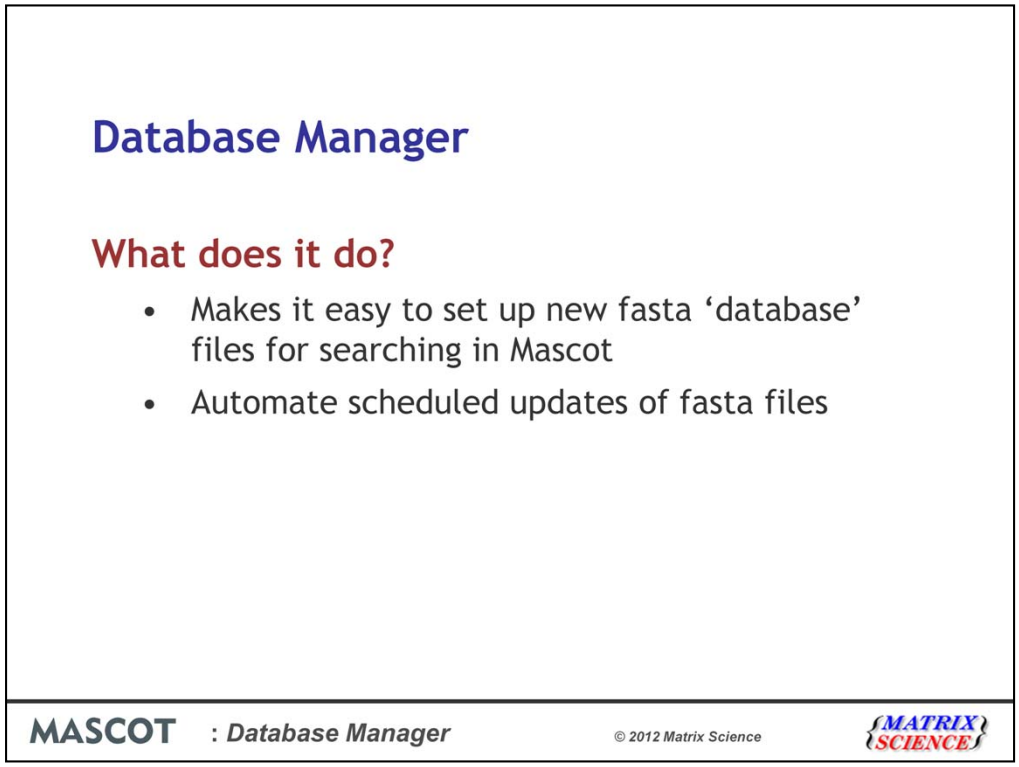
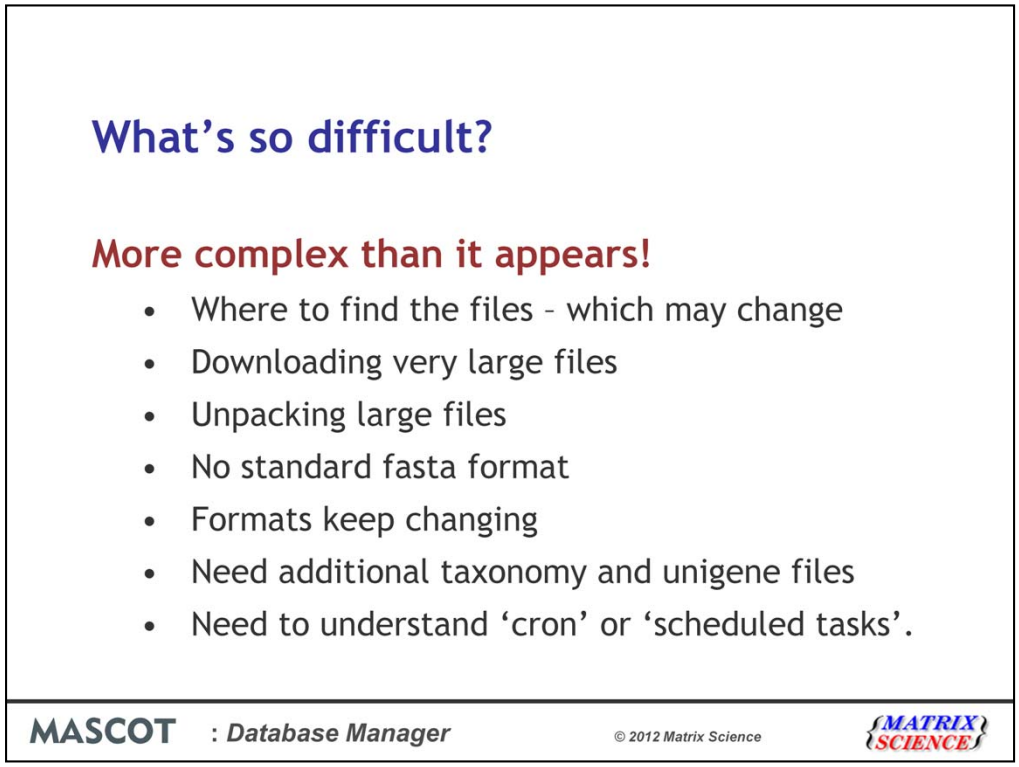
Database Manager
in
Mascot 2.4

MASCOT

MATRIX SCIENCE



Database manager does something that sounds trivial. It makes it easy to setup a new database for searching with Mascot. It also makes it easy to automate regular updates of these databases.



However, it's not quite as simple as that in practice for a number of reasons.

In most cases the databases are frequently updated and you want to get the latest version, but download locations and file names can change over time.

There are also issues with downloading very large files, so we need to check that the whole file has been downloaded. Running out of disk space or a lost connection are common problems. Also, it's quite possible for the file to be updated on the server while it is being downloaded which can result in a corrupt file.

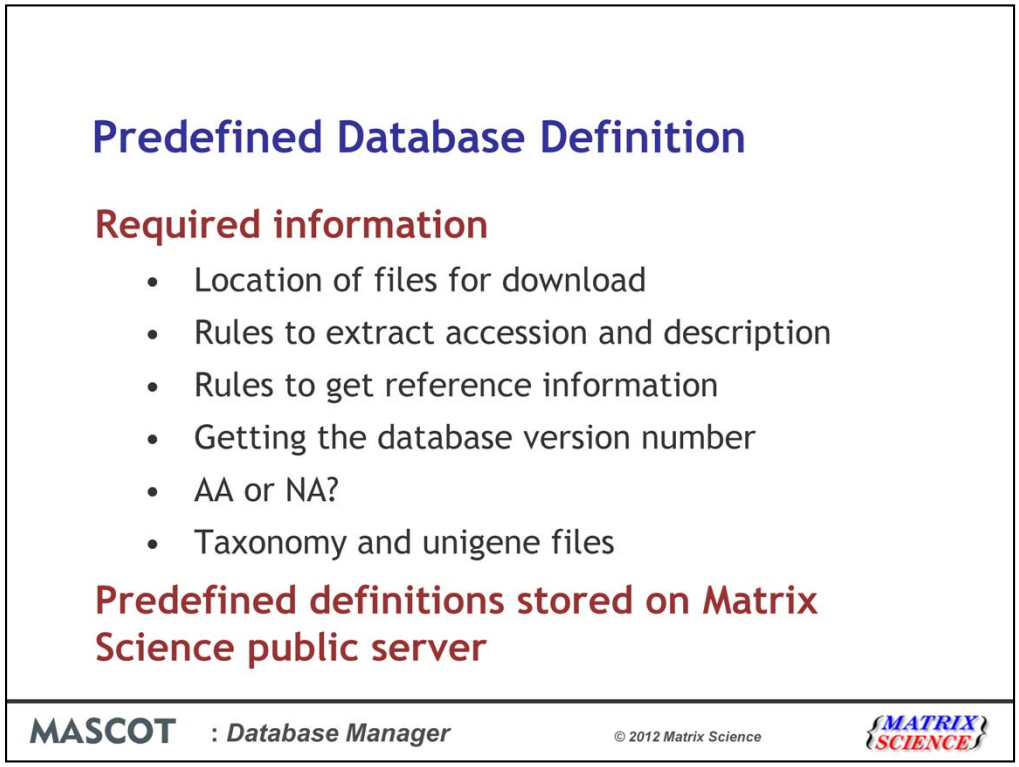
Some tools, for example the built in unzip utility in Windows 7 will fail to unpack files larger than 4Gb

For each sequence, we need a unique identifier or accession but there is no standard fasta format.

Once we've got it right, it can change…

We also need a collection of taxonomy files so that we can filter searches based on taxonomy. Locations, names and formats of these files can change

It's also not trivial to automate regular updating of the fasta files. We suspect that many of our users just have very old versions of the databases.



We use the term Predefined Database Definition quite often in the software and I'd just like to explain what we mean by this.

Database manager needs this information to be able to configure a database:

•The location and the names of the files for download. If the files are compressed, it also needs to know how they are compressed.

•The fasta format isn't really a standard, and we need to be able to extract a unique identifier, or accession for each sequence. The predefined database definition includes rules to do this.

• It's also useful to be able to show reference information for the sequences. For SwissProt and Trembl, this can come from a local reference file or from an internet resource. Again, we need some rules for this.

