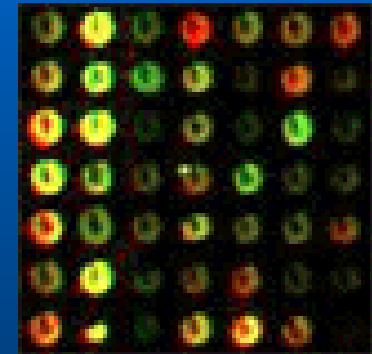


Infrastruttura computazionale per l'archiviazione e l'analisi dei dati da microarray

Silvia Giuliani

Andrew Emerson

Elda Rossi



Storage and analysis of micro-array data

- Currently most researchers use personal workstations and spreadsheet programs for storage and analysis



	A	B	C	D	E	F	G	H
1	1	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
2	2	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
3	3	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
4	4	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
5	5	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
6	6	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
7	7	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
8	8	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
9	9	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
10	10	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
11	11	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
12	12	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
13	13	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
14	14	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
15	15	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
16	16	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
17	17	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
18	18	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
19	19	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
20	20	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
21	21	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
22	22	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
23	23	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
24	24	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528
25	25	0.015834	0.005833	0.011528	0.005833	0.011528	0.005833	0.011528

Storage and analysis of micro-array data

- ❑ Latest chips now contain ~ 20,000-30,000 probes → beyond capabilities of desktop systems
- ❑ Requirement for more sophisticated infrastructure based on industry standard database systems (e.g Oracle) running on powerful (UNIX) servers.



Microarray data storage and analysis

- ❑ **Data Storage – raw image data, post-processing, experimental details → essential for comparing experiments**
- ❑ **Software tools – free + commercial**
- ❑ **Interpretation of results – e.g. text mining**

Archiviazione di esperimenti microarray

□ Storage

- Annotazione e Standards
- Repositories pubblici
- Repositories locali

Annotazione

- ❑ Riproducibilità dell'esperimento
- ❑ Valutazione dei risultati finali
- ❑ Confronto con esperimenti
microarray depositati in repositories
pubblici

Annotazione

- **Informazioni necessarie per riprodurre ed analizzare un esperimento microarray: dati e metadati**

The minimum information about a published microarray based gene expression experiment:

The minimum information about a published microarray based gene expression experiment should include:

1. expression level measurement results, in particular:
 - a. the TIFF image file from the hybridised microarray scanning;
 - b. the image analysis output (of the particular image analysis software) for each spot, for each channel;
 - c. a derived value summarising each spot in the authors' interpretation (e.g., a background subtracted intensity typically used for Stanford or Incyte technologies);
2. the following annotations:
 - a. array (e.g., platform type, substrate, number of spots, provider);
 - b. each element (spot) on the array (e.g., sequence or clone and relevant accession numbers);
 - c. sample source and treatment (e.g., organism, development stage, tissue, drug treatment);
 - d. controls in the sample and on the array;
 - e. hybridisation extract preparation (e.g., cell rupture method, nucleic acid extraction and labelling protocol);
 - f. hybridisation procedure (e.g., time, concentration, volumes, washes);
 - g. scanning procedure (e.g., hardware, output TIFF file loader);
 - h. image analysis and quantification (e.g., software, version, parameters);
 - also, we would like to encourage the image analysis...

Miame (minimum information about microarray experiment)

- ❑ a standard for describing microarray experiments
- ❑ array design description:
 - design dell'array, spot dell'array
- ❑ gene expression experiment description
 - disegno dell'esperimento, campioni impiegati, preparazione dell'estratto e labeling, procedure di ibridazione, misurazione di dati e processamento dei dati (raw data, image quantitation table, gene expression data matrix).

Annotazione

- ❑ **Le specifiche di Miame richiedono un'adatta terminologia:**
 - **ontology del gruppo di lavoro MGED**

