Proteomics Standard Group

Bioinformatique et protéomique
Montbonnot – 1er juin 2006

yves.vandenbrouck@cea.fr

Proteomics : the state of play

- The volume of generated proteome data is rapidly increasing
  - Movement towards high–throughput approaches
  - New experimental techniques and analyses (DiGE, ICAT, etc.)

- Publicly available proteomics data is rather limited
  - 2D–Gel image databases (e.g. SWISS–2DPAGE) contain little information about sample preparation, or analysis of results
  - No widely used databases of mass spectrometry data or analyses
  - Data fragmentation (not synchronized) and data formats incompatibility (PID)
  - Incomplete data and noisy
  - Need for cross-validation and comparative analysis

A robust, future proofed, standard representation of both methods and data from proteomics experiments is required (HUPO-PSI, Hermjakob and Apweiler, 2002)

The Proteomics Standard initiative (PSI-HUPO)

http://psidev.sourceforge.net/

•The HUPO Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparision, exchange and verification. The PSI was founded at the HUPO meeting in Washington, April 28-29, 2002 (see Science 296, 827).

¬Develop data model and format standards

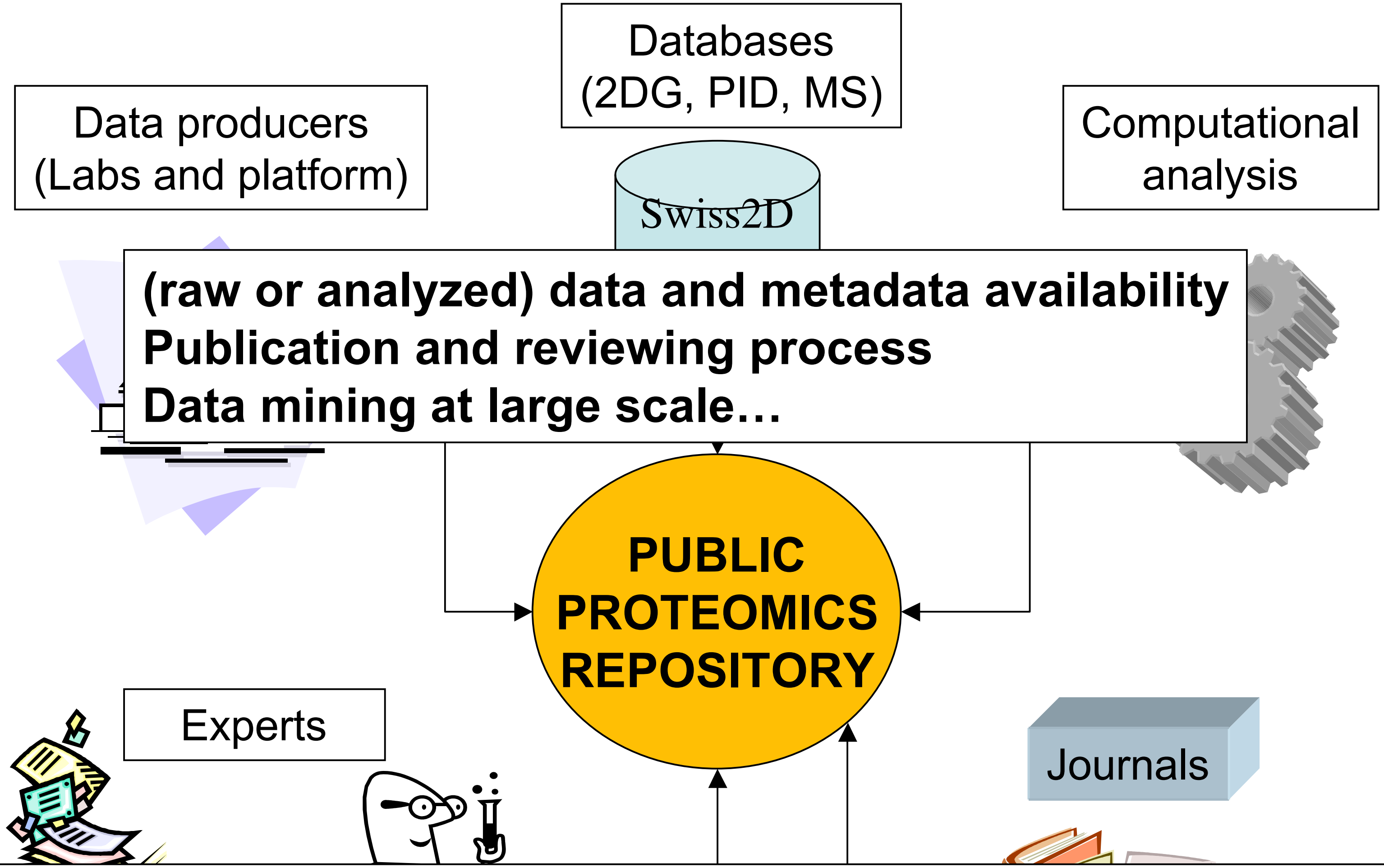
¬Guidelines (check-list, reporting how, GP)

¬Data representation and annotations standards : Controlled vocabularies, ontologies….

•Involve data producers, database providers, software designers, publishers, vendors

MIAPE : The Minimum Information About Proteomics Experiment
Analogous to MIAME (MGED)

Proteomics data dissemination : players and stakes



(raw or analyzed) data and metadata availability
Publication and reviewing process
Data mining at large scale…

PUBLIC PROTEOMICS REPOSITORY

French Genopole Network Context (1/06/04) :
Data exchange : necessity (and constraints), agreement?
Model & Data format : involvement about current attempts?

