



# Presentation of MGED: lessons from the microarray community

Catherine Ball  
Stanford University  
[ball@genome.stanford.edu](mailto:ball@genome.stanford.edu)

# Why have data standards?

- Generation of large-scale data sets are costing the public big bucks.
- Researchers or regulatory agencies must be able to understand, validate/contradict conclusions and re-use data.
- The value of large-scale data sets are cumulative -- we want to combine data sets to make novel scientific discoveries.
- The scientific community has a responsibility to share data in a meaningful way.

# Data standards...

- Are NO substitute for exercising scientific judgment or critical thinking.
- Should NOT be used to standardize what is actually done (experimentally, technically or during analysis).
- ARE useful for describing what was actually done so that others can apply scientific judgment and critical thinking to your data.

# What is MGED?

- An international organization of biologists, computer scientists, and data analysts that aims to facilitate the sharing of data generated by large-scale biological experiments.
- The current focus is to establish 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.

# www.mged.org

## 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:

<b>New MISFISHIE WG</b>	<b>16/12/2004</b>
The MISFISHIE (Minimum Information Specification For In Situ Hybridization and Immunohistochemistry Experiments) working group has been formed and will be headed by Eric Deutsch. More details from <a href="http://mged.sourceforge.net/misfishie">http://mged.sourceforge.net/misfishie</a> .	
<b>MGED 8 Announcement - Date Change</b>	<b>15/11/2004</b>
The MGED 8 meeting will take in Bergen, Norway from <b>11th to 13th</b> of September, 2005. The MGED Programming Jamboree will follow from 14th to 17th of September, 2005. More details and registration will be available later from <a href="#">here</a> .	

**MGED Sponsors:**

-  Axon Instruments
-  Ipsogen  
CancerProfiler
-  BioDiscovery, Inc.
-  sas
-  iobion  
www.iobion.com
-  IBM
-  AFFYMETRIX
-  Agilent Technologies
-  Sun

**Publications**  
**News Archive**  
**Sponsors**  
**MIAME**  
**MAGE**  
**Ontologies (OWG)**  
**Transformations**  
**RSBI WGs**  
**MISFISHIE**  
**OMG**  
**Sourceforge**  
**MGED 8 Meeting**

# MGED History

- Nov 1999 - MGED was founded as a grass roots movement by many groups, including Affymetrix, Stanford and the EBI.
- Dec 1999 - The MGED web-page and e-mail discussion groups were established, and first-draft proposals for standards posted
- November 2000 - A proposal for a microarray data exchange format was submitted to the Object Management Group (OMG).

- Mar 2001 - The development of the MAGE standard began in cooperation between many academic and commercial groups (including Rosetta, Affymetrix and Agilent).
- Dec 2001 - A paper describing MIAME was published in Nature Genetics.
- Jan 2002 - The MAGE standard became an Adopted Specification by the OMG.
- June 2002 - MGED became a non-profit organization.

- Oct 2002 - Several major journals, including Nature, The Lancet, Cell and EMBO Journal adopted MIAME recommendations as a requirement for publication of microarray experiments.
- Oct 2002 - MAGE became the 'Available Specification for Gene Expression' at the OMG. A number of implementations have already been developed, including implementations by Affymetrix, EBI, TIGR, U Penn, Agilent and Stanford.
- Apr 2004 - Letter to journal editors about sequences used as microarray features published by several journals, including PLoS Biology.



# MGED meetings

- MGED 8: Sept 2005, Bergen, Norway.
- MGED 7: Sept 2004, Toronto, Canada.
- MGED 6: Sept 2003, Aix-En-Provence, France.
- MGED 5: Sept 2002, Tokyo, Japan.
- MGED 4: Feb 2002, Boston, USA.
- MGED 3: Mar 2001, Stanford, CA, USA.
- MGED 2: May, 2000, Heidelberg, Germany
- MGED 1: Nov, 1999, Cambridge, UK



What standards are currently  
accepted by the microarray  
community?

# MGED Standards

- MIAME - Minimal Information Annotating a Microarray Experiment
- MAGE-ML - MicroArray Gene Expression Markup language
- MGED Ontology - ontology that can be used to construct a MAGE document

# MIAME



- A list of information that researchers should strive to share in order to fully describe their experiments.
- Include information about experimental design, biological samples, features on microarrays, experimental protocols, data acquisition and processing.

