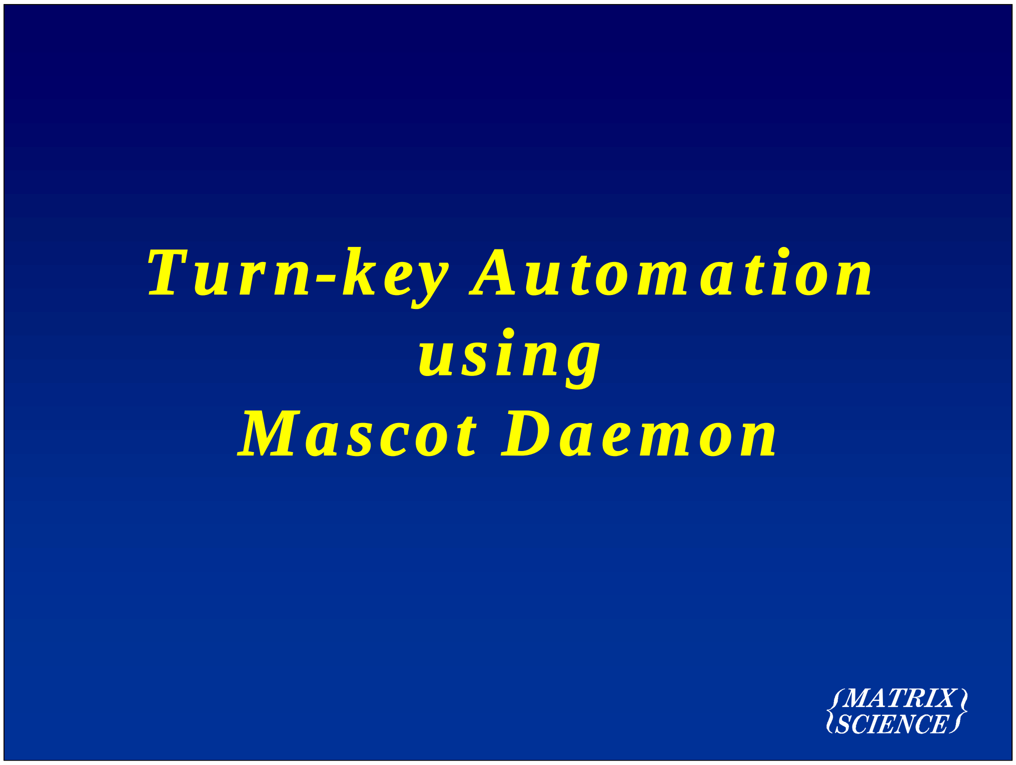
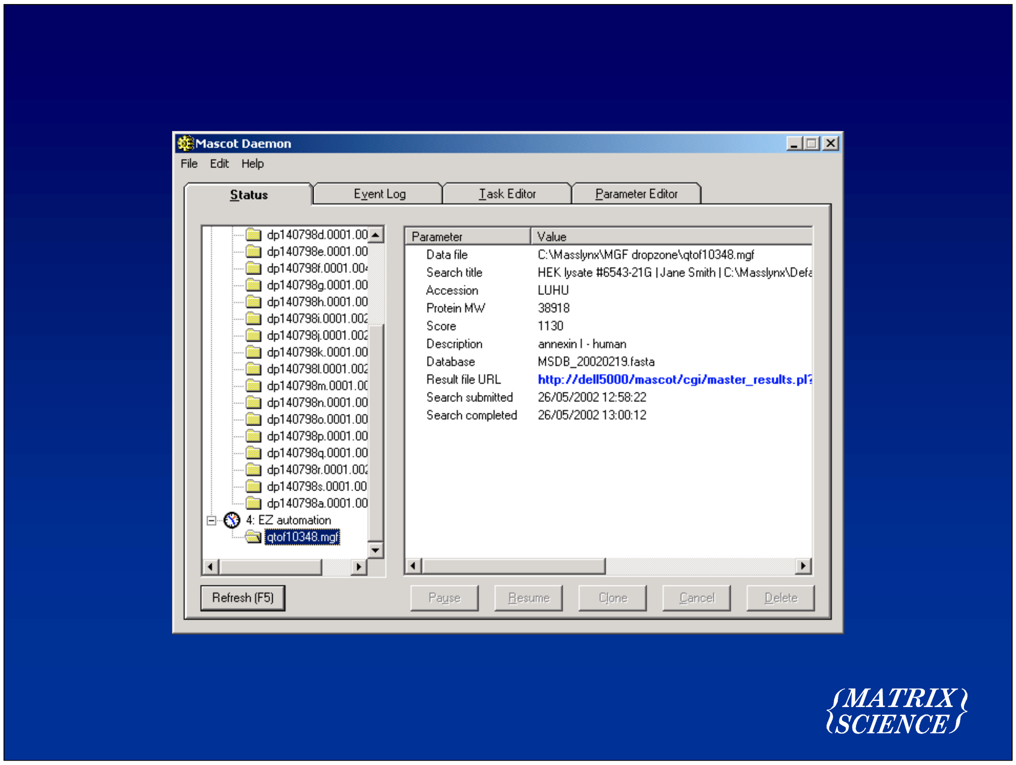


Turn-key Automation using Mascot Daemon

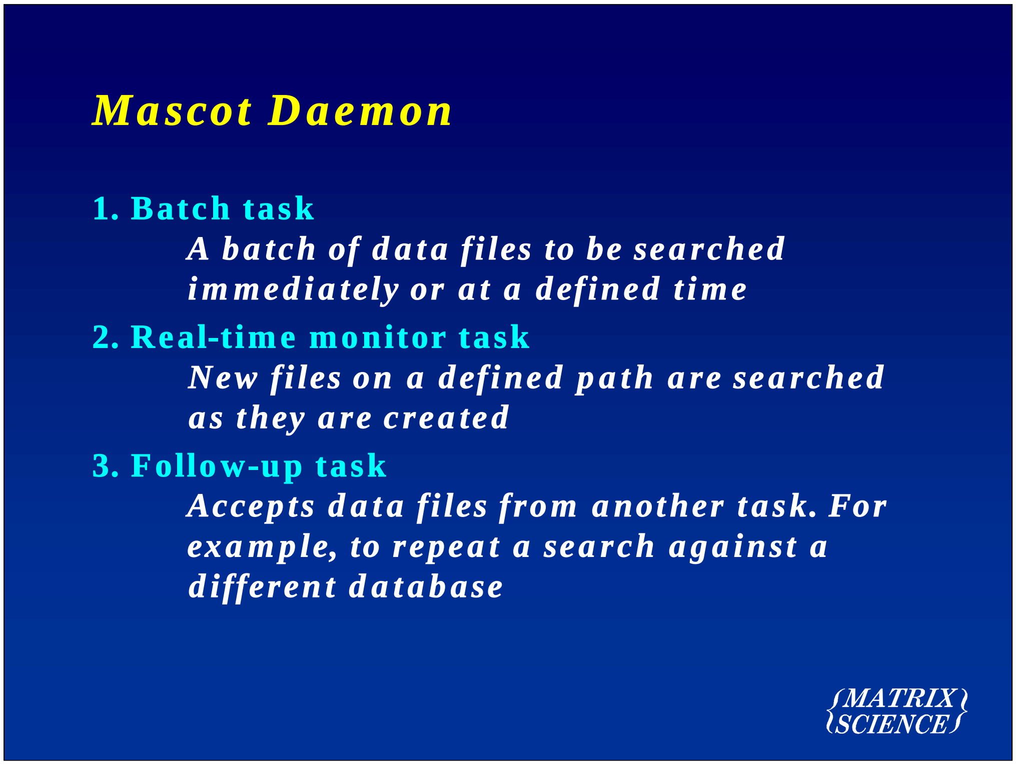
MATRIX SCIENCE



dp140798d.0001.00
dp140798e.0001.00
dp140798f.0001.00
dp140798g.0001.00
dp140798h.0001.00
dp140798i.0001.002
dp140798j.0001.002
dp140798k.0001.00
dp140798l.0001.002
dp140798m.0001.00
dp140798n.0001.00
dp140798o.0001.00
dp140798p.0001.00
dp140798q.0001.00
dp140798r.0001.002
dp140798s.0001.00
dp140798a.0001.00
4: EZ automation
qtof10348.mgf

Parameter	Value
Data file	C:\Masslynx\MGF dropzone\qtof10348.mgf
Search title	HEK lysate #6543-21G | Jane Smith | C:\Masslynx\Defa
Accession	LUHU
Protein MW	38918
Score	1130
Description	annexin I - human
Database	MSDB_20020219.fasta
Result file URL	http://dell5000/mascot/cgi/master_results.pl?
Search submitted	26/05/2002 12:58:22
Search completed	26/05/2002 13:00:12

Mascot Daemon is our automation client. It's a part of the Mascot package that is only available when you have Mascot on an in-house server



Mascot Daemon

1. **Batch task**

   A batch of data files to be searched immediately or at a defined time

2. **Real-time monitor task**

   New files on a defined path are searched as they are created

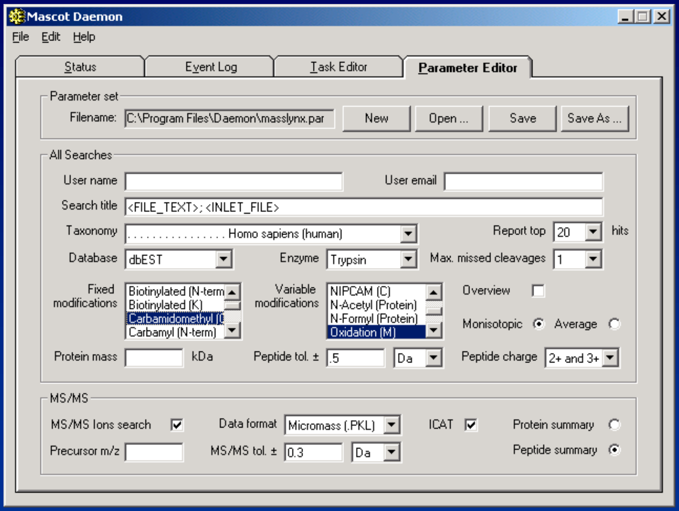
3. **Follow-up task**

   Accepts data files from another task. For example, to repeat a search against a different database

MATRIX SCIENCE

Mascot Daemon runs on any Win32 platform and supports three kinds of tasks.

The follow-up task is very powerful because it allows searches to be chained together to implement complex decision paths. For example, as batch of data files might be screened against a contaminants database containing entries for keratins, BSA, trypsin, etc. Those data files which fail to find a match can then be automatically searched against a non-redundant protein database. Spectra which are still unmatched can then be searched against a large EST database, etc., etc.

