatomy
 - ChemID – compounds
 - ICD-9 – diseases classification
- **More is needed.....**