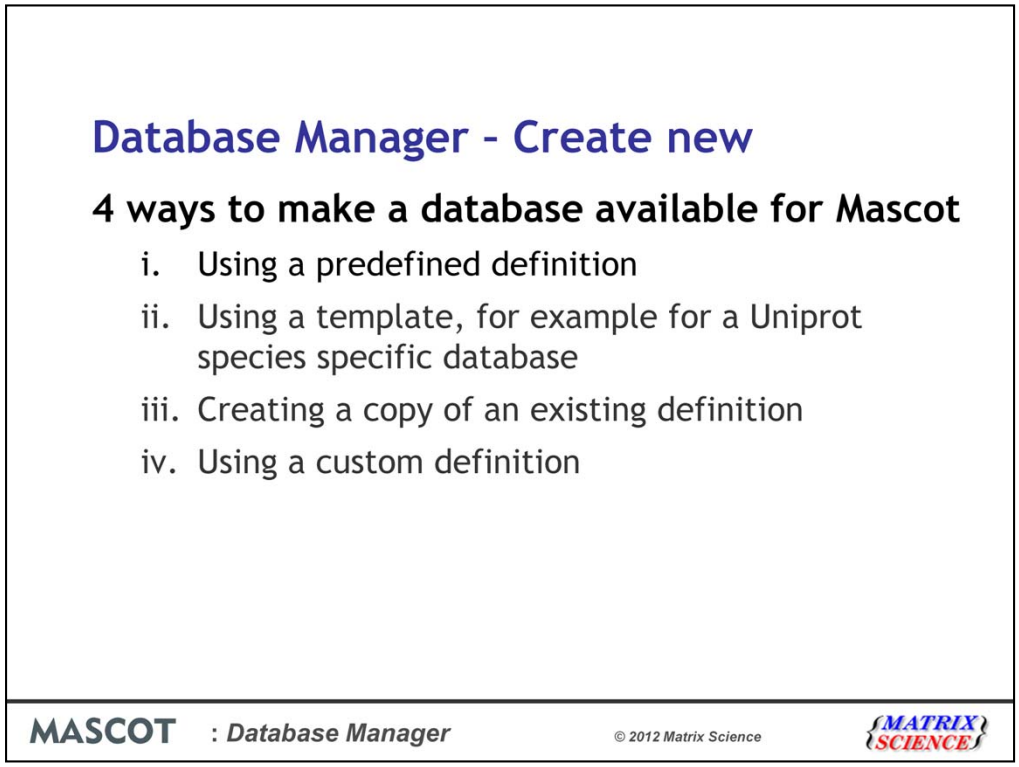
• Most sequence database providers always give the same filename for the fasta file to make it easy to download with a fixed url. However, some provide a separate file with the version number in. Again, we need rules for this

• We need to know whether the fasta file is a protein database or a nucleic acid database. If you get that wrong, your searches clearly won't work.

• If the sequence database comes from multiple species, it's useful to be able to filter searches based on the taxonomy of your sample. For this to work, we need to know what taxonomy files to download.

• And finally, for the EST databases, it’s really useful to be able to group the fragments using unigene indexes. Patrick will describe this in a later talk.

However, it’s no good us just putting this information for each of the common databases on the CD we send out because the database providers often change some of these things and that prevents Mascot from working with the latest download. To cope with this, we now have a file on our public web server that contains all the rules. In future, when the NCBI or EBI change the format of a database, we can just change the rules and everyone’s Mascot server will get updated automatically. Or, if someone produces a new database that is commonly used, we’ll just add that to our predefined database definitions file.



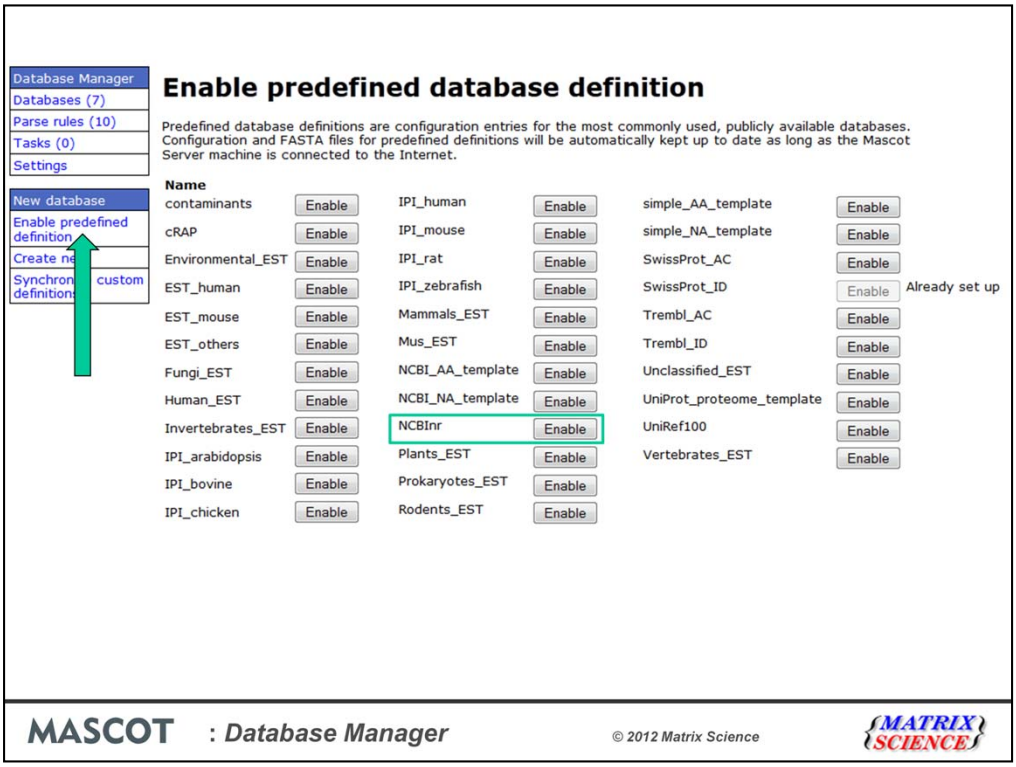
When you start database manager for the first time it will import your existing configuration. After that you'll probably want to make new databases available and automate scheduled downloads for them.

There are four ways to make a new database available for Mascot.

I'll start by describing how to get NCBInr online using a predefined definition.

Next, I'll describe how to use a template for a uniprot species specific database.

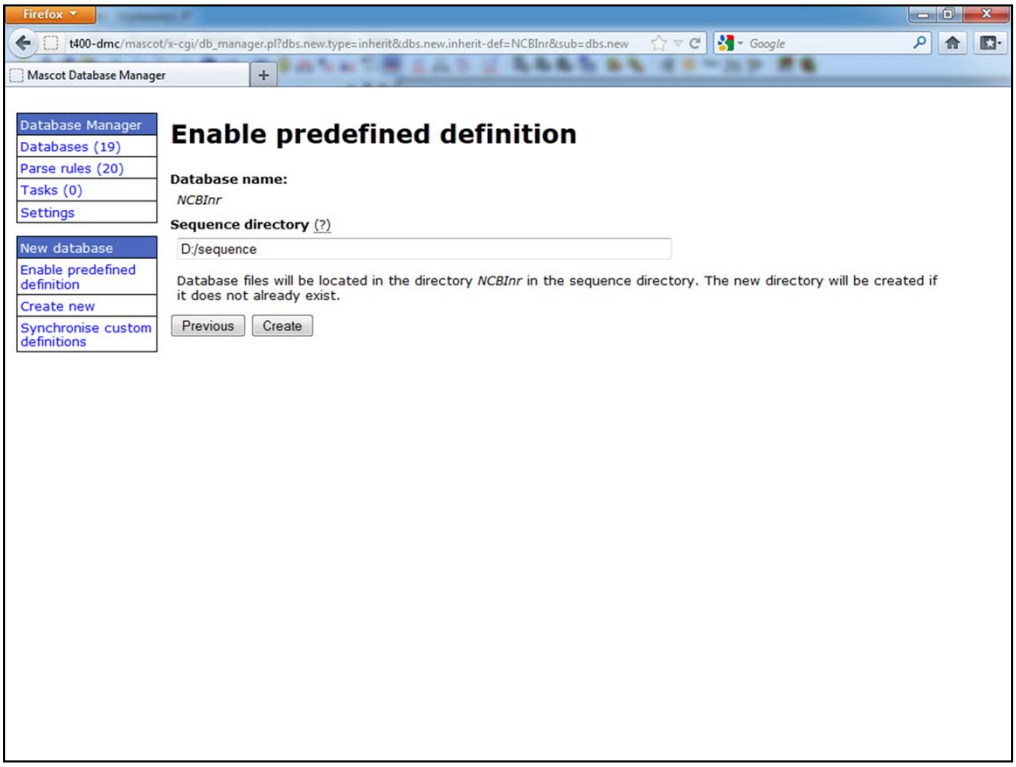
I'll just touch briefly on the last two options because they are fairly obvious.