French Proteomics Standard Group

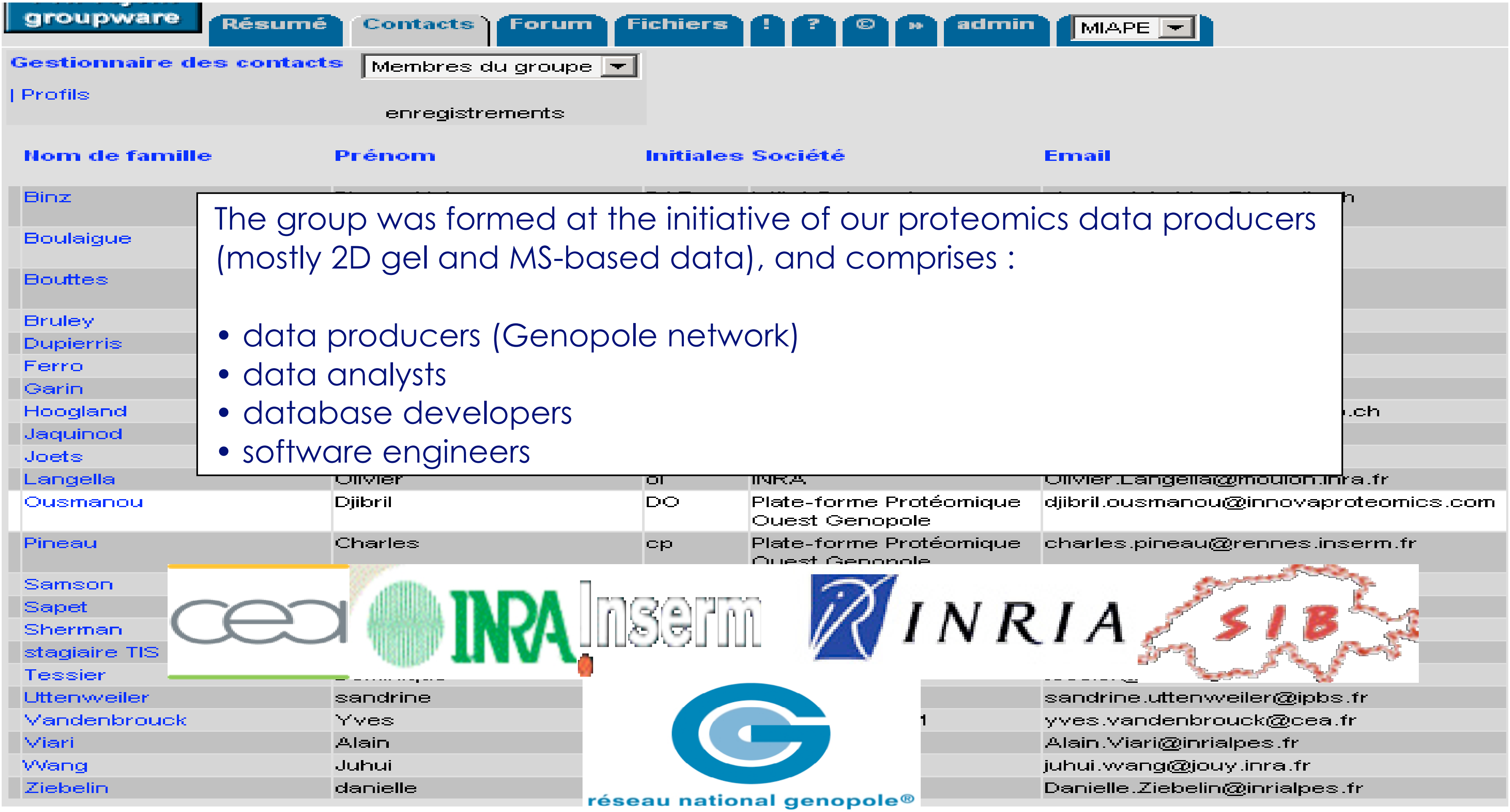
Created in Nov 2004 (Grenoble)
Presented at the HUPO-France meeting 5/01/2005
Supported by the National Genopole Network (RNG)

Workgroup objectives :

- To participate in establishing guidelines, defining data models, and elaborating terminological resources in the context of existing initiatives including PSI.
- To provide framework for data model design and tools (data models, integrity rules, representation formats, translations between models) towards end-users for evaluation purpose (practical, efficient, realistic).
- Mid-term objective : to set up a national repository for proteomics data that respects the MIAPE recommendations.

French Proteomics Standard Group : Missions (1)

- **Federate know-how and expertise of labs and plat forms in the MIAPE context**
  - **⇒ Web server http://www-miape.inrialpes.fr/**



Nom de famille	Prénom	Initiales	Société	Email
Binz				h
Boulaigue				
Bouttes				
Bruley				
Dupierris				
Ferro				
Garin				
Hoogland				.ch
Jaquinod				
Joets				
Langella	Olivier	OL	INRA	Olivier.Langella@moulon.inra.fr
Ousmanou	Djibril	DO	Plate-forme Protéomique Ouest Genopole	djibril.ousmanou@innovaproteomics.com
Pineau	Charles	cp	Plate-forme Protéomique Ouest Genopole	charles.pineau@rennes.inserm.fr
Samson				
Sapet				
Sherman				
stagiaire TIS				
Tessier				
Uttenweiler	sandrine			sandrine.uttenweiler@ipbs.fr
Vandenbrouck	Yves			yves.vandenbrouck@cea.fr
Viari	Alain			Alain.Viari@inrialpes.fr
Wang	Juhui			juhui.wang@jouy.inra.fr
Ziebelin	danielle			Danielle.Ziebelin@inrialpes.fr

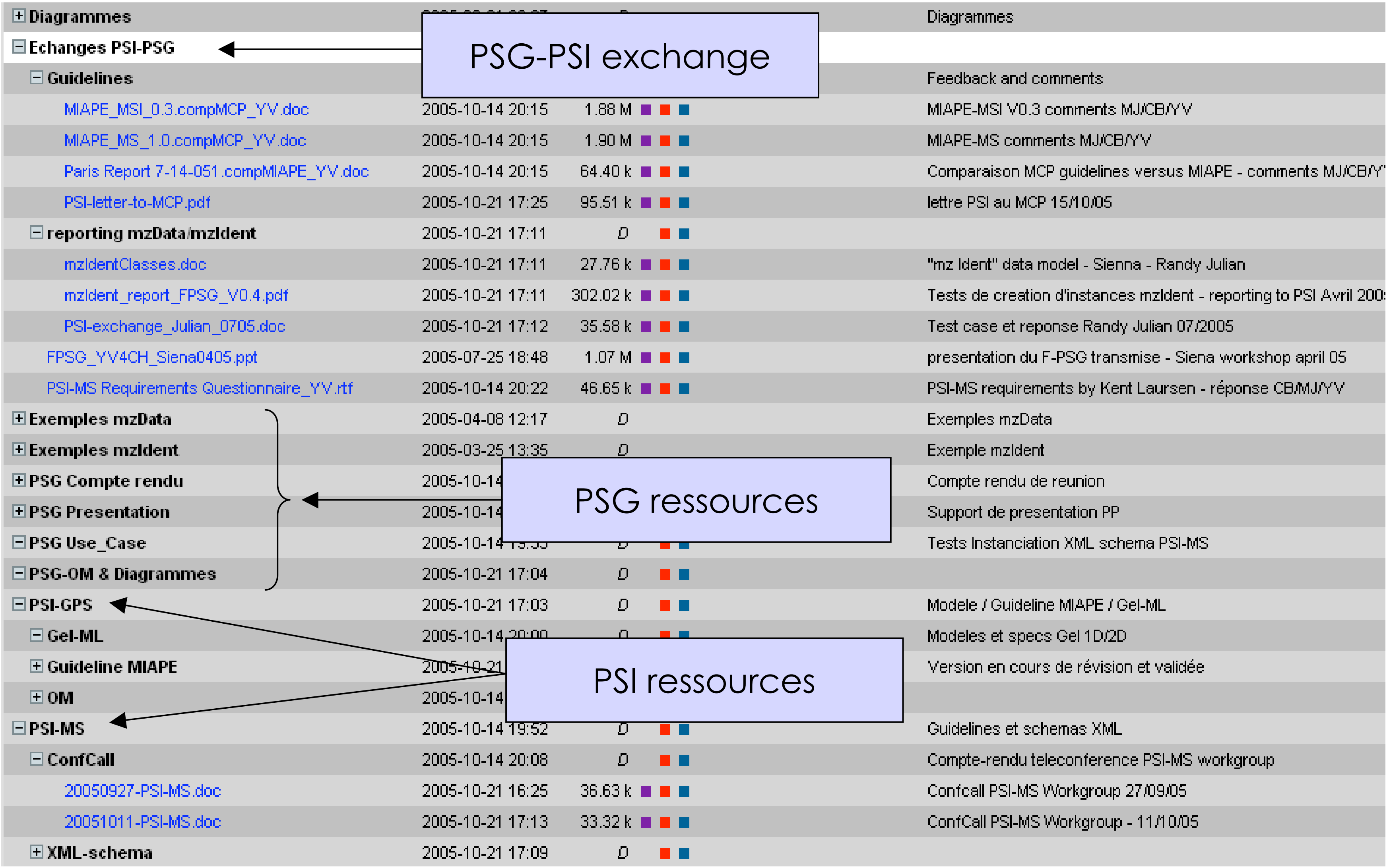
The group was formed at the initiative of our proteomics data producers (mostly 2D gel and MS-based data), and comprises :