# MAGE

- MAGE-OM is an object model describing the workflow of microarray experiments (can be applied to many types of high throughput experiments).
- MAGE-ML is a markup language used to describe microarray experiments (files can be very large).
- MAGE-stk is an open-source software toolkit that helps one construct and use MAGE files.

# MGED Ontology

- Provides a controlled vocabulary to describe microarray experiments using MAGE.
- Does not re-invent the wheel -- MO refers to existing ontologies/controlled vocabularies whenever possible.



How have microarray  
standards emerged and been  
accepted?

# Community, communication, cooperation

- Input from many groups was solicited very early in process.
- Detractors are actively sought out and recruited to be part of the solution.
- Small working groups devote considerable efforts toward specific goals.
- Results are disseminated for comment through publications, letters to editors, website, conferences, workshops and tutorials.



# Corporate sponsors ensure that communication with industry goes two ways



# MGED board of directors is a diverse group

<a href="#">Home</a>	<a href="#">Meetings</a>	<a href="#">Workgroups</a>	<a href="#">Mission</a>	<a href="#">MGED Board</a>	<a href="#">Site Map</a>
<a href="#">Home : MGED Board</a>					
<b>MGED Board Of Directors:</b>					
Cathy Ball ( <b>President</b> )	Stanford University				
Alvis Brazma	EMBL-EBI				
Helen Causton	Clinical Sciences Centre, Imperial College, London.				
Steve Chervitz	Affymetrix				
Ron Edgar	NCBI, NIH				
Michael Miller	Rossetta Biosoftware				
John Quackenbush	TIGR				
Susanna-Assunta Sansone	EMBL-EBI				
Paul Spellman ( <b>Treasurer</b> )	University of California at Berkeley				
Gavin Sherlock	Stanford Univeristy				
Chris Stoeckert	University of Pennsylvania				
Ronald Taylor ( <b>Secretary</b> )	University of Colorado				
Yoshio Tateno	DDBJ				
Neil Winegarten	University Health Network, Toronto				



# MGED advisory board keeps us honest

**MGED Advisory Board**

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

[Home](#) : [MGED Board](#) : [MGED Advisory Board](#)


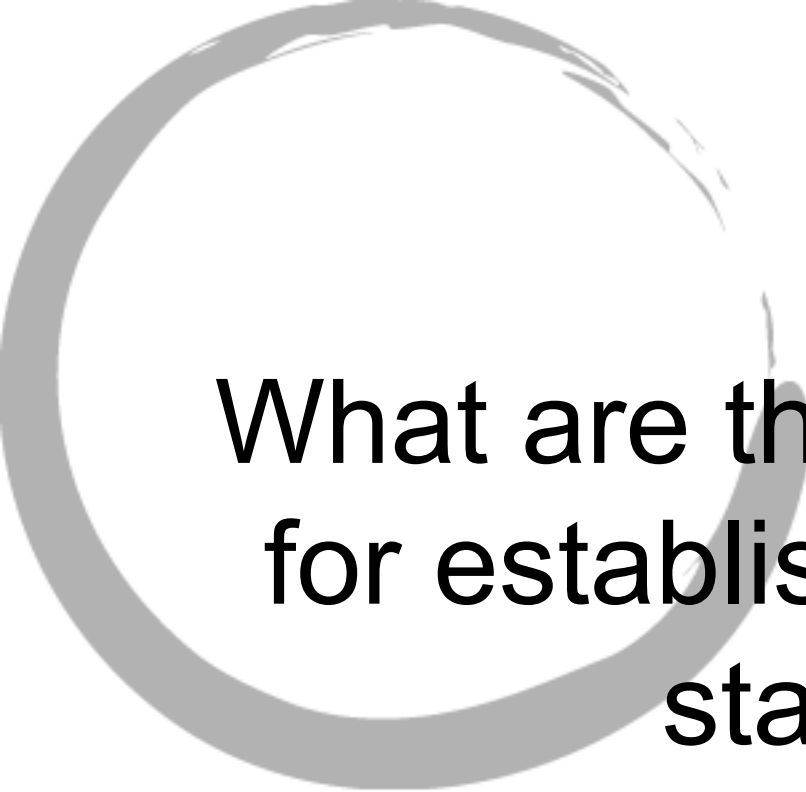
Virginia Barbour	PLOS
Ken Buetow	NCICB, NCI, NIH
Eric Deutsch	Institute For Systems Biology
Terry Gaasterland	Rockefeller University
Jason Goncalves	IOBION
Chris Gunter	Nature
Pascal Hingamp	Universite D'Aix-Marseille II
Frank Holstege	UMC Utrecht
Barbara Jasny	Science
Inge Jonassen	Bergen University
John Matese	Princeton University
Mark Miller	UCSD/San Diego Supercomputer Center
Mark Musen	Stanford University
Kousaku Okubo	Kyushu University
Helen Parkinson	EMBL-EBI
Martin Ringwald	The Jackson Laboratories
Kate Rice	Sanger Institute
Robert Stevens	University of Manchester
Jordan Stockton	Silicon Genetics
Charles Troup	Agilent
Mike Waters	NIEHS
Joseph White	TIGR