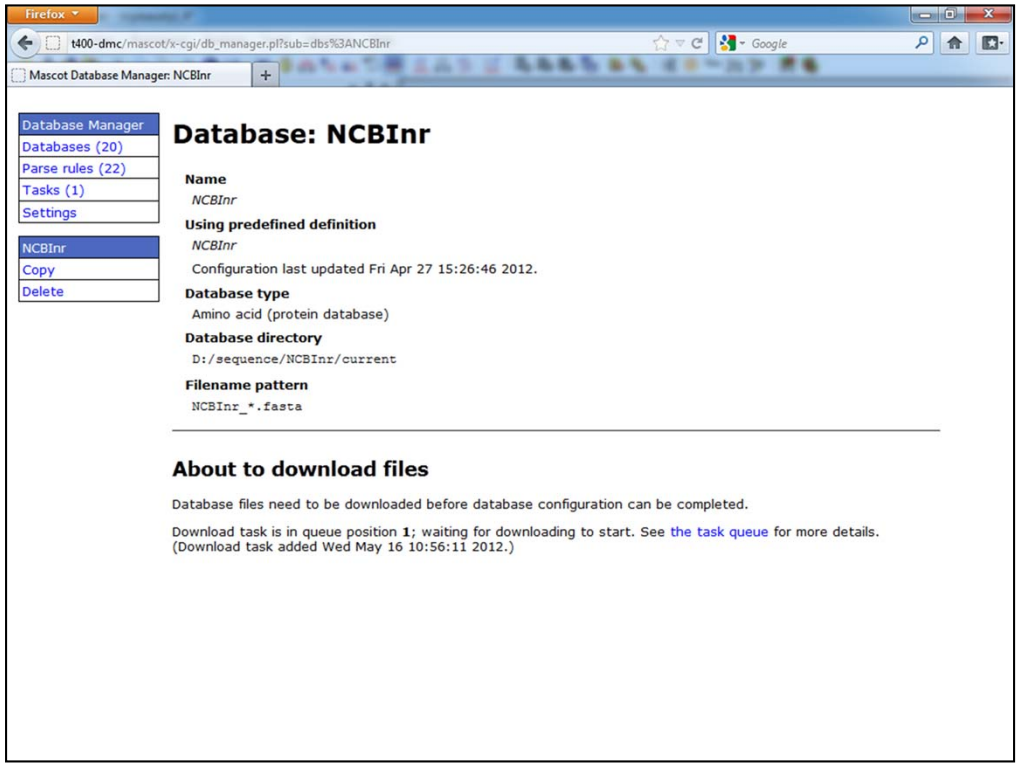
So, lets start with making NCBInr available. We do this by Enabling a predefined definition and this is ridiculously easy!

Simply click on the link on the left, and choose one of the available definitions. There are over 20 definitions available.

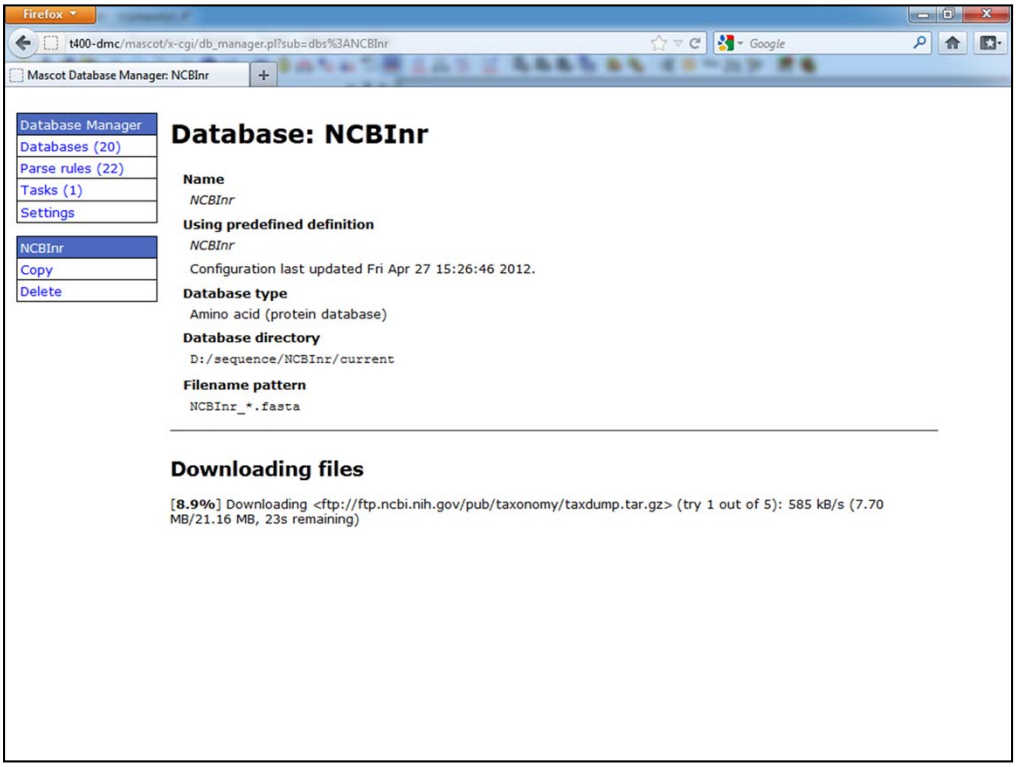
Just click on the enable button next to NCBInr.



Database manager then asks you to confirm the sequence directory under which the NCBInr directory will be created. You can change the default directory in the settings page, or just change it here. Mascot is installed on the C: drive on my computer, but I've decided to put all my databases on D: because there's plenty of spare space there. Click on create, and you could just go off and get a cup of coffee, or you could sit and watch the screen update. I don't recommend watching an update for EST_others, which is a 10Gb download and over 20Gb uncompressed. I'll just show you what happens next.

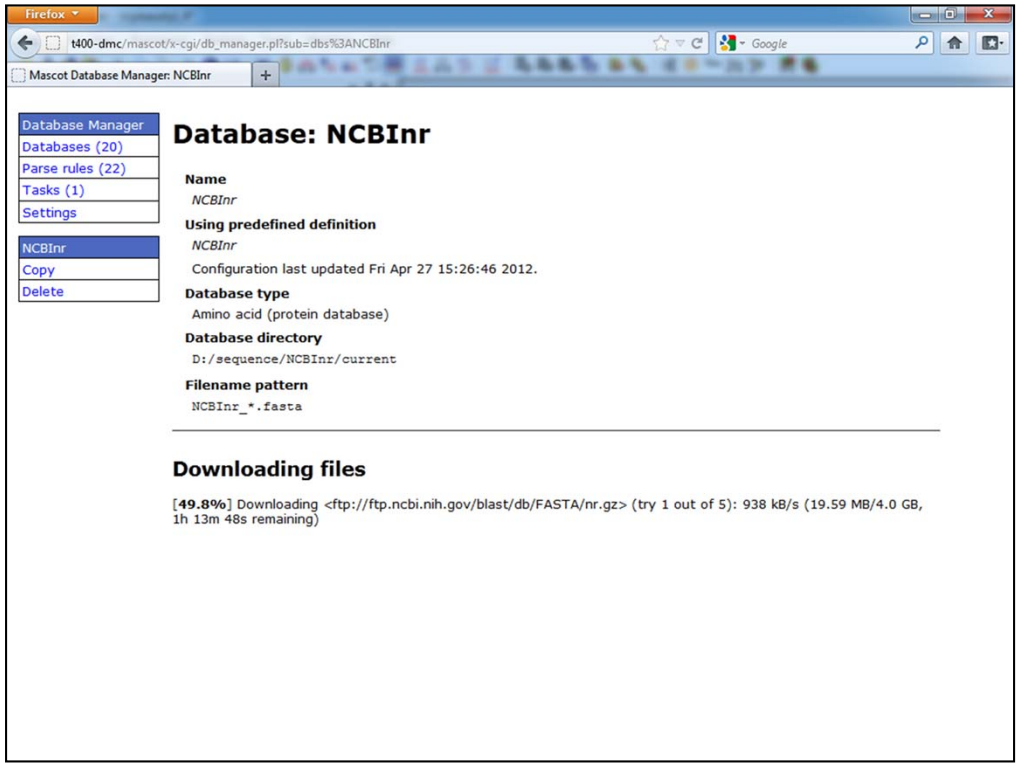


Database manager lets you know that it's going to start downloading files. You could click on the link to see the task queue if nothing happens in a minute or so,

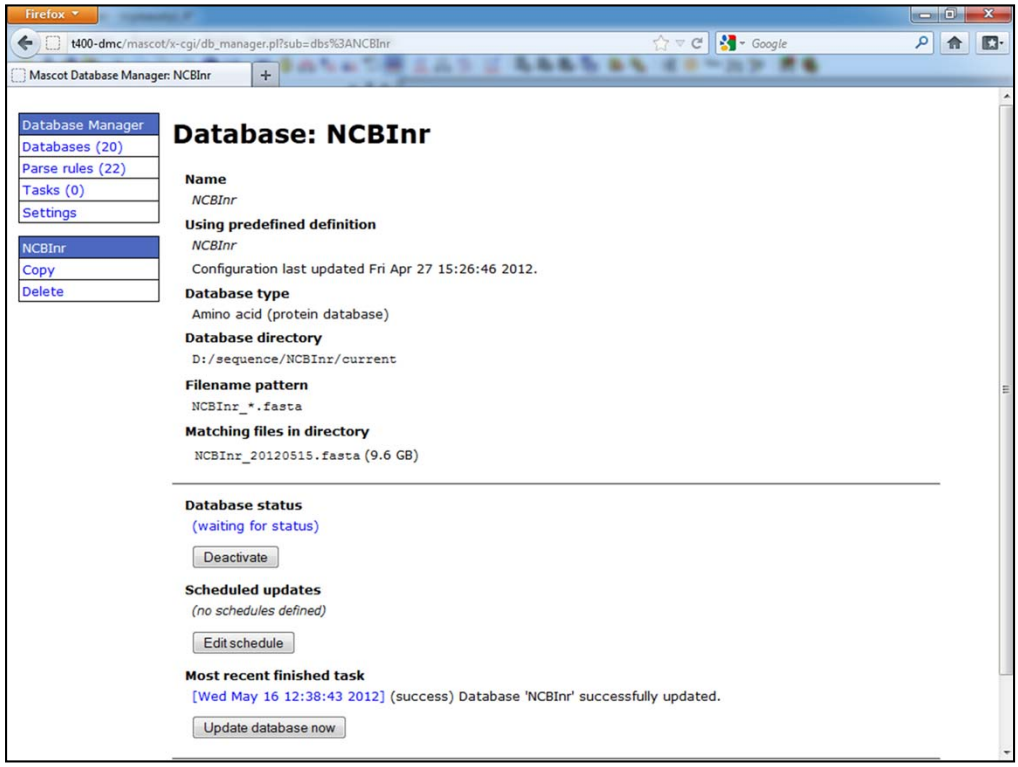


But most likely you'll see it starts to download pretty quickly.

