New features in Mascot 2.1

MATRIX SCIENCE

There's a very minor change from the published program: I'll be presenting the changes to Mascot 2.1 before the session on Mascot security.

Administration . . .

- On the public web site on 7th May 2005
- Windows and Linux CDs were sent to all customers on support contracts on 16th May
- AIX, Irix, Tru64 and Solaris due within the next month.

ASMS 2005

MATRIX SCIENCE

A few minor administration points - Mascot 2.1 was available on our public web site on 7th May this year.

The Linux and Windows update CDs were sent out to all our customers in the middle of May, so if you haven't received yours yet, and you believe that you should have done, please send an email to info@matrixscience.com

We aim to release the remaining Unix builds by the end of this month.

This talk will be divided into these four sections:

We will start with the new export utility.

Export script

- Provides export to a number of formats:
  - CSV (comma separated values)
  - XML
  - pepXML (from ISB)
  - DTASelect.

ASMS 2005

When a Mascot search is run, the results for the search are saved in a mime format text file on the Mascot server.A perl script reads that results file, and displays the html in a nice friendly way in your browser. The results text file itself could never be described as bedtime reading - even for me.

In Mascot version 2.0 and later, these perl scripts use a toolkit that we call Mascot parser. If you are developing a relational database product, then you should also license a copy of Mascot Parser from us to extract the data from the results files - you definitely shouldn't screen scrape the html. Mascot Integra, which we will hear about later, uses Mascot Parser.

However, what if you want to just put the results into Excel or if you don't have the necessary skills to write some code using Mascot Parser?

The new export facility is designed to fulfil your requirements.

It will output the results in a number of formats.

Starting the export
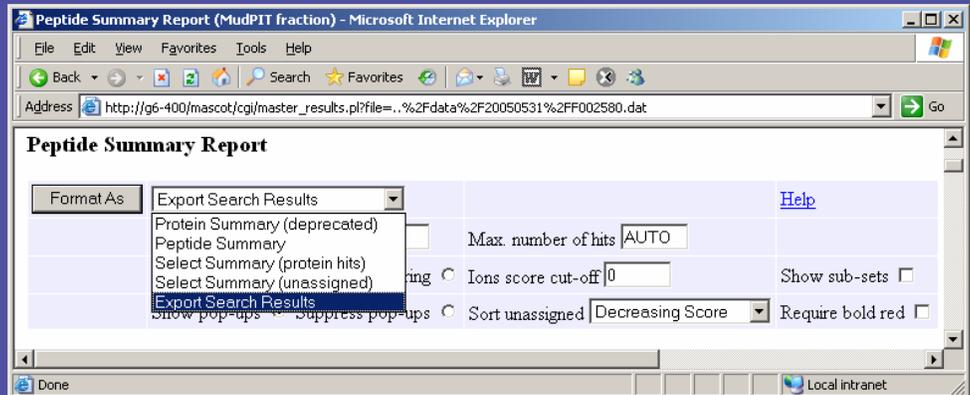
ASMS 2005

You may not have noticed in Mascot 2.1 that there is an extra item in the drop down list - Export Search results. Select this, and then press the "Format As" button

Choose output format:

You now have a page with lots of formatting options - the first choice is the output format.

If you want the XML format, you probably know that this is what you want. If you've no idea what XML is, chances are you don't want it.

Choose CSV if you want to export to Excel - I'll show an example in a moment.

Choose pepXML if you want to export to Protein Prophet from ISB. We would recommend that you use this exporter rather than ISB's own Mascot2XML converter.

Finally, if you are using Dave Tabb's DTASelect then you will want to choose this last option.
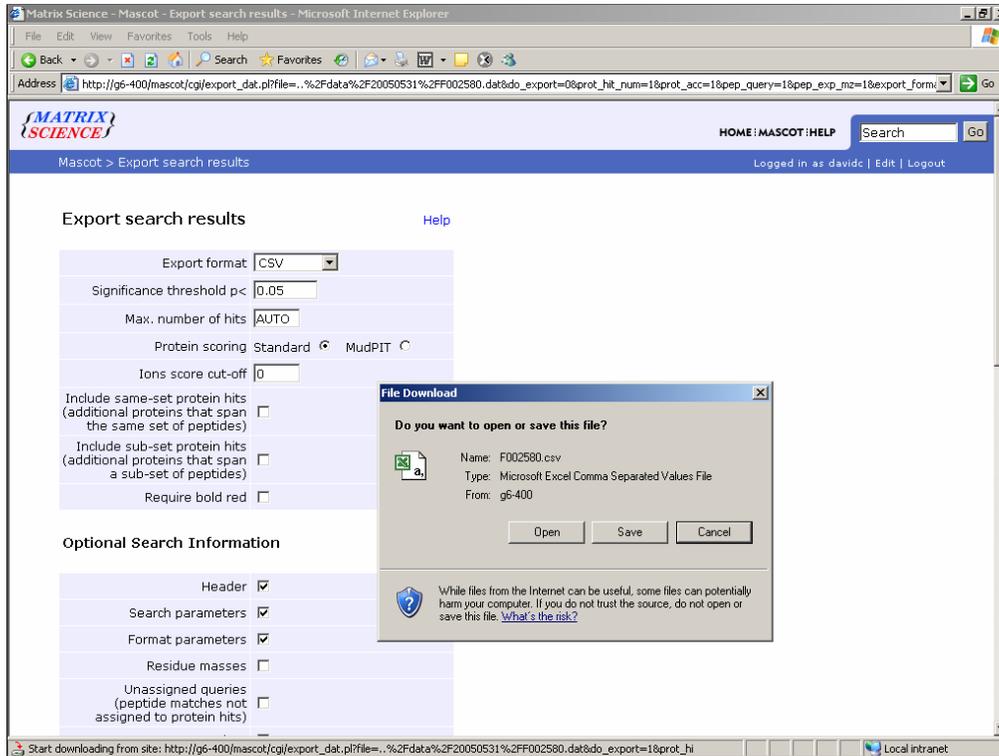
To export to Excel, simply select CSV as the format, and click on the Export Search Results button at the bottom of the page.

You can then click on the Open button to open it into Excel:

Microsoft Excel - F002580[1].csv

File Edit View Insert Format Tools Data Window Help

Arial   10   B I U   %   ,   Prompt

A1   =

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
38	Max. numb	0													
39	Use MudP	0													
40	Ions score	0													
41	Include same-set proteins														
42	Include sul	0													
43	Include unassigned														
44	Require bo	0													
45															
46	Protein hit:---														
47															
48	prot_hit_nu	prot_acc	prot_desc	prot_score	prot_mass	prot_matc	pep_query	pep_exp_r	pep_exp_r	pep_exp_z	pep_calc_	pep_delta	pep_miss	pep_score	pep_expec	pep
49	1	Q5SQB0_I	Nucleopho	100	29677	4	143	911.09	1820.16	2	1818.84	1.33	0	53.12	0.14	
50	1						144	912	1821.99	2	1818.84	3.16	0	31.83	18	
51	1						203	1235.16	2468.31	2	2467.39	0.93	1	40.46	2.1	
52	1						204	824.41	2470.21	3	2467.39	2.83	1	47.32	0.44	
53	2	Q8N1C8_I	Heat shoc	77	74093	1	146	930.09	1858.16	2	1855.9	2.26	0	79.45	0.00033	
54	3	1A12A	regulator o	73	43209	2	152	949.79	1897.57	2	1897.91	-0.34	0	50.8	0.23	
55	3						153	950.86	1899.7	2	1897.91	1.79	0	73.04	0.0014	
56	4	Q5TBN3_I	Lymphocy	72	27575	2	228	1385.96	2769.91	2	2769.2	0.71	0	71.79	0.0014	
57	4						229	925.37	2773.09	3	2769.2	3.89	0	42.32	1.2	
58	5	1C4ZD	ubiquitin c	71	16967	5	207	828.86	2483.55	3	2482.32	1.24	0	48.87	0.31	
59	5						208	1243.25	2484.48	2	2482.32	2.17	0	71.49	0.0017	
60	5						210	1243.56	2485.1	2	2482.32	2.79	0	66.97	0.0047	
61	5						211	829.51	2485.51	3	2482.32	3.2	0	48.01	0.37	
62	5						212	1244.16	2486.3	2	2482.32	3.99	0	12.48	1.30E+03	
63	6	AAC32447	CPU51572	67	13072	1	148	936.54	1871.06	2	1869	2.06	0	67.07	0.0057	
64	7	IBHUN	hydroxyme	66	39552	2	177	1081.2	2160.38	2	2160.14	0.24	0	66.26	0.0063	
65	7						178	1082.11	2162.2	2	2160.14	2.06	0	45.48	0.75	
66	8	EF1D_HUI	Elongation	65	31086	1	180	1093.44	2184.87	2	2184.04	0.83	0	66.95	0.0052	
67	9	Q5R7E0_I	Hypothetic	64	25119	3	149	943.52	1885.02	2	1883.99	1.03	1	61.72	0.019	
68	9						150	943.86	1885.71	2	1883.99	1.72	1	63.6	0.012	
69	9						151	944.96	1887.91	2	1883.99	3.92	1	41.88	1.8	
70	10	AAG43129	AF060511	60	10281	1	147	936.07	1870.13	2	1868.88	1.26	0	59.91	0.029	
71																
72																

.csv]F002580[1]

Ready

It is likely that you will then want to go back to the previous page and select some different options.

It's as easy as that.

Altova XMLSpy - [F002580.xml]

File   Edit   XML   DTD/Schema   Schema design   XSL/XQuery   Authentic   View   Browser   Tools   ADVANCED   Window   Help