**How are microarray standards  
improved, implemented and  
made to serve the community  
better?**

# Standards are being modified by those who have to use them

- MIAME working group includes people from databases, repositories, journals, companies and laboratories.
- MAGE working group includes biologists and computer scientists from industry and academics.
- Ontology working group includes people from databases, repositories and laboratories.
- All working groups have open mailing lists.



**What are the main problems  
for establishing microarray  
standards?**


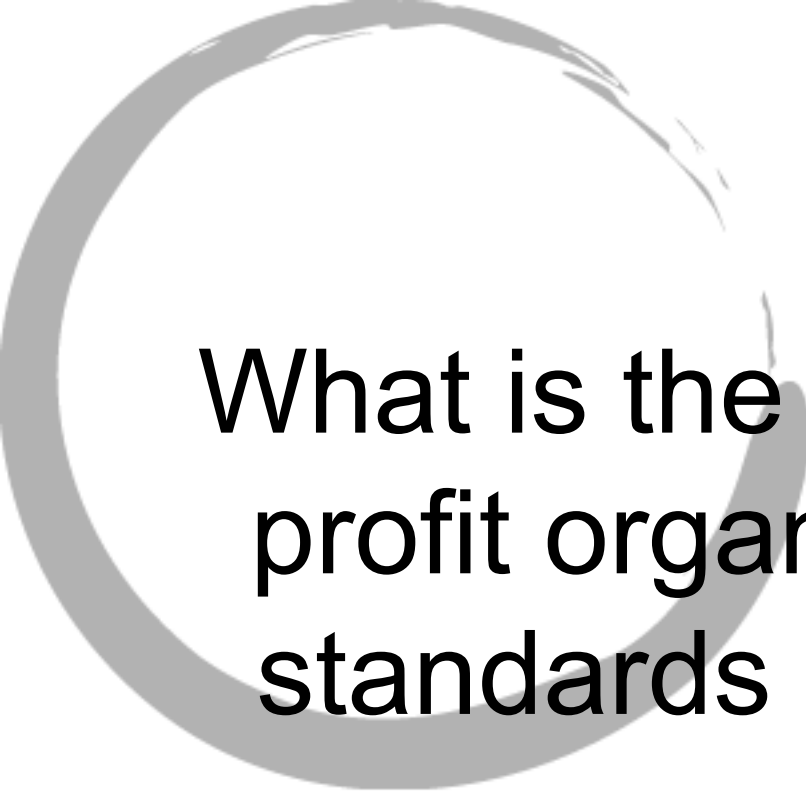
# Time and Money

- NO MGED standards developed thus far have been explicitly funded.
- Standards have usually been established by the informal donation of time and resources (weekends and evenings).
- This has also been a blessing, since it has required us to rely on corporate sponsors and accept the help of all comers -- standards are truly the creation of the community.
- A proposal for explicit funding is in progress.

# Current shortcomings

- Combining data sets from different sources still not trivial.
- Public data repositories (GEO at NCBI, ArrayExpress at EBI and Cibex at DDBJ) are do not represent data in identical formats, nor are data sharing processes in place (yet).
- MAGE-OM is free enough that there are multiple ways to record the same data -- MAGE-ML files from different groups are not identical.
- Data quality metrics are nowhere close to useful.







What is the attitude of the for-profit organizations towards standards and open source software?