- data producers (Genopole network)
- data analysts
- database developers
- software engineers

réseau national genopole®

French Proteomics Standard Group : Missions (2)

- **Share documentation, reporting and ressources**



PSG-PSI exchange

PSG ressources

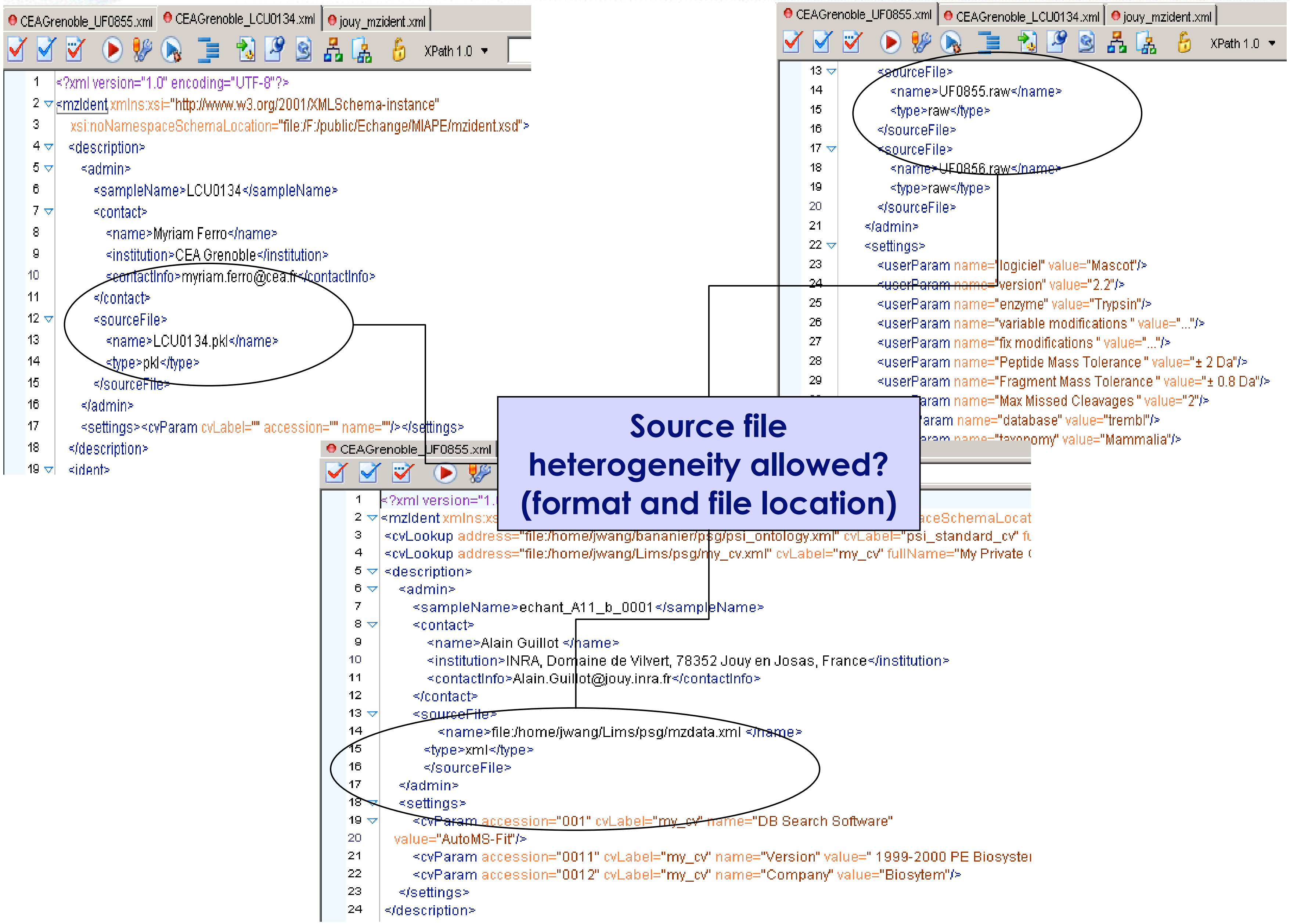
PSI ressources

⊞ **Diagrammes**				Diagrammes
⊟ **Echanges PSI-PSG**				
⊟ **Guidelines**				Feedback and comments
MIAPE_MSI_0.3.compMCP_YV.doc	2005-10-14 20:15	1.88 M	🟪 🟥 🟦	MIAPE-MSI V0.3 comments MJ/CB/YV
MIAPE_MS_1.0.compMCP_YV.doc	2005-10-14 20:15	1.90 M	🟪 🟥 🟦	MIAPE-MS comments MJ/CB/YV
Paris Report 7-14-051.compMIAPE_YV.doc	2005-10-14 20:15	64.40 k	🟪 🟥 🟦	Comparaison MCP guidelines versus MIAPE - comments MJ/CB/Y
PSI-letter-to-MCP.pdf	2005-10-21 17:25	95.51 k	🟪 🟥 🟦	lettre PSI au MCP 15/10/05
⊟ **reporting mzData/mzIdent**	2005-10-21 17:11	D	🟥 🟦	
mzIdentClasses.doc	2005-10-21 17:11	27.76 k	🟪 🟥 🟦	"mz Ident" data model - Sienna - Randy Julian
mzIdent_report_FPSG_V0.4.pdf	2005-10-21 17:11	302.02 k	🟪 🟥 🟦	Tests de creation d'instances mzIdent - reporting to PSI Avril 200
PSI-exchange_Julian_0705.doc	2005-10-21 17:12	35.58 k	🟪 🟥 🟦	Test case et reponse Randy Julian 07/2005
FPSG_YV4CH_Siena0405.ppt	2005-07-25 18:48	1.07 M	🟪 🟥 🟦	presentation du F-PSG transmise - Siena workshop april 05
PSI-MS Requirements Questionnaire_YV.rtf	2005-10-14 20:22	46.65 k	🟪 🟥 🟦	PSI-MS requirements by Kent Laursen - réponse CB/MJ/YV
⊞ **Exemples mzData**	2005-04-08 12:17	D		Exemples mzData
⊞ **Exemples mzIdent**	2005-03-25 13:35	D		Exemple mzIdent
⊞ **PSG Compte rendu**	2005-10-14			Compte rendu de reunion
⊞ **PSG Presentation**	2005-10-14			Support de presentation PP
⊟ **PSG Use_Case**	2005-10-14		🟥 🟦	Tests Instanciation XML schema PSI-MS
⊟ **PSG-OM & Diagrammes**	2005-10-21 17:04	D	🟥 🟦	
⊟ **PSI-GPS**	2005-10-21 17:03	D	🟥 🟦	Modele / Guideline MIAPE / Gel-ML
⊟ **Gel-ML**	2005-10-14 20:00	D	🟥 🟦	Modeles et specs Gel 1D/2D
⊞ **Guideline MIAPE**	2005-10-21			Version en cours de révision et validée
⊞ **OM**	2005-10-14			
⊟ **PSI-MS**	2005-10-14 19:52	D	🟥 🟦	Guidelines et schemas XML
⊟ **ConfCall**	2005-10-14 20:08	D	🟥 🟦	Compte-rendu teleconference PSI-MS workgroup
20050927-PSI-MS.doc	2005-10-21 16:25	36.63 k	🟪 🟥 🟦	Confcall PSI-MS Workgroup 27/09/05
20051011-PSI-MS.doc	2005-10-21 17:13	33.32 k	🟪 🟥 🟦	ConfCall PSI-MS Workgroup - 11/10/05
⊞ **XML-schema**	2005-10-21 17:09	D	🟥 🟦	