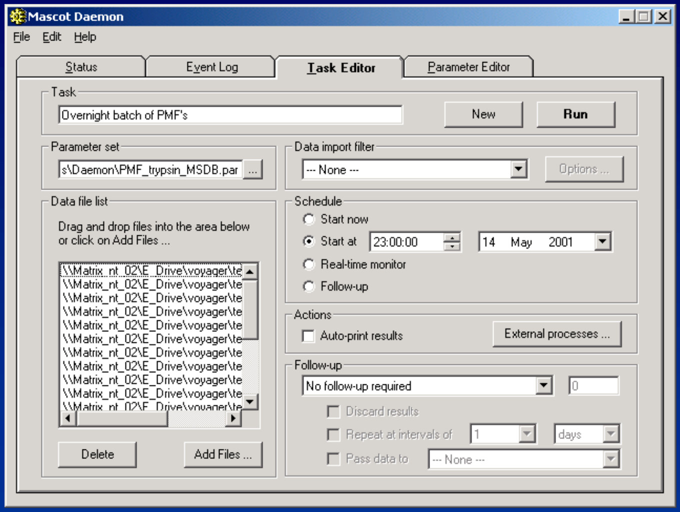


Parameter set
Filename: C:\Program Files\Daemon\masslynx.par

All Searches
User name:
User email:
Search title <FILE_TEXT>; <INLET_FILE>
Taxonomy: . . . . Homo sapiens (human)
Report top: 20 hits
Database: dbEST
Enzyme: Trypsin
Max. missed cleavages: 1
Fixed modifications: Biotinylated (N-term, Biotinylated (K), Carbamidomethyl (C, Carbamyl (N-term)
Variable modifications: NIPCAM (C), N-Acetyl (Protein), N-Formyl (Protein), Oxidation (M)
Overview ☐
Monisotopic ● Average ○
Protein mass: kDa
Peptide tol. ± .5 Da
Peptide charge: 2+ and 3+

MS/MS
MS/MS Ions search ☑
Data format: Micromass (.PKL)
ICAT ☑
Protein summary ○
Precursor m/z:
MS/MS tol. ± 0.3 Da
Peptide summary ●

The parameter editor allows sets of search parameters to be defined and saved to disk, so that they can be used over and over again. The search parameters define *how* the data will be searched.



Task: Overnight batch of PMF's

Parameter set: s\Daemon\PMF_trypsin_MSDB.par

Data import filter: --- None ---

Data file list

Drag and drop files into the area below or click on Add Files ...

```
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
\\Matrix_nt_02\E_Drive\voyager\te
```

Schedule

○ Start now
● Start at 23:00:00 14 May 2001
○ Real-time monitor
○ Follow-up

Actions

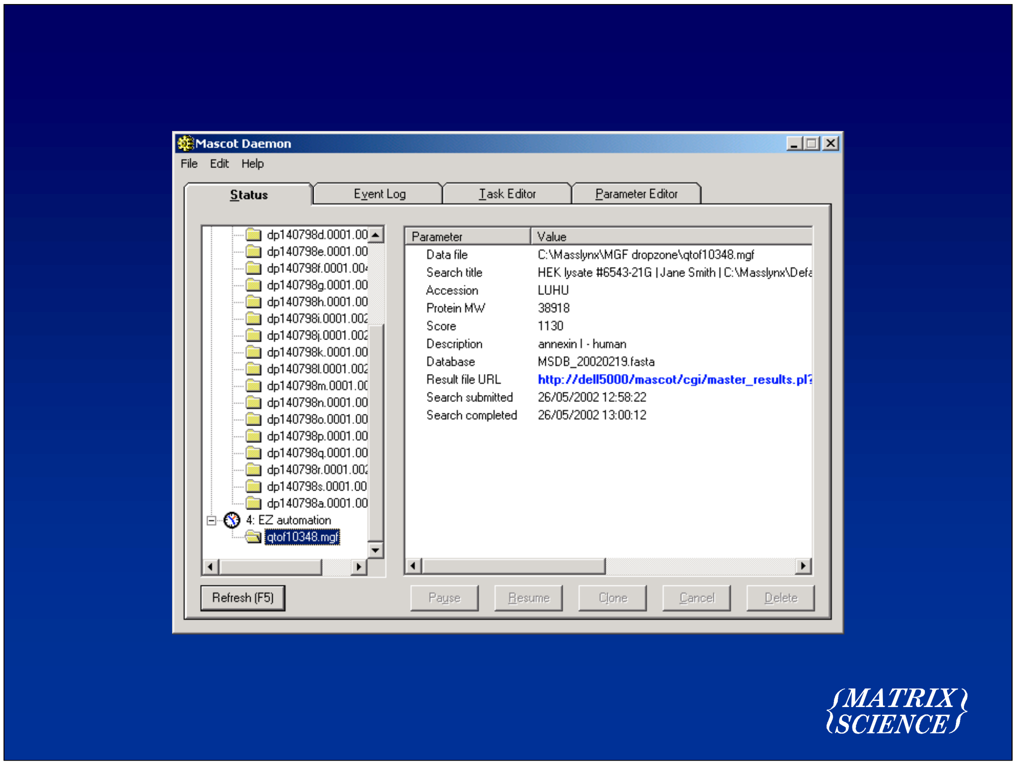
□ Auto-print results

Follow-up

No follow-up required 0
□ Discard results
□ Repeat at intervals of 1 days
□ Pass data to --- None ---

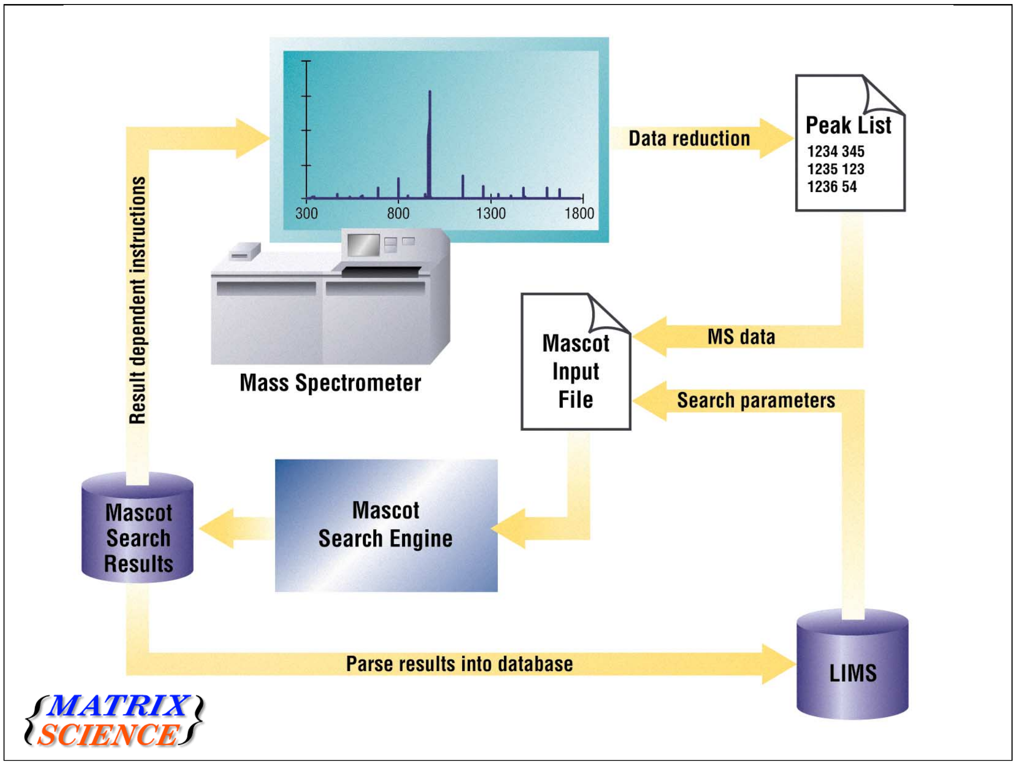
The Task Editor tab is used to define each task. A task defines *what* data will be searched and and *when* the search will take place.

Here we have a simple batch task. A set of data files has been created, a parameter set has been chosen, and the task will run all the searches as a batch at a predefined time.



Parameter	Value
Data file	C:\Masslynx\MGF dropzone\qtof10348.mgf
Search title	HEK lysate #6543-21G | Jane Smith | C:\Masslynx\Defa
Accession	LUHU
Protein MW	38918
Score	1130
Description	annexin I - human
Database	MSDB_20020219.fasta
Result file URL	http://dell5000/mascot/cgi/master_results.pl?
Search submitted	26/05/2002 12:58:22
Search completed	26/05/2002 13:00:12

Each completed search is added to an Explorer-like tree. There is a minimum of summary information, and a hyperlink to the full result report



Data reduction

Peak List
1234 345
1235 123
1236 54

Mass Spectrometer

MS data

Mascot Input File

Search parameters

Result dependent instructions

Mascot Search Results

Mascot Search Engine

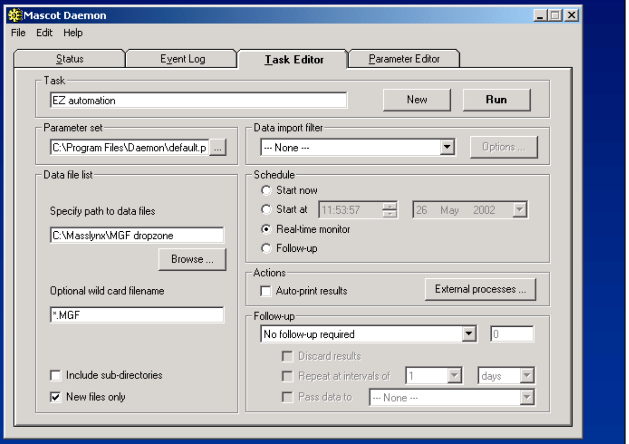
Parse results into database

LIMS

However, automation is more than just searching batches of files. This block diagram illustrates the workflow.

Automation: zero coding

Use Mascot Daemon in real-time monitor mode to search files as they are created



Task: EZ automation

Parameter set: C:\Program Files\Daemon\default.p

Data import filter: --- None ---

Data file list

Specify path to data files: C:\Masslynx\MGF dropzone

Optional wild card filename: *.MGF

□ Include sub-directories

☑ New files only

Schedule

○ Start now

○ Start at 11:53:57 26 May 2002

● Real-time monitor

○ Follow-up

Actions

□ Auto-print results

Follow-up: No follow-up required 0

□ Discard results

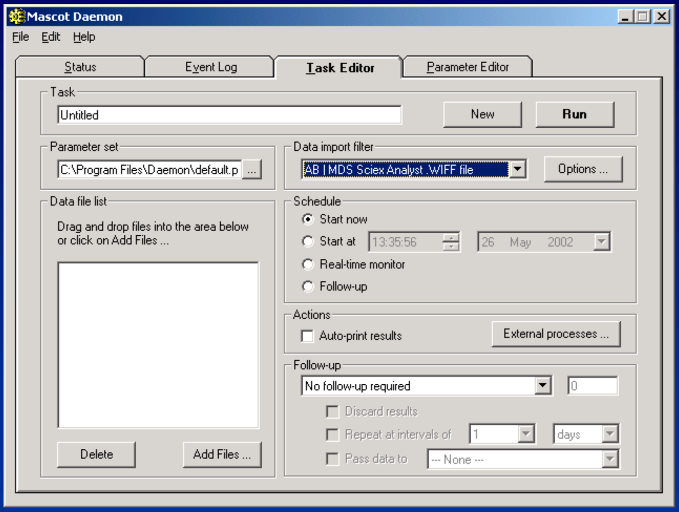
□ Repeat at intervals of 1 days

□ Pass data to --- None ---

MATRIX SCIENCE

How far can you get in implementing this workflow with a minimum of coding? In fact, what can you do without any coding?