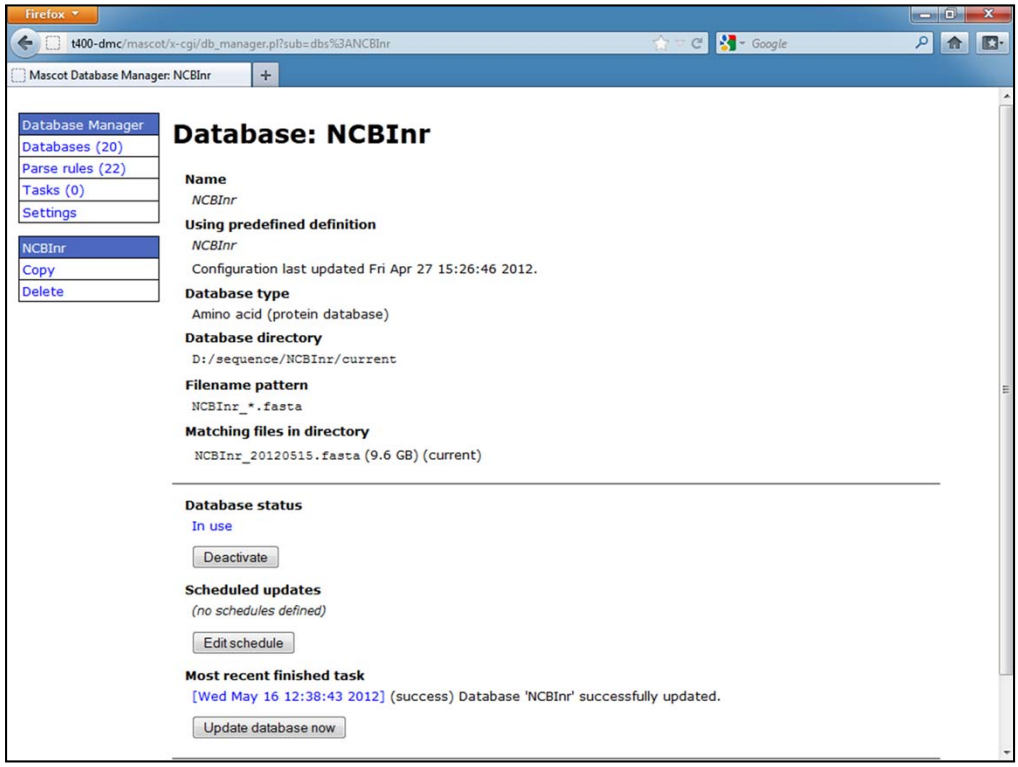
You can see here that it's started by downloading the taxonomy files that it needs



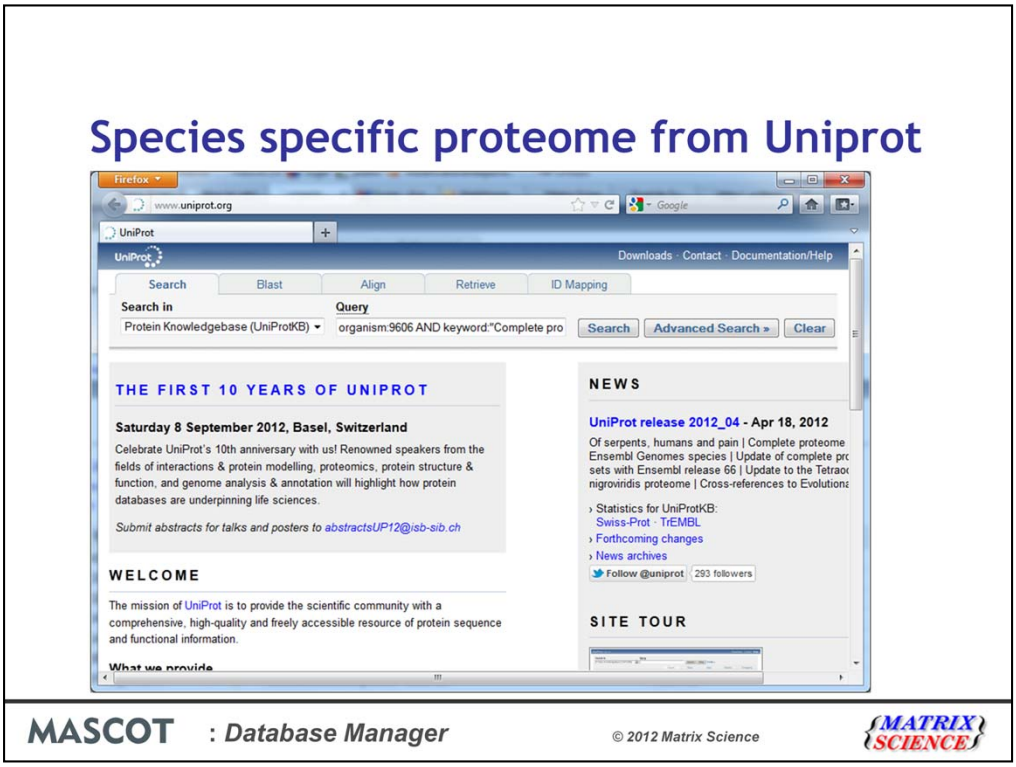
And it swiftly moves on to downloading the compressed fasta file which will probably take about an hour depending on the speed of your internet connection.



Once it's finished downloading and decompressing there will be a pause before it's available for searching, but you don't need to do anything.



And finally, it will come in use. If you want to automatically check for updates, say every week, you can do that here, but I'll come back to that later. There's also a manual "update database now" button if you want to just do this manually at any time.

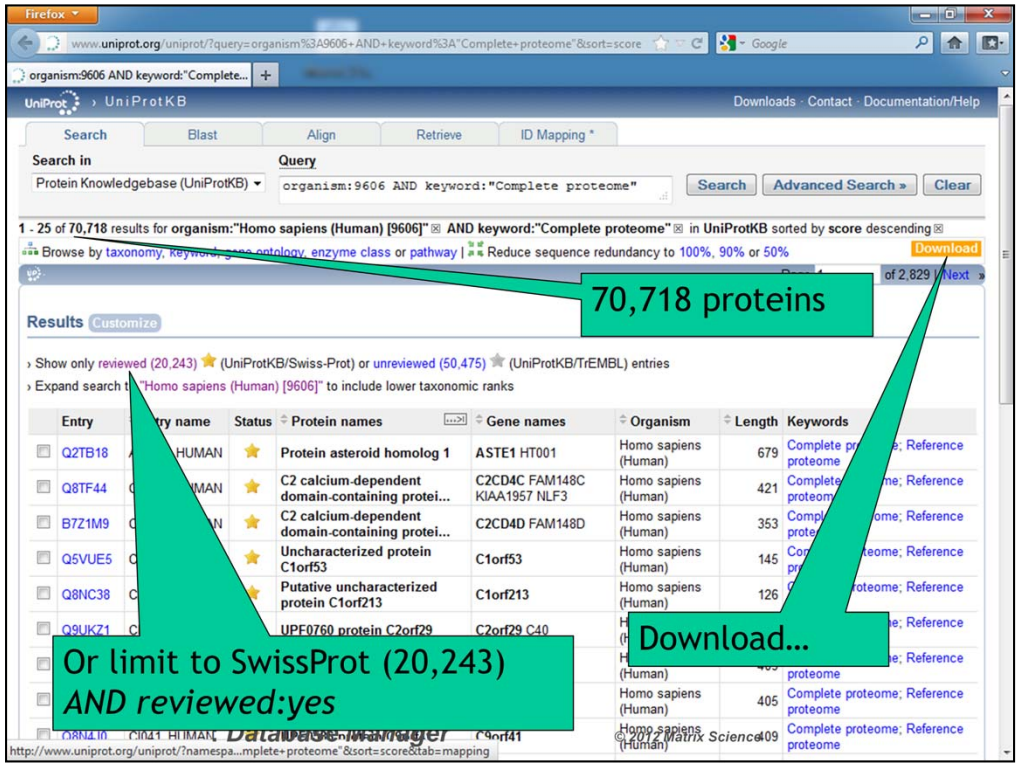


The next example is using a template, and in this case I'll generate a species specific proteome.

Last September saw the final update to the IPI databases which have become very popular in our community, and this is the recommended replacement. There are no downloadable fasta files, but rather a query interface and the option to produce fasta files for any species on the fly.