# Standards benefit all groups

- Most corporate groups recognize that their products are more valuable if they use community standards (academics can publish, pharmaceutical companies can get FDA approval, etc.).
- Open-source software toolkits (MAGE-stk) have been used by corporate groups when developing their proprietary tools.





What are the main concerns when establishing microarray standards and how these concerns can be addressed?

# Challenges

- Standards should be as complete and accurate as possible.
- Microarray technology is being (and will continue to be) adapted to new and sometimes unanticipated uses.
- Microarray standards should be accessible to normal laboratories.

# Solutions


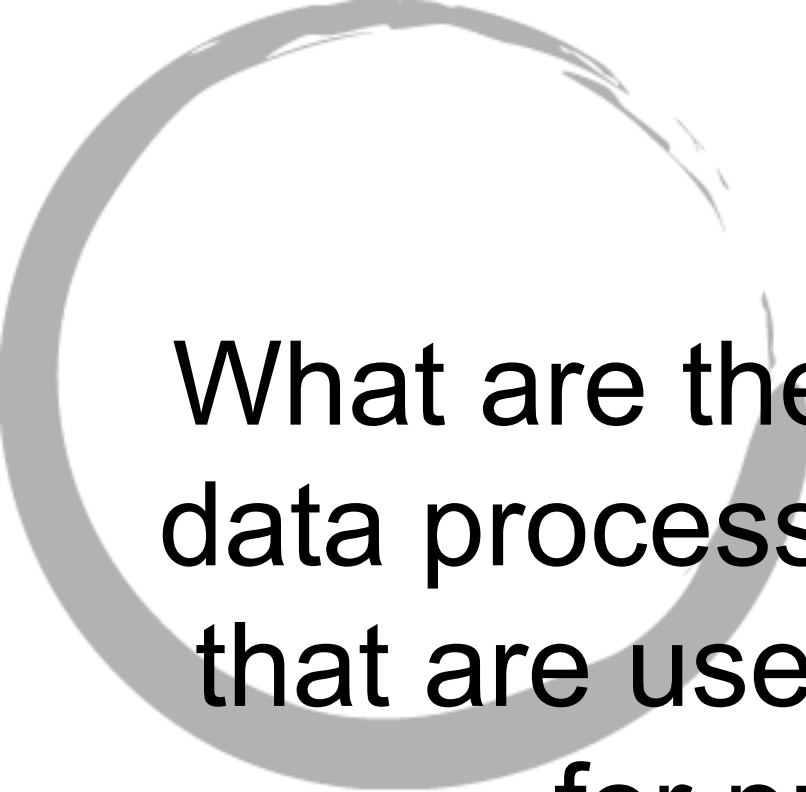
- In response to these challenges, we do not disband our working groups, but continuously adapt and improve using input from those “pushing the envelope.”
- Self-appointed MGED members come from many backgrounds, use microarray data in many ways and work at many institutions, so we have a reasonable cross-section of the community.
- Communication with the research community is key -- meetings, web site, sourceforge for software and PUBLICATIONS.



Should we start developing a  
proteomic dictionary for  
facilitating standardization  
(semantic approach)?

# Semantic solutions

- MAGE and the MGED ontology provided terrific semantic solutions.
- Have introduced a new (and neutral) vocabulary so we can understand each other.
- Not unlike Esperanto, it can be a little awkward and non-obvious to novices.