The starting point is to use Mascot Daemon in real-time monitor mode. This will pick up peak lists as they are created, and search them automatically.



Task: Untitled

Parameter set: C:\Program Files\Daemon\default.p

Data import filter: AB | MDS Sciex Analyst .WIFF file

Data file list

Drag and drop files into the area below or click on Add Files ...

Schedule

- ● Start now
- ○ Start at 13:35:56 26 May 2002
- ○ Real-time monitor
- ○ Follow-up

Actions

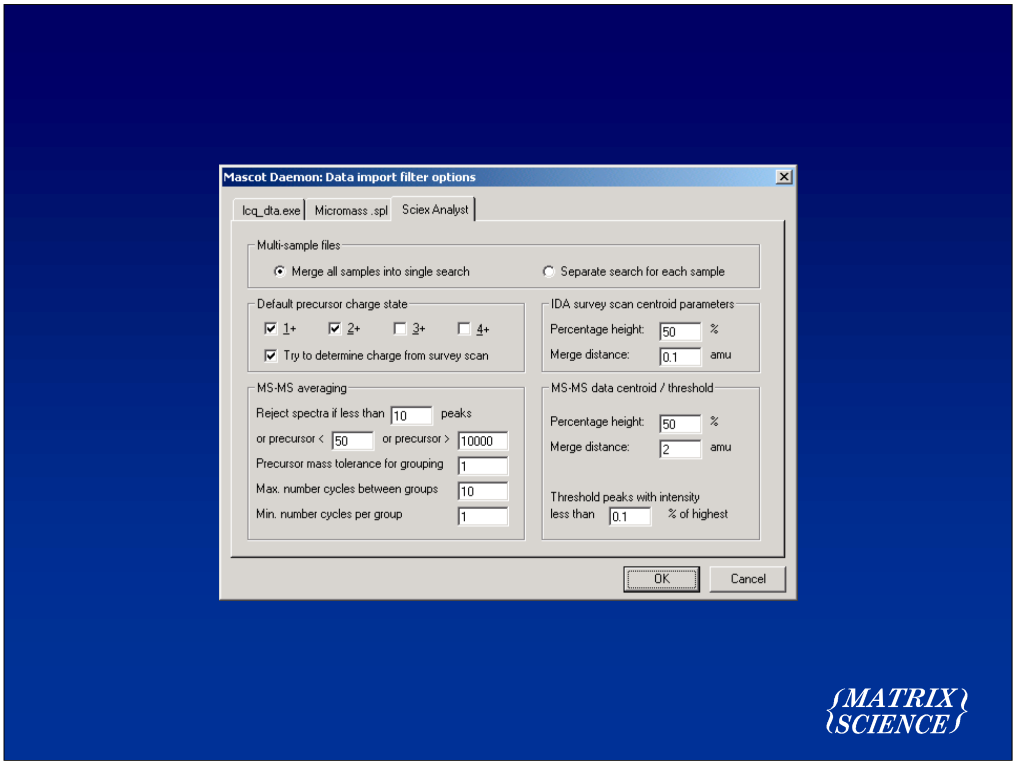
- □ Auto-print results

Follow-up: No follow-up required 0

- □ Discard results
- □ Repeat at intervals of 1 days
- □ Pass data to --- None ---

{MATRIX}
{SCIENCE}

For some file formats, we have data import filters. These allow Daemon to generate peak lists directly from the binary files created by the MS data system.



Mascot Daemon: Data import filter options

Multi-sample files

● Merge all samples into single search ○ Separate search for each sample

Default precursor charge state

☑ 1+ ☑ 2+ □ 3+ □ 4+

☑ Try to determine charge from survey scan

IDA survey scan centroid parameters

Percentage height: 50 %

Merge distance: 0.1 amu

MS-MS averaging

Reject spectra if less than 10 peaks

or precursor < 50 or precursor > 10000

Precursor mass tolerance for grouping 1

Max. number cycles between groups 10

Min. number cycles per group 1

MS-MS data centroid / threshold

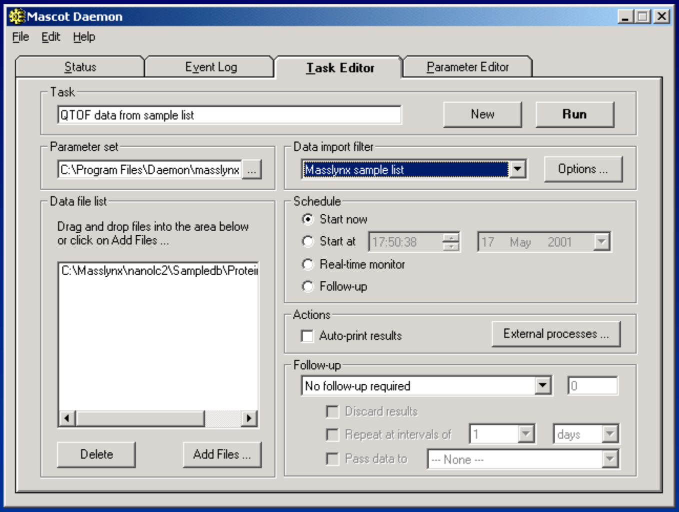
Percentage height: 50 %

Merge distance: 2 amu

Threshold peaks with intensity less than 0.1 % of highest

For example, the import filter for Sciex wiff files is fairly sophisticated.

We can also work directly from Xcalibur raw files and Applied Biosystems voyager data files



Task: QTOF data from sample list

Parameter set: C:\Program Files\Daemon\masslynx

Data import filter: Masslynx sample list

Data file list: C:\Masslynx\nanolc2\Sampledb\Protei

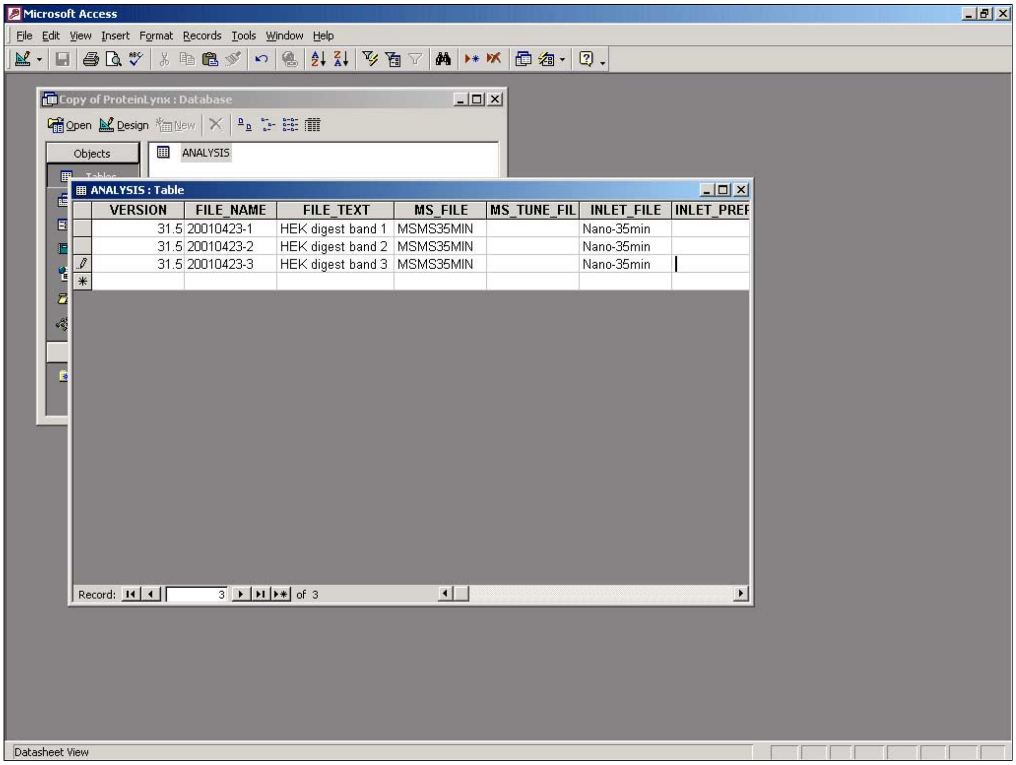
Schedule:
- ● Start now
- ○ Start at 17:50:38 17 May 2001
- ○ Real-time monitor
- ○ Follow-up

Actions:
- □ Auto-print results

Follow-up: No follow-up required 0
- □ Discard results
- □ Repeat at intervals of 1 days
- □ Pass data to --- None ---

MATRIX SCIENCE

The MassLynx sample list import filter is a little different.



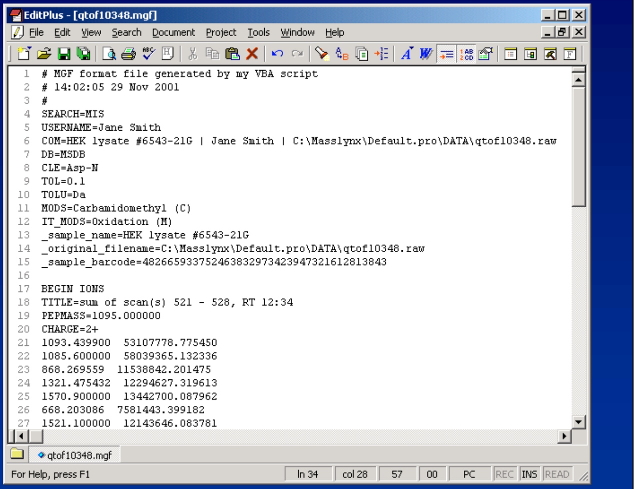
VERSION	FILE_NAME	FILE_TEXT	MS_FILE	MS_TUNE_FIL	INLET_FILE	INLET_PREF
31.5	20010423-1	HEK digest band 1	MSMS35MIN		Nano-35min	
31.5	20010423-2	HEK digest band 2	MSMS35MIN		Nano-35min	
31.5	20010423-3	HEK digest band 3	MSMS35MIN		Nano-35min	

The aim is to extract filenames and sample information from the sample list. We still depend on Masslynx to generate PKL peak lists.

Tags in search title field can be substituted at run time by fields from the MassLynx sample list. For example, “HEK digest band 1”, or any text from any of the columns, could appear at the top of the Mascot search report.