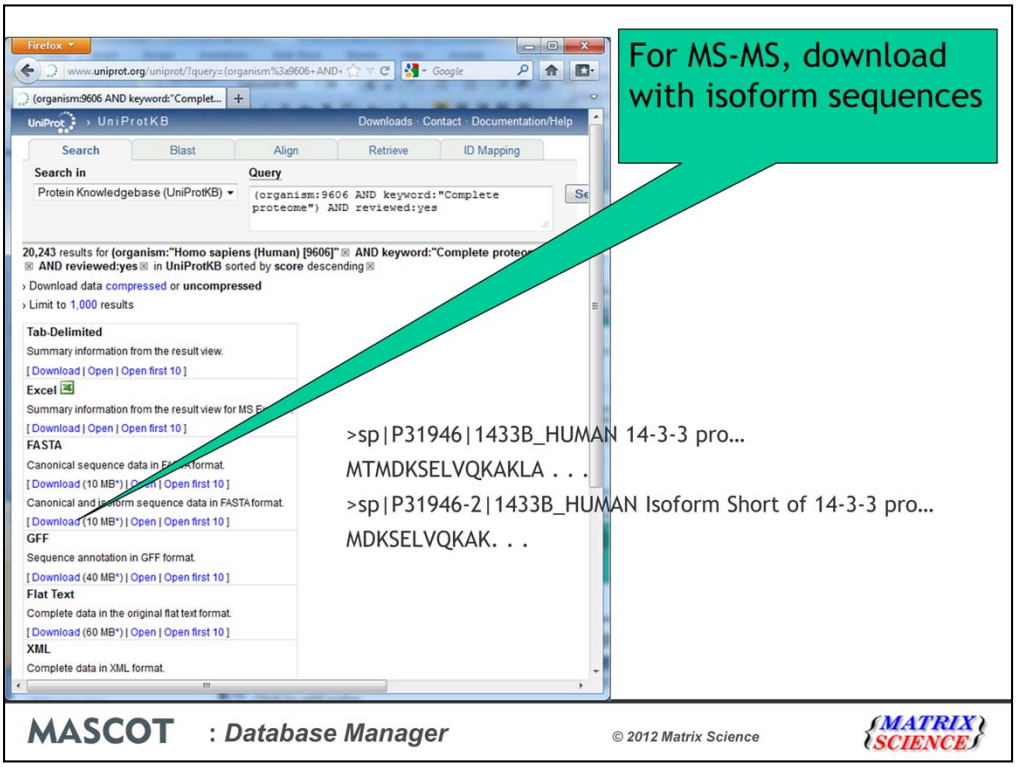
Firstly, I'll show you the uniprot 'interface' for producing the fasta files.

Go to www.uniprot.org, and in the search box enter the organism or taxonomy you are interested in and also the keyword "Complete proteome". I've typed organism:9606 and "Complete Proteome" to find all the human proteome proteins.



There are 70,718 proteins in SwissProt and Trembl, but I can limit it to just 20,243 proteins in SwissProt if I want, and this adds the text “reviewed:yes” to the URL

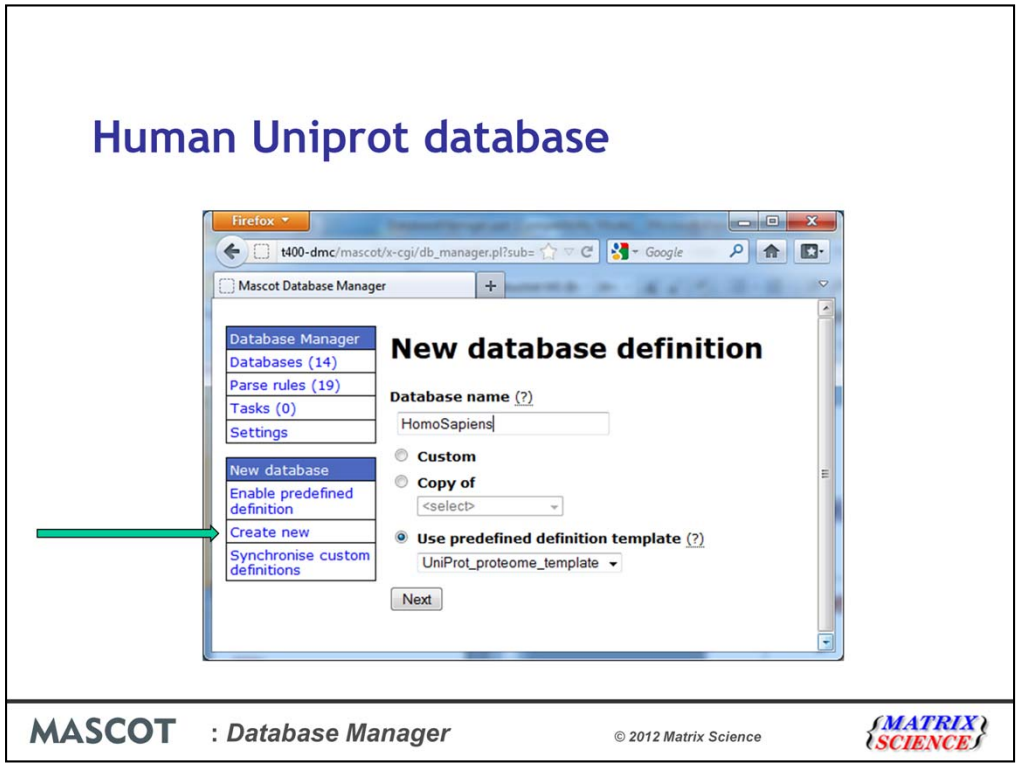
Clicking on the download link takes me to some choices:



For searching ms-ms data with Mascot, we need to choose the fasta format and ideally should include with isoforms.

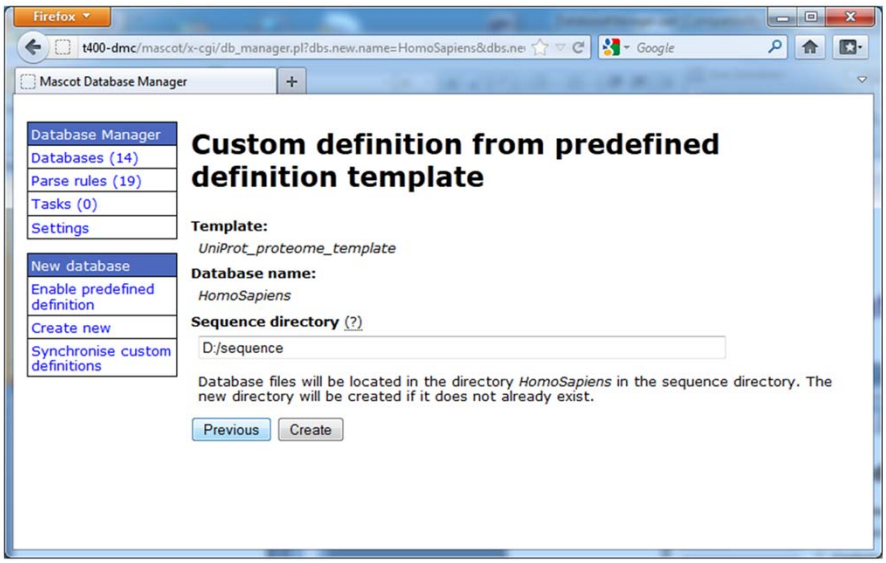
Isoforms means that multiple copies of some sequences are provided with the different isoforms. A -2, -3 -4 etc is appended to the accession. Of course, this means that the accession rather than the ID needs to be used for the Mascot accessions.

However, there's actually a better way of downloading if we want to keep the fasta file updated automatically.



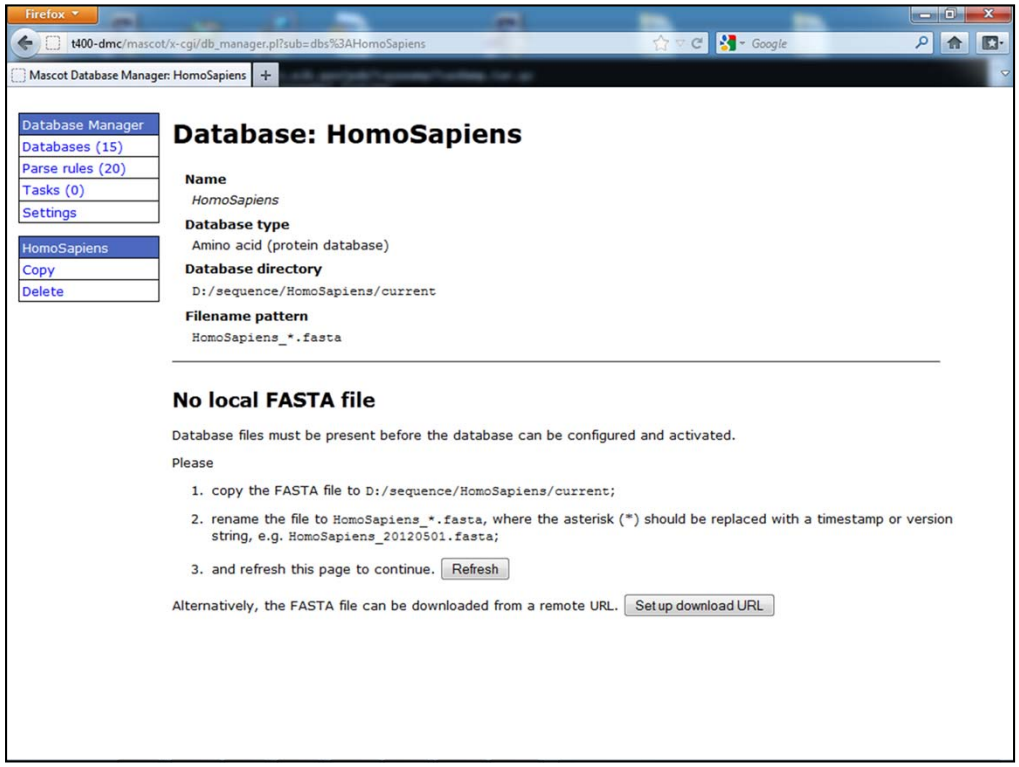
Back in Database Manager, click on Create new, and then enter the name and say that we want to use a template.