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<mascot_search_results xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns="http://www.matrixscience.com/xmlns/schema/mascot_search_results_1"
    xsi:schemaLocation="http://www.matrixscience.com/xmlns/schema/mascot_search_results_1
    http://g6-
    400/mascot/xmlns/schema/mascot_search_results_1/mascot_search_results_1.xsd"
    majorVersion="1" minorVersion="0">
  <hits>
    <hit number="1">
      <protein accession="Q5SQB0_MOUSE">
        <prot_desc>Nucleophosmin 1.- Mus musculus (Mouse).</prot_desc>
        <prot_score>100</prot_score>
        <prot_mass>29677</prot_mass>
        <prot_matches>4</prot_matches>
        <peptide query="143">
          <pep_exp_mz>911.09</pep_exp_mz>
          <pep_exp_mr>1820.16</pep_exp_mr>
          <pep_exp_z>2</pep_exp_z>
          <pep_calc_mr>1818.84</pep_calc_mr>
          <pep_delta>1.33</pep_delta>
          <pep_miss>0</pep_miss>
          <pep_score>53.12</pep_score>
          <pep_expect>0.14</pep_expect>
          <pep_rank>1</pep_rank>
          <pep_res_before>R</pep_res_before>
          <pep_seq>MTDQEAIQDLWQWR</pep_seq>
          <pep_res_after>K</pep_res_after>
          <pep_var_mod />
        </peptide>
        <peptide query="144">
          <pep_exp_mz>912.00</pep_exp_mz>
          <pep_exp_mr>1821.99</pep_exp_mr>
          <pep_exp_z>2</pep_exp_z>
          <pep_calc_mr>1818.84</pep_calc_mr>
          <pep_delta>3.16</pep_delta>
          <pep_miss>0</pep_miss>
          <pep_score>31.83</pep_score>
```

Attribute

Entity

Text   Schema/WSDL   Authentic   Browser

F002580.xml

Ln 2, Col 24   CAP NUM SCRL

For those of you into XML, here is a sample XML file. The schema is available from our web site or your local Mascot installation.

Please read the help for details.

If you have the right tools, then you can obviously view the XML data in a 'database' format.
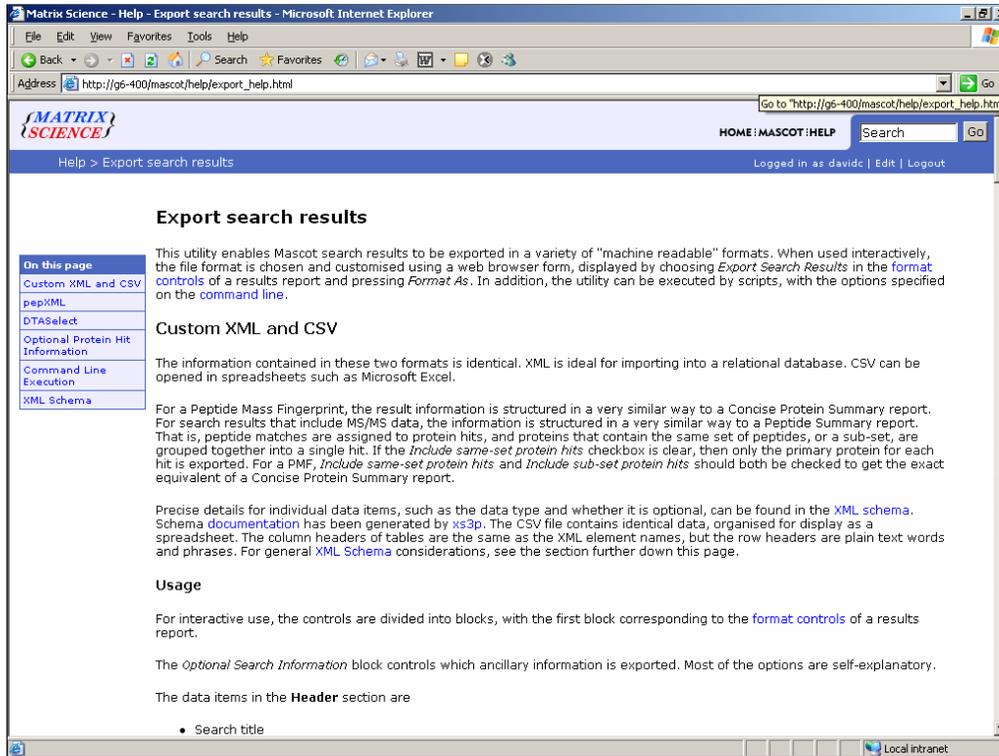
Before I go any further, I'd like to gently point out that there is a help page for this. Simply click on the help link and read...

Finally, with an in-house server, the export script can be called from the command line or a shell prompt.

I won't go into any detail here, but this means that it is possible to set up a script that will, for example, automatically convert all of your Mascot results to XML files.

Obviously, if you do this automatically, please make sure that you keep the original .dat files or your won't be able to view the results in a browser.

New features in Mascot 2.1

- Export results in various formats
- Search engine and report enhancements
- Mascot Daemon enhancements
- Mascot Security

ASMS 2005

MATRIX SCIENCE

Lets move on now to look at some of the search engine and report enhancements. These items are not in any particular order.

mzData 1.05 (with CV 1.0)

HUPO Proteomics Standards Initiative - Microsoft Internet Explorer

File   Edit   View   Favorites   Tools   Help

Back ▾ ▾ 🔍 Search ⭐ Favorites

Address http://psidev.sourceforge.net/   ➔ Go

HUPO

Proteomics Standards Initiative

HU♦O
Human Proteome
Organisation

Mission Statement

The HUPO Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparision, exchange and verification.

The Proteomics Standards Initiative was founded at the HUPO meeting in Washingt (see Science 296, 827). PSI currently develops standards for two key areas of prote spectrometry and protein-protein interaction data, as well as a standardised general

Data file ciex\Beta Galactosidase 1.xml

Data format Mascot generic ▾

Instrument

Overview

Mascot generic
Sequest (.DTA)
Finnigan (.ASC)
Micromass (.PKL)
PerSeptive (.PKS)
Sciex API III
Bruker (.XML)
mzData (.XML)

ASMS 2005

MATRIX SCIENCE

In Mascot 2.1, you can search data in the new mzData format.

mzData is a peak list format, just like mgf, pkl and dta. For those who are interested, it is XML. The good thing is that mzData files will contain additional information about the instrument settings and parameters that were used to generate the data.

This format has been defined by the HUPO Proteomics Standards Initiative - and is the result of a large number of meetings with representatives from most of the mass spec instrument manufacturers and software vendors.

Over the coming year, expect most of the instrument data systems to start exporting peak lists in mzData format.

We only support release 1.05 - earlier releases are not supported. We also support the controlled vocabulary version 1.0

For MS-MS searches, simply select from the list.

For PMF searches, the format is detected automatically.

Unfortunately there isn't much more to show - it just works.

Negative ions

Charge = 1-

Bruker XML

Optional new instrument with added 'c' ions

Need to enter precursor m/z for Bruker XML

ASMS 2005

Mascot 2.1 now supports data from negative ions experiments.

It should be very obvious how to use this. Simply enter the charge state (singly charged in this case).

For this example, I'll also illustrate that Bruker XML files are now also supported in Mascot 2.1. Unfortunately, the Bruker XML format doesn't include the precursor mass for ms-ms data, so you will need to enter this manually.

For this example, I've also created a new instrument definition. I chose to add support for 'c' ions to the standard tof-tof definition. This isn't strictly necessary, but it can help scores slightly.

Peptide Summary Report (../data/20050602/F004224.dat) - Microsoft Internet Explorer

File   Edit   View   Favorites   Tools   Help

Back ... | Search | Favorites

Address http://t41-dmc/mascot/cgi/master_results.pl?file=../data/20050602/F004224.dat

{MATRIX SCIENCE} Mascot Search Results

```
User            : David Creasy
Email           : dcreasy@matrixscience.com
Search title    :
MS data file    : Y:\mascot_test_data\bruker_negative_ions\ACTH1-17_LIFTnegativ\0_G11\0\2091.0705.LIFT_1.LIFT\fast\pdata\1\p
Database        : Sprot 20040705 (153871 sequences; 56608159 residues)
Timestamp       : 2 Jun 2005 at 10:27:54 GMT
Significant hits: COLI_BALPH  (P01195) Corticotropin-lipotropin (Pro-opiomelanocortin) (POMC) [Contains: Corticotropin (Adre
```

Peptide Summary Report

Format As	Peptide Summary ▼		Help
	Significance threshold p< 0.05	Max. number of hits 20	
	Standard scoring ○ MudPIT scoring ●	Ions score cut-off 0	Show sub-sets ☐
	Show pop-ups ○ Suppress pop-ups ●	Sort unassigned Decreasing Score ▼	Require bold red ☐

Select All   Select None   Search Selected   ☐ Error tolerant   Archive Report

Local intranet

And, we get a match to a Corticotropin protein

16

If we look at the details, we can see an accurate peptide mass as you would expect from this instrument. Note that the Mr value is 1 Dalton higher than the observed mass because the negative charge means that we lose rather than gain a proton.

Monoisotopic mass of neutral peptide Mr(calc): 2092.0788
Ions Score: 38 Expect: 0.031
Matches (**Bold Red**): 34/288 fragment ions using 61 most intense peaks

#	Immon.	a	a*	a0	b	b*	b0	c	d	Seq.	v	w	y	y*	y0	#
1	58.0298	58.0298		40.0193	86.0248		68.0142	103.0513		S						17
2	134.0611	221.0932		203.0826	249.0881		231.0775	**266.1146**		Y	1895.9820		2004.0395	1987.0130	1986.0290	16
3	58.0298	308.1252		290.1146	336.1201		**318.1095**	353.1467		S	1808.9500	1807.9547	1840.9762	1823.9497	**1822.9656**	15
4	102.0383	439.1657		**421.1551**	467.1606		449.1500	484.1871		M	1677.9095	1676.9143	**1753.9442**	**1736.9176**	1735.9336	14
5	100.0404	568.2083		550.1977	596.2032		578.1926	613.2297		E	1548.8669	1547.8717	**1622.9037**	1605.8771	**1604.8931**	13
6	108.0567	705.2672		687.2566	733.2621		715.2515	750.2886		H	1411.8080		**1493.8611**	1476.8346		12
7	118.0662	852.3356		834.3250	880.3305		862.3199	897.3571		F	1264.7396		**1356.8022**	1339.7756		11
8	127.0989	1008.4367	991.4101	990.4261	**1036.4316**	1019.4051	1018.4210	**1053.4582**	923.3727	R	1108.6385	1107.6432	**1209.7338**	1192.7072		10
9	157.0771	1194.5160	1177.4895	1176.5054	1222.5109	1205.4844	1204.5004	1239.5375		W	922.5592		**1053.6327**	**1036.6061**		9
10	28.0193	1251.5375	1234.5109	1233.5269	1279.5324	1262.5058	1261.5218	1296.5589		G			**867.5534**	850.5268		8
11					1390.6008	1389.6168	1424.6539	1322.5746		K	737.4428	736.4475	**810.5319**	793.5054		7
					1487.6536	1486.6695	1521.7067	1450.6695		P	640.3900	639.3948	**682.4369**	665.4104		6
					1586.7220	1585.7379	1620.7751	1561.7380		V	541.3216	554.3420	**585.3842**	568.3576		5
					1643.7434	1642.7594	1677.7965			G			486.3158	469.2892		4
					1771.8384	1770.8544	**1805.8915**	1703.8122		K	356.2052	355.2099	429.2943	412.2678		3
					1899.9333	1898.9493	1933.9864	1831.9071		K	228.1102	227.1150	301.1994	284.1728		2
										R	72.0091	71.0139	173.1044	156.0778		1

Ion types same as for +ve mode, but with one proton per charge subtracted rather than added.

Sometimes see some 'c' ions

Seq	ya	yb	Seq	ya	yb	Seq	ya	yb
YS	221.0932	249.0881	YSM	352.1336	380.1286	YSME	481.1762	509.1712
YSMEH	618.2351	646.2301	SM	189.0703	217.0652	SME	**318.1129**	346.1078
SMEH	455.1718	483.1667	SMEHF	602.2402	630.2351	ME	231.0809	259.0758
MEH	368.1398	396.1347	MEHF	515.2082	543.2031	MEHFR	671.3093	699.3042
EH	237.0993	265.0942	EHF	384.1677	412.1626	EHFR	540.2688	568.2637
HF	255.1251	283.1200	HFR	**411.2262**	439.2212	HFRW	597.3055	625.3005
HFRWG	654.3270	**682.3219**	FR	274.1673	302.1622	FRW	460.2466	488.2416

Finally, looking at the ions table, those of you with a great memory for residue masses and exceptionally good eyesight will instantly notice that the ion types are the same as in positive mode, but the masses are all 2 Dalton out because we have to subtract rather than add a proton.

It's also worth noting that we do indeed see some 'c' ions

Peptide Summary Report (../data/20050602/F004225.dat) - Microsoft Internet Explorer

File   Edit   View   Favorites   Tools   Help

Back   ·   Search   Favorites

Address   http://t41-dmc/mascot/cgi/master_results.pl?file=../data/20050602/F004225.dat

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
1	2091.0710	2092.0783	2092.0788	-0.0005	3	38	0.031	1	-.SYSMEHFRWGKPVGKKR.R

Top scoring peptide matches to query 1
Score greater than 23 indicates homology
Score greater than 36 indicates identity
Status bar shows all hits for this peptide

Score	Delta	Hit	Protein	Peptide
38.1	-0.00	1	COLI_BALPH	SYSMEHFRWGKPVGKKR
8.6	-0.04	2	CICK_RABIT	R.AQLVQALQAEAPARASGQQR.C
8.1	-0.08	3	FDHF_ECOLI	K.RLGIEDEALVWVHSRKGK.I
5.9	-0.06	4	ARGC_AQUAE	K.ERVGIESVIVHALSGVSGAGR.K
4.9	-0.04	5	RS7_FUSNN	K.ADRQQTLAIRWLTTYTR.A
4.6	-0.08	6	CWF5_SCHPO	R.LKVQWG...
3.6	0.00	7	SYO2_MOUSE	K.REVCAD...
3.4	-0.10	8	SFRA_MOUSE	R.RPAAAI...
3.0	-0.08	9	CUL3_HUMAN	K.KMQHNV...
2.8	0.05	10	RF3_ENTFA	K.EFSGFV...

COLI_MACNE
Matched peptides shown in **Bold Red**