Storage dei dati

- ❑ **Format di scambio dei dati**
- ❑ **Repositories pubblici**

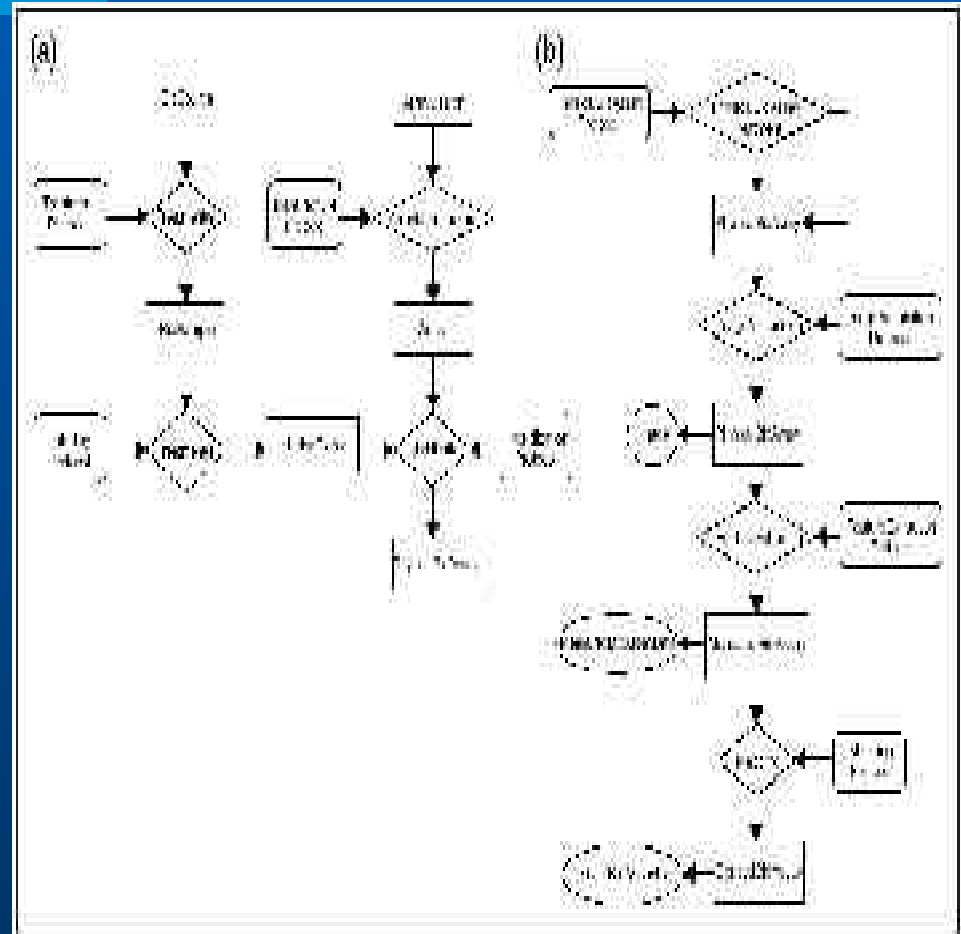
Format di scambio dei dati

□ MAGE (MicroArray Gene Expression)

- MAGE-OM (microarray gene expression-object model)
- MAGE-ML (microarray gene expression Markup language)

MAGE-OM

- **Modello entità-relazione che rappresenta tutte le fasi di un esperimento descritte da Miame**



MAGE-ML

- Infine il MAGE-OM viene trasformato nel formato di dati MAGE-ML (basato sul linguaggio XML) attraverso un software disponibile gratuitamente chiamato MAGE-STK.

Repositories pubblici

□ I Miame-compliant databases di riferimento sono:

- ⤴ GEO di NCBI
- ⤴ Array Express di EBI
- ⤴ CIBAX (Japan)

Database	Organization	Description
AYAL	Stanford University, University of California at Berkeley, University of Illinois at San Francisco (UCSF)	local data store
ArrayExpress	European Bioinformatics Institute (EBI)	public data repository and public queries (tracking only)
CAGE	Cardiff University	local installation (public queries coming soon)
ChIP-DB	Genitabase Institute for Human Genome Research, The Center for Genome Research	public queries
ChIPDB	Cornell University	public queries
ExpressDB	Harvard University	public queries of E. coli and yeast data
Expresso	Stanford University, Genitabase Institute for Human Genome Research	public queries of yeast data
GeneTractor	Trinity Microarray	local data store
Genet	Bioinformatics	local installation, public data repository, and public queries
Genes	NCBI	local installation, public data repository, and public queries of E. coli and yeast data
GEO	National Center for Biotechnology Information (NCBI)	local data repository and public queries
GSI	The Jackson Laboratory	public data repository and public queries of mouse data (coming soon)
Janelia	NIH	local data store
NCBI	National Center for Biotechnology Information (NCBI)	local data store
ncs300	the University of Manchester	local data store
NCYS	UCSF	local data store
RAI	University of Pennsylvania	public queries
SIL	Stanford University	local data store and public queries
SIBEX	Genitabase Institute for Human Genome Research	public queries of yeast data

Requisiti di un database in locale

- **Storage di ogni tipo di dato (dati e metadati)**
- **Indipendenza dalla specie, dalla tecnologia microarray, dal software dello scanner**
- **MIAME compliance**
- **Impiego di un format standard per lo scambio dei dati**

Requisiti di un database

- **Disponibilità pubblica attraverso un'interfaccia**
- **OS Unix/linux**
- **riservatezza dei dati:**
 - **possibilità di rendere pubblici i dati sperimentali, lasciando privati i dati di espressione genica.**