Next.

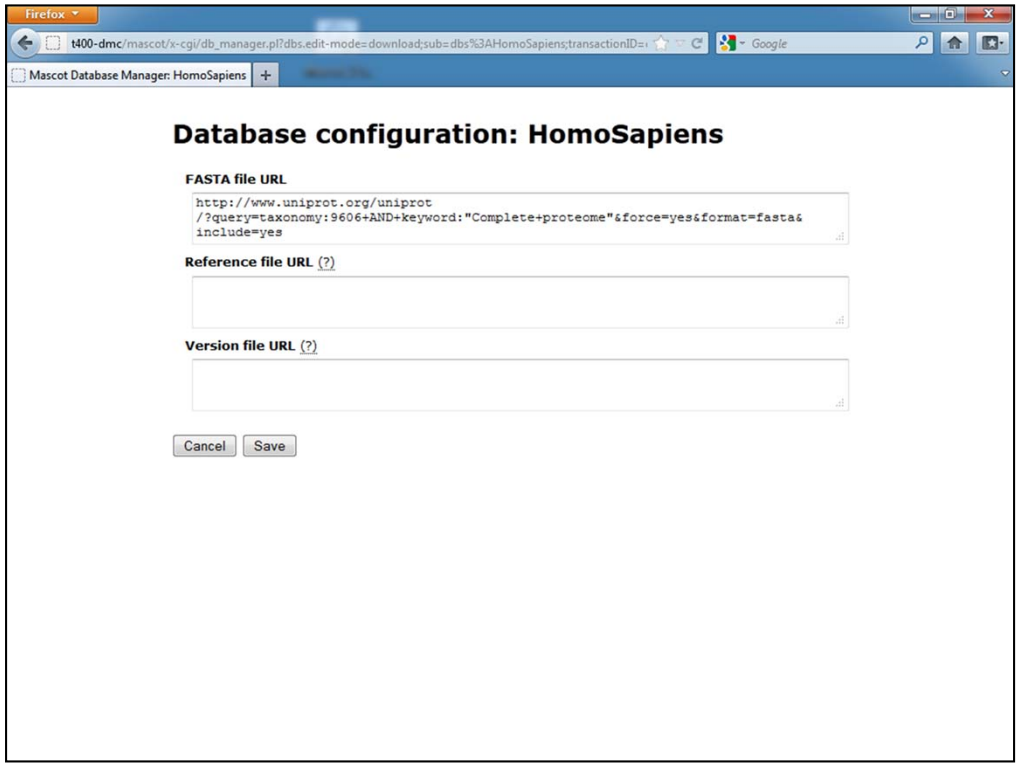


MASCOT : Database Manager MATRIX SCIENCE

The next screen asks us to confirm the directory and we then click on Create



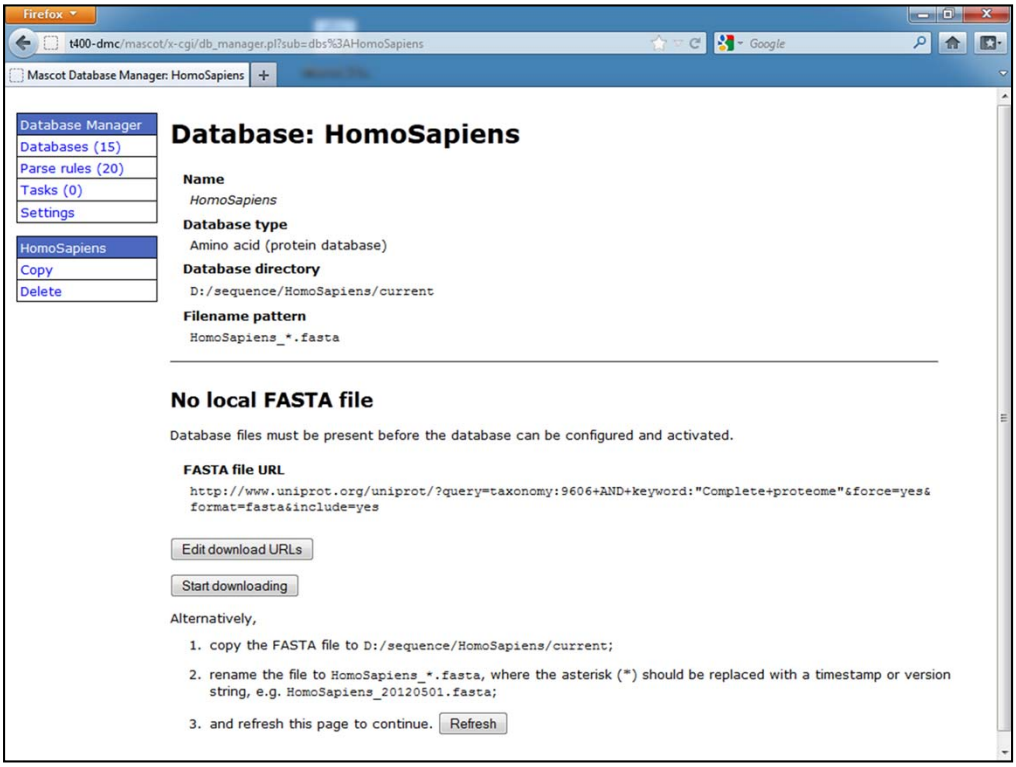
We now have a choice of downloading the file manually from uniprot.org as and copying it into the specified folder, or we can database manager to download it for us which is what I will do.



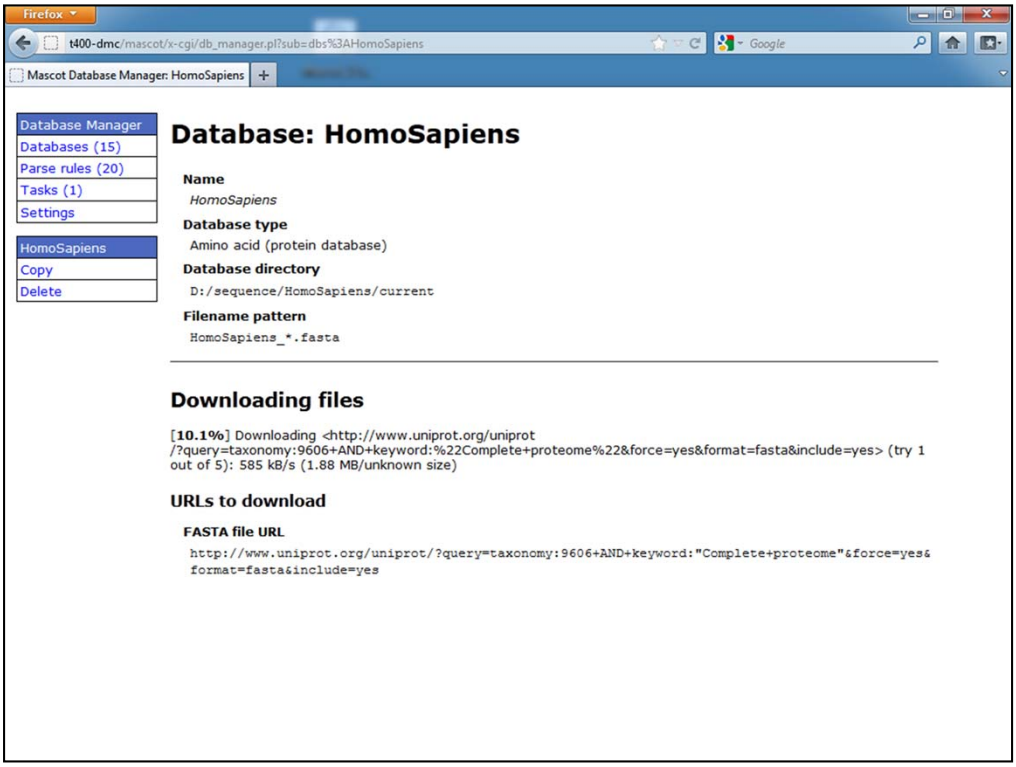
I don’t expect you to remember this URL, but it’s in the help for Database Manager on our public web site or on the uniprot web site.

The include=yes is for the isoforms.

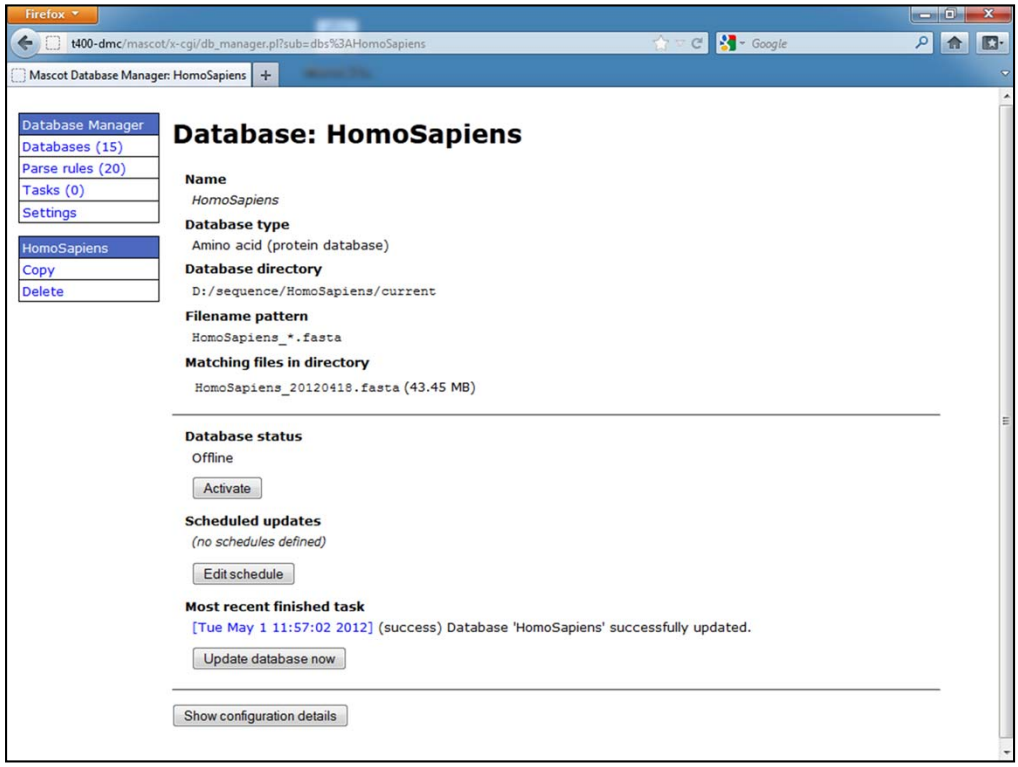
If we use a URL here, we can come back to database manager later and just click a single button to force an update of the local database. Alternatively, we can set scheduled updates automatically. If we had chosen to download manually now, we would always need to download manually, so it’s worth the extra effort here.