```
  1 MPRSCCSRSG ALLLALLLQA SMEVRGWCLE SSQCQDLTTE SNLLECIRAC
 51 KPDLSAETPV FPGNGDEQPL TENPRKYVMG HFRWDRFGRR NSSSGSAHQK
101 REDVAAGEDR GLLPEGGPEP RGDGAGPGPR EGKRSYSMEH FRWGKPVGKK
151 RRPVKVYPNG AEDESAEAFP LEFKRELTGQ RPRAGDGPDG PADDGAGPRA
201 DLEHSLLVAA EKKDEGPYRM EHFRWGSPPK DKRYGGFMTS EKSQTPLVTL
251 FKNAIIKNAY KKGQ
```

1:COLI_BALPH

ASMS 2005

MATRIX SCIENCE

In Mascot 2.1 we now show the residues each side of the peptide. So, for example, still looking at this negative ions data, we see that this was an 'N' terminus peptide, indicated by the dash.

Looking at the popup, you can see the terminus residues - except for the first peptide. However, and this isn't a bug (or what we in the software industry commonly refer to as a 'feature') . The reason is that this peptide was found in a number of proteins. The top scoring protein was the Corticotropin from the finback whale, and we can see that this is at the N terminus

However, you still just see behind the yellow popup that this peptide was present in a number of proteins - if we look at the protein view for the pig-tailed macaque protein, we can see that it's in the middle of a (larger) protein and therefore does have a residue before the n terminus.

Clearly, showing the flanking residues would be very misleading in cases like this.

Where the peptide was only found in one protein, the flanking residues are displayed.

Detecting phospho peptides and phosphorylation sites in search engines is notoriously difficult. One of the reasons for the difficulty is that there are three different fragmentation channels. We can see intact fragments, loss of 80 and loss of 98. Some peptides just have one of these channels, others, in the same sample will have another channel, and some will have 2 or 3 channels in the same peptide.

In Mascot 2.0 and earlier, it was necessary to select a different modification for each neutral loss.

Furthermore, when using, for example, two modifications, one looking for no neutral loss and one looking for a loss of 98, only one fragmentation channel would be considered at a time - the combination of both losses wouldn't be considered together.

Modifications can now have multiple neutral losses

```
Title:Phospho (STY)
Residues:S 166.998359 167.0572
Residues:T 181.014010 181.0838
Residues:Y 243.029660 243.1532
NeutralLoss:97.976896 97.9952 0
NeutralLoss:0 0 0
*
```

ASMS 2005

This is a new definition in the Mascot 2.1 mod_file.

Firstly, note that we now choose to group all three possible sites into the same modification. Secondly, note that we have two of three possible neutral losses.

If you download the latest mod_file from unimod, then you won't get this in the mod_file - because this format wasn't supported in earlier versions of Mascot, and not everyone will have upgraded yet.

There's a good reason for not including the loss of 80 Daltons in the standard modification. Since the neutral loss is the same as the precursor delta, if there are multiple potential phosphorylation sites in a peptide, there will be ambiguity as to which site is phosphorylated. See the example in the Post translational modifications help page on our web site.
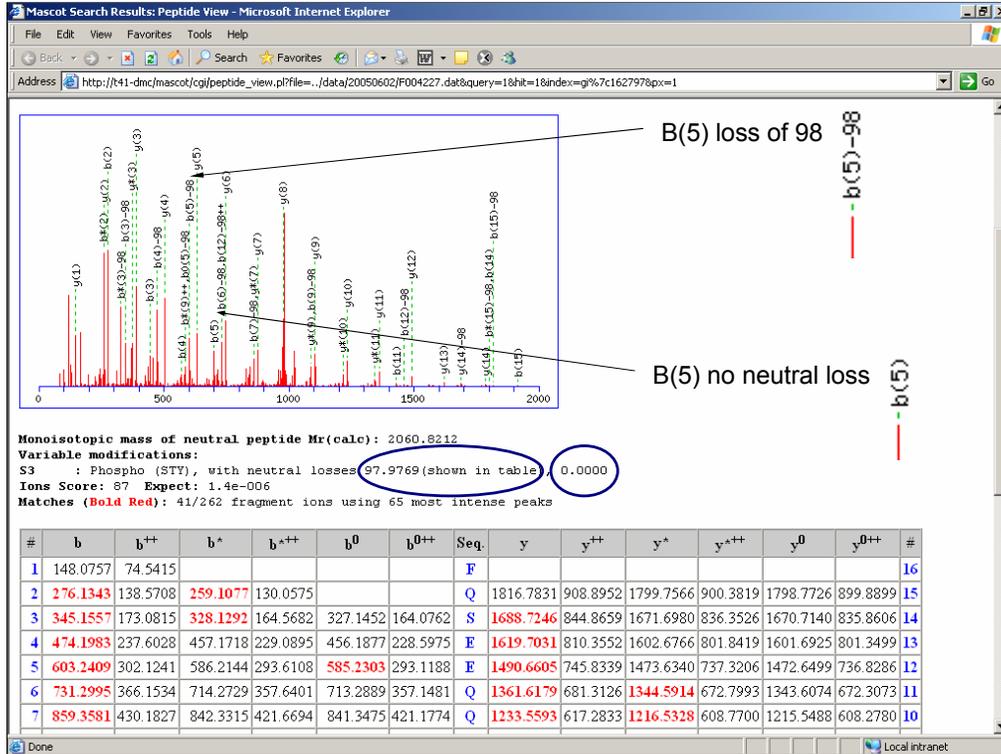
Just looking at an example, we can see a case here with two different neutral losses.

We now show the different neutral losses more clearly on the spectrum and also make it clear what is shown in the table

Neutral loss from precursor

- Neutral loss from precursor can be a guide to the identity of the phosphorylated residue
  - loss of 98 Da, then the expectation is phosphoserine or phosphothreonine
  - no neutral loss, then probably phosphotyrosine
- PepNeutralLoss - neutral loss from precursor so that this peak is not treated as noise
- ReqPepNeutralLoss performs same function but the peak *must* be present.

MATRIX
SCIENCE

We have added two new keywords:

Neutral loss from precursor