French Proteomics Standards Group : actions

- January 05 : EBI contact (PSI coordinators)
- March 05 : Test case with mzdata and mzIdent standard
- April 05 : feedback and suggestion available for community mzIdent test case report and sample files

Mzident instance : few question remarks (1)



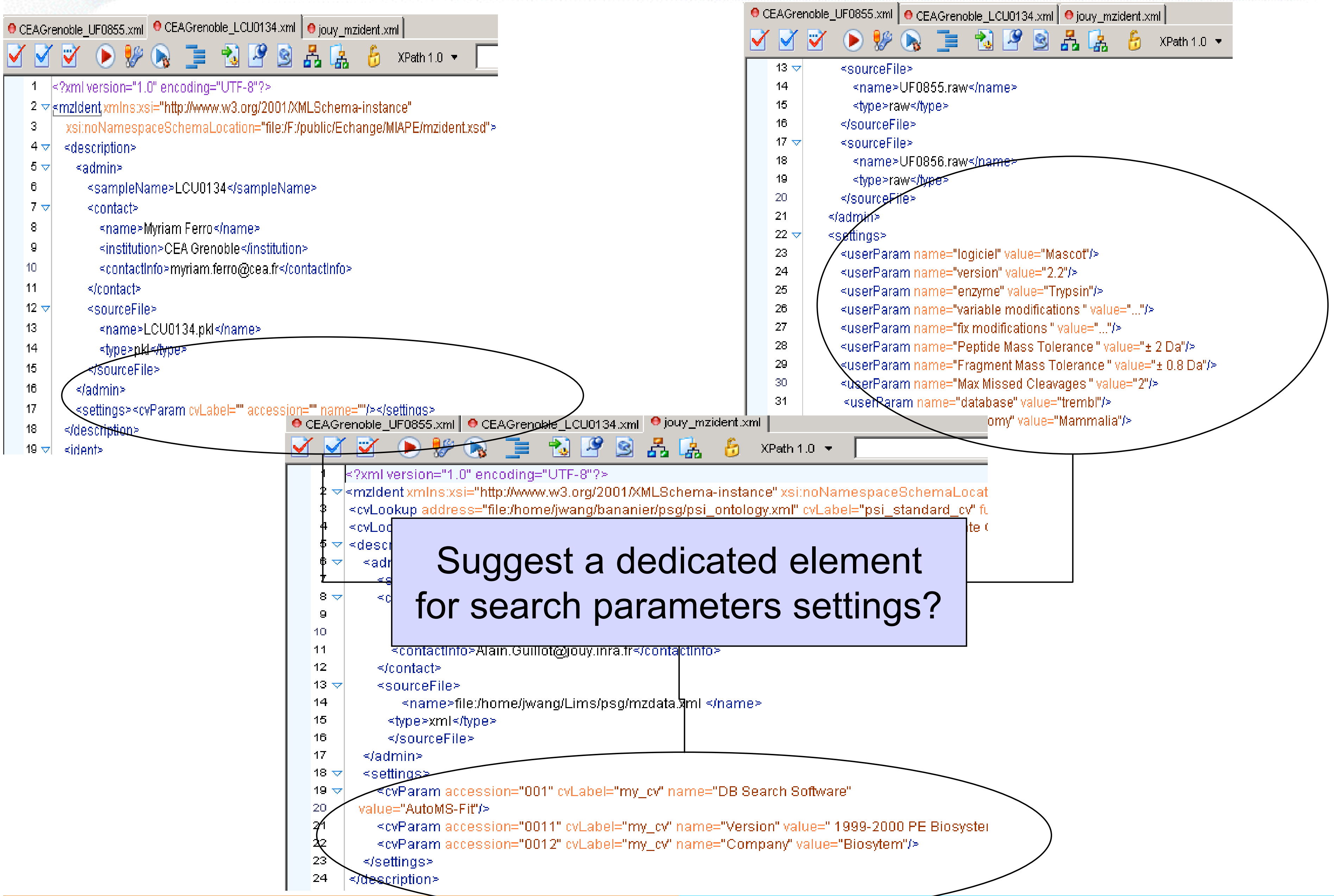
```
<?xml version="1.0" encoding="UTF-8"?>
<mzIdent xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="file:/F:/public/Echange/MIAPE/mzident.xsd">
  <description>
    <admin>
      <sampleName>LCU0134</sampleName>
      <contact>
        <name>Myriam Ferro</name>
        <institution>CEA Grenoble</institution>
        <contactInfo>myriam.ferro@cea.fr</contactInfo>
      </contact>
      <sourceFile>
        <name>LCU0134.pkl</name>
        <type>pkl</type>
      </sourceFile>
    </admin>
    <settings><cvParam cvLabel="" accession="" name=""/></settings>
  </description>
  <ident>
```

```
        <sourceFile>
          <name>UF0855.raw</name>
          <type>raw</type>
        </sourceFile>
        <sourceFile>
          <name>UF0856.raw</name>
          <type>raw</type>
        </sourceFile>
      </admin>
      <settings>
        <userParam name="logiciel" value="Mascot"/>
        <userParam name="version" value="2.2"/>
        <userParam name="enzyme" value="Trypsin"/>
        <userParam name="variable modifications " value="..."/>
        <userParam name="fix modifications " value="..."/>
        <userParam name="Peptide Mass Tolerance " value="± 2 Da"/>
        <userParam name="Fragment Mass Tolerance " value="± 0.8 Da"/>
            aram name="Max Missed Cleavages " value="2"/>
            aram name="database" value="trembl"/>
            aram name="taxonomy" value="Mammalia"/>
```