Alcuni Databases utilizzati

- ❑ BASE
- ❑ Miamexpress ed Array Express
- ❑ Maxdload
- ❑ Commercial e.g. GeNET (GeneSpring)

	BASE	ArrayExpress	Maxdload	GeNET
Reference data type	ADTbase	ADTbase	ADTbase	ADTbase
Species limitation	16	16	16	16
Design flexibility	Yes	Yes (partial)	Yes	Yes
Software availability	Yes	Yes (data export)	Yes	Commercial (GeNET)
Raw data management	Yes	Yes	Yes	Yes
Normalized data management	Yes	Yes	Yes	Yes
Image management	Yes	Yes	Yes	Yes
Example data management	Yes	Yes	Yes	Yes
Full-text search capabilities	Yes	Yes	Yes	Yes
Public databases links	Yes	Yes	Yes	GeNET only
OS	Windows and Mac OS X only	Linux only	Mac OS X only	Windows
DBMS	Microsoft SQL Server	Oracle	Microsoft SQL Server	Oracle
Search API	Yes	Yes (limited)	Yes (limited)	Yes
Integration	Yes (via ODBC)	Yes (via ODBC)	Yes	Yes
Backend format	Protein	MSB-XML	MSB-XML	Yes
Partner exchange format			MSB-XML	Yes
Input format	Yes	MSB-XML	MSB-XML	GeNET only
Output format	MSB-XML	MSB-XML	MSB-XML	Yes

BASE: installazione

- ❑ **Installazione di BASE 1.2.14 su Vega**
 - un web server Apache
 - un interprete PHP versione 4.3
 - un database relazionale MySQL o Postgre SQL
 - tools e librerie
 - java

<http://base.thep.lu.se/documentation>

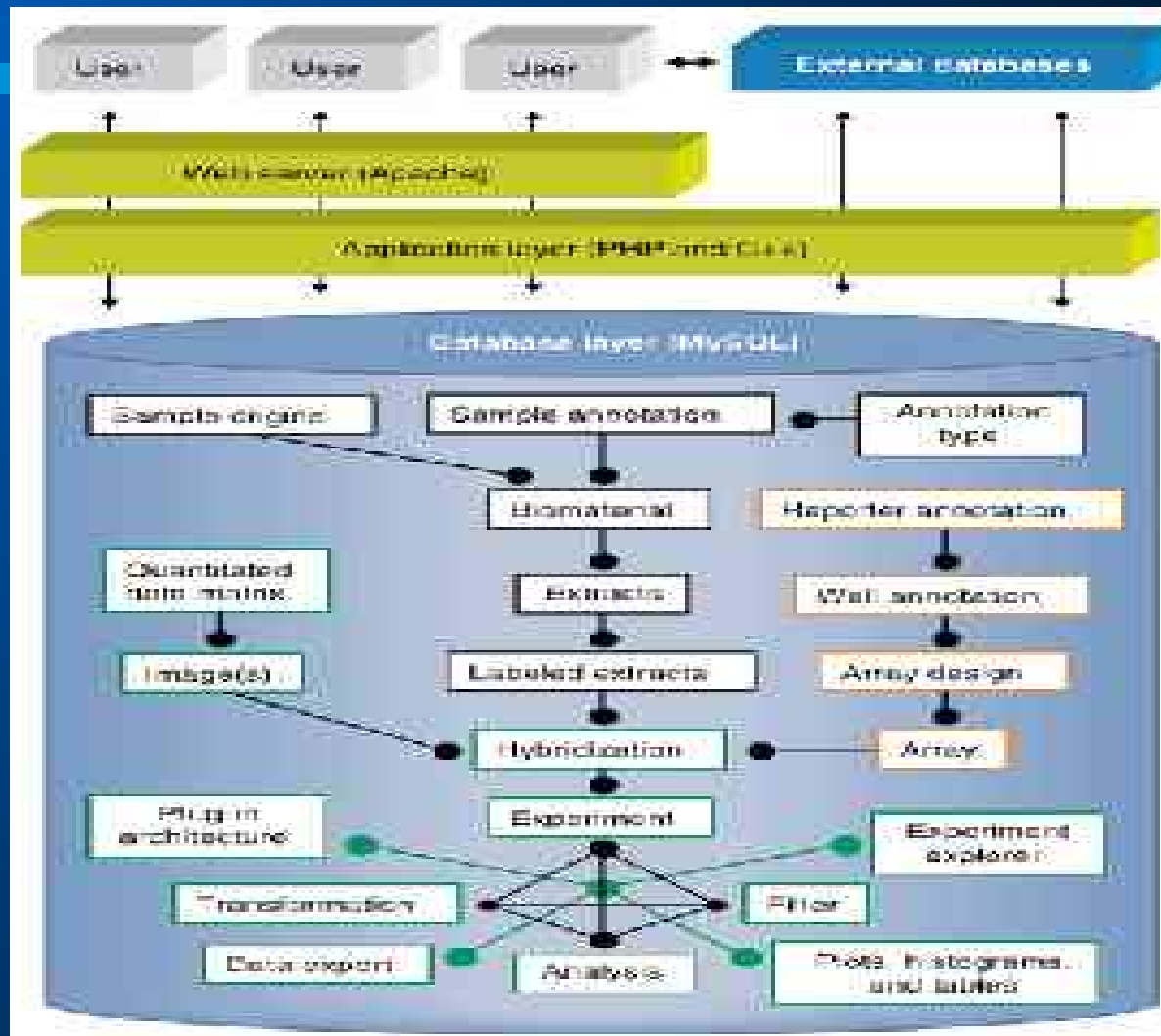
BASE (BioArray Software Environment)