**What are the requirements for data processing software tools that are used to prepare data for publication?**



# Software tools used for microarray publications

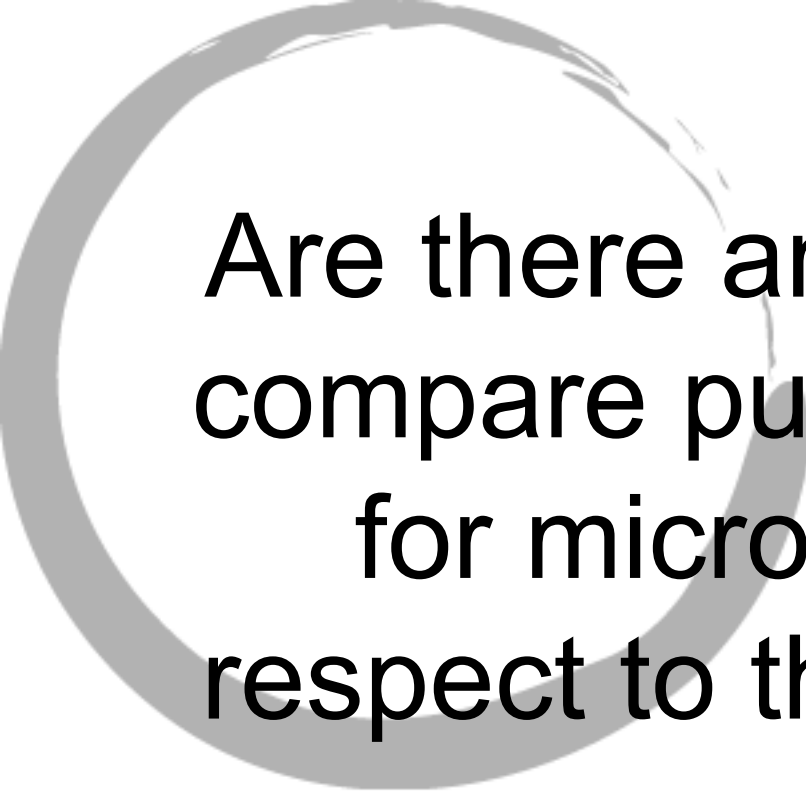
- Should not be “black box” algorithms
- Need not be open source
- Enough information should be provided such that a different group can reproduce the results without buying the software (might have to do some hard work, though)



How much and in which form  
should microarray data be  
accessible to reviewers and  
readers?

# Data publication

- All data should be released
- All raw data
- All processed data
- Names and versions of all software packages used or written
- All the steps used to process data
- All biological data about the samples used
- All sequence data about the reporters on the microarrays



Are there any mechanisms to compare public software tools for microarray data with respect to their performance?

- Not yet, sadly.
- 



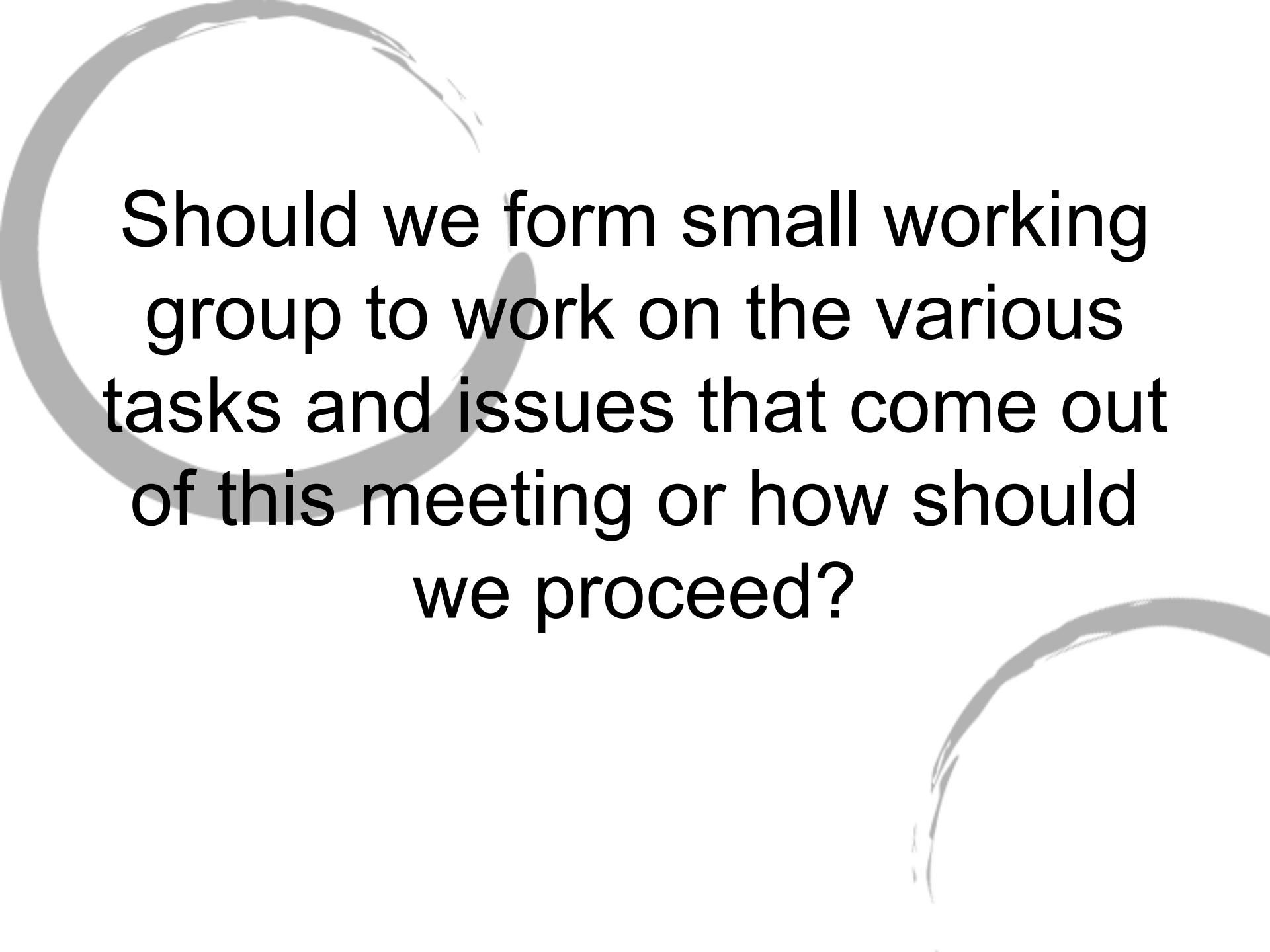
**Should or can journals enforce  
submission of microarray  
data?**