Source file heterogeneity allowed? (format and file location)

```
<?xml version="1.
<mzIdent xmlns:xs                                                aceSchemaLocat
<cvLookup address="file:/home/jwang/bananier/psg/psi_ontology.xml" cvLabel="psi_standard_cv" fu
<cvLookup address="file:/home/jwang/Lims/psg/my_cv.xml" cvLabel="my_cv" fullName="My Private (
<description>
  <admin>
    <sampleName>echant_A11_b_0001</sampleName>
    <contact>
      <name>Alain Guillot </name>
      <institution>INRA, Domaine de Vilvert, 78352 Jouy en Josas, France</institution>
      <contactInfo>Alain.Guillot@jouy.inra.fr</contactInfo>
    </contact>
    <sourceFile>
        <name>file:/home/jwang/Lims/psg/mzdata.xml </name>
      <type>xml</type>
      </sourceFile>
  </admin>
  <settings>
    <cvParam accession="001" cvLabel="my_cv" name="DB Search Software"
 value="AutoMS-Fit"/>
    <cvParam accession="0011" cvLabel="my_cv" name="Version" value=" 1999-2000 PE Biosyste
    <cvParam accession="0012" cvLabel="my_cv" name="Company" value="Biosytem"/>
  </settings>
</description>
```

Mzident instance : few question remarks (2)



```
<?xml version="1.0" encoding="UTF-8"?>
<mzIdent xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="file:/F:/public/Echange/MIAPE/mzident.xsd">
  <description>
    <admin>
      <sampleName>LCU0134</sampleName>
      <contact>
        <name>Myriam Ferro</name>
        <institution>CEA Grenoble</institution>
        <contactInfo>myriam.ferro@cea.fr</contactInfo>
      </contact>
      <sourceFile>
        <name>LCU0134.pkl</name>
        <type>pkl</type>
      </sourceFile>
    </admin>
    <settings><cvParam cvLabel="" accession="" name=""/></settings>
  </description>
  <ident>
```

```
      <sourceFile>
        <name>UF0855.raw</name>
        <type>raw</type>
      </sourceFile>
      <sourceFile>
        <name>UF0856.raw</name>
        <type>raw</type>
      </sourceFile>
    </admin>
    <settings>
      <userParam name="logiciel" value="Mascot"/>
      <userParam name="version" value="2.2"/>
      <userParam name="enzyme" value="Trypsin"/>
      <userParam name="variable modifications " value="..."/>
      <userParam name="fix modifications " value="..."/>
      <userParam name="Peptide Mass Tolerance " value="± 2 Da"/>
      <userParam name="Fragment Mass Tolerance " value="± 0.8 Da"/>
      <userParam name="Max Missed Cleavages " value="2"/>
       <userParam name="database" value="trembl"/>
                                        omy" value="Mammalia"/>
```

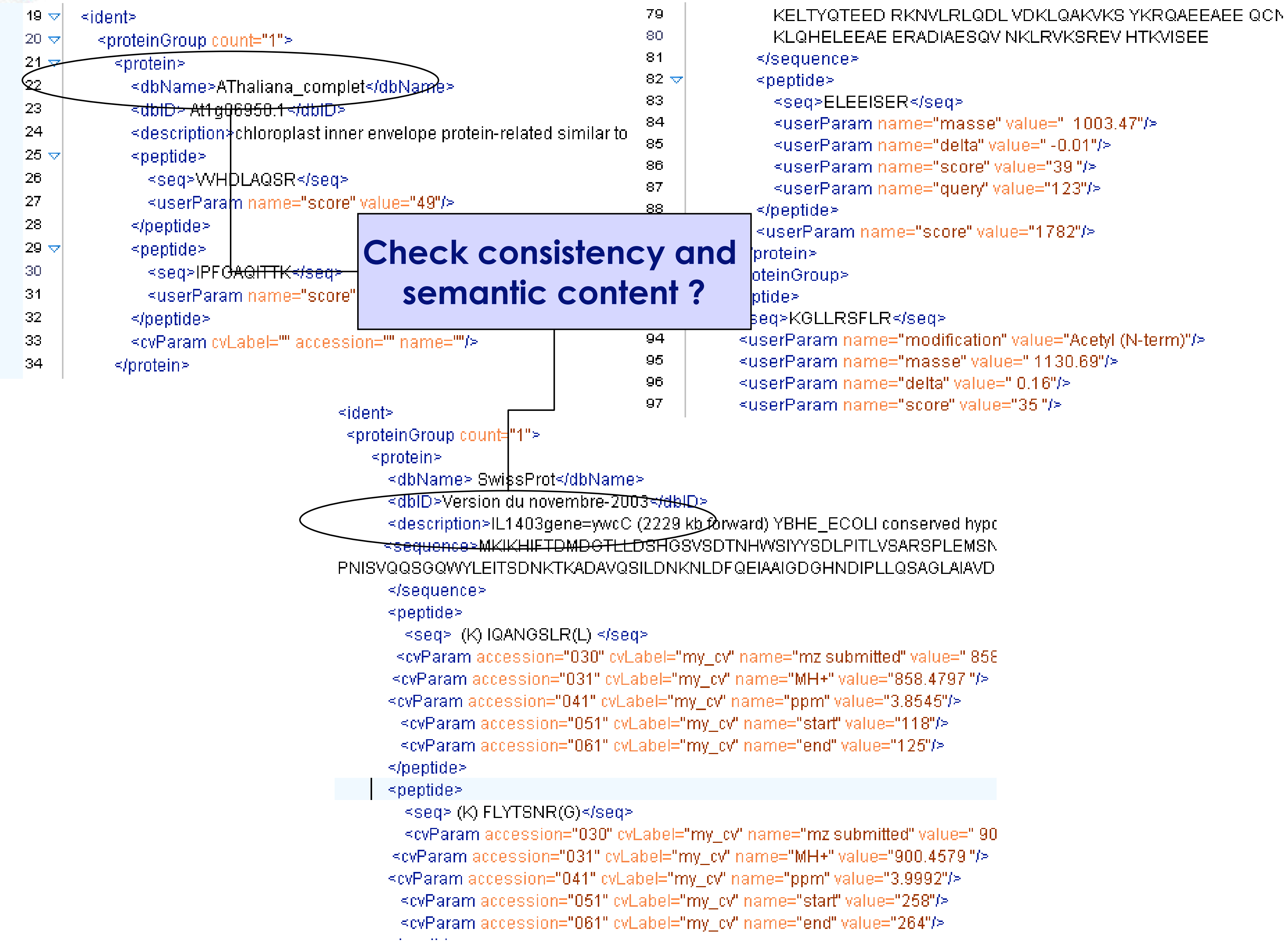
```
<?xml version="1.0" encoding="UTF-8"?>
<mzIdent xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSchemaLocat
 <cvLookup address="file:/home/jwang/bananier/psg/psi_ontology.xml" cvLabel="psi_standard_cv" fu
 <cvLoc                                                                                      te (
 <descr
   <adr
     <s
     <c



         <contactInfo>Alain.Guillot@jouy.inra.fr</contactInfo>
     </contact>
     <sourceFile>
         <name>file:/home/jwang/Lims/psg/mzdata.xml </name>
       <type>xml</type>
       </sourceFile>
   </admin>
   <settings>
     <cvParam accession="001" cvLabel="my_cv" name="DB Search Software"
 value="AutoMS-Fit"/>
     <cvParam accession="0011" cvLabel="my_cv" name="Version" value=" 1999-2000 PE Biosyste
     <cvParam accession="0012" cvLabel="my_cv" name="Company" value="Biosytem"/>
   </settings>
 </description>
```

Suggest a dedicated element for search parameters settings?