```
Title:Phospho (ST)
Residues:S 166.998359 167.0572
Residues:T 181.014010 181.0838
NeutralLoss:97.976896 97.9952 0
NeutralLoss:0 0 0
ReqPepNeutralLoss:97.976896 97.9952
*
```

MATRIX
SCIENCE

And here is an example that you could include in your mod_file if you
want.

Ambiguous charge states

- For many data, particularly ion trap, cannot determine the precursor charge state reliably
- Mascot allows "2+ and 3+" etc.
- In Mascot 2.0 and earlier, this creates 2 or more separate queries per spectrum
- At best, this leads to a lot of unmatched spectra
- At worst, it can result in false positives.

ASMS 2005

MATRIX SCIENCE

The next new feature that I want to describe is the new way in which we deal with unknown precursor charge states.

To summarise the problem, I'll describe what happens with Mascot 2.0 and earlier:

 - We often cannot determine the precursor charge state, but we know for example, that we are only likely to see 1+ 2+ or 3+ charge states from an ion trap

- So, Mascot 2.0 allows you to specify 2+ and 3+

- In Mascot 2.0 and earlier, this creates two or three separate queries for each spectrum. In effect, it is just duplicating the spectrum.

- If you had a data file with 100k spectra, and specified 1+,2+ and 3+ for all them, you would get 200 thousand spectra which couldn't possibly give a match.

- As a worst case, you can also get some false positives - I'll just show an example:

Peptide Summary Report

Format As   Peptide Summary

Significance threshold p< 0.05   Max. number of hits 20   Help

Standard scoring ○ MudPIT scoring ●   Ions score cut-off 0   Show sub-sets ☐

Show pop-ups ● Suppress pop-ups ○   Sort unassigned Decreasing Score   Require bold red ☐

Select All   Select None   Search Selected   ☐ Error tolerant

1.   Q08211   Mass: 142099   Score: 136   Queries matched: 2
     (DHX9_HUMAN) ATP-dependent RNA helicase A (Nuclear DNA helicase II) (NDH II) (DEAH-box protein 9) ATP-dependent RNA he
     ☐ Check to include this hit in error tolerant search

     Query Observed Mr(expt) Mr(calc) Delta Miss Score Expect Rank Peptide
   ☑   1   1243.87   2485.73   2485.19   0.54   0   97   4.3e-05   1   K.NELMYQLEQDHDLQAILQER.E
   ☑   2   1243.87   3728.60   3730.78 -2.18   0   76   0.0056   1   F.ATPEQISMDLKNELMYQLEQDHDLQAILQER.E + 2 Oxidation

2.   Q13207   Mass: 742_   Score: 9   Queries matched: 1
     (TBX2_HUMAN) T-box transcription _tor TBX2 (T-box protein 2) T-box transcription factor TBX2 (T-box protein 2)
     ☐ Check to include this hit in error t___ search

     Query Observed Mr(expt) Mr(calc) Delta M   Expect Rank Peptide
      2   1243.87   3728.60   3725.75   2.85   0   _02   2   G.AFSAMGMGHLLASVAGGGNGGGGGPGTAAGLDAGGLGPAASAAST.A

3.   P00751-01-05-00   Mass: 69911   Score: 9
     (CFAB_HUMAN) Splice isoform 2; Variant allel_   actor B precursor (EC 3.4
     ☐ Check to include this hit in error tolerant s

Same spectrum, 2+ and 3+

Internet

Here is an example from Mascot 2.0 where the 2+ and 3+ option has been applied to a single spectrum. The correct charge state in this case is 2+ , but you can see that there are significant matches for both 2+ and 3+

if you look carefully, you will see that the 3++ peptide contains the 2++ peptide. It just has more residues on the n terminus. If we look at the peptide view, then

#														
7	727.36	364.18	710.34	355.67	709.35	355.18	S	3092.47	1546.74	3075.44	1538.22	3074.46	1537.73	25
8	874.40	437.70	857.37	429.19	856.39	428.70	M	3005.43	1503.22	2988.41	1494.71	2987.42	1494.22	24
9	989.42	495.22	972.40	486.70	971.41	486.21	D	2858.40	1429.70	2841.37	1421.19	2840.39	1420.70	23
10	1102.51	551.76	1085.48	543.24	1084.50	542.75	L	2743.37	1372.19	2726.35	1363.68	2725.36	1363.18	22
11	1230.60	615.81	1213.58	607.29	1212.59	606.80	K	2630.29	1315.65	2613.26	1307.13	2612.28	1306.64	21
12	1344.65	672.83	1327.62	664.31	1326.64	663.82	N	2502.19	1251.60	2485.17	1243.09	2484.18	1242.59	20
13	1473.69	737.35	1456.66	728.83	1455.68	728.34	E	2388.15	1194.58	2371.12	1186.07	2370.14	1185.57	19
14	1586.77	793.89	1569.75	785.38	1568.76	784.88	L	2259.11	1130.06	2242.08	1121.54	2241.10	1121.05	18
15	1733.81	867.41	1716.78	858.89	1715.80	858.40	M	2146.02	1073.52	2129.00	1065.00	2128.01	1064.51	17
16	1896.87	948.94	1879.85	940.43	1878.86	939.93	Y	1998.99	1000.00	1981.96	991.48	1980.98	990.99	16
17	2024.93	1012.97	2007.90	1004.46	2006.92	1003.96	Q	1835.92	918.47	1818.90	909.95	1817.91	909.46	15
18	2138.01	1069.51	2120.99	1061.00	2120.00	1060.51	L	1707.87	854.44	1690.84	845.92	1689.86	845.43	14
19	2267.06	1134.03	2250.03	1125.52	2249.05	1125.03	E	1594.78	797.89	1577.76	789.38	1576.77	788.89	13
20	2395.12	1198.06	2378.09	1189.55	2377.10	1189.06	Q	1465.74	733.37	1448.71	724.86	1447.73	724.37	12
21	2510.14	1255.57	2493.12	1247.06	2492.13	1246.57	D	1337.68	669.34	1320.65	660.83	1319.67	660.34	11
22	2647.20	1324.10	2630.17	1315.59	2629.19	1315.10	H	1222.65	611.83	1205.63	603.32	1204.64	602.83	10
23	2762.23	1381.62	2745.20	1373.10	2744.22	1372.61	D	1085.59	543.30	1068.57	534.79	1067.58	534.30	9
24	2875.31	1438.16	2858.29	1429.65	2857.30	1429.15	L	970.57	485.79	953.54	477.27	952.56	476.78	8
25	3003.37	1502.19	2986.34	1493.68	2985.36	1493.18	Q	857.48	429.25	840.46	420.73	839.47	420.24	7