<http://base.thep.lu.se/>

- Submission database (flat file)
- Query database (MAGE-ML, BaseFile)



BASE technology



BASE: hybridization

The screenshot shows the BASE web interface. On the left is a navigation menu with items like Home, Reports, Analyze DMS, etc. The main content area is titled "View hybridization" and displays details for a specific hybridization scan. Below the details is a table of scans and an "Add scan" form.

View hybridization

Parent: [Parent](#)

Full hybridization

Name	HH3-27
Arrive slide	2009
Labelled set (and ch1)	1x27 set P 3-7 5 mg
Labelled set (and ch2)	1x27 set P 3-7 5 mg
Hybridization protocol	10 min
Description	
Hybridized	2009-11-27 11:11
Date added	2009-11-27 11:11
Owner	guest
Group	Users (100)
World access	no

Add scan

Name &	Scanner	Scan date	Owner	[A/H] group	World	Images	Raw data	sets
HH3-24 scan 1	3000 scan ch1	2009-11-27 11:11	guest	[]	Users (100)			HH3-24 scan 1

Add scan

Update name: Group: with: Access:

[Deleted \(Deleted\)](#)

BASE: hybridization

BASE

Legend: [array](#) [chip](#) [array](#) [chip](#)
Department: [Department](#)
Users: [Users](#)

- Reports
- Array LIMS
- Biomaterials
- Hybridizations
- Hybridization
- Experiments
- Scanners
- Image processors
- Result file format
- Protocols
- Uploads
- Analyze data
- Users
- GUI settings
- Site info
- Report a bug
- BASE project site
- Event log

New hybridization

12.000

Name:

Labeled extract chip: Qty:

Labeled extract chip: Qty:

Labeled extract chip:

Hybridization protocol:

Description:

Hybridization date (Y-M-D):

Date added: 2004-11-10

Owner: guest:147

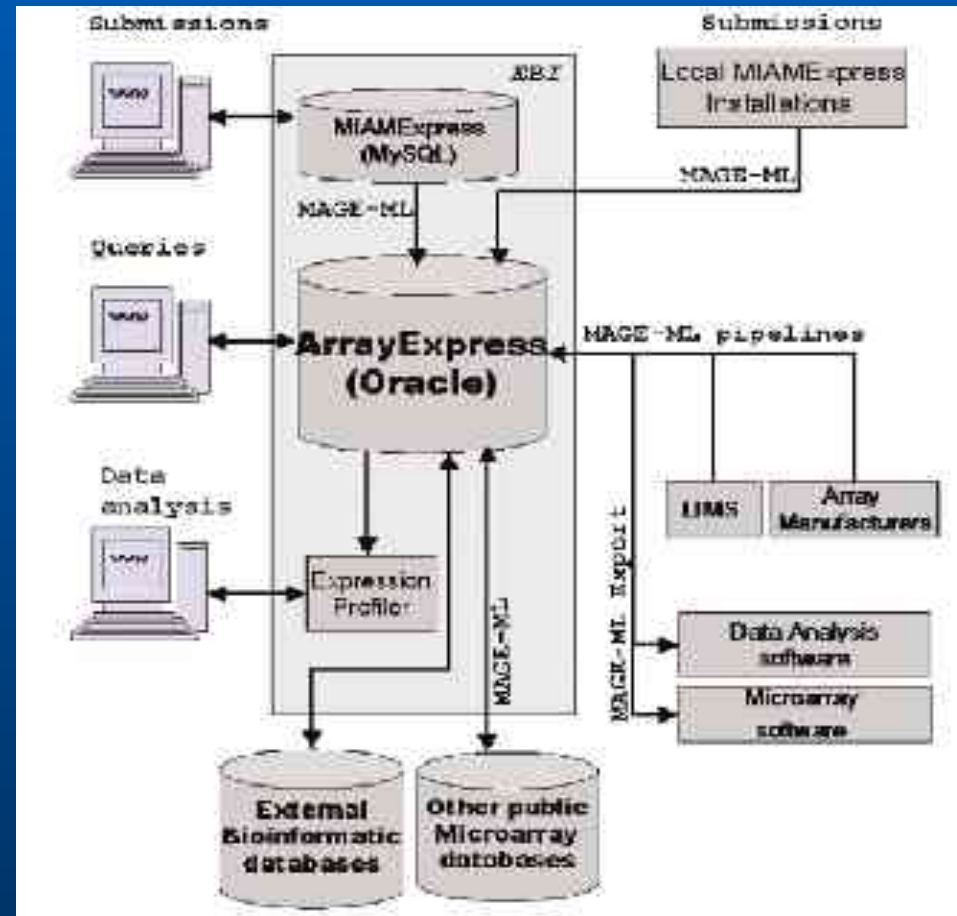
group:

World access: --

Use Array LIMS: No Yes

Array Express

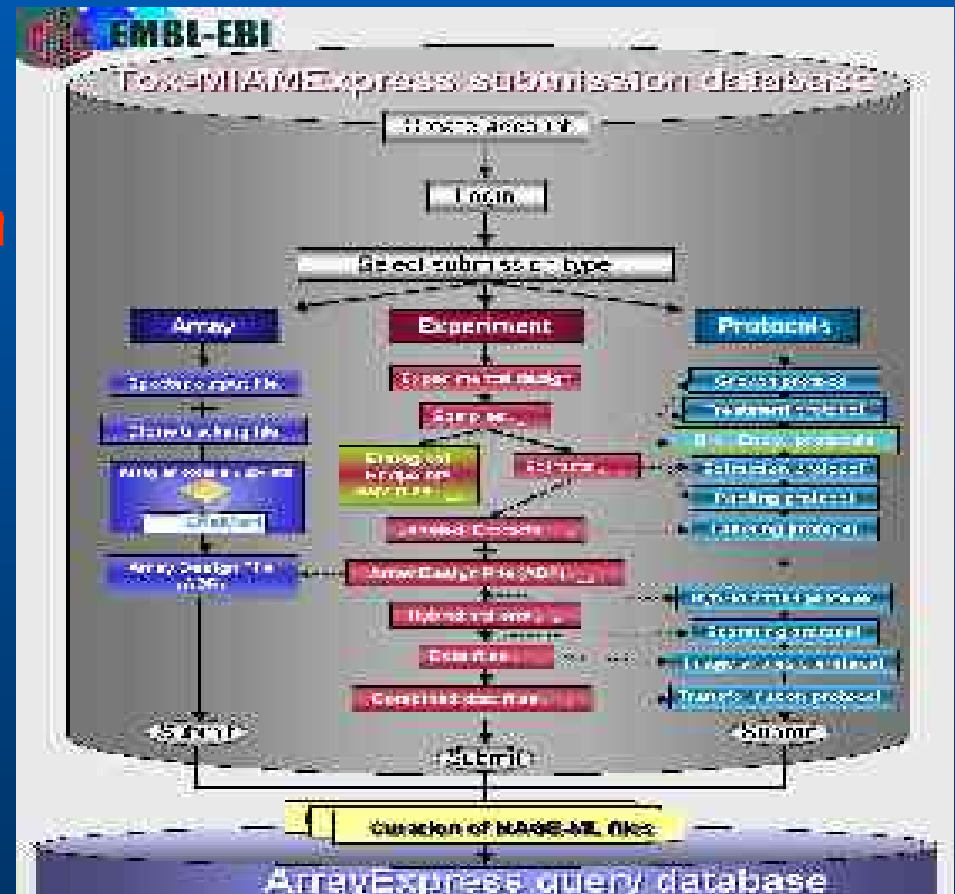
- ❑ Miamexpress (submission database)
- ❑ Array Express (query database)
- ❑ Expression Profiler (tool d'analisi)



MIAMExpress

- Submission (flat file) database via web:
<http://www.ebi.ac.uk/mia>

- Interfaccia Perl-CGI
- Database mySQL
- MAGE-stk



Array Express

- ❑ Query database via web (Mage-ML)

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

- Database Oracle
- Server Oracle
- Perl, Java

The screenshot displays the Array Express web interface, which is organized into three main sections, each with a green header bar:

- Query Experiment:** This section contains a search form with a dropdown menu for 'Experiment ID' and a 'Go' button. Below the dropdown are several input fields for 'Accession', 'Accession', 'Accession', 'Accession', and 'Accession'.
- Query Profile:** This section contains a search form with a dropdown menu for 'Profile ID' and a 'Go' button. Below the dropdown are two input fields for 'Accession' and 'Accession'.
- Query Process:** This section contains a search form with a dropdown menu for 'Process ID' and a 'Go' button. Below the dropdown is one input field for 'Accession'.