Mzident instance : few question remarks (3)



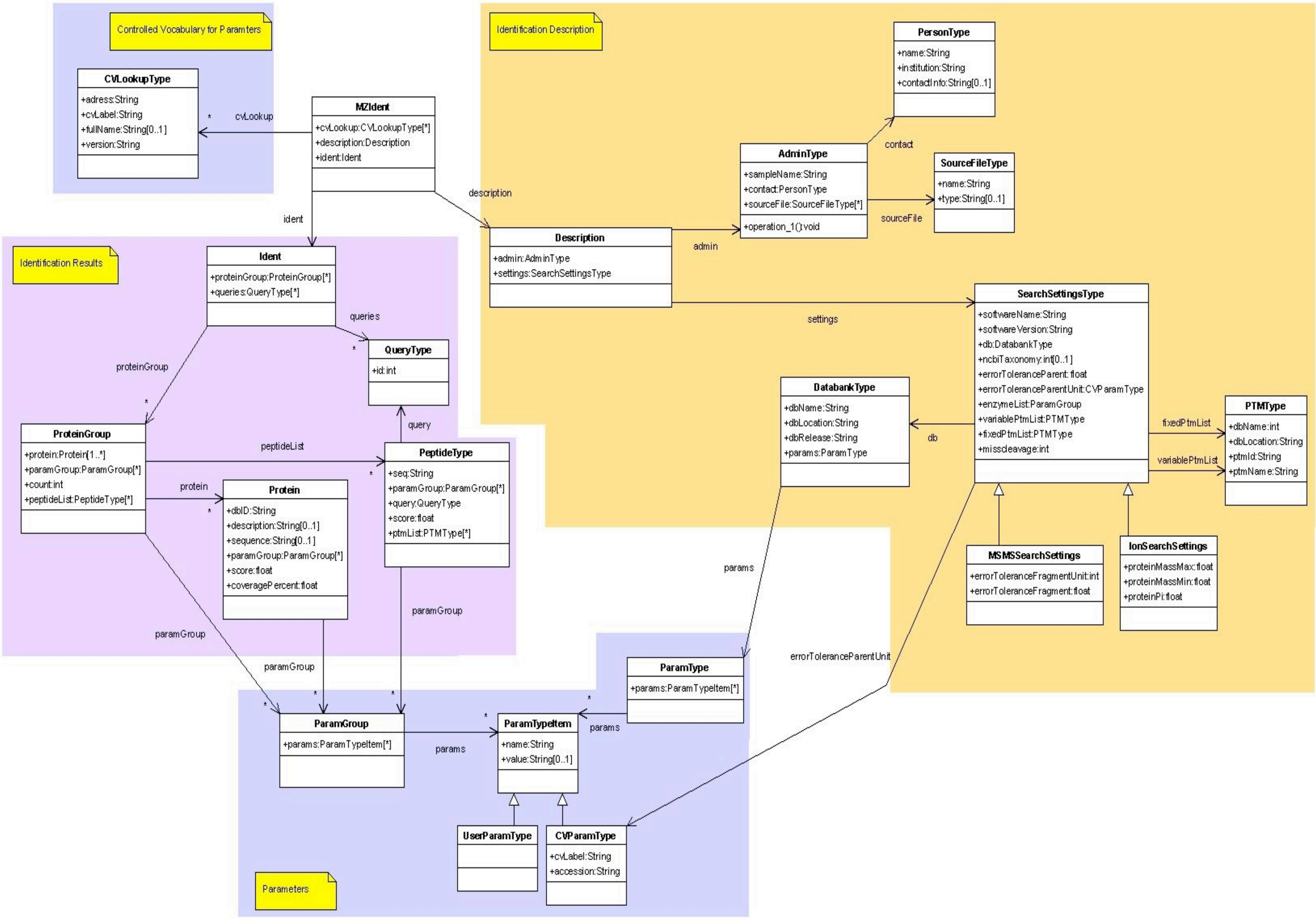
Check consistency and semantic content ?

```
<ident>
  <proteinGroup count="1">
    <protein>
      <dbName>AThaliana_complet</dbName>
      <dbID> At1g06950.1 </dbID>
      <description>chloroplast inner envelope protein-related similar to
      <peptide>
        <seq>VVHDLAQSR</seq>
        <userParam name="score" value="49"/>
      </peptide>
      <peptide>
        <seq>IPFGAQITTK</seq>
        <userParam name="score"
      </peptide>
      <cvParam cvLabel="" accession="" name=""/>
    </protein>
```

```
      KELTYQTEED RKNVLRLQDL VDKLQAKVKS YKRQAEEAEE QCN
      KLQHELEEAE ERADIAESQV NKLRVKSREV HTKVISEE
    </sequence>
    <peptide>
      <seq>ELEEISER</seq>
      <userParam name="masse" value="  1003.47"/>
      <userParam name="delta" value=" -0.01"/>
      <userParam name="score" value="39 "/>
      <userParam name="query" value="123"/>
    </peptide>
    <userParam name="score" value="1782"/>
  /protein>
oteinGroup>
ptide>
seq>KGLLRSFLR</seq>
  <userParam name="modification" value="Acetyl (N-term)"/>
  <userParam name="masse" value=" 1130.69"/>
  <userParam name="delta" value=" 0.16"/>
  <userParam name="score" value="35 "/>
```

```
<ident>
 <proteinGroup count="1">
     <protein>
       <dbName> SwissProt</dbName>
       <dbID>Version du novembre-2003</dbID>
       <description>IL1403gene=ywcC (2229 kb forward) YBHE_ECOLI conserved hypo
      <sequence>MKIKHIFTDMDGTLLDSHGSVSDTNHWSIYYSDLPITLVSARSPLEMSN
PNISVQQSGQWYLEITSDNKTKADAVQSILDNKNLDFQEIAAIGDGHNDIPLLQSAGLAIAVD
       </sequence>
       <peptide>
         <seq>  (K) IQANGSLR(L) </seq>
        <cvParam accession="030" cvLabel="my_cv" name="mz submitted" value=" 858
       <cvParam accession="031" cvLabel="my_cv" name="MH+" value="858.4797 "/>
       <cvParam accession="041" cvLabel="my_cv" name="ppm" value="3.8545"/>
         <cvParam accession="051" cvLabel="my_cv" name="start" value="118"/>
         <cvParam accession="061" cvLabel="my_cv" name="end" value="125"/>
       </peptide>
       <peptide>
         <seq> (K) FLYTSNR(G)</seq>
         <cvParam accession="030" cvLabel="my_cv" name="mz submitted" value=" 90
       <cvParam accession="031" cvLabel="my_cv" name="MH+" value="900.4579 "/>
       <cvParam accession="041" cvLabel="my_cv" name="ppm" value="3.9992"/>
         <cvParam accession="051" cvLabel="my_cv" name="start" value="258"/>
         <cvParam accession="061" cvLabel="my_cv" name="end" value="264"/>
```