26	3074.41	1537.71	3057.38	1529.19	3056.40	1528.70	A	729.43	365.22	712.40	356.70	711.41	356.21	6
27	3187.49	1594.25	3170.47	1585.74	3169.48	1585.24	I	658.39	329.70	641.36	321.18	640.38	320.69	5
28	3300.58	1650.79	3283.55	1642.28	3282.57	1641.79	L	545.30	273.16	528.28	264.64	527.29	264.15	4
29	3428.63	1714.82	3411.61	1706.31	3410.62	1705.82	Q	432.22	216.61	415.19	208.10	414.21	207.61	3
30	3557.68	1779.34	3540.65	1770.83	3539.67	1770.34	E	304.16	152.58	287.13	144.07	286.15	143.58	2
31							R	175.12	88.06	158.09	79.55			1

For the 3++ results, we will see that the reason is simply that the y ions were dominant, and that these of course all still match. This is always likely to occur for some spectra in a large data set when using wide peptide tolerances with no enzyme and several variable modifications.

Before Mascot 2.1, there wasn't much that you could do about this.

Mascot 2.1 with multiple charges

Select
1+, 2+ and 3+

(or 2+ and 3+)

MASCOT MS/MS Ions Search

ASMS 2005

Following the same example with Mascot 2.1, I've selected 1+, 2+ and 3+

And we see now that we have just one query and one significant match. All that Mascot is doing is to iterate through the charges, and just keep the charges with the highest score.

As I've already explained, The other advantage is that the list of unmatched peptides will be much smaller.

Miscellaneous changes

- Allow fixed and variable mod for same residue.
- Remove MaxAccessionLen
- Multiple accessions for identical sequences
- Parent / precursor cut-out window in MS/MS
- Only compress one database at a time
- Can load larger results files
- minPepLenInSearch.

ASMS 2005

MATRIX SCIENCE

There are number of smaller, but still significant changes that were introduced in the Mascot 2.1 search engine:

In cases where you want to specify that a residue should have one of two possible modifications, but shouldn't be allowed to be unmodified, you can now select one as fixed and one as variable. This would for example, be sensible for ICAT.

We've removed the MaxAccessionLen parameter that has caused some grief for Mascot system administrators. If you've not had an issue with this, then you never will.

We have also done something to address the issue of the database providers changing accessions. For a non-redundant database such as the NCBInr database, identical sequences are only included in the fasta once, but the accession for each of the source sequences are available on the description line. The problem occurs when the database provider changes the order of the accessions - Mascot 2.0 and earlier would only look at the first accession. This could mean that if you looked at an old result file with a newer database, then Mascot wouldn't have been able to find the protein and display the protein view.

You can now configure the precursor cut-out window. It now removes a smaller region depending on the charge state.

Start up of a Mascot server should now be faster if you a lot of databases - particularly if they all need updating.

We've decreased the memory requirements when loading large results files

Some searches, particularly when using a large modification such as ITRAQ could be very slow because it would try to match 3mers. This is now configurable - default is 5 residues.

New features in Mascot 2.1

- Export results in various formats
- Search engine and report enhancements
- Mascot Daemon enhancements
- Mascot Security.

ASMS 2005

So, we have looked at exporting results and search engine and report enhancements. Moving on to Mascot Daemon:

Merge a batch of data files

- For example, multiple fractions from a MudPIT run

You can now specify that you want to merge a whole group of raw files into one big search file. To do this simply drag all the files into the box here, click the check box and then Run.

For obvious reasons, you can't do this with a monitor task. For a monitor task, it doesn't know when it has finished, so it would never know when to stitch all of the files together and submit them.
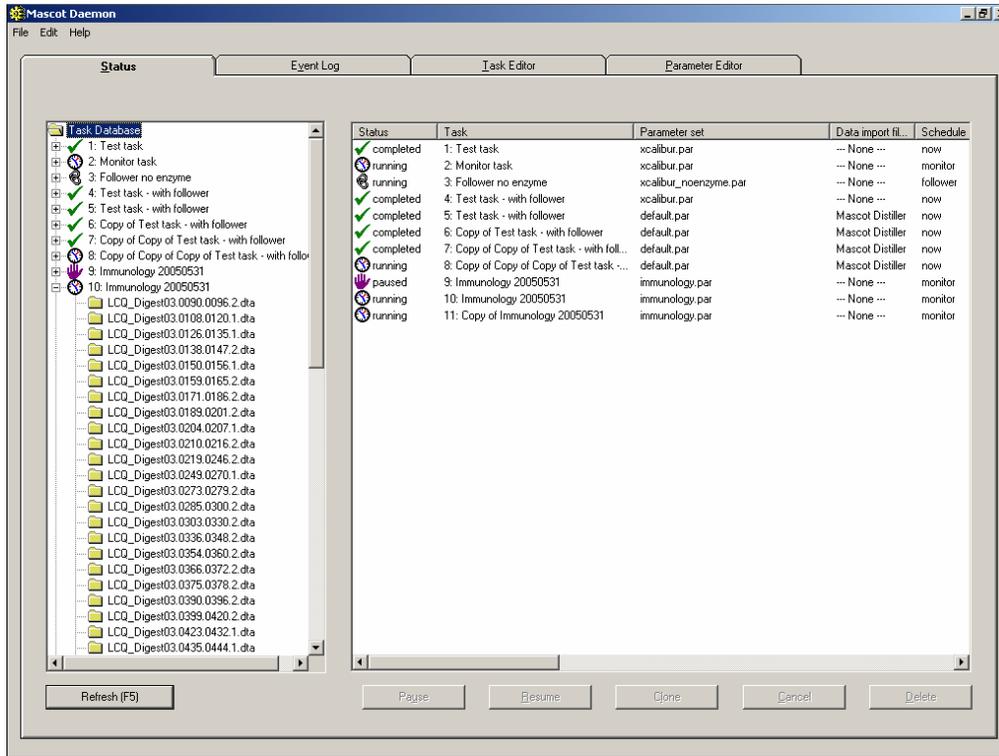
I don't need to say anything about this slide. (Daemon Window is now resizable)

If you haven't used Mascot Daemon before, it won't mean anything to you anyway. If you use Daemon regularly, then you will probably appreciate it - ah do I hear some subdued applause from the back?

Peak lists saved to disk

- In Mascot 2.0 and earlier, when an import filter (for example Mascot.dll for Sciex, or Distiller) created a peak list this was a temporary file
- Creating the peak list often takes time. . . . . .
- Peak list and project file now saved in a directory

MATRIX SCIENCE

34

Peak list and Distiller projects
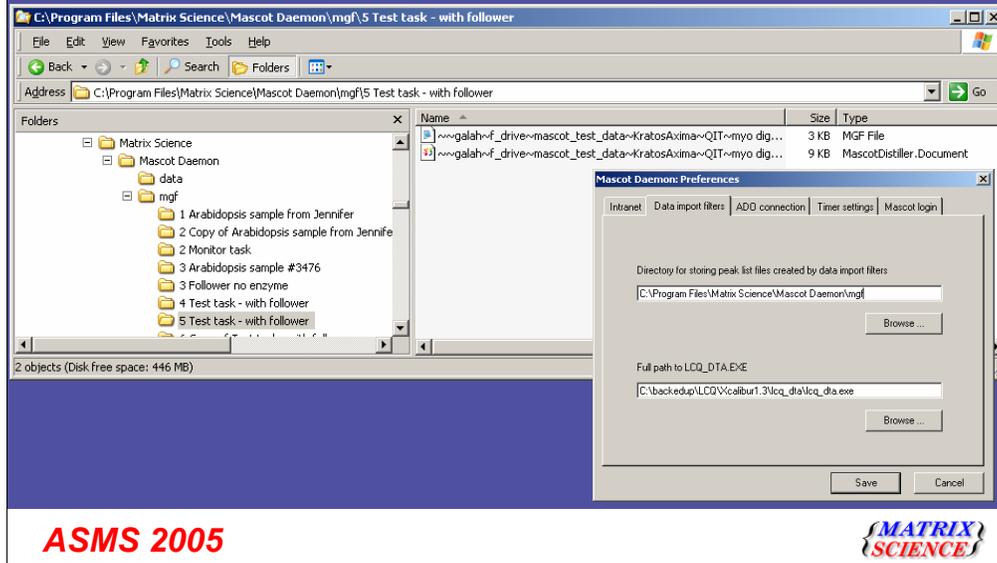
ASMS 2005

Of course, the directory that it saves the files in is configurable.

However, what is really smart is if you use Mascot Distiller to do the processing:

Integrating Distiller and Daemon
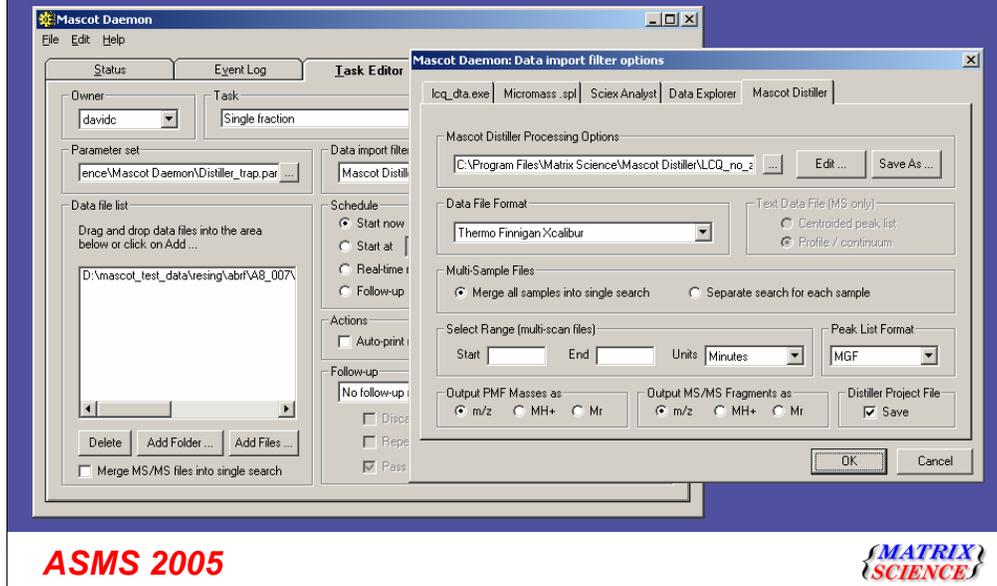
ASMS 2005

I'll just show a simple example - one raw file. I've chosen to use Mascot Distiller for the import filter, and we now have a few extra options here, including the one for saving the project file.

New 'link' to Distiller
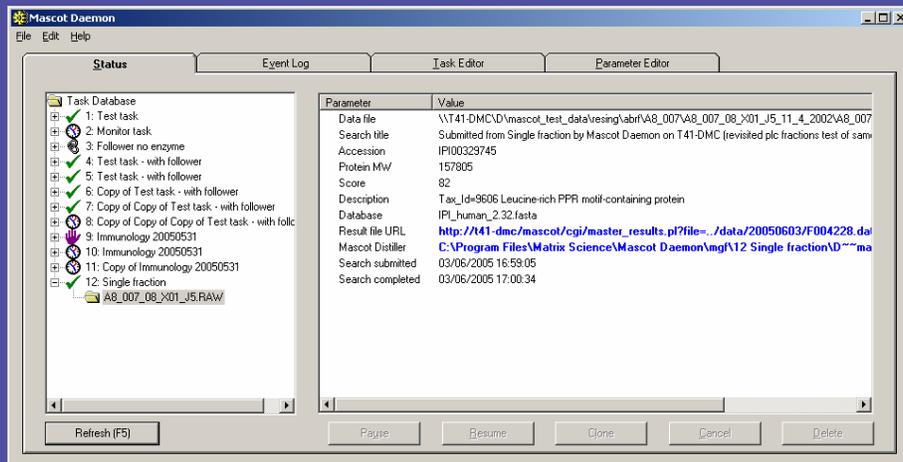
ASMS 2005

The processing is done, and we now have an extra link in the results.
