

HUPO PSI PI WG Teleconference 1500 GMT 01 August 06

AGENDA

1. Search engine output parameter analysis spreadsheet
2. Powerpoint representation of draft AnalysisXML UML model
3. AOB:
 - 3.i. SourceForge space quota
 - 3.ii. Agenda item for next teleconference
4. Dates of future teleconferences

MINUTES

PRESENT

Angel Pizarro, University of Pennsylvania (AP) Chair
Lennart Martens, EBI (LM) Teleconference organiser & minutes
Kent Larsen, Indigo Biosystems (KL)
Ruth McNally, ESRC CESAGen (RM) Minutes
Puneet (Pete) Souda, UCLA (PS)

APOLOGIES

Philip Jones, EBI (PJ) Secretary

1. SEARCH ENGINE OUTPUT PARAMETER ANALYSIS SPREADSHEET

Since last Teleconference (6 June 06), David Creasy (DC) has posted an updated spreadsheet on the docstore. (PI WG Teleconferences scheduled for 4 & 11 July 06 were postponed.)

http://psidev.sourceforge.net/docstore/view.php?&action=xls_show&id=84

This spreadsheet will be the basis for drafting the requirements document for AnalysisXML.

AP asked the Teleconference whether any major features were missing from the spreadsheet. None was identified.

LM suggested that once the first draft of the spreadsheet is finalised, the PI WG should send it to Jimmy Eng (SEQUENT) and Ron Beavis (X!Tandem) (and other search engine experts) for their critical evaluation. The Teleconference agreed with the suggestion.

Action:

AP to communicate with DC by email re progress towards finalisation of the first draft of the spreadsheet.

2. POWER POINT REPRESENTATION OF DRAFT AnalysisXML UML MODEL

The PowerPoint representation of the model can be viewed at:

<http://psidev.sourceforge.net/proteomics-informatics/documents/analysisUML.ppt>

AP explained that the PowerPoint representation is just illustrative, and apologized for not yet circulating the actual UML model. He invited the Teleconference to discuss the UML model, as represented by Power Point.

KL asked **AP** to explain the model for the benefit of the psi ms wg, which is designing mzData v. 2.0.

AP explained the three PowerPoint slides.

Slide 1 represents the structure of an external spectral data file, indicating what AnalysisXML will require from mzData 2.0.

Slide 2 represents a UML model of the relationships between the elements in the AnalysisXML schema that was derived from a merger of ISB's PepXML and ProtXML.

The root element is 'AnalysisXML' (on LHS). This is linked to:

- 'Polypeptide' which is a lookup table of what was found;
- '*ProtocolApplication*' which captures: 'We did this 'Search' and our 'SearchHypothesis' (e.g. proposed identification) is that we found peptide X at spectra no. Y.
- '*Protocol*' where you can reference the 'SearchProtocol' used, e.g. the SEQUEST params file, or one's own customised computational search algorithms.
- 'SequenceDatabase' – **AP** thinks this captures the database that was searched – it may contain a lookup table of databases.

Slide 3 is a subschema that can call in superstructures of protocols.

PJ had sent email requesting clarification on what information was mandatory. This email, together with responses from **DC**, can be found at:

http://sourceforge.net/mailarchive/forum.php?forum_id=48874&max_rows=25&style=nested&viewmonth=200607

AP responded to **PJ**'s requests regarding what information provision was mandatory as follows:

1. PolyPeptide

- Sequence?

AP: Yes

- Database name /version?

AP: Yes, by way of association with FuGE

- Accession / accession version?

AP: Is possible to specify within the schema, but not sure if they should be mandatory because not always available, e.g. cases where the protein is a synthetic construct.

2. SearchProtocol

- Search engine identity
- version
- input parameters / settings?

AP: Yes - via FuGE protocol subclass that gives a generic parameters attribute as well as a software attribute

3. Protein modifications – CustomModifications class

- monoisotopicMass & averageMass values?

AP: Needs to be worked out with psi pm wg

4. Protein modifications – Modifications class

- Position?

AP: Perhaps should be possible to include where available, but not mandatory. It is likely that AnalysisXML will use FuGE identifiers to reference (big) items in other FuGE-derived documents.

KL asked for comment regarding DC's email comment about minimising the use of CVs in the schema. **AP** responded that DC's spreadsheet should permit hard-coding of the schema, but unlikely that it would ever be possible to do away with the use of a CV entirely.

Action:

AP To solicit off-line work on the UML model via email

3. ANY OTHER BUSINESS

3.i. SOURCEFORGE SPACE QUOTA

LM reported that the PSI is currently evaluating different options for the PSI website. **AP** is of the opinion that memory is cheap and Sourceforge is slow, and so PSI should seek alternative host to Sourceforge.

3.ii. AGENDA ITEM FOR NEXT TELECONFERENCE

PS suggested that the next Teleconference should discuss plans for HUPO Congress 2006. It was agreed.

4. DATES OF NEXT TELECONFERENCES

MS WG: 1500 GMT 15 August 2006; PI WG: 1500 GMT 29 August 2006