Automation: more flexibility

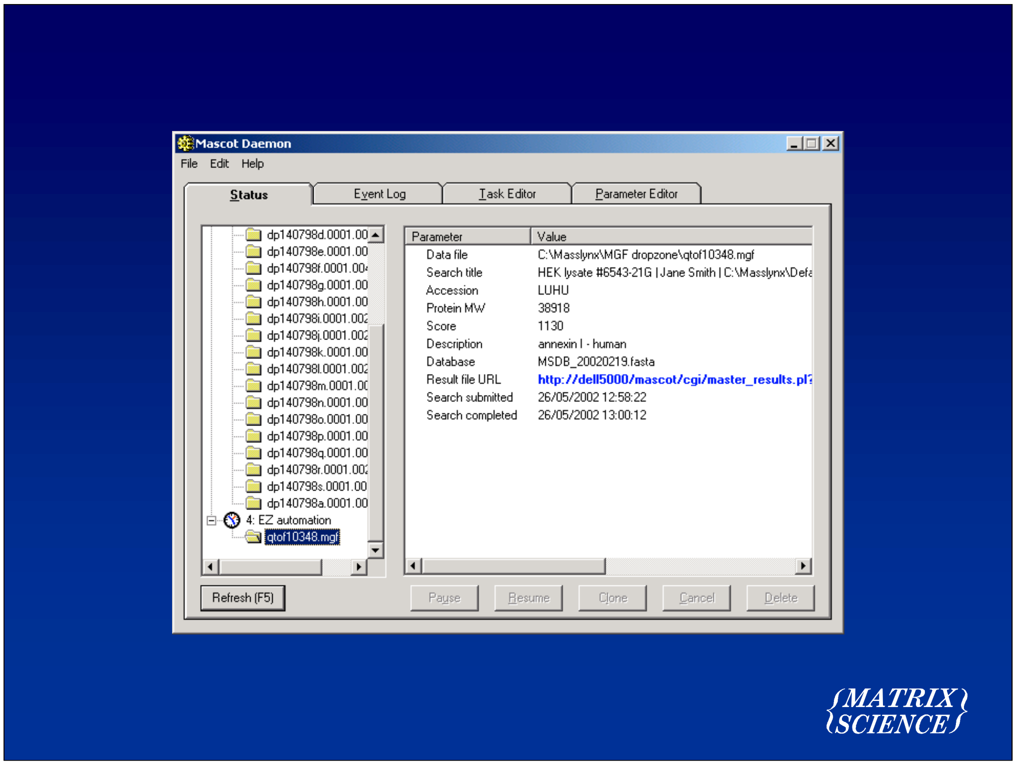
Use a script or macro to write peak lists in MGF format, including search parameters


EditPlus - [qtof10348.mgf]
File Edit View Search Document Project Tools Window Help
qtof10348.mgf
For Help, press F1
ln 34
col 28
57
00
PC
REC
INS
READ

```
# MGF format file generated by my VBA script
# 14:02:05 29 Nov 2001
#
SEARCH=MIS
USERNAME=Jane Smith
COM=HEK lysate #6543-21G | Jane Smith | C:\Masslynx\Default.pro\DATA\qtof10348.raw
DB=MSDB
CLE=Asp-N
TOL=0.1
TOLU=Da
MODS=Carbamidomethyl (C)
IT_MODS=Oxidation (M)
_sample_name=HEK lysate #6543-21G
_original_filename=C:\Masslynx\Default.pro\DATA\qtof10348.raw
_sample_barcode=4826659337524638329734239473216128l3843

BEGIN IONS
TITLE=sum of scan(s) 521 - 528, RT 12:34
PEPMASS=1095.000000
CHARGE=2+
1093.439900  53107778.775450
1085.600000  58039365.132336
868.269559  11538842.201475
1321.475432  12294627.319613
1570.900000  13442700.087962
668.203086  7581443.399182
1521.100000  12143646.083781
```

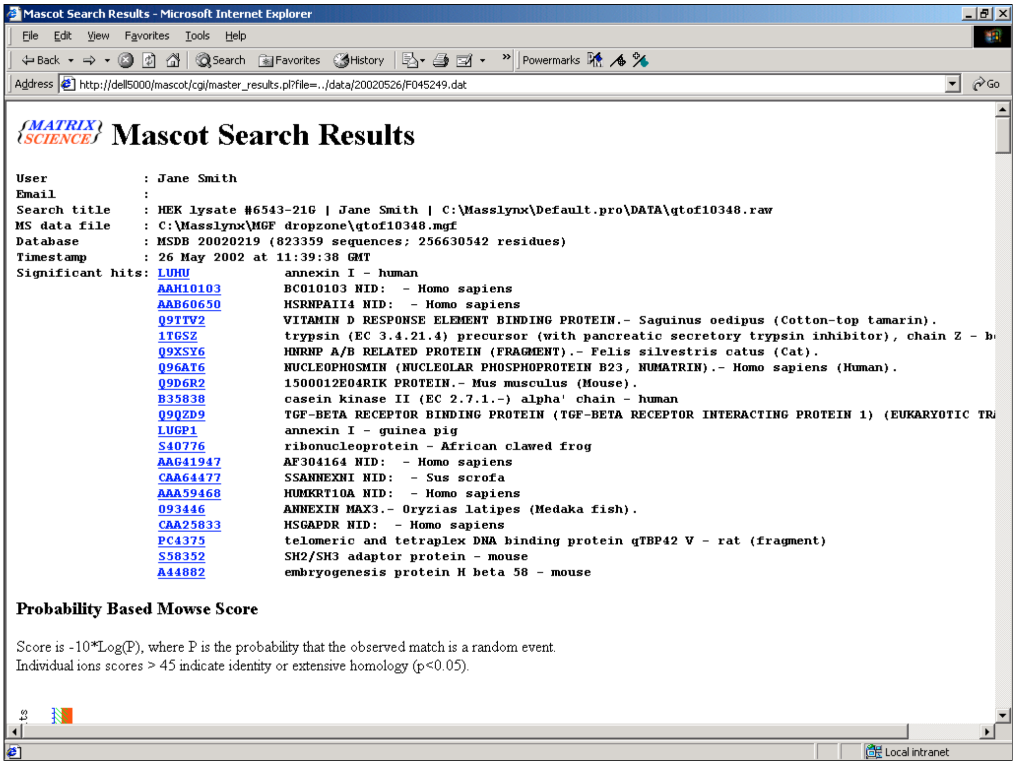
MATRIX SCIENCE

That is about as far as we can get with zero coding. Some simple coding is required to achieve more flexibility. Maybe the MS data system supports a macro language, such as VBA. If so, you can write out peak lists in the Mascot Generic Format (MGF). This allows the search parameters to be embedded into the data file, avoiding the need to set up search parameters manually in Daemon. You can also add your own parameters, which will be passed through the search engine into the results file. Any parameter that starts with an underscore is a 'user' parameter.



Parameter	Value
Data file	C:\Masslynx\MGF dropzone\qtof10348.mgf
Search title	HEK lysate #6543-21G | Jane Smith | C:\Masslynx\Defa
Accession	LUHU
Protein MW	38918
Score	1130
Description	annexin I - human
Database	MSDB_20020219.fasta
Result file URL	http://dell5000/mascot/cgi/master_results.pl?
Search submitted	26/05/2002 12:58:22
Search completed	26/05/2002 13:00:12

It is also important to make full use of the search title field. This is displayed in Daemon


Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Powermarks
Go
Local intranet

Address: http://dell5000/mascot/cgi/master_results.pl?file=../data/20020526/F045249.dat

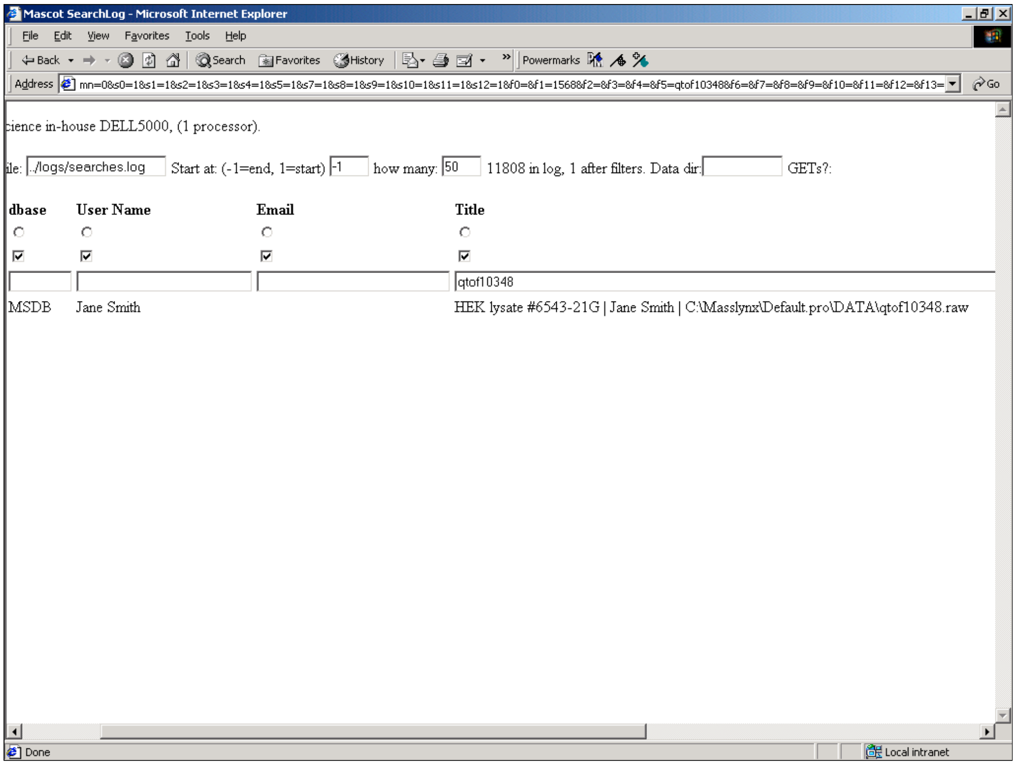
Mascot Search Results

```
User             : Jane Smith
Email            :
Search title     : HEK lysate #6543-21G | Jane Smith | C:\Masslynx\Default.pro\DATA\qtof10348.raw
MS data file     : C:\Masslynx\MGF dropzone\qtof10348.mgf
Database         : MSDB 20020219 (823359 sequences; 256630542 residues)
Timestamp        : 26 May 2002 at 11:39:38 GMT
Significant hits: LUHU            annexin I - human
                  AAH10103        BC010103 NID:  - Homo sapiens
                  AAB60650        HSRNPAII4 NID:  - Homo sapiens
                  Q9TTV2          VITAMIN D RESPONSE ELEMENT BINDING PROTEIN.- Saguinus oedipus (Cotton-top tamarin).
                  1TGSZ           trypsin (EC 3.4.21.4) precursor (with pancreatic secretory trypsin inhibitor), chain Z - b
                  Q9XSY6          HNRNP A/B RELATED PROTEIN (FRAGMENT).- Felis silvestris catus (Cat).
                  Q96AT6          NUCLEOPHOSMIN (NUCLEOLAR PHOSPHOPROTEIN B23, NUMATRIN).- Homo sapiens (Human).
                  Q9D6R2          1500012E04RIK PROTEIN.- Mus musculus (Mouse).
                  B35838          casein kinase II (EC 2.7.1.-) alpha' chain - human
                  Q9QZD9          TGF-BETA RECEPTOR BINDING PROTEIN (TGF-BETA RECEPTOR INTERACTING PROTEIN 1) (EUKARYOTIC TR
                  LUGP1           annexin I - guinea pig
                  S40776          ribonucleoprotein - African clawed frog
                  AAG41947        AF304164 NID:  - Homo sapiens
                  CAA64477        SSANNEXNI NID:  - Sus scrofa
                  AAA59468        HUMKRT10A NID:  - Homo sapiens
                  O93446          ANNEXIN MAX3.- Oryzias latipes (Medaka fish).
                  CAA25833        HSGAPDR NID:  - Homo sapiens
                  PC4375          telomeric and tetraplex DNA binding protein qTBP42 V - rat (fragment)
                  S58352          SH2/SH3 adaptor protein - mouse
                  A44882          embryogenesis protein H beta 58 - mouse
```

Probability Based Mowse Score

Score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 45 indicate identity or extensive homology ($p<0.05$).