Click on the link, and this will open Distiller:
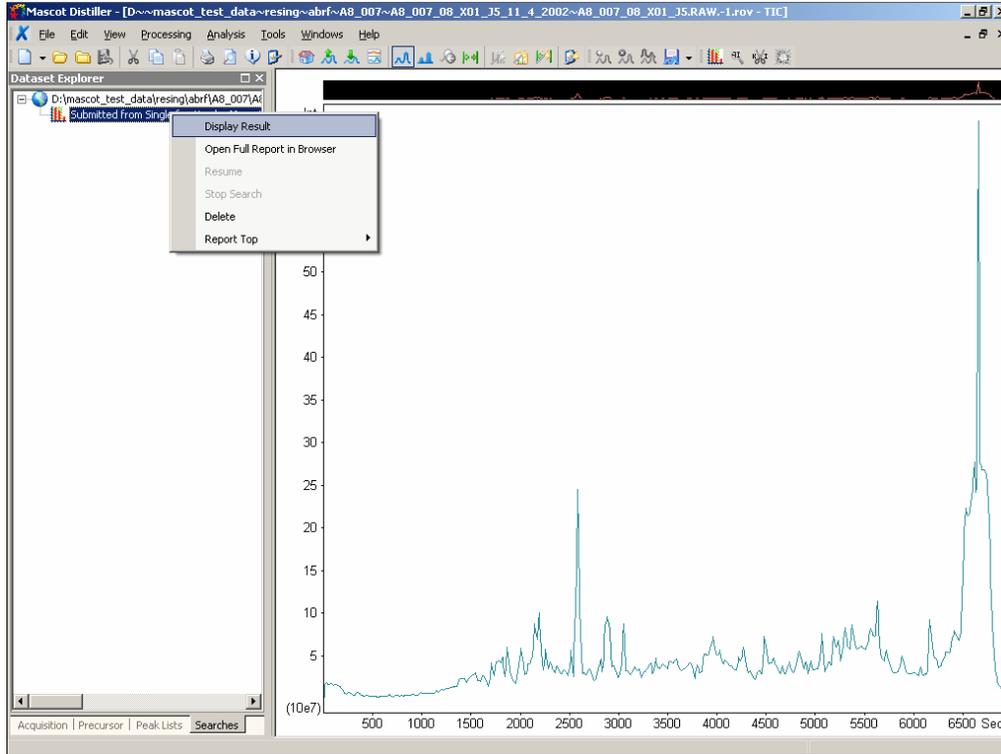
Which opens the project and the raw file in Mascot Distiller

There is an extra tab now for searches and you can see that there is currently just one search against this raw file saved in the project.

Next, you can choose to display the results and
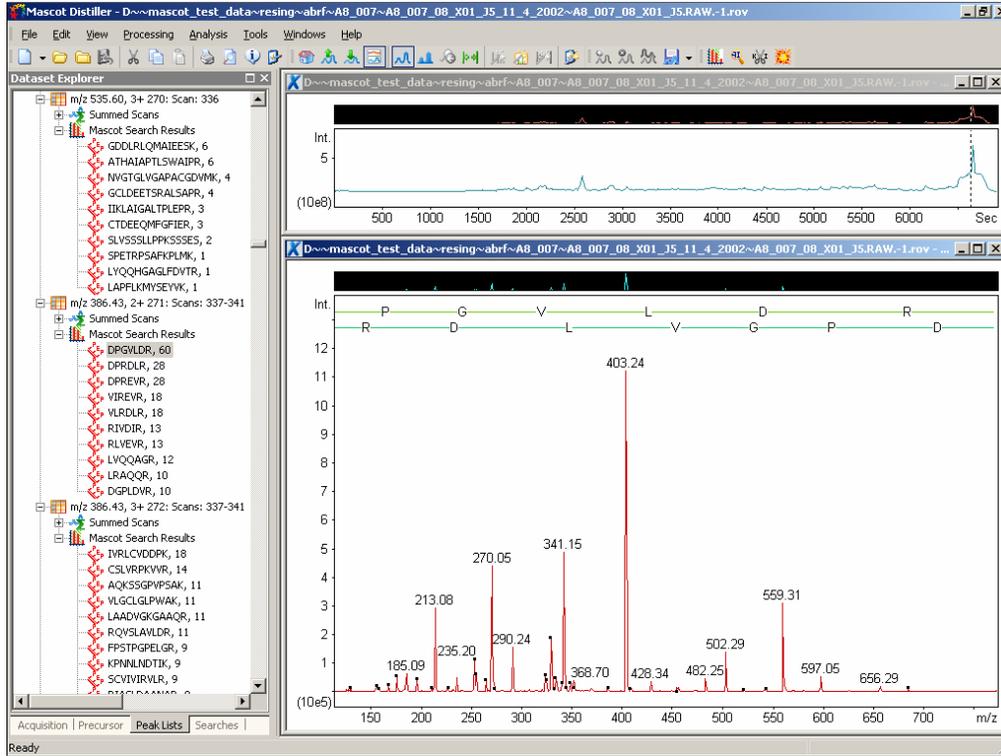
You can then see the matches.

If you see a good looking spectrum with no significant matches, you can then use Denovo, as John explained earlier.

- Merge a batch of data files into single search
- Window is re-sizeable
- Peak lists from data import filters saved to disk
- Option to save Mascot Distiller project file
- No limits on the size of data files
- Work-around if problem with running as a service.

MATRIX SCIENCE

We have now seen most of the new functionality in Daemon. We can now merge a batch of data files into a single search, the window is re-sizeable, peak lists from all data import filters are now saved to disk, and, with Distiller 2.0, there is the option to save project files.

Two other significant changes are that there is no longer a limit on the file size of data files. Mascot Daemon version 2.0 used to run out of memory in some cases.

In addition, in some organisations with particularly over-zealous IT departments, it was difficult for some people to run Daemon as a service, so there is now an option to run it just as a normal application.

New features in Mascot 2.1

- Export results in various formats
- Search engine and report enhancements
- Mascot Daemon enhancements
- Mascot Security.

MATRIX SCIENCE

Mascot Security

- Why?
- The 'end user experience'
- Basic principles - groups, users and tasks
- Enabling / disabling security
- Using the administration utility
- Example - a core lab
- Workarounds for third party applications
- Hints and tips.

MATRIX SCIENCE

42

Why?

DB confidentiality - Mozilla Thunderbird

File Edit View Go Message Tools Help

Get Mail   Write   Address Book   Reply   Reply All   Forward   Delete   Junk   Print   Stop

Subject: DB confidentiality
From:
Date: 22/04/2005 11:52
To: David Creasy

Hello David,

I am wondering if there is a simple way to configure Mascot to achieve the following :
One of our user in the institute would like to use a confidential database for his searches – he would therefore would like to have this database hidden from other users across the site.
Thanks in advance for your help,

Marielle
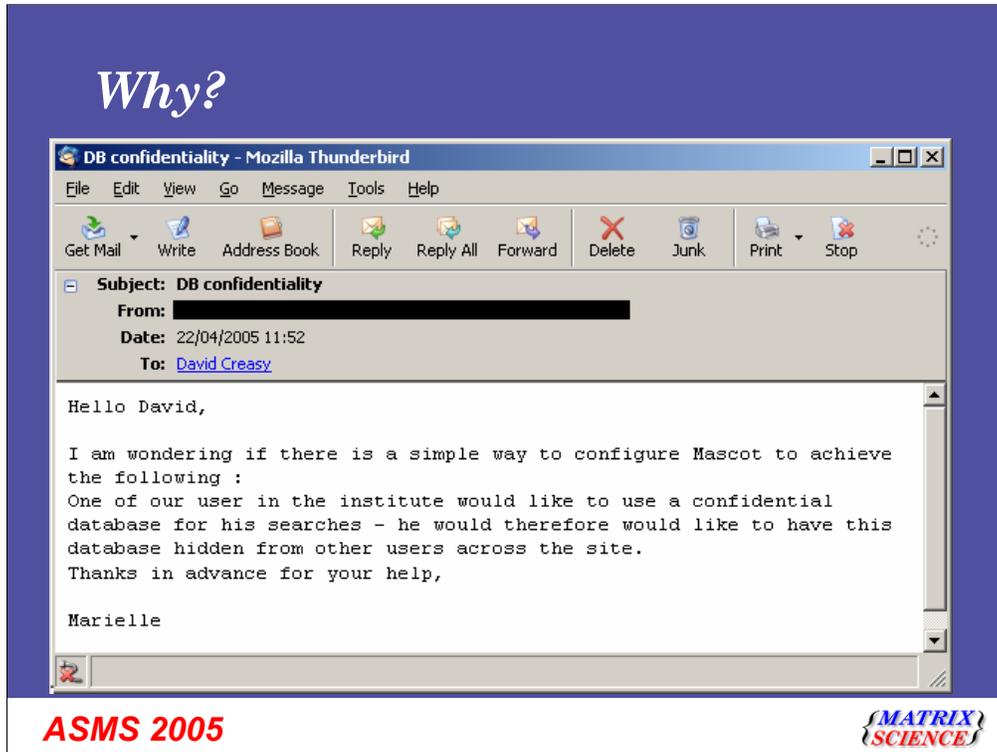
ASMS 2005   *MATRIX SCIENCE*

We have had quite a number of requests like this one over the years. To be honest, I had assumed that as databases have seemed to become more public,the number of these requests would be reduced. However, we still get asked this quite often.

Why?

Re: Large mascot server - Mozilla Thunderbird

File   Edit   View   Go   Message   Tools   Help

Get Mail   Write   Address Book   Reply   Reply All   Forward   Delete   Junk   Print   Stop

Subject: **Re: Large mascot server**
From: John Cottrell <jcottrell@matrixscience.com>
Date: 22/10/2004 13:58
To:

Dear John,

I am trying to float the idea of having a large mascot server centralised here that disparate groups could use on a contribution basis. One immediate thing is data security, how would I allow access for separate groups to get their data without them seeing others data and vice versa. This would also include users without daemon, or would

ASMS 2005

MATRIX SCIENCE

This is a much more common request - I'll just let you read it.

The end user experience

ASMS 2005

Just to give you an idea of what the end user will see when they use a Mascot 2.1 server with security enabled.

Firstly, they can access all of the help pages - they are freely available on our public web site, so there seems no point in blocking them.

However, when the user comes to click on one of the Mascot Search forms, they are given a requested to log in:

Their name and email address is then filled into the form, and they can see that they are logged in up here.

Users can edit their own profile

Mascot Security

User ID:	1002
Name:	davidc
Full Name:	David Creasy
Email address:	dcreasy@matrixscience.com

Save Changes

Change password

And change their password

ASMS 2005

Obviously a user can edit their own profile by clicking on the edit link, and they can edit their own name and email address.

The administrator can disable this functionality for groups of people if they wish.

Obviously a user can change their password if they want.

Search log - only see own searches

ASMS 2005

One of the requirements is to be able to see the search log - in this case I have only been given rights to see my own searches.

Basic principles - Users

- Login name
- Password, password expiry
- Full name, email address
- Account enabled / disabled
- Member of one or more groups.

MATRIX
SCIENCE

I'll quickly cover some basic principles:

Basic principles - Groups

- Name
- List of members
- List of allowed tasks.

MATRIX SCIENCE

A group is just a collection of users.

Note that for simplicity, allowed tasks can only be applied to groups - they cannot be applied to individual users.

49

Basic principles - Tasks