Maxd (Manchester University)

<http://bioinf.man.ac.uk/microarray/maxd/index.html>

- ❑ MaxdLoad2 (database per lo storage dei dati)
- ❑ MaxdView (sistema di visualizzazione ed analisi dei dati)
- ❑ MaxdSetup (gestione dell'installazione)

MaxdLoad2

- Home Page

<http://bioinf.man.ac.uk/microarray/maxc>

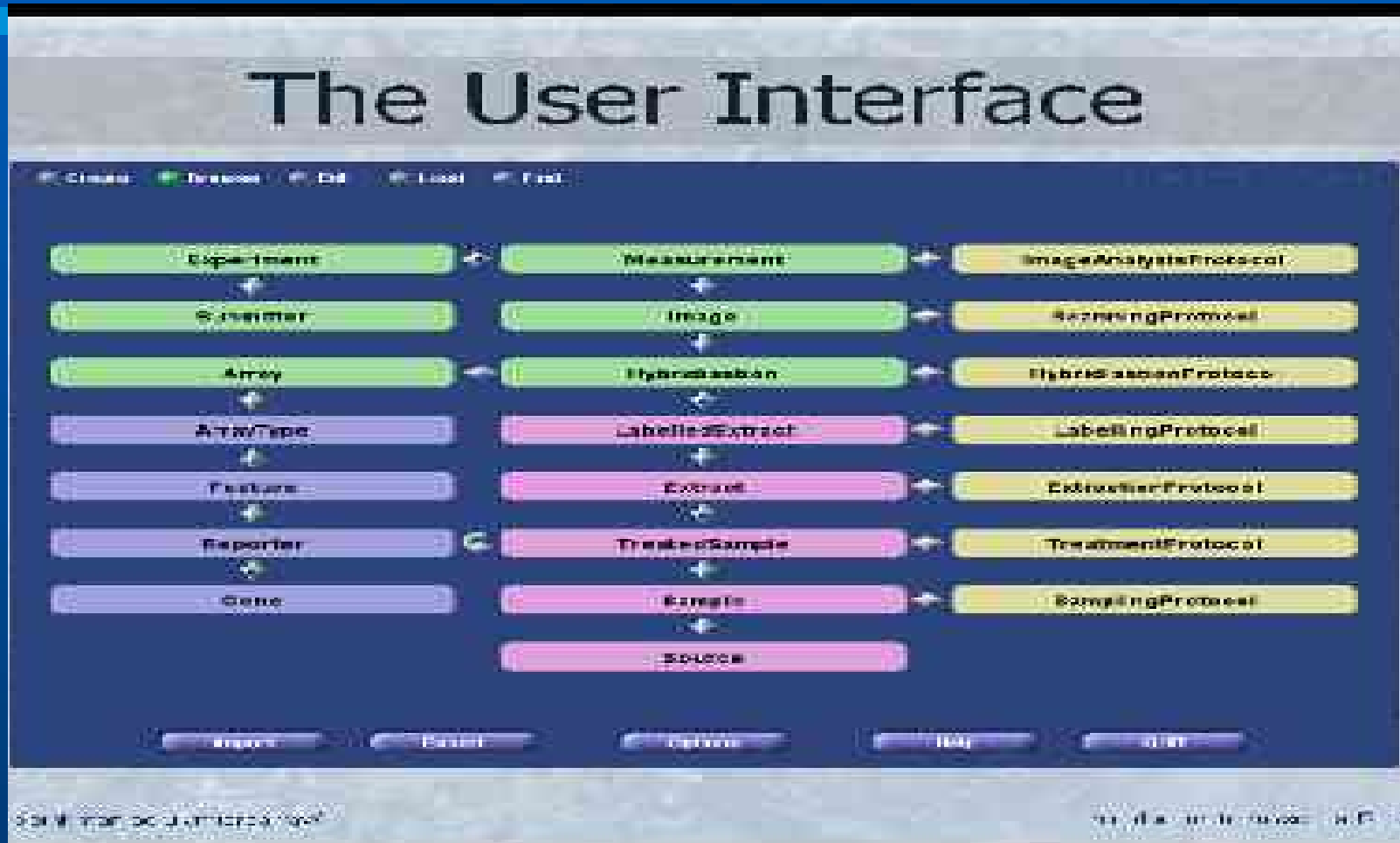
- Submission database (Text file, MaxdML)

- Query database (Mage-ML, MaxdML)

MaxdLoad2

- ❑ Database relazionale Oracle, MySQL, Postgres and Interbase.
- ❑ Java application
- ❑ Server JDBC (java database connectivity)

MaxdLoad2: interfaccia utente



MaxdLoad2: load/measurement

- ❑ Data entry forms
- ❑ Data entry fields
 - XML based
 - personalizzazione

... Measurement ...