And also at the top of the results report. So, it is worth taking the trouble to include key sample tracking information in the title field



Mascot SearchLog - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Powermarks
Address
Go
Done
Local intranet

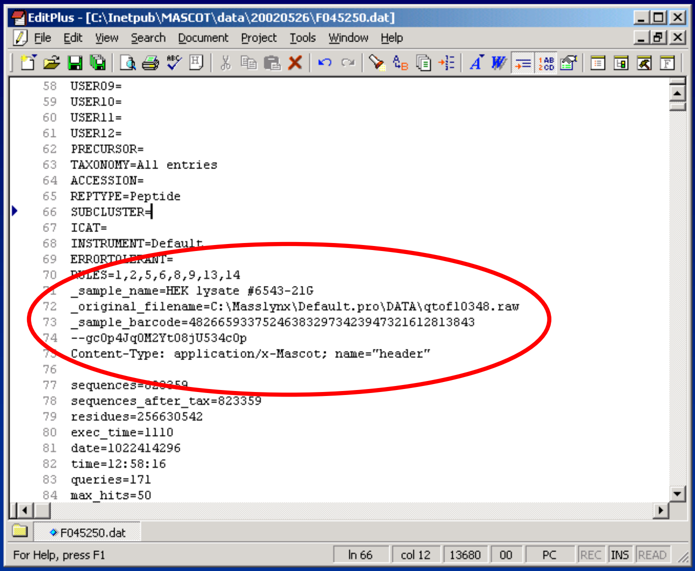
mn=0&s0=1&s1=1&s2=1&s3=1&s4=1&s5=1&s7=1&s8=1&s9=1&s10=1&s11=1&s12=1&f0=&f1=1568&f2=&f3=&f4=&f5=qtof10348&f6=&f7=&f8=&f9=&f10=&f11=&f12=&f13=

cience in-house DELL5000, (1 processor).

ile: ../logs/searches.log Start at: (-1=end, 1=start) -1 how many: 50 11808 in log, 1 after filters. Data dir: GETs?:

dbase	User Name	Email	Title
○	○	○	○
☑	☑	☑	☑
			qtof10348
MSDB	Jane Smith		HEK lysate #6543-21G | Jane Smith | C:\Masslynx\Default.pro\DATA\qtof10348.raw

Especially when you need to find an old search from among the tens of thousands on the Mascot server

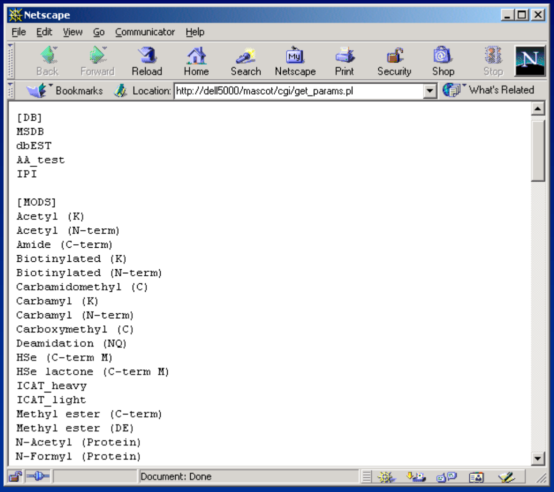

EditPlus - [C:\Inetpub\MASCOT\data\20020526\F045250.dat]
File Edit View Search Document Project Tools Window Help
F045250.dat
For Help, press F1
In 66
col 12
13680
00
PC
REC
INS
READ

```
USER09=
USER10=
USER11=
USER12=
PRECURSOR=
TAXONOMY=All entries
ACCESSION=
REPTYPE=Peptide
SUBCLUSTER=
ICAT=
INSTRUMENT=Default
ERRORTOLERANT=
RULES=1,2,5,6,8,9,13,14
_sample_name=HEK lysate #6543-21G
_original_filename=C:\Masslynx\Default.pro\DATA\qtof10348.raw
_sample_barcode=4826659337524638329734239473216128138 43
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="header"

sequences=823359
sequences_after_tax=823359
residues=256630542
exec_time=1110
date=1022414296
time=12:58:16
queries=171
max_hits=50
```

As mentioned earlier, any parameters with an underscore will be passed through to the result file. So you can include structured sample tracking information.

Automation: configuration

Use get_params.pl to query Mascot server configuration


Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://dell5000/mascot/cgi/get_params.pl
What's Related
Document: Done

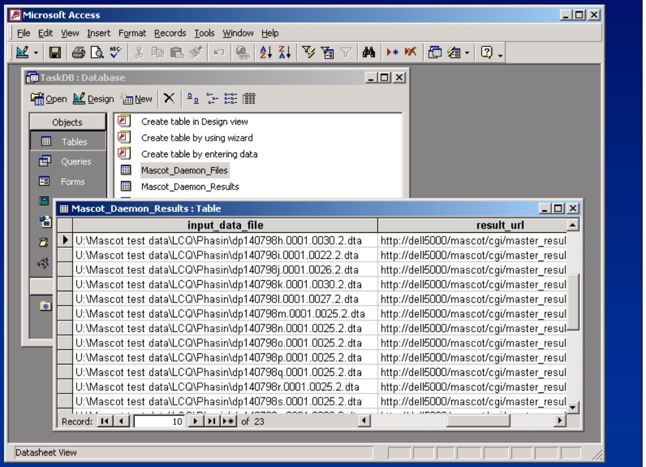
```
[DB]
MSDB
dbEST
AA_test
IPI

[MODS]
Acetyl (K)
Acetyl (N-term)
Amide (C-term)
Biotinylated (K)
Biotinylated (N-term)
Carbamidomethyl (C)
Carbamyl (K)
Carbamyl (N-term)
Carboxymethyl (C)
Deamidation (NQ)
HSe (C-term M)
HSe lactone (C-term M)
ICAT_heavy
ICAT_light
Methyl ester (C-term)
Methyl ester (DE)
N-Acetyl (Protein)
N-Formyl (Protein)
```

{MATRIX SCIENCE}

Another improvement to the workflow is to query the Mascot server to find out what databases are available,as well as the choice of enzymes, modifications, taxonomy, etc. This is achieved using a utility called get_params.pl

Automation: search results

Query Mascot Daemon's local database



Create table in Design view
Create table by using wizard
Create table by entering data
Mascot_Daemon_Files
Mascot_Daemon_Results

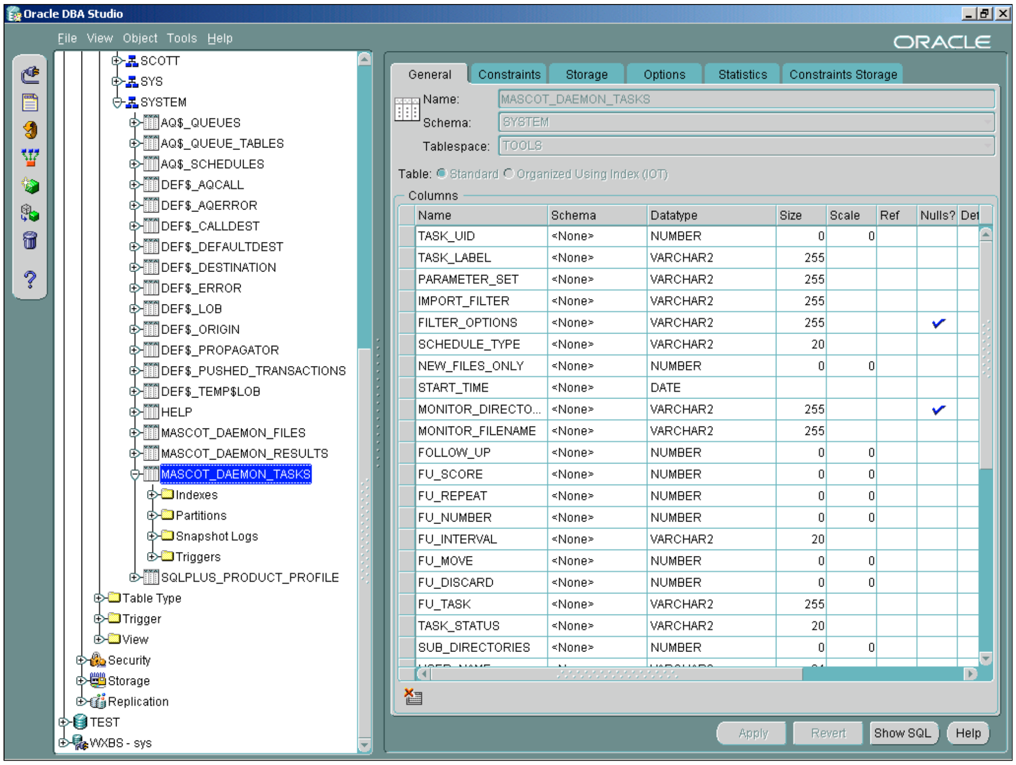
input_data_file	result_url
U:\Mascot test data\LCQ\Phasin\dp140798h.0001.0030.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798i.0001.0022.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798j.0001.0026.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798k.0001.0030.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798l.0001.0027.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798m.0001.0025.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798n.0001.0025.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798o.0001.0025.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798p.0001.0025.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798q.0001.0025.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798r.0001.0025.2.dta	http://dell5000/mascot/cgi/master_resul
U:\Mascot test data\LCQ\Phasin\dp140798s.0001.0025.2.dta	http://dell5000/mascot/cgi/master_resul

Record: 10 of 23

{MATRIX}
{SCIENCE}

One thing we haven't addressed yet is how to tie up the result file to the input data.

This information can be found in the Mascot Daemon database. By default, this is an Access database.



Name: MASCOT_DAEMON_TASKS

Schema: SYSTEM

Tablespace: TOOLS

Table: ● Standard ○ Organized Using Index (IOT)

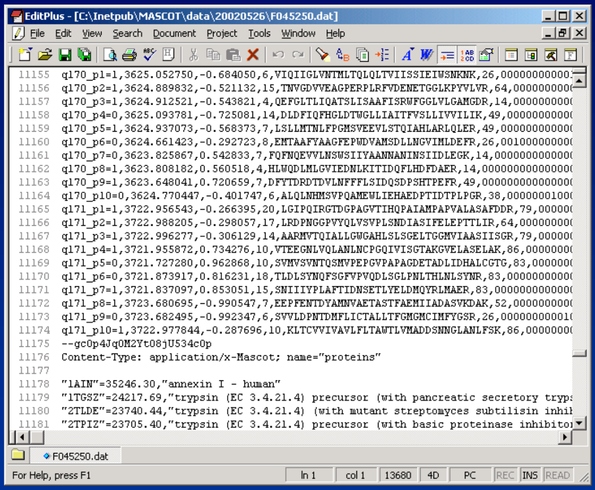
Columns

Name	Schema	Datatype	Size	Scale	Ref	Nulls?	Def
TASK_UID	<None>	NUMBER	0	0			
TASK_LABEL	<None>	VARCHAR2	255				
PARAMETER_SET	<None>	VARCHAR2	255				
IMPORT_FILTER	<None>	VARCHAR2	255				
FILTER_OPTIONS	<None>	VARCHAR2	255			✓	
SCHEDULE_TYPE	<None>	VARCHAR2	20				
NEW_FILES_ONLY	<None>	NUMBER	0	0			
START_TIME	<None>	DATE					
MONITOR_DIRECTO...	<None>	VARCHAR2	255			✓	
MONITOR_FILENAME	<None>	VARCHAR2	255				
FOLLOW_UP	<None>	NUMBER	0	0			
FU_SCORE	<None>	NUMBER	0	0			
FU_REPEAT	<None>	NUMBER	0	0			
FU_NUMBER	<None>	NUMBER	0	0			
FU_INTERVAL	<None>	VARCHAR2	20				
FU_MOVE	<None>	NUMBER	0	0			
FU_DISCARD	<None>	NUMBER	0	0			
FU_TASK	<None>	VARCHAR2	255				
TASK_STATUS	<None>	VARCHAR2	20				
SUB_DIRECTORIES	<None>	NUMBER	0	0			

However, the Mascot Daemon tables can live in any ODBC compliant database engine, such as Oracle or SQL Server. So, if you have an Oracle based LIMS, the Daemon tables can be inside the LIMS.