- For example
  - Allow PMF search
  - Allow MS/MS search
  - Maximum number of queries
  - Can view the search log
  - Can search specific databases
  - Can view other peoples results.

MATRIX
SCIENCE

There are 30 different tasks that members of a group can be allowed to perform - for example:

Enable / disable security

ASMS 2005
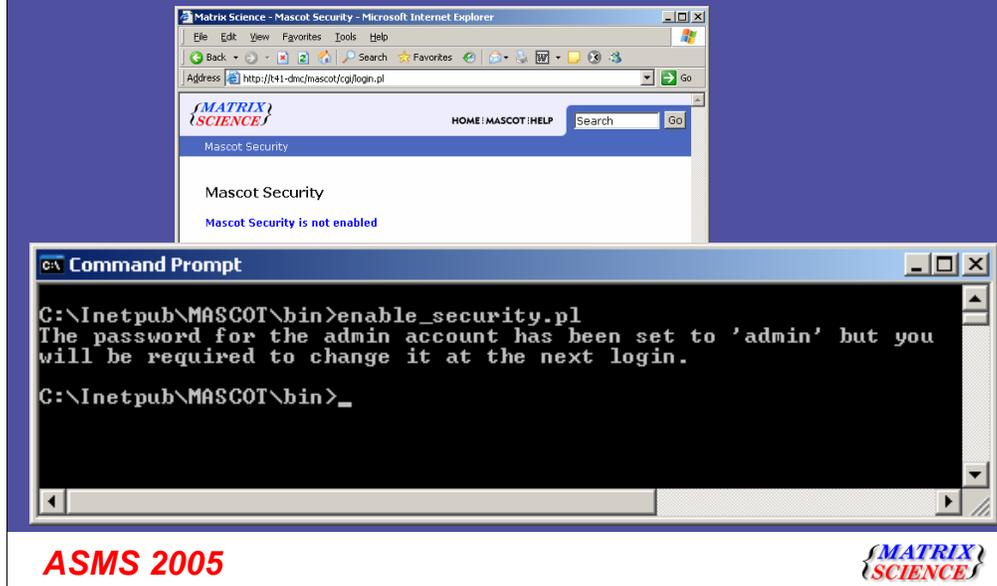
When Mascot 2.1 is first installed, Mascot security is disabled. So, if you try and login you will see this message.

To enable security, open a command prompt or a shell on the Mascot server, and change to the mascot/bin directory. Then type enable_security.pl.

This takes a few seconds. If you forget the administrator password, it can always be reset to admin by running this script again.
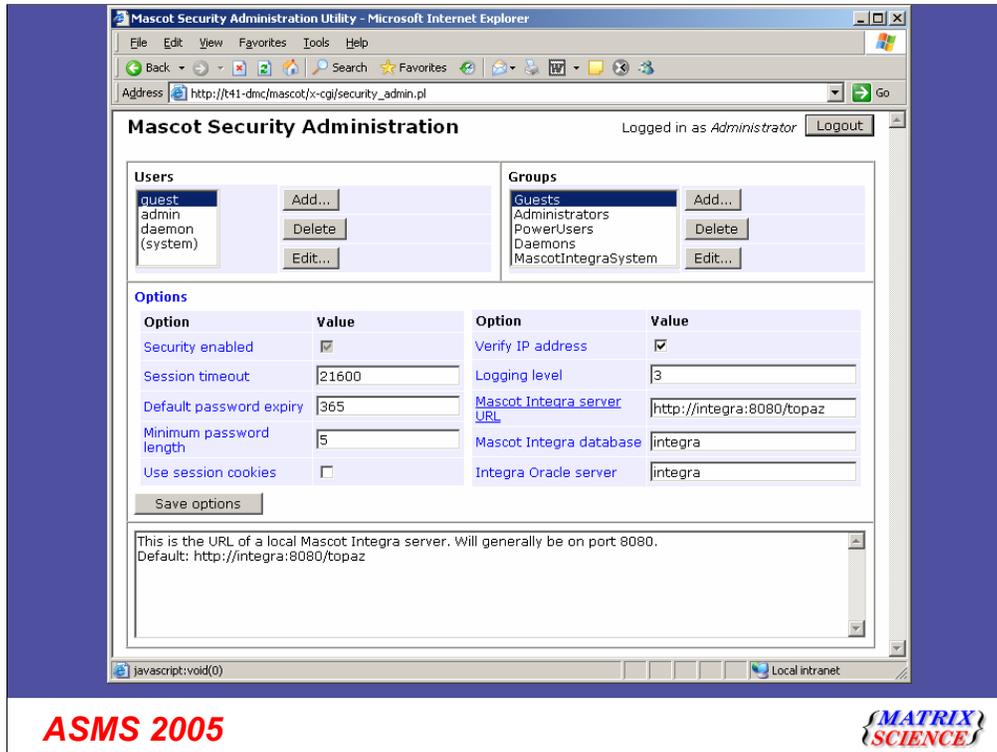
The security administration utility looks like this.

There is a list of users, a list of groups and some options.

If you hold the mouse over any of the blue text, help will appear at the bottom of the screen. I'll quickly run through the options.

You can't disable security from here - it has to be done on the server using the disable security command

The session timeout is in seconds. After this period of inactivity, the user will be required to login again.

The password expiry time is in days. After this period of time, the user will be required to enter a new password.

Any new password must be at least this length. Maximum length is 50. Um, think I'll set the minimum length to 25 on our server back in the office - that should keep people on their toes.

Session cookies are automatically destroyed when the browser is closed. With some browsers, session cookies are not shared when a new instance of the browser is opened, which might mean that a user has to login again for each new window opened.

If verify the IP address is set, then any request to perform a privileged action will compare the IP address that the request is coming from with the one originally used to login.

Logging level should normally be left at 3

To add a user, simply click on the Add User button

ASMS 2005
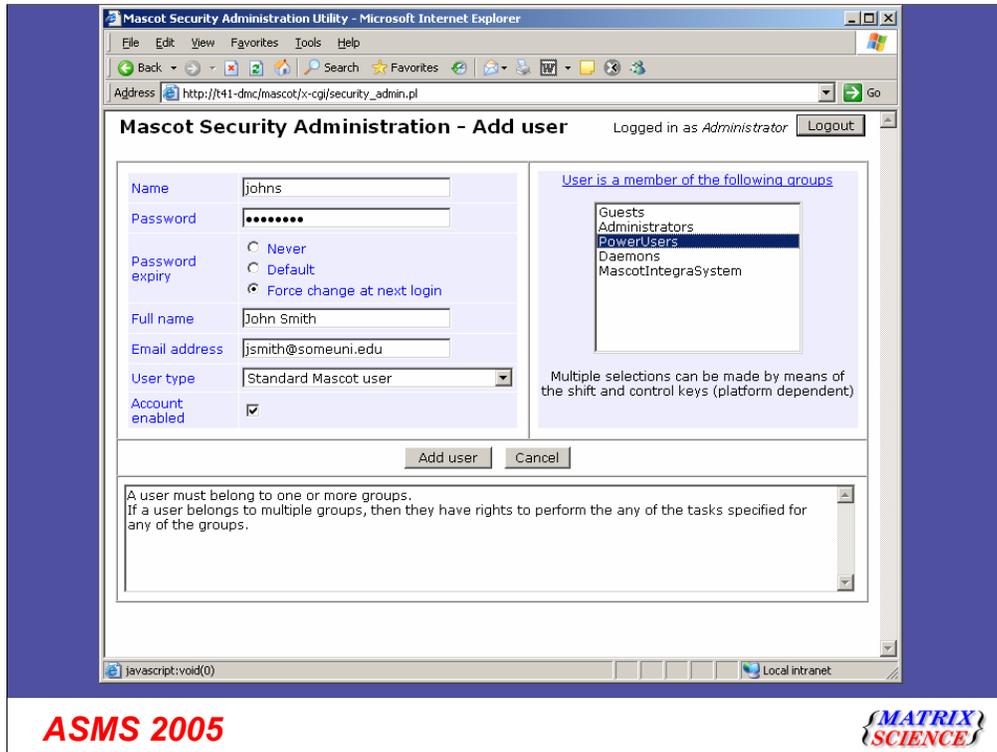
Enter a name and password.

It's often a good idea to force the user to enter a new password when they first login

Enter their full name and email address.

I'll return to the choice of user types later - most users should just be standard Mascot users.

Make sure that the account is enabled, and then select one or more groups for the user to belong to.

ASMS 2005

ASMS 2005

MATRIX SCIENCE

When editing a group, you can change the name - it is the unique ID that remains constant. Users can be added to, or removed from the group like this

This is the list of tasks that members of the group cannot perform. So, for example, if I want to add restrict the number of searches that members of this group can perform, I simply click on the task, then Add Task. In this case I need to enter the number as well.

Finally, nothing is saved until you click on Save Changes.

Mascot Security

- Why?
- The end user 'experience'
- Basic principles - groups, users and tasks
- Enabling / disabling security
- Using the administration utility
- Example - a core lab
- Workarounds for third party applications
- Hints and tips

MATRIX SCIENCE

I'd now like to take you through an example - in this case for a core lab.

Example 1 - Core lab

- 2 people in the core lab: Jack and Jill
- Many departments submit samples
- Most departments have > 1 person
- Each department should not be able to see other departments searches
- Immunology department: Ian and Mary

ASMS 2005

MATRIX
SCIENCE

Core lab - add a group for each 'customer'

Mascot Security Administration Utility - Microsoft Internet Explorer

File   Edit   View   Favorites   Tools   Help

Back   Search   Favorites

Address   http://t41-dmc/mascot/x-cgi/security_admin.pl   Go

Mascot Security Administration – Add group   Logged in as *Administrator*   Logout

Name   Immunology   Add...   Cancel

Help window.
Type a new group name, and press the 'Add...' button.
You will then be able to assign users and permitted tasks to the new group.

Done   Local intranet

ASMS 2005   *MATRIX SCIENCE*

We can add users or groups first - doesn't matter. I'll add the group first
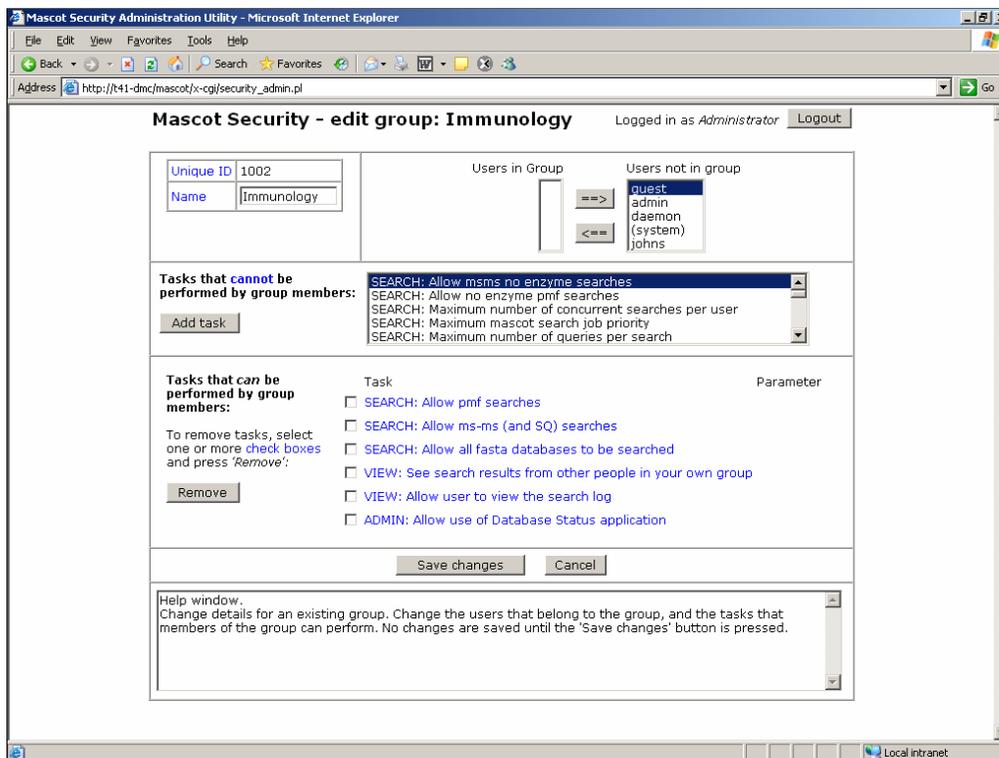
And I'll set up some rights for the group.

As you can see, they are allowed to perform PMF and MS-MS searches - I've added this so that they repeat searches with different parameters if they want, but I'm not allowing them to do no enzyme searches. They are allowed to use all the different databases, can see each others searches. Finally, they can use the search logs and status screens to find their own searches.
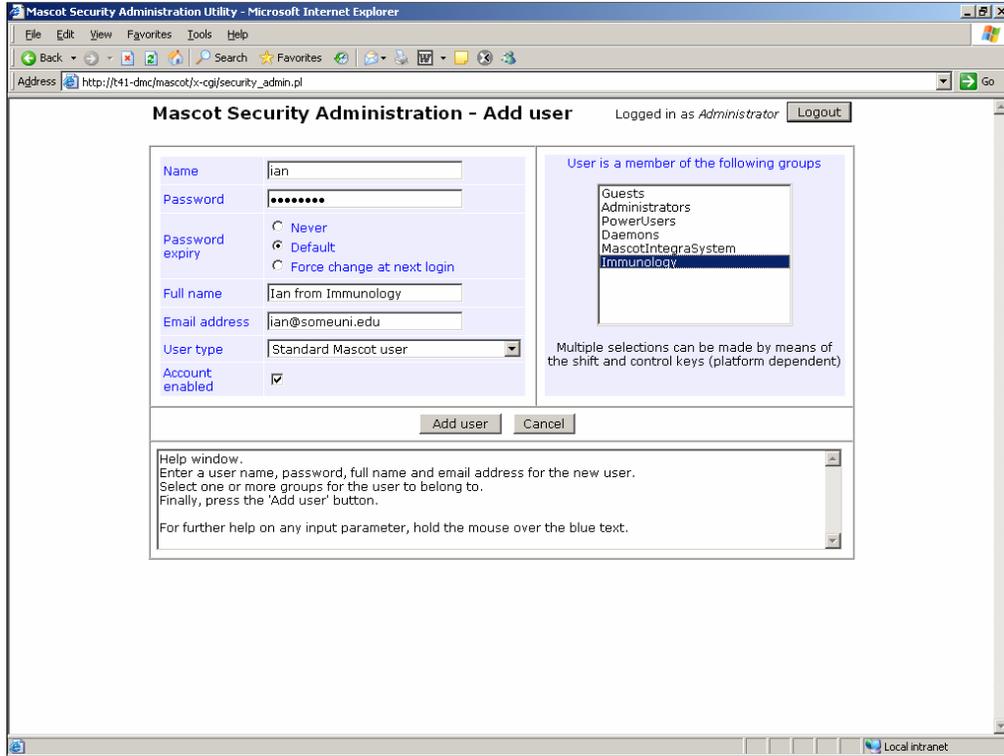
We now add Ian as a user, and assign him to the immunology group.
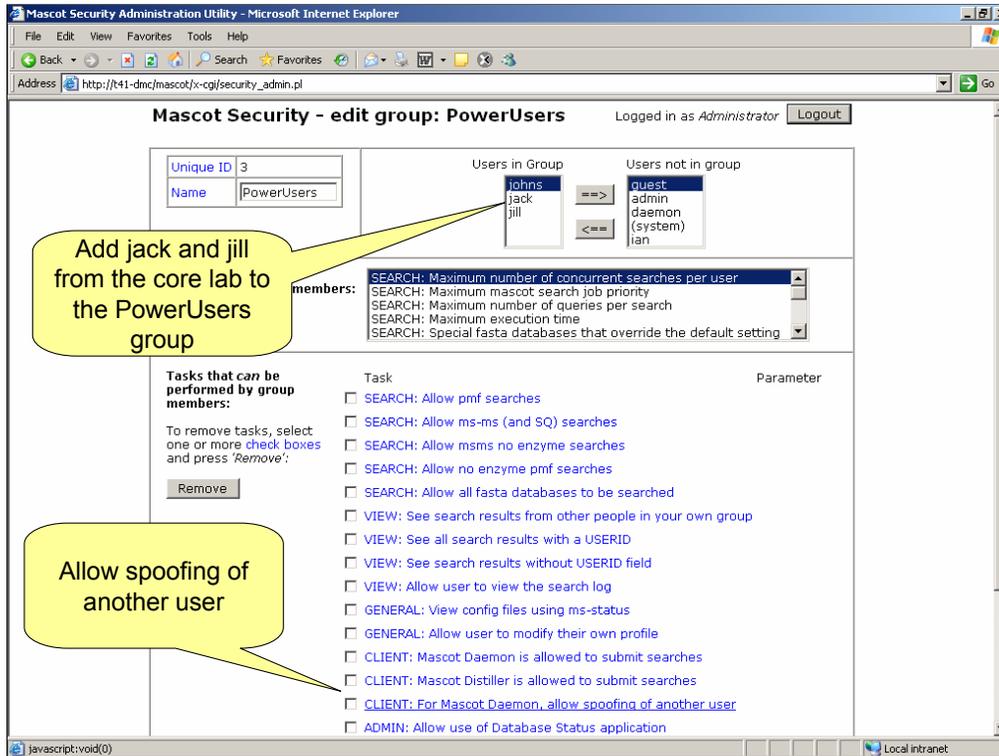
Next we need to set up the hard workers in the core lab - jack and jill. We simply create the two users and add them to the group.

Also, note that I have added this task which allows 'spoofing' of another user- we will come to that again in a minute.

Use Daemon to run searches
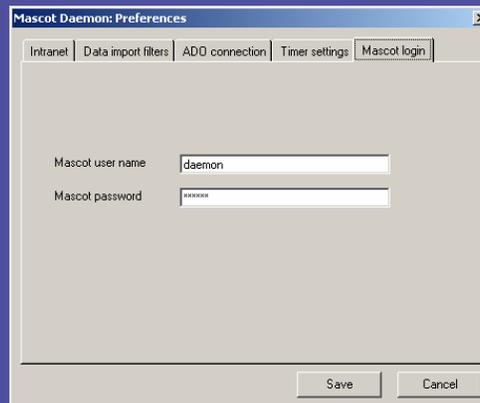