To measure and monitor starting points, Measurement (Fig. 1) allows from the `Measurement` data table to linked `config` data table in `maxdSQL`.

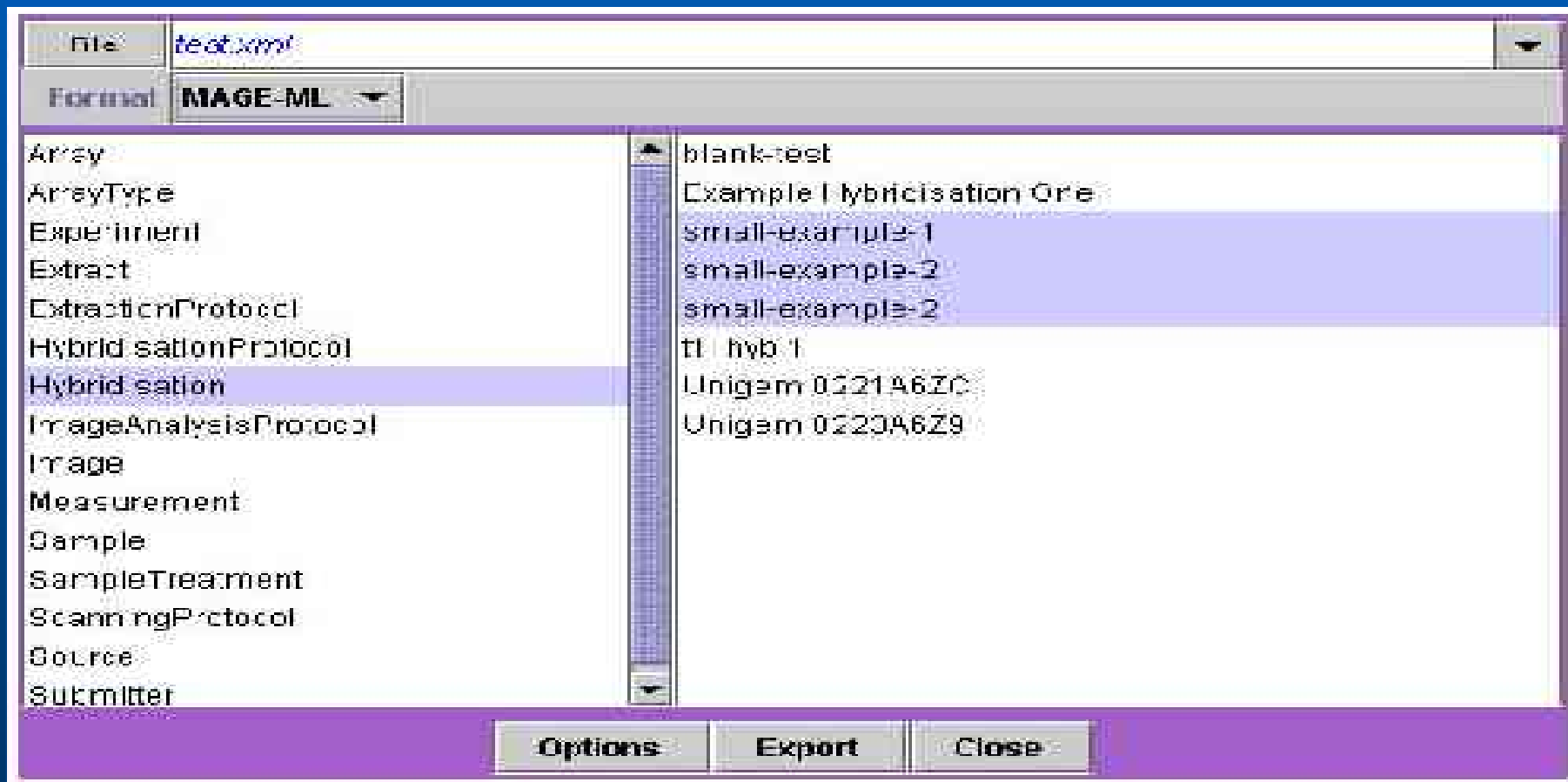
```
graph TD; Measurement[Measurement] --> StartingRecord[Starting record]; Measurement --> Stop[Stop]; Measurement --> Log[Log (begin/end)]; Measurement --> ProcessCount[Process count];
```

Fig. 1. `Measurement` data table.

`Measurement` represents a single column of expressions data tables in a `connection` file with format file. It links the `config` table in `maxdSQL` protocol and is connected to a large collection of `Spreadsheets` records in which are stored the `Measurement` entries.