Proposition de schéma pour l'identification des protéines

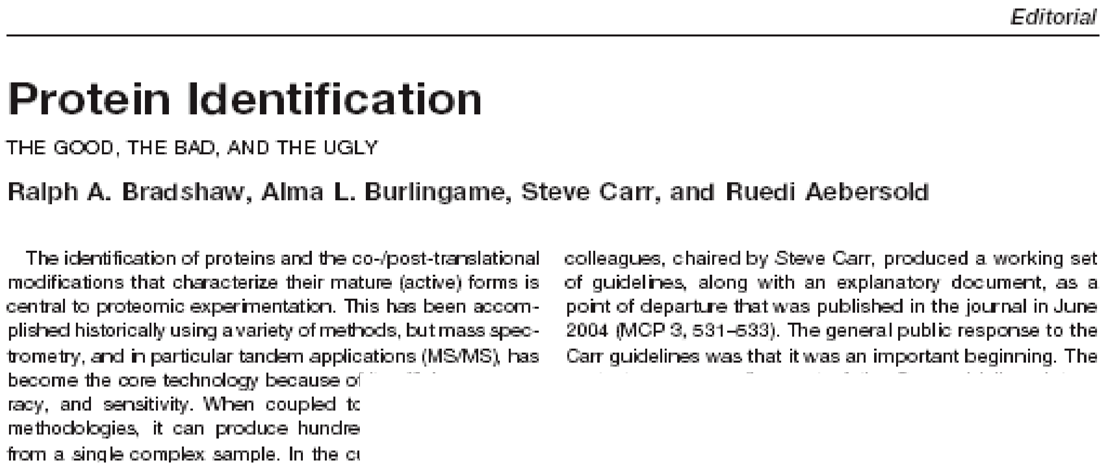


Created with Poseidon for UML Community Edition. Not for Commercial Use.

French Proteomics Standards Group : actions

- January 05 : EBI contact (PSI coordinators)
- March 05 : Test case with mzdata and mzIdent standard
- April 05 : feedback and suggestion available for community mzIdent test case report and sample files
- Sept 05 : Geneva PSI meeting
- Oct-Dec 05 : MCP and MIAPE guideline mapping (coordinated to PA Binz)

Mol. Cell. Proteomics and MIAPE comparison



Editorial

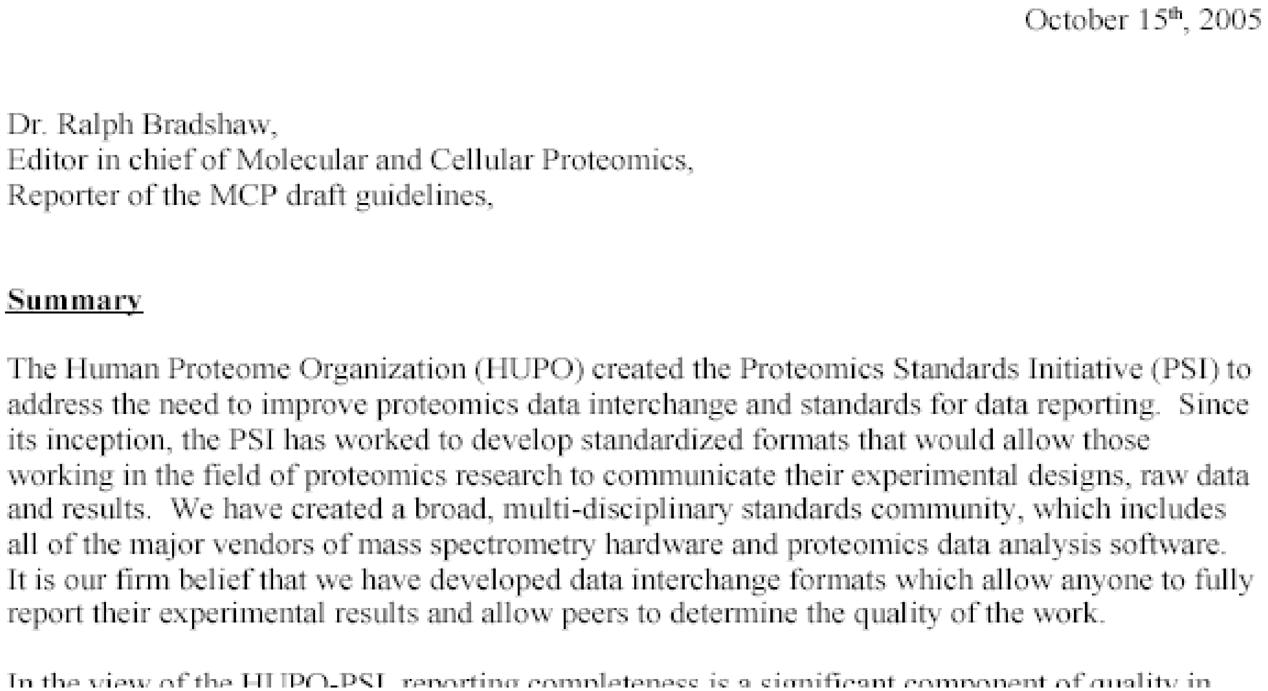
Protein Identification

THE GOOD, THE BAD, AND THE UGLY

Ralph A. Bradshaw, Alma L. Burlingame, Steve Carr, and Ruedi Aebersold

The identification of proteins and the co-/post-translational modifications that characterize their mature (active) forms is central to proteomic experimentation. This has been accomplished historically using a variety of methods, but mass spectrometry, and in particular tandem applications (MS/MS), has become the core technology because o racy, and sensitivity. When coupled t methodologies, it can produce hundre from a single complex sample. In the c

colleagues, chaired by Steve Carr, produced a working set of guidelines, along with an explanatory document, as a point of departure that was published in the journal in June 2004 (MCP 3, 531–533). The general public response to the Carr guidelines was that it was an important beginning. The



October 15th, 2005

Dr. Ralph Bradshaw,
Editor in chief of Molecular and Cellular Proteomics,
Reporter of the MCP draft guidelines,

Summary

The Human Proteome Organization (HUPO) created the Proteomics Standards Initiative (PSI) to address the need to improve proteomics data interchange and standards for data reporting. Since its inception, the PSI has worked to develop standardized formats that would allow those working in the field of proteomics research to communicate their experimental designs, raw data and results. We have created a broad, multi-disciplinary standards community, which includes all of the major vendors of mass spectrometry hardware and proteomics data analysis software. It is our firm belief that we have developed data interchange formats which allow anyone to fully report their experimental results and allow peers to determine the quality of the work.