We are nearly ready to start searches. Just before we start, we need to setup Daemon to login as somebody. By default, there is a 'daemon' user, and you can use this user on all the daemons. Before you do that, you will need to enable the daemon account and set a password using the security administration utility. Then, simply enter this name and password in the Daemon settings.

Alternatively, jack and Jill could put their own user name and passwords in here.
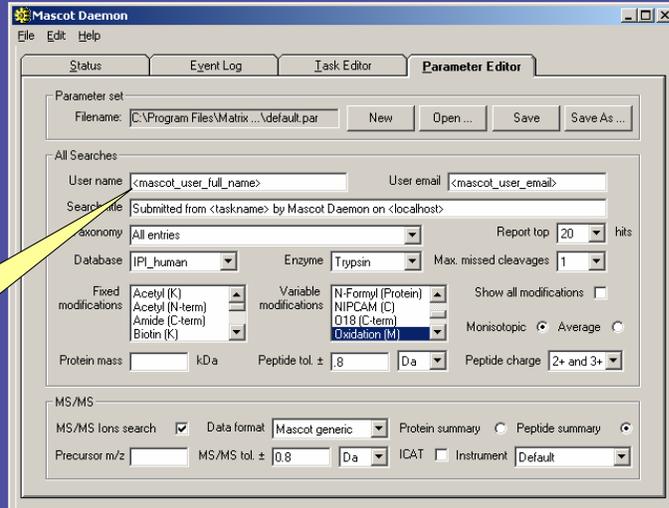
When you set up parameters, use these special tags for the user name and email address. They will then be substituted by the spoofed name.
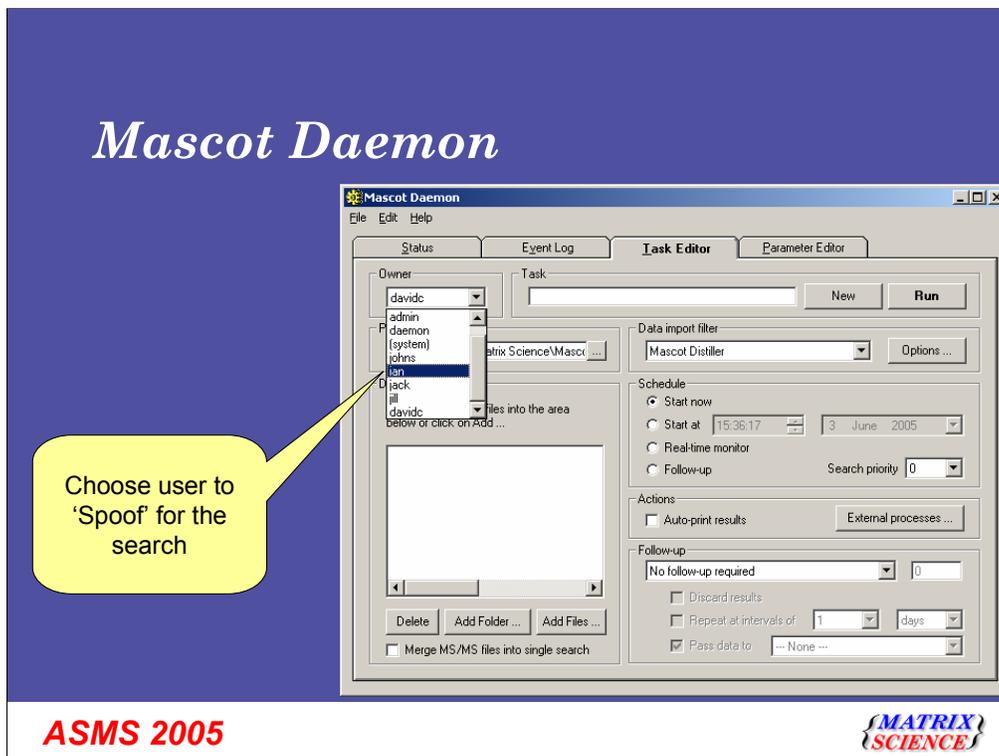
In the Task tab, we set up the 'owner' for the searches. In this case, Jill will choose ian as the owner.

Jack Core now runs all his searches for Ian. When Jack or Jill are logged in, because they are power users, they can see all of the searches.

But when Ian from immunology is logged in, he can only see his own searches, or those submitted by Jack on his behalf.

Core lab - less access for users

- If you don't want the customers to be able to do any searches, but just view the results, then just give them the access to tasks:
  - SEARCH: Allow all fasta databases to be searched
  - VIEW: See search results from other people in your own group
  - VIEW: Allow user to view the search log
  - ADMIN: Allow use of Database Status application

MATRIX SCIENCE

As I said earlier, you may want some groups to just be able to view results. In this case, only give them rights to perform these tasks:

 - You must allow all fasta databases to be searched otherwise they won't be able to see the protein view. However, they can't perform PMF or MS/MS searches.
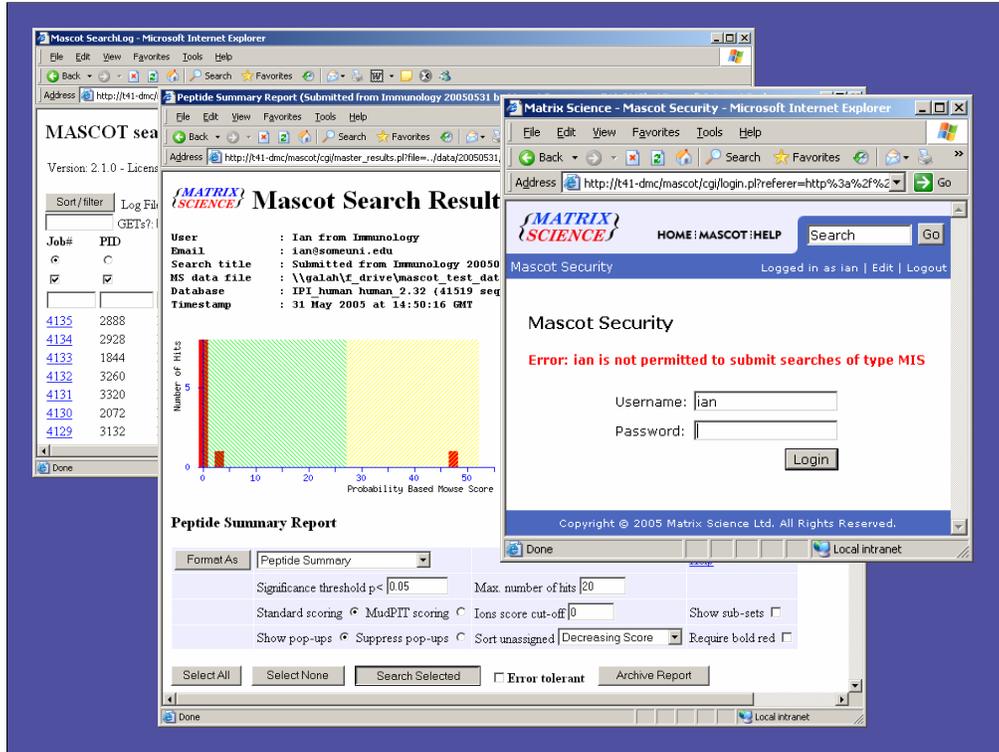
So in this case, Ian from immunology can view the search log, see his results, but when he tries to do a repeat search, he is denied access

Supporting 3rd party applications

- Older applications that submit searches to Mascot won't have their own login
- Try logging in using Internet Explorer before running the application - cookies
- Use one of the 'special' user types:

User type	Standard Mascot user ▼
Account enabled	Standard Mascot user
	Mascot Integra user
	Computer name
	IP address
	Agent string
	User authenticated using web server

ASMS 2005

MATRIX SCIENCE

There is a potential issues with other applications that interface with Mascot but don't yet have code to support the login functionality.

Since session ids are saved as cookies, and since most Windows applications that access web sites use internet explorer libraries, it is worth just trying to login in an Internet explorer window before starting the application.

Alternatively, you can use one of the special user types.

Special user types

- Computer name / IP address
  - Never have to log in from that computer
  - Use the computer name / IP address as the 'name'
- Agent string
  - Can determine the agent string from the web server logs
  - Not secure because someone could create another app to use this agent string
- Web server authentication.

ASMS 2005

MATRIX SCIENCE

70

I can't stress enough that you should plan what you intend to do before you start. Think carefully about what groups you want to create

As any Unix administrator will tell you, it's always best to separate your administration and user tasks. However, we can't force you, but that is why the default admin user cannot submit searches.

Check Session - Microsoft Internet Explorer

File   Edit   View   Favorites   Tools   Help

Back ▾ ◯ ▾ ⤫ ⟳ ⟲ | Search ⭐ Favorites ⟳ | ⟳ ▾ ⟳ W ▾ ⟳ ⟳

Address http://t41-dmc/mascot/cgi/security_session.pl   ▾ → Go

This utility shows session information for when Mascot Security is enabled.
You may be asked to give this information to a support engineer if you have security problems

Retrieved cookie value is: ian_207301854233611

ID	ian_207301854233611
security enabled	1
last accessed	31 May 2005 15:24:56
ip address	192.168.8.136
user	ian
user ID	1004
full username	Ian from Immunology
email address	ian@someuni.edu
valid	1

Permitted tasks

Task	Param type	Parameter
SEARCH: Allow all fasta databases to be searched	None	
VIEW: See search results from other people in your own group	User list	1004
VIEW: Allow user to view the search log	None	
ADMIN: Allow use of Database Status application	None	

Done   Local intranet

ASMS 2005

MATRIX SCIENCE

You can see the list of tasks that the user can perform

72

Hints and tips - 2

- Be careful when adding someone to multiple groups
- Don't forget to enable the Daemon account and set a new password
- View the log file periodically
- Moving Mascot to another PC? Just copy user.xml, group.xml and security_options.xml.

MATRIX
SCIENCE

If you added Jack and Jill to the immunology group, then any search that they perform under their name would be visible to members of the immunology group. Mascot would have no way of knowing who the search was intended for.

New features in Mascot 2.1

- Export results in various formats
- Search engine and report enhancements
- Mascot Daemon enhancements
- Mascot Security
- Request for version 2.2 always welcome
- Coffee break - restart at 10:45.

MATRIX
SCIENCE

We have covered quite a lot in the last hour -