Automation: parsing results

Mascot result files are simple text files. No need for screen scraping.

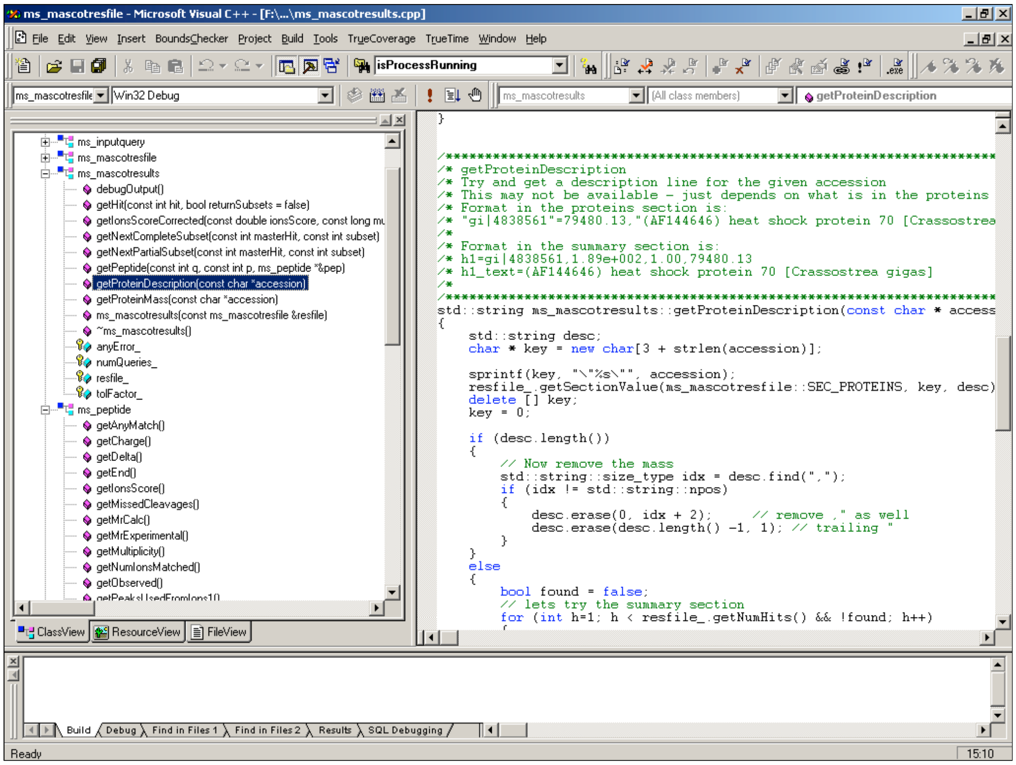

EditPlus - [C:\Inetpub\MASCOT\data\20020526\F045250.dat]
File Edit View Search Document Project Tools Window Help
F045250.dat
For Help, press F1
ln 1
col 1
13680
4D
PC
REC
INS
READ

```
q170_p1=1,3625.052750,-0.684050,6,VIQIIGLVNTMLTQLQLTVIISSIEIWSNKNK,26,0000000000
q170_p2=1,3624.889832,-0.521132,15,TNVGDVVEAGPERPLRFVDENETGGLKPYVLVR,64,000000000
q170_p3=1,3624.912521,-0.543821,4,QEFGLTLIQATSLISAAFISRWFGGLVLGAMGDR,14,000000000
q170_p4=0,3625.093781,-0.725081,14,DLDFIQFHGLDTWGLLIAITFVSLLIVVILIK,49,0000000000
q170_p5=1,3624.937073,-0.568373,7,LSLLMTNLFPGMSVEEVLSTQIAHLARLQLER,49,00000000000
q170_p6=0,3624.661423,-0.292723,8,EMTAAFYAAGFEPWDVAMSDLLNGVIMLDEFR,26,0010000000
q170_p7=0,3623.825867,0.542833,7,FQFNQEVVLNSWSIIYAANNANINSIIDLEGK,14,000000000000
q170_p8=1,3623.808182,0.560518,4,HLWQDLMLGVIEDNLKITIDQFLHDFDAER,14,00000000000000
q170_p9=1,3623.648041,0.720659,7,DFYTDRDTDVLNFFFLSIDQSDPSHTPEFR,49,00000000000000
q170_p10=0,3624.770447,-0.401747,6,ALQLNHMSVPQAMEWLIEHAEDPTIDTPLPGR,38,000000010
q171_p1=1,3722.956543,-0.266395,20,LGIPQIRGTDGPAGVTIHQPAIAMPAPVALASAFDDR,79,00000
q171_p2=1,3722.988205,-0.298057,17,LRDPNGGPVYQLVSVPLSNDIASIFELEPTTLIR,64,00000000
q171_p3=1,3722.996277,-0.306129,14,AARMVTQIALLGWGAHLSLSGELTGGMVIAASIISGR,79,00000
q171_p4=1,3721.955872,0.734276,10,VTEEGNLVQLANLNCPGQIVISGTAKGVELASELAK,86,0000000
q171_p5=0,3721.727280,0.962868,10,SVMVSVNTQSMVPEPGVPAPAGDETADLIDHALCGTG,83,000000
q171_p6=0,3721.873917,0.816231,18,TLDLSYNQFSGFVPVQDLSGLPNLTHLNLSYNR,83,0000000000
q171_p7=1,3721.837097,0.853051,15,SNIIIYPLAFTIDNSETLYELDMQYRLMAER,83,000000000000
q171_p8=1,3723.680695,-0.990547,7,EEPFENTDYAMNVAETASTFAEMIIADASVKDAK,52,000000000
q171_p9=0,3723.682495,-0.992347,6,SVVLDPNTDMFLICTALLTFGMGMCIMFYGSR,26,00000000001
q171_p10=1,3722.977844,-0.287696,10,KLTCVVIVAVLFLTAWTLVMADDSNNGLANLFSK,86,0000000
--gc0p4Jq0M2Yt08jU534c0p
Content-Type: application/x-Mascot; name="proteins"

"1AIN"=35246.30,"annexin I - human"
"1TGSZ"=24217.69,"trypsin (EC 3.4.21.4) precursor (with pancreatic secretory tryp
"2TLDE"=23740.44,"trypsin (EC 3.4.21.4) (with mutant streptomyces subtilisin inhi
"2TPIZ"=23705.40,"trypsin (EC 3.4.21.4) precursor (with basic proteinase inhibito
```

{MATRIX}
{SCIENCE}

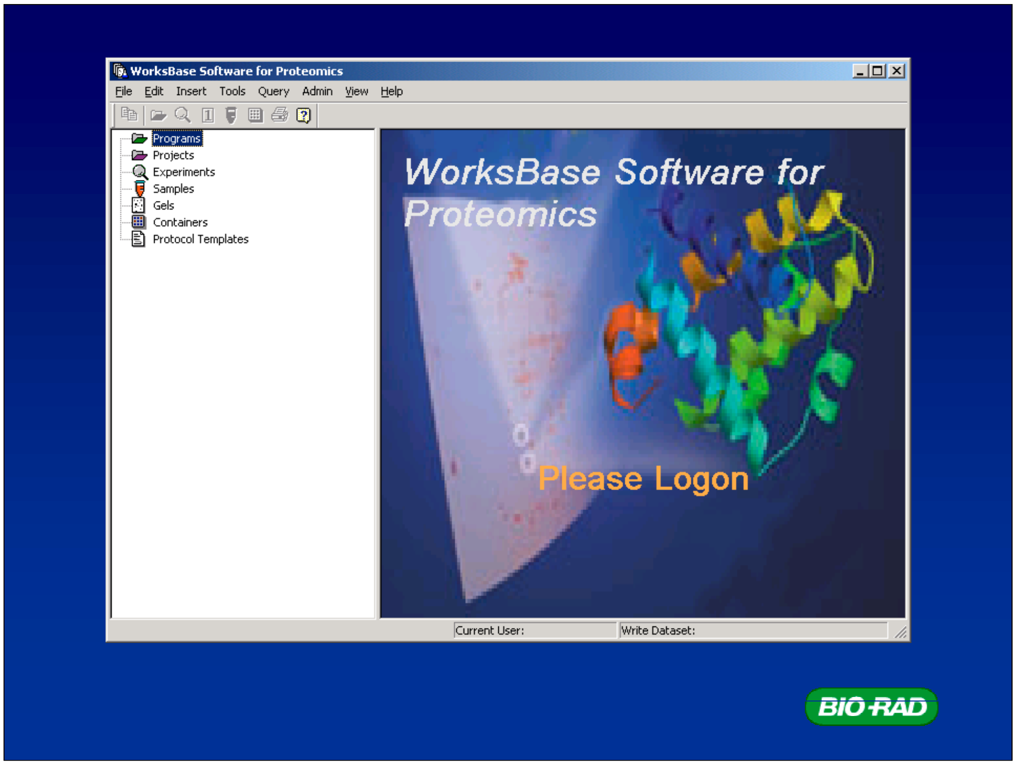
The next link in the chain is to export the results to a relational database, possibly the LIMS.