# Many journals require submission of data

- An incomplete list of journals (there are dozens) that require data release include:
  - Nature journals, Cell, EMBO Journal, PLoS journals, New England Journal of Medicine
- Many reviewers require submission as a requirement for publication (like me).
- Importantly, the microarray community has two public data repositories (GEO at NCBI and ArrayExpress at EBI).

# In what form and where should data be archived?

- Currently, this is determined by the data repositories, largely due to their own resource limitations.
- My personal conviction is that all primary data (images) as well as derived data (raw and transformed measurements) should be recorded and released in MAGE, but this is currently beyond the abilities of the data repositories.



Should we form small working group to work on the various tasks and issues that come out of this meeting or how should we proceed?



# MGED working groups

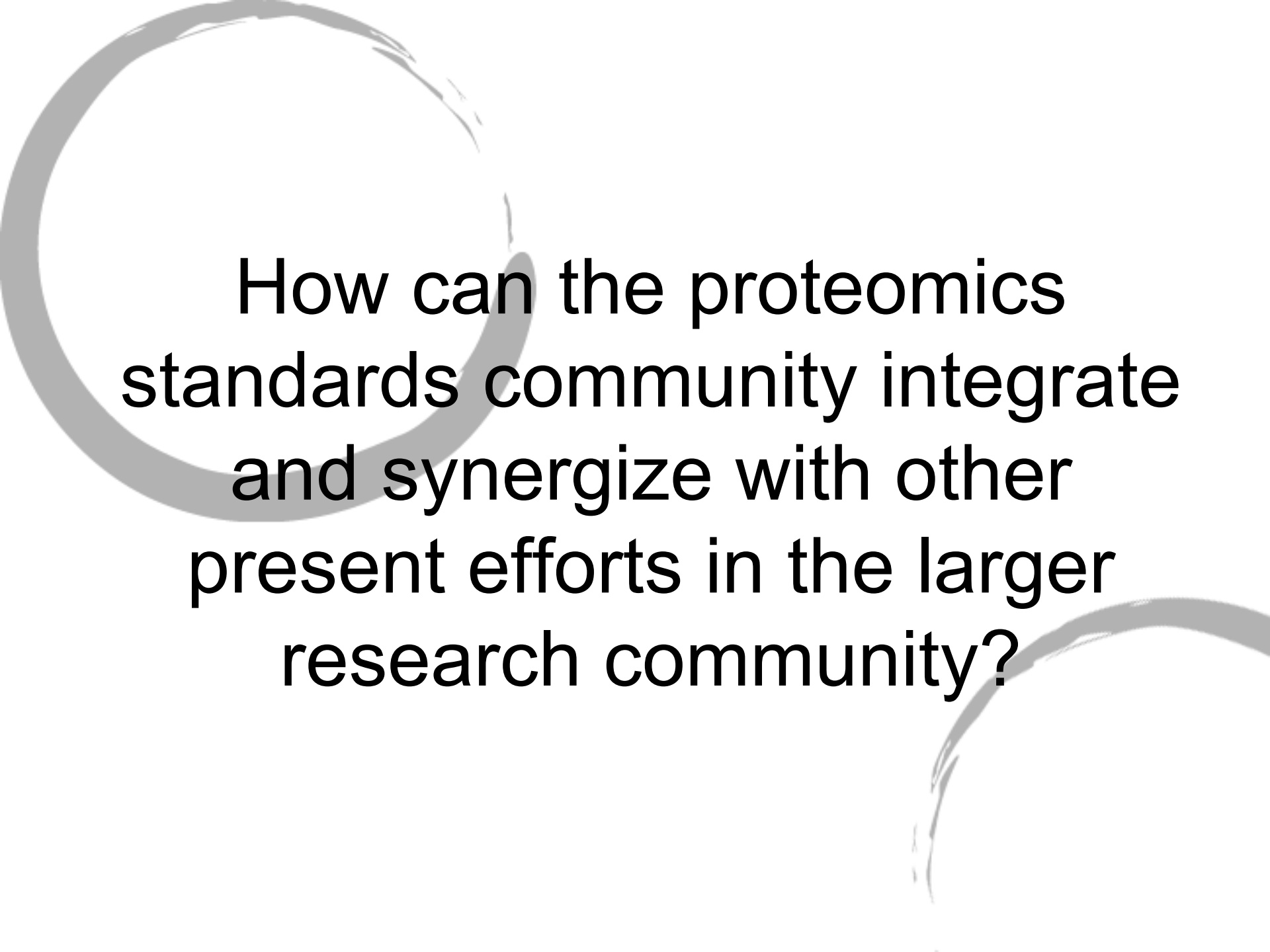
- MIAME (what information should be shared?)
- MAGE working group (how to communicate data?)
- Programming jamborees (twice-yearly long weekends spent writing open-source code)
- Ontology working group (what terms are needed?)
- Data transformation working group (what happens to data, how can its quality be assessed?)
- RSBI working group (applications of MGED standards to other technologies or disciplines)
- MISFISHE working group (standards for immunohistochemistry and *in situ* hybridizations)

# MGED working groups work hard!

- Small groups of dedicated people who donate their time, expertise and effort.
- Work together at intense “jamborees” that last 2-3 days.
- Use the internet for virtual meetings to share work.