Click on “Start downloading”.

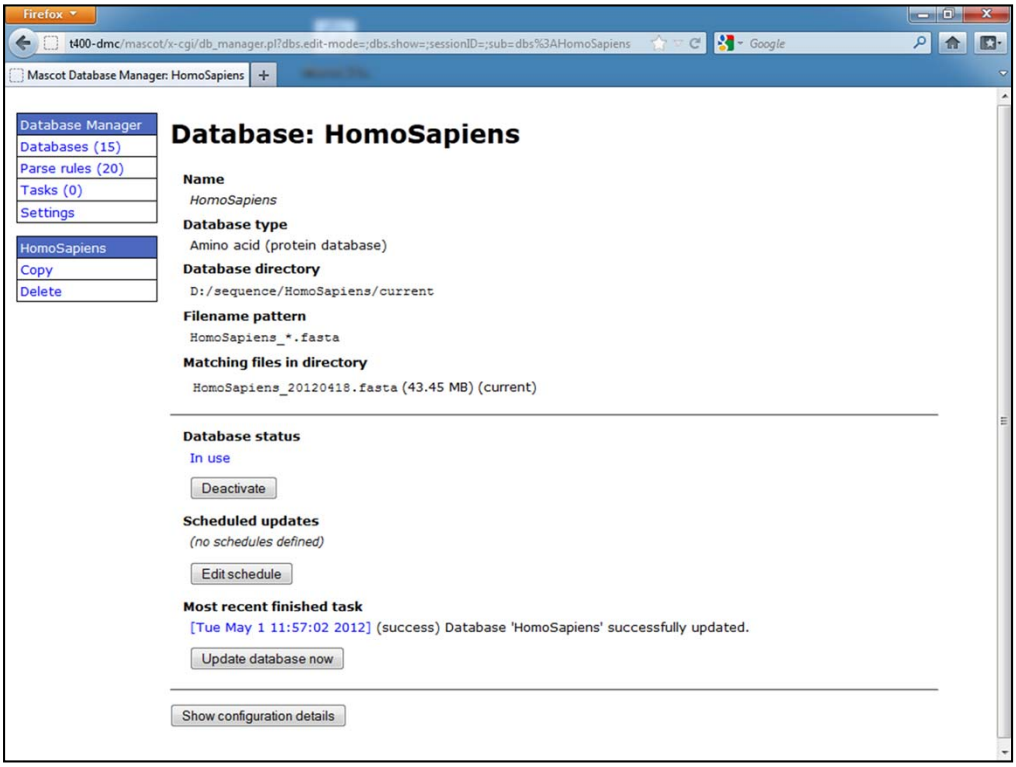


And you'll see some progress

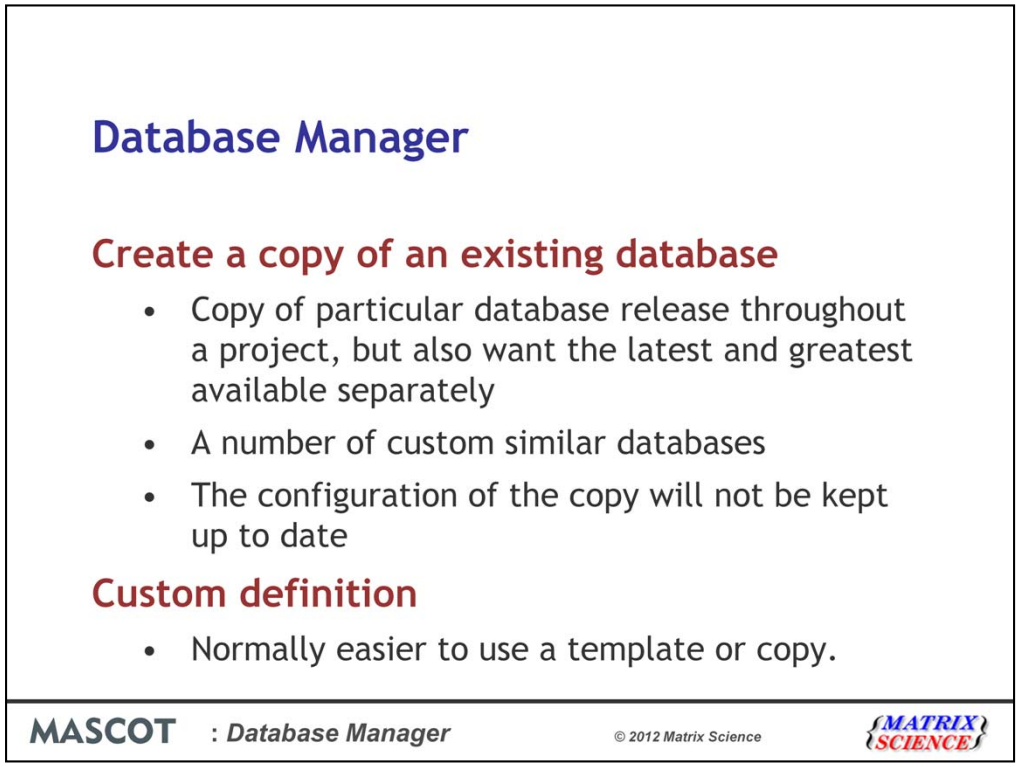


And finally you will need to click on Activate.

We don’t need to worry about parse rules because it’s all coped with in the predefined definition. If you remember, from a few screens ago, we needed to use the accession rather than the ID, and this is the default for this predefined definition.



And the database is now in use.



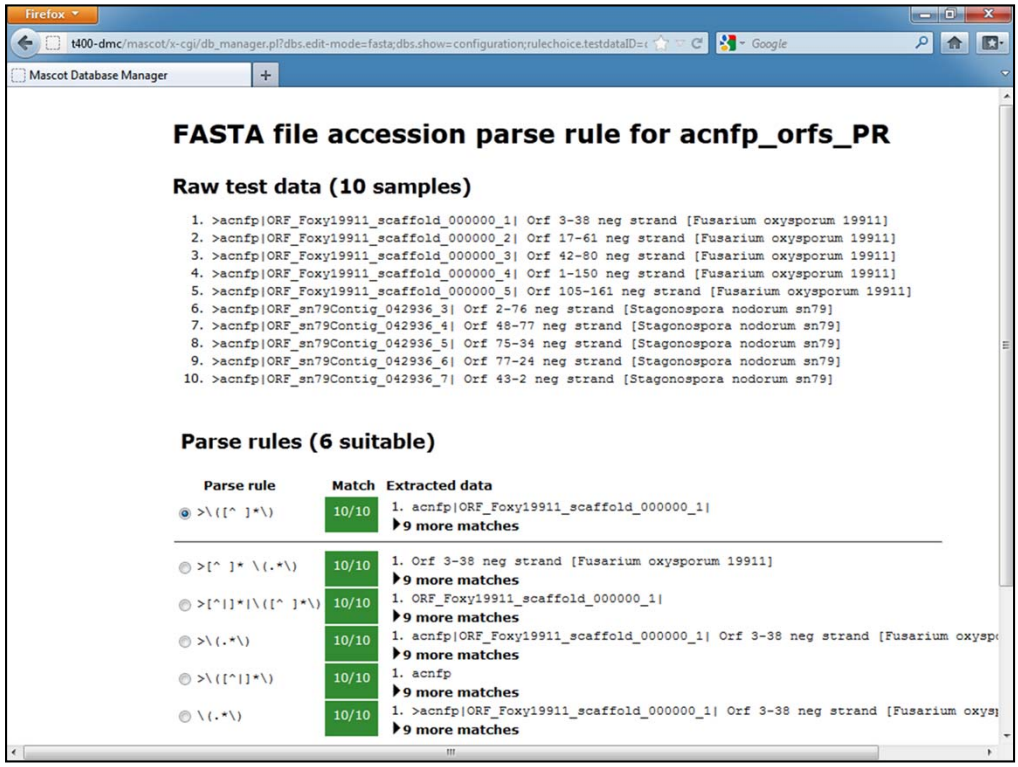
There are two other ways of installing a new database.

The first is to create a copy of an existing database. For example, you may wish to use a specific release of SwissProt throughout the duration of a lengthy project, but also want to have an up to date SwissProt available for searching. This is easy to do, and database manager will even make a copy of the files for you.