Mascot result files are structured text files, and it isn't difficult to write Perl scripts to fish out the information of interest

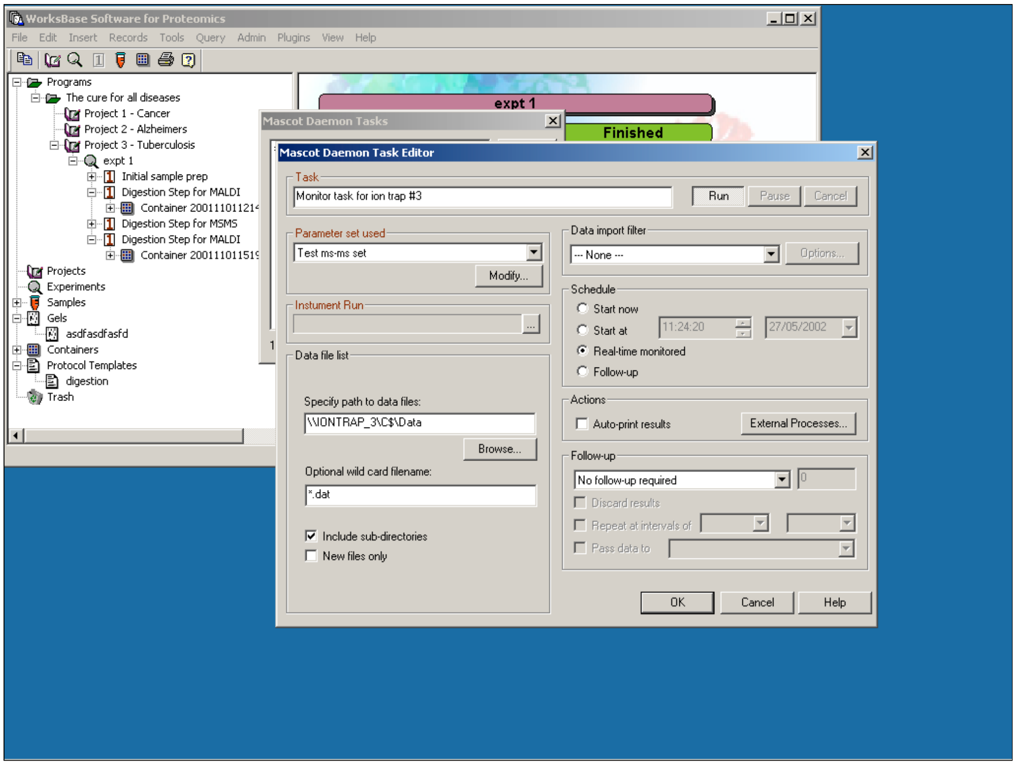

ms_mascotresfile - Microsoft Visual C++ - [F:\...\ms_mascotresults.cpp]
File Edit View Insert BoundsChecker Project Build Tools TrueCoverage TrueTime Window Help
isProcessRunning
ms_mascotresfile
Win32 Debug
ms_mascotresults
(All class members)
getProteinDescription
ms_inputquery
ms_mascotresfile
ms_mascotresults
debugOutput()
getHit(const int hit, bool returnSubsets = false)
getIonsScoreCorrected(const double ionsScore, const long mu
getNextCompleteSubset(const int masterHit, const int subset)
getNextPartialSubset(const int masterHit, const int subset)
getPeptide(const int q, const int p, ms_peptide *&pep)
getProteinDescription(const char *accession)
getProteinMass(const char *accession)
ms_mascotresults(const ms_mascotresfile &resfile)
~ms_mascotresults()
anyError_
numQueries_
resfile_
tolFactor_
ms_peptide
getAnyMatch()
getCharge()
getDelta()
getEnd()
getIonsScore()
getMissedCleavages()
getMrCalc()
getMrExperimental()
getMultiplicity()
getNumIonsMatched()
getObserved()
getPeaksUsedFromIons1()
ClassView
ResourceView
FileView
Build
Debug
Find in Files 1
Find in Files 2
Results
SQL Debugging
Ready
15:10

```
}


/***********************************************************************
/* getProteinDescription
/* Try and get a description line for the given accession
/* This may not be available - just depends on what is in the proteins
/* Format in the proteins section is:
/* "gi|4838561"=79480.13,"(AF144646) heat shock protein 70 [Crassostrea
/*
/* Format in the summary section is:
/* h1=gi|4838561,1.89e+002,1.00,79480.13
/* h1_text=(AF144646) heat shock protein 70 [Crassostrea gigas]
/*
/***********************************************************************
std::string ms_mascotresults::getProteinDescription(const char * access
{
    std::string desc;
    char * key = new char[3 + strlen(accession)];

    sprintf(key, "\"%s\"", accession);
    resfile_.getSectionValue(ms_mascotresfile::SEC_PROTEINS, key, desc)
    delete [] key;
    key = 0;

    if (desc.length())
    {
        // Now remove the mass
        std::string::size_type idx = desc.find(",");
        if (idx != std::string::npos)
        {
            desc.erase(0, idx + 2);     // remove ," as well
            desc.erase(desc.length() -1, 1); // trailing "
        }
    }
    else
    {
        bool found = false;
        // lets try the summary section
        for (int h=1; h < resfile_.getNumHits() && !found; h++)
```

However, many programmers prefer to work in C++, so we are developing a class library that will support an object oriented approach to extracting the result information



A good illustration of extensive integration between Mascot and a relational database is the WorksBase package from Bio-Rad. WorksBase is a proteomics LIMS



Task: Monitor task for ion trap #3

Parameter set used: Test ms-ms set

Instument Run:

Data import filter: --- None ---

Schedule:
○ Start now
○ Start at 11:24:20 27/05/2002
● Real-time monitored
○ Follow-up

Data file list

Specify path to data files: \\IONTRAP_3\C$\Data

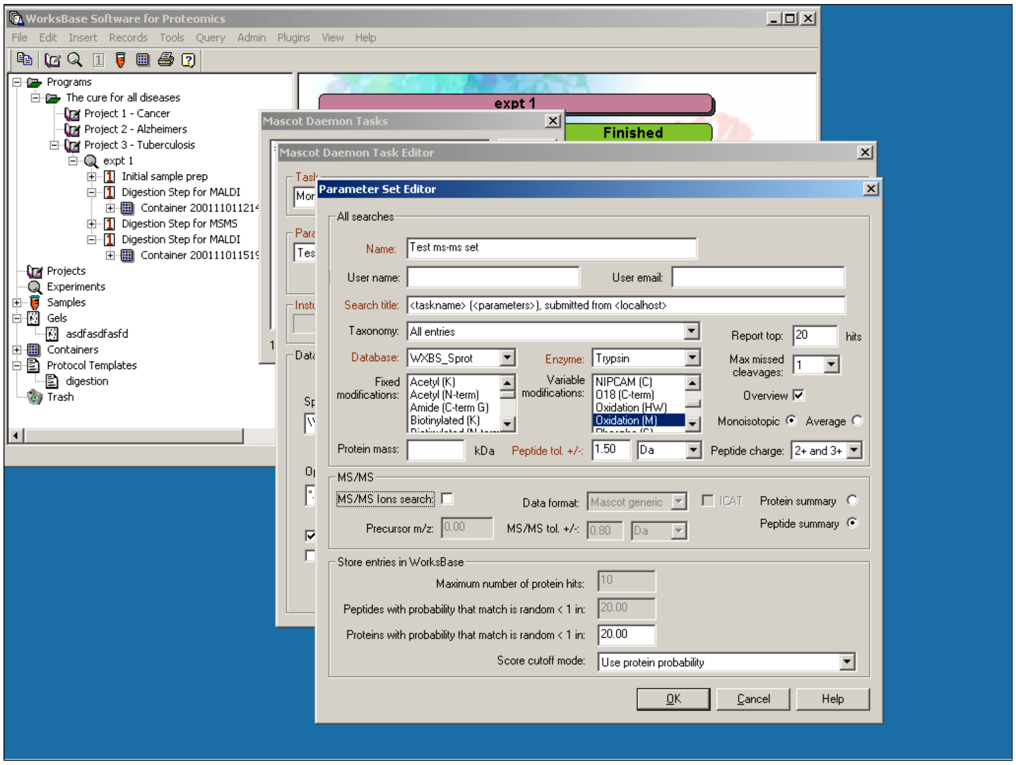
Optional wild card filename: *.dat

☑ Include sub-directories
☐ New files only

Actions: ☐ Auto-print results

Follow-up: No follow-up required 0
☐ Discard results
☐ Repeat at intervals of
☐ Pass data to

WorksBase provides the user interface for all aspects of Mascot searching. For example, tasks are defined in WorksBase dialogs and saved to WorksBase tables



All searches

Name: Test ms-ms set

User name:

User email:

Search title: <taskname> (<parameters>), submitted from <localhost>

Taxonomy: All entries

Report top: 20 hits

Database: WXBS_Sprot

Enzyme: Trypsin

Max missed cleavages: 1

Fixed modifications: Acetyl (K), Acetyl (N-term), Amide (C-term G), Biotinylated (K)

Variable modifications: NIPCAM (C), O18 (C-term), Oxidation (HW), Oxidation (M)

Overview ☑

Monoisotopic ● Average ○

Protein mass: kDa

Peptide tol. +/-: 1.50 Da

Peptide charge: 2+ and 3+

MS/MS

MS/MS Ions search: □

Data format: Mascot generic

□ ICAT

Protein summary ○

Peptide summary ●

Precursor m/z: 0.00

MS/MS tol. +/-: 0.80 Da

Store entries in WorksBase

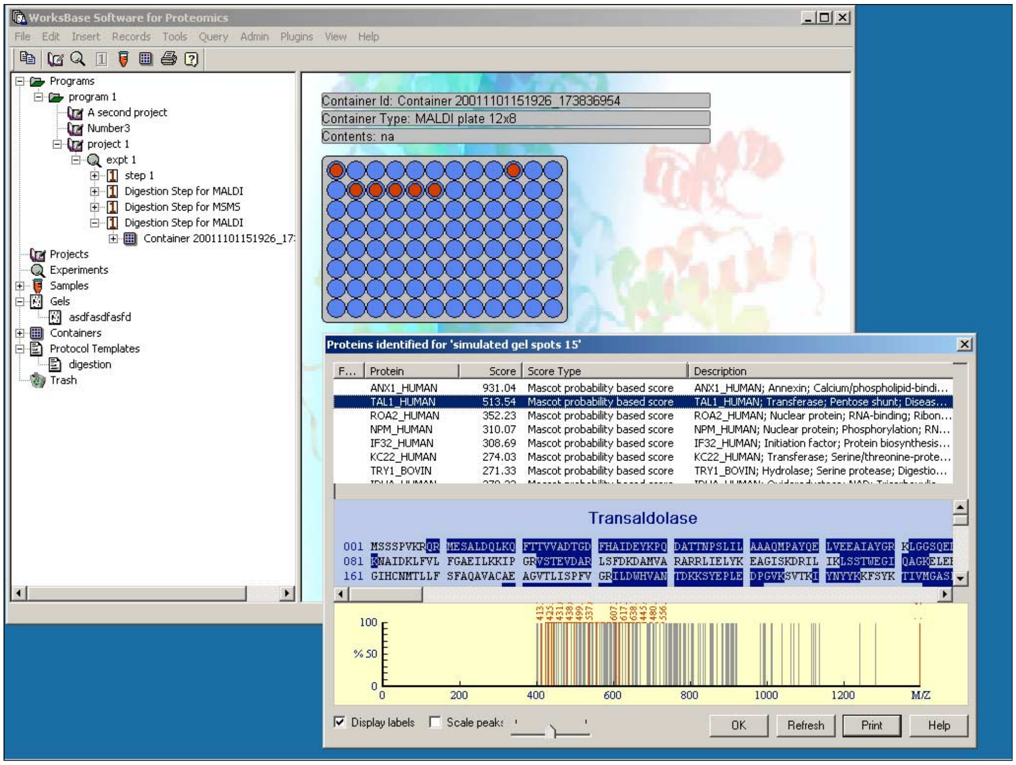
Maximum number of protein hits: 10

Peptides with probability that match is random < 1 in: 20.00

Proteins with probability that match is random < 1 in: 20.00

Score cutoff mode: Use protein probability

As are sets of search parameters



Container Id: Container 20011101151926_173836954
Container Type: MALDI plate 12x8
Contents: na