In the view of the HUPO-PSI, reporting completeness is a significant component of quality in

French Proteomics Standards Group : actions

- January 05 : EBI contact (PSI coordinators)
- March 05 : Test case with mzdata and mzIdent standard
- April 05 : feedback and suggestion available for community mzIdent test case report and sample files
- Sept 05 : Geneva PSI meeting
- Oct-Dec 05 : MCP and MIAPE guideline mapping (coordinated by PA Binz)
- Oct 05- janv06 : PSI-MS requirements (sent to K. Laursen)
- Dec05-Mars06 : Miape compliancy & proteomics server prototype (D. Ziebelin)
- Juin06-Oct06 : mzData tools evaluation (BIM)

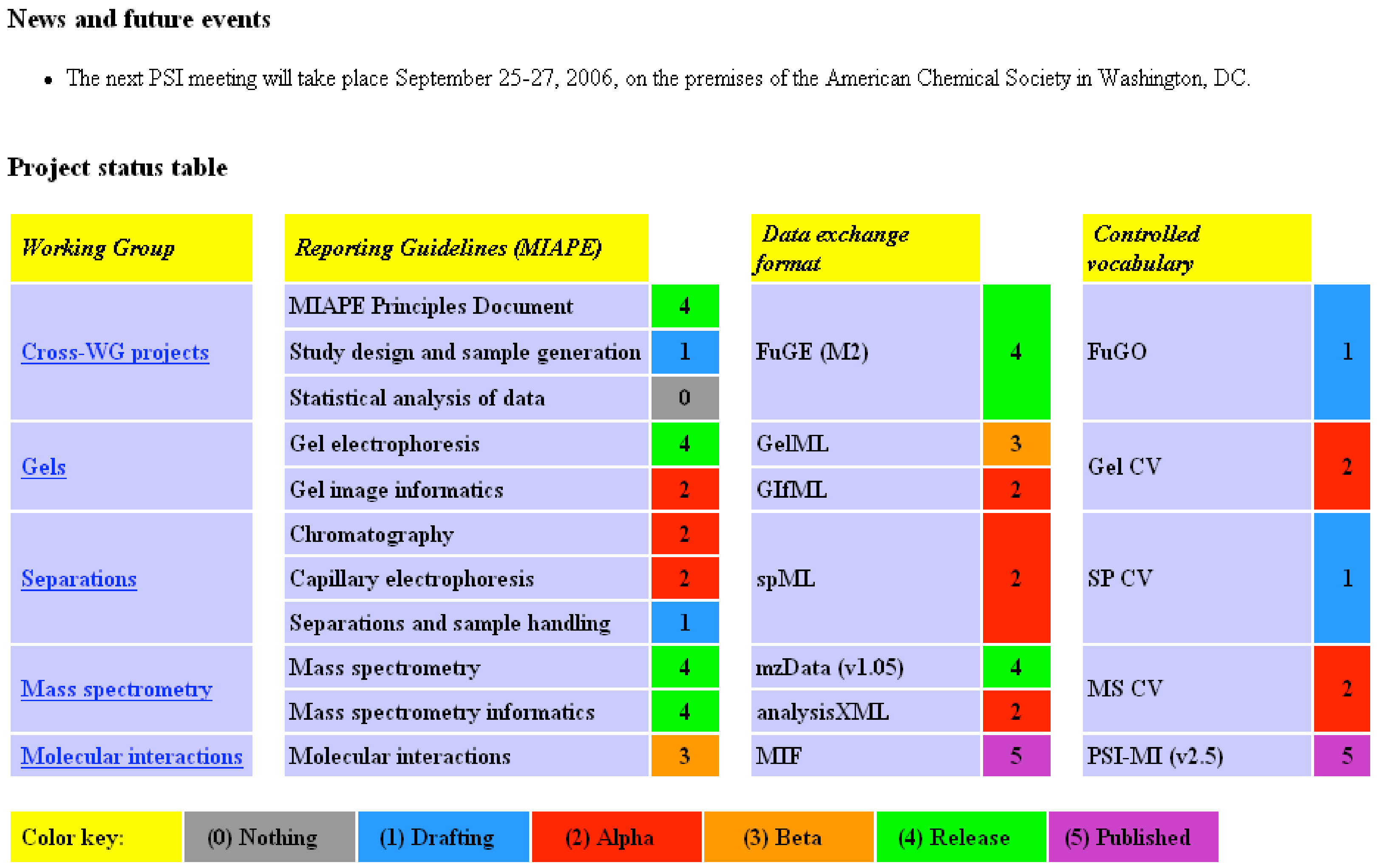
mzData implementation : converters, parsers, viewers, analysis…

Support for mzData (listed alphabetically by organisation). *Last updated 2006-05-23.*

	Project status key:	*U* = Unknown	*A* = Alpha	*B* = Beta	*R* = Release

	Organisation	*Product*	*Version (date)*	*Contact person*
U	Agilent Technologies	Spectrum Mill	? (?)	?
R	Applied Biosystems | MDS Sciex	QSTAR and Q TRAP wiff converters	? (?)	Support link
	.wiff format converters for the QSTAR and Q TRAP. [N.B. For 4700/4800 support, see proteomecommons.org*]*			
B	Babraham Bioinformatics	mzViewer	0.4 (2005-12-20)	Support link
	Viewer for mzData-encoded mass spectra			
B	Bioinformatics Solutions Inc.	Peaks	4.1 (*t.b.a.*)	Support link
	MS/MS analysis software for qualitative/quantitative proteomics. Any complete analysis involves de novo.			
R	Bruker Daltonics / Protagen	ProteinScape	1.3 (2005-12-01)	Herbert Thiele
R	EBI / University of Gent	PRIDE - PRoteomics IDEntifications DB	core: 2.131 web: 2.161 (2005-07-29)	pride-support@ebi.ac.uk
	The PRIDE PRoteomics IDEntifications database is a centralized, standards compliant, public data repository for proteomics data. It has been developed to provide the proteomics community with a public repository for protein and peptide identifications together with the evidence supporting these identifications			

PSI : news and future (april 2006)



News and future events

- The next PSI meeting will take place September 25-27, 2006, on the premises of the American Chemical Society in Washington, DC.

Project status table

Working Group	*Reporting Guidelines (MIAPE)*		*Data exchange format*		*Controlled vocabulary*	
Cross-WG projects	MIAPE Principles Document	4	FuGE (M2)	4	FuGO	1
	Study design and sample generation	1				
	Statistical analysis of data	0				
Gels	Gel electrophoresis	4	GelML	3	Gel CV	2
	Gel image informatics	2	GIfML	2		
Separations	Chromatography	2	spML	2	SP CV	1
	Capillary electrophoresis	2				
	Separations and sample handling	1				
Mass spectrometry	Mass spectrometry	4	mzData (v1.05)	4	MS CV	2
	Mass spectrometry informatics	4	analysisXML	2		
Molecular interactions	Molecular interactions	3	MIF	5	PSI-MI (v2.5)	5

Color key:	(0) Nothing	(1) Drafting	(2) Alpha	(3) Beta	(4) Release	(5) Published

Structured process, correspondants well-identified, transparency, mailing lists, documents availability, call for contributions…

French Proteomics Standards Group : Perspectives

PSG : Groupe identifié et fonctionnel avec une force de proposition et des actions en cours

1. La définition de sous-groupe thématique au sein du PSG de manière à mener des actions concrètes
2. L'identification de relais au sein du PSG vis-à-vis des différents groupes de travaux PSI

PAB : les attentes immédiates du PSI se traduisent par des retours circonstanciés sur les standards existants permettant de les valider/amender/améliorer et une participation accrue, en particulier au sein des mailing-lists, confcall ou encore lors du prochain Workshop PSI.

=> Nature Methods editorial, June 06

Table ronde : attentes, intérêts et volonté de contributions actives…?