You may also have a number of databases with the same format. If you've set up one correctly, just create a copy of the configuration. With copies of any configuration, any changes to the parent database configuration won't trickle through to the copy.

Secondly, you can create a custom definition. To be honest, it's normally slightly easier to use a template such as the simple_AA_template or to make a copy of a similar working database.

In every case, you can still edit the rules (we call them parse rules) to extract the unique accession and also the sequence description. For most cases, this will follow the EBI or NCBI standards, but here's a different case:



Database manager searches all the rules that it knows about and finds which give any match at all and rejects the others. In this case, it found 6 possible rules. Database manager shows the first and last 5 sequences in the file. In this case, we can see that all sequences start with >acnfp, so it may make sense to exclude that.

The first suggestion includes everything up to the first space which is OK, but not ideal.

The second suggestion is the free form text. We'd be better off using that as a description.

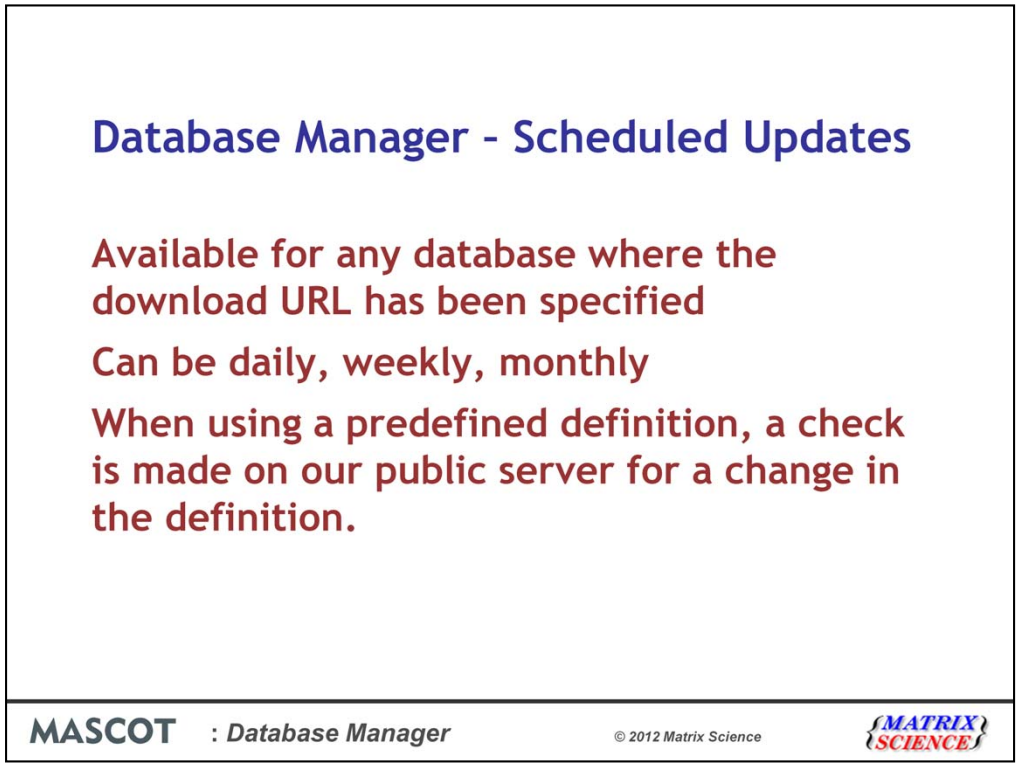
The third suggestion excludes the acnfp and the pipe symbol, so is pretty much what we want

The fourth suggestion looks to be the whole description line which is too long

The fifth suggestion is not useful because all the accessions will be identical

The sixth suggestion is the whole line including the >

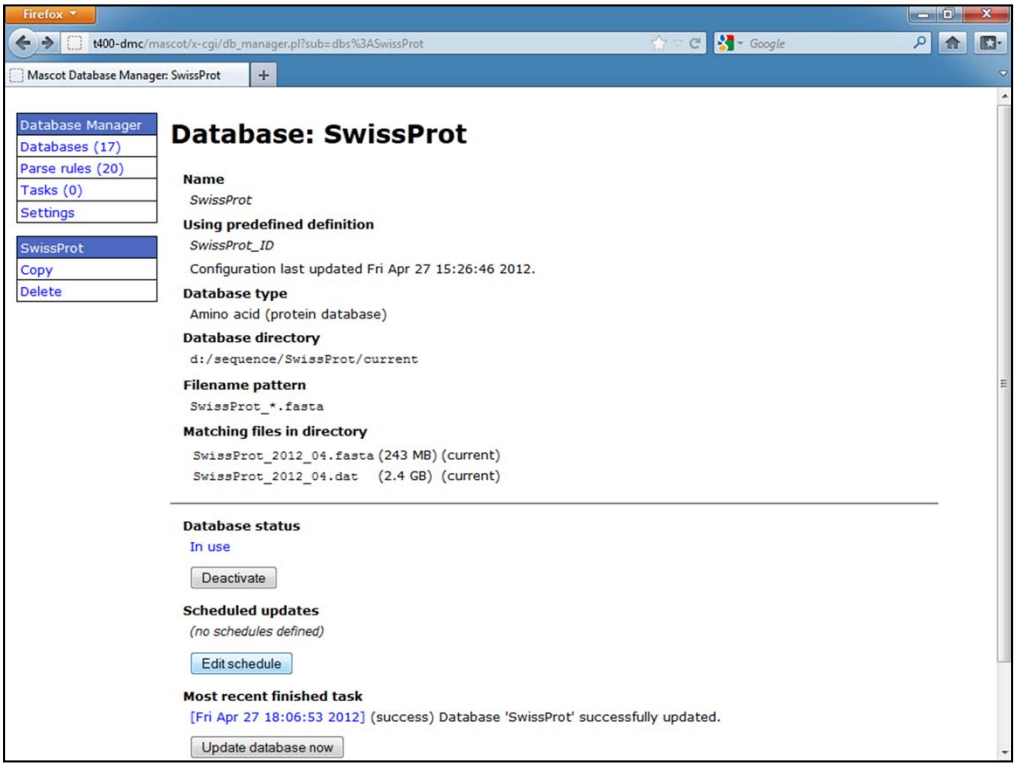
So, could choose the third suggestion. It's not quite perfect because it includes the trailing pipe symbol. Alternatively, we could define a new parse rule here which is much easier than in the older database maintenance utility.



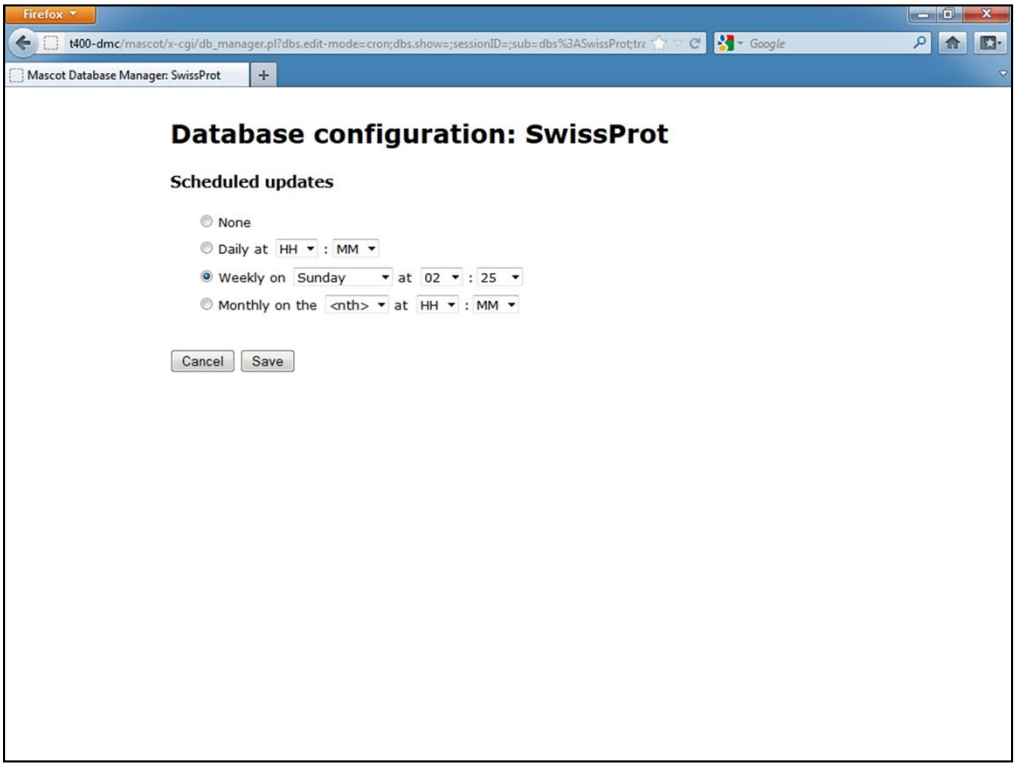
Most people, most of the time want to keep their sequence databases up to date. In previous versions of Mascot, we provided a utility called db_update.pl to do this. You needed to modify the script, test it at a command prompt and then run set up scheduled tasks or a cron job to automate it. Unfortunately, because this wasn't trivial, lots of people didn't bother. Also, for people who have set it up, they would sometimes find it stopped working because the database supplier moved the files, or changed the format.

It's much easier in Mascot 2.4:

- As long as you can specify a download URL, it can be automated.
- There's some simple choices for daily, weekly or monthly which should be sufficient for most cases.
- If the database supplier changes something, we will change our public definitions, and the updates should continue to work. I'll give an example of how to do it for SwissProt.