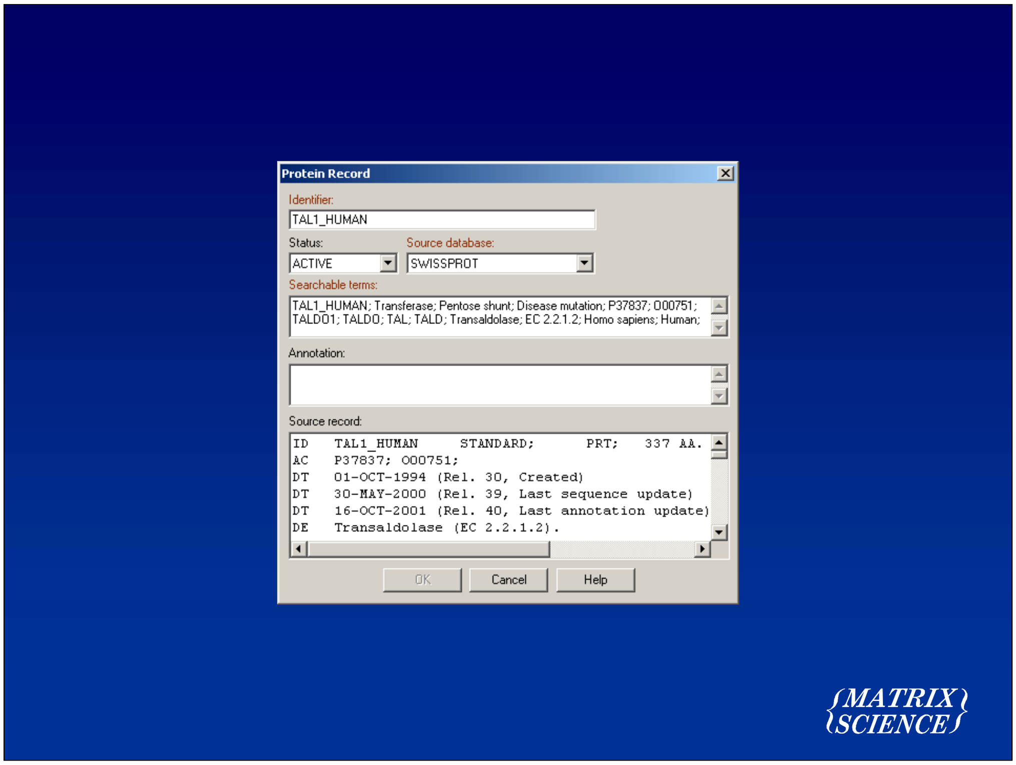
Proteins identified for 'simulated gel spots 15'

F...	Protein	Score	Score Type	Description
	ANX1_HUMAN	931.04	Mascot probability based score	ANX1_HUMAN; Annexin; Calcium/phospholipid-bindi...
	TAL1_HUMAN	513.54	Mascot probability based score	TAL1_HUMAN; Transferase; Pentose shunt; Diseas...
	ROA2_HUMAN	352.23	Mascot probability based score	ROA2_HUMAN; Nuclear protein; RNA-binding; Ribon...
	NPM_HUMAN	310.07	Mascot probability based score	NPM_HUMAN; Nuclear protein; Phosphorylation; RN...
	IF32_HUMAN	308.69	Mascot probability based score	IF32_HUMAN; Initiation factor; Protein biosynthesis...
	KC22_HUMAN	274.03	Mascot probability based score	KC22_HUMAN; Transferase; Serine/threonine-prote...
	TRY1_BOVIN	271.33	Mascot probability based score	TRY1_BOVIN; Hydrolase; Serine protease; Digestio...

Transaldolase

```
001 MSSSPVKRQR MESALDQLKQ FTTVVADTGD FHAIDEYKPQ DATTNPSLIL AAAQMPAYQE LVEEAIAYGR KLGGSQE
081 KNAIDKLFVL FGAEILKKIP GRVSTEVDAR LSFDKDAMVA RARRLIELYK EAGISKDRIL IKLSSTWEGI QAGKELE
161 GIHCNMTLLF SFAQAVACAE AGVTLISPFV GRILDWHVAN TDKKSYEPLE DPGVKSVTKI YNYYKKFSYK TIVMGAS
```

The results are automatically exported to WorksBase, and reports can be generated by querying the database



Identifier: TAL1_HUMAN

Status: ACTIVE

Source database: SWISSPROT

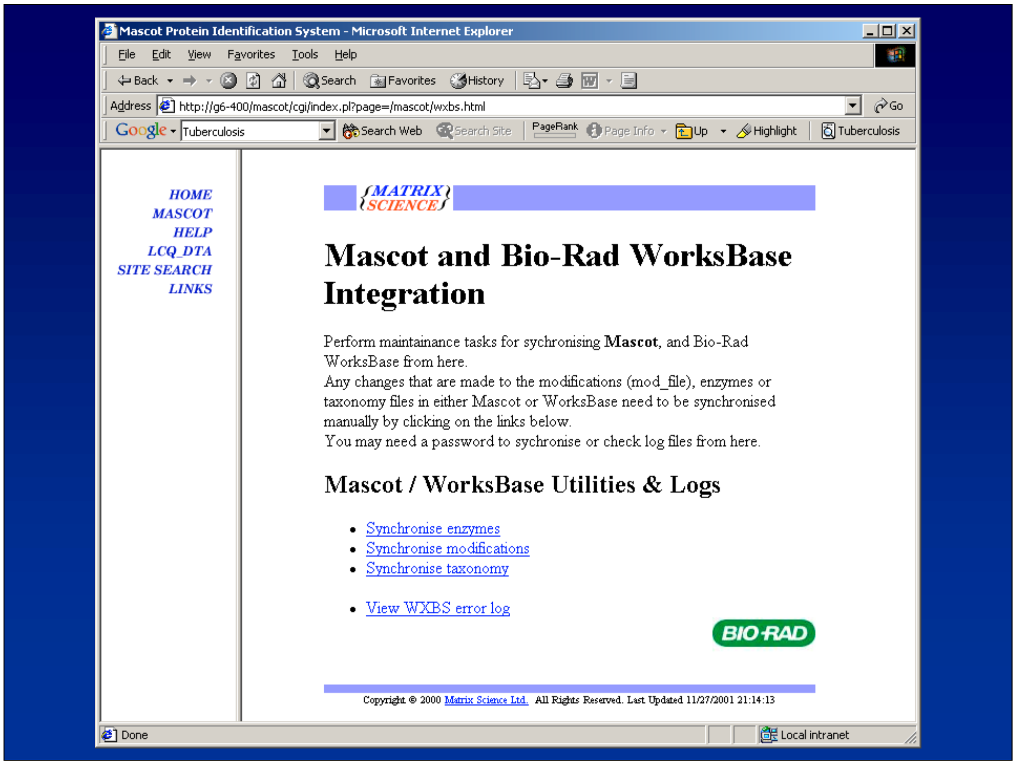
Searchable terms: TAL1_HUMAN; Transferase; Pentose shunt; Disease mutation; P37837; O00751; TALDO1; TALDO; TAL; TALD; Transaldolase; EC 2.2.1.2; Homo sapiens; Human;

Annotation:

Source record:

```
ID   TAL1_HUMAN     STANDARD;      PRT;   337 AA.
AC   P37837; O00751;
DT   01-OCT-1994 (Rel. 30, Created)
DT   30-MAY-2000 (Rel. 39, Last sequence update)
DT   16-OCT-2001 (Rel. 40, Last annotation update)
DE   Transaldolase (EC 2.2.1.2).
```

Even the sequence database is contained within WorksBase tables


Mascot Protein Identification System - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://g6-400/mascot/cgi/index.pl?page=/mascot/wxbs.html
Go
Google Tuberculosis Search Web Search Site PageRank Page Info Up Highlight Tuberculosis
HOME
MASCOT
HELP
LCQ_DTA
SITE SEARCH
LINKS
MATRIX SCIENCE
BIO-RAD
Done
Local intranet

Mascot and Bio-Rad WorksBase Integration

Perform maintainance tasks for sychronising **Mascot**, and Bio-Rad WorksBase from here.
Any changes that are made to the modifications (mod_file), enzymes or taxonomy files in either Mascot or WorksBase need to be synchronised manually by clicking on the links below.
You may need a password to sychronise or check log files from here.

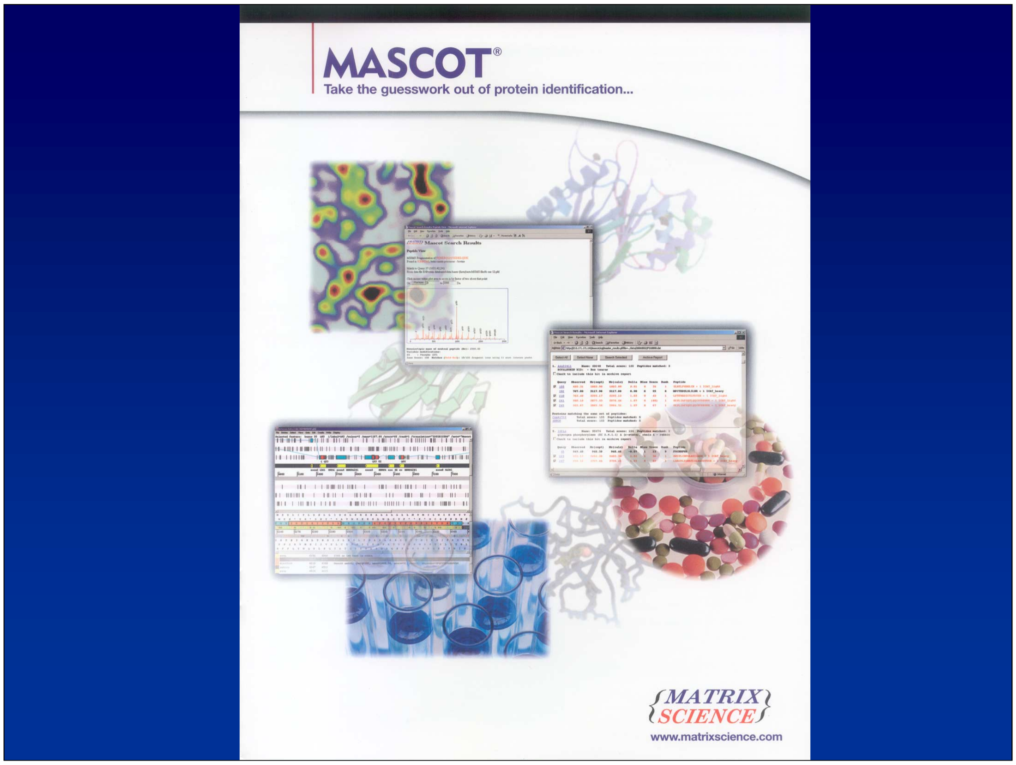
Mascot / WorksBase Utilities & Logs

- Synchronise enzymes
- Synchronise modifications
- Synchronise taxonomy

- View WXBS error log

Copyright © 2000 Matrix Science Ltd. All Rights Reserved. Last Updated 11/27/2001 21:14:13

And there are web page links for administration tasks



MASCOT®

Take the guesswork out of protein identification...

MATRIX SCIENCE

www.matrixscience.com