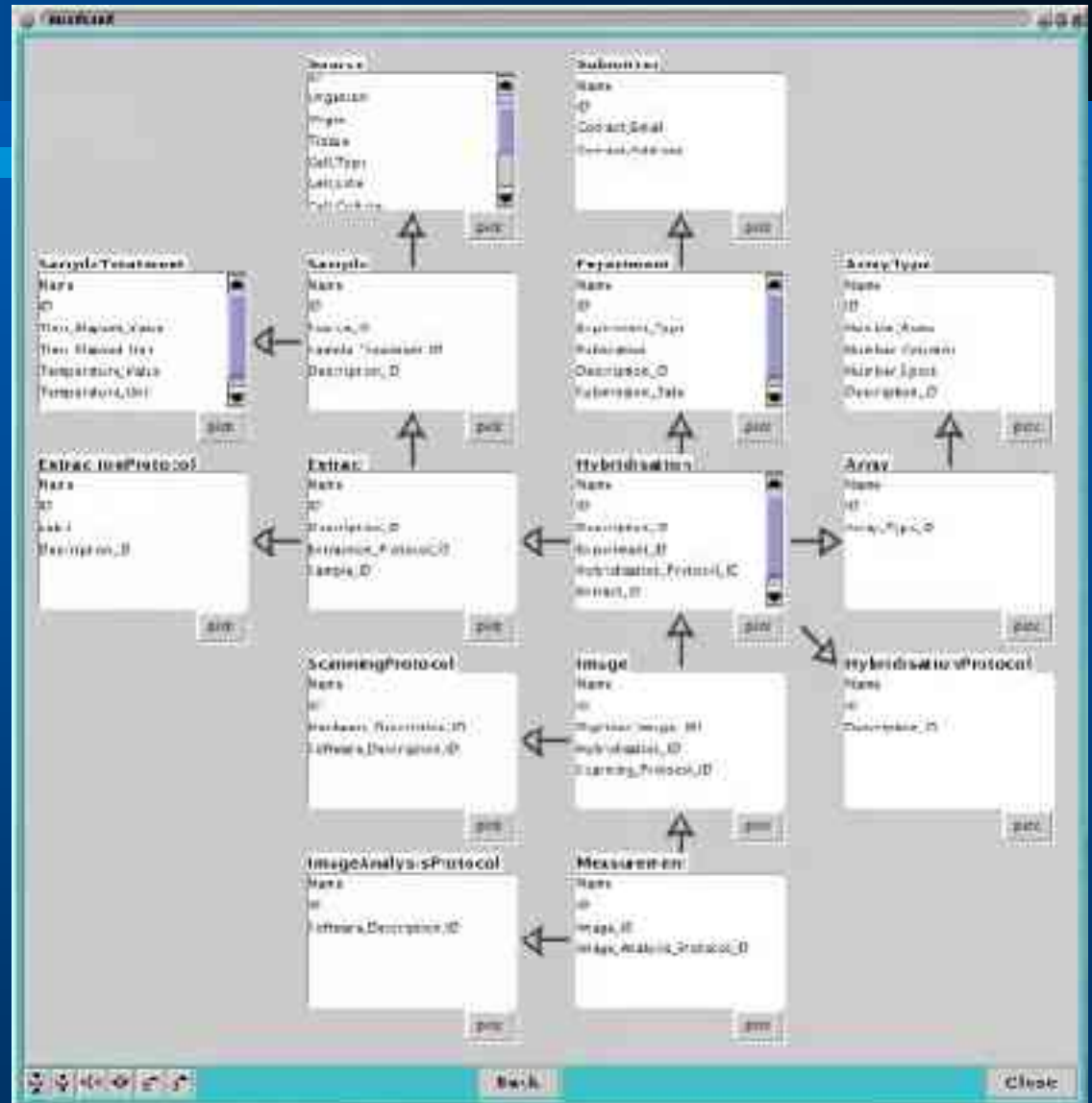
Fig. 2. `Measurement` form.

MaxdLoad2: export



MaxdLoad2: browser

- Il browser mostra un subset del database ed i possibili links tra gli elementi.
- Permette anche di editare i valori dei dati



Analysing expression data with data (text)-mining

- ❑ **Microarrays is a powerful method for studying gene function producing large amounts of data.**
- ❑ **However difficult to determine the biological causes for differential gene expression.**
- ❑ **Solution: use the extensive medical literature (MEDLINE, ~15 M citations)**

Analysing expression data with data (text)-mining

- ❑ **Text mining – automatic procedure for finding new, previously unknown information from written sources.**
- ❑ **Can be used in the biomedical literature to look for, for example, patterns of gene association.**

Analysing expression data with data (text)-mining – MedMOLE

MedMOLE - Mining On-Line Expert on MedLine.

Collaboration with CINECA + R. Calogero (Torino)

For each document in MEDLINE, MedMOLE analyzes the text to identify keywords. Using a GENE vocabulary potential gene names are analyzed.

The user interface allows a query of one or more keywords and is shown graphical description of document clusters.