- MGED working group members and leaders communicate frequently via e-mail and monthly conference calls.
- Actively participate and present work at yearly MGED meetings.

# What lessons can be learned from MGED?

- Community input is NOT enough.
- Standards need to be driven by the science and those who will use the standards.
- Small working groups can make more rapid progress than large committees.
- All work should be widely and frequently disseminated for criticism and comment.
- Public data repositories and other resources should exist.
- Re-use existing resources.
- Have meetings in fun places, ensure interesting scientific talks to provide context.
- Since science is a moving target, we must expect standards to evolve.

The background features two large, overlapping gray brushstroke circles. One circle is on the left side, and another is on the right side, both rendered with a soft, painterly texture.

How can the proteomics standards community integrate and synergize with other present efforts in the larger research community?

# Suggestions for proteomics community

- Join MGED!!
- Evaluate MGED standards
- Identify shortcomings
- Suggest improvements
- Develop extensions or complementary standards
- Come to MGED meetings (Sept 2005, Bergen, Norway)

# Come to MGED 8



**8th International Meeting of the Microarray Gene Expression Data Society  
11-13 September 2005**

[home](#)

[local](#)

[contact](#)

## Paper work

- ▶ [Registration](#)
- ▶ [Agenda](#)
- ▶ [Abstract submission](#)
- ▶ [Key dates](#)

## Local info

- ▶ [Conference Venue](#)
- ▶ [Hotel Info](#)
- ▶ [Travel](#)
- ▶ [Bergen](#)
- ▶ [Norway](#)

## Links

## Welcome to MGED 8 and Bergen in 2005!



## MGED 7

- ▶ [This years meeting](#)

Visit: [MGED 7](#)



[www.mged.org](http://www.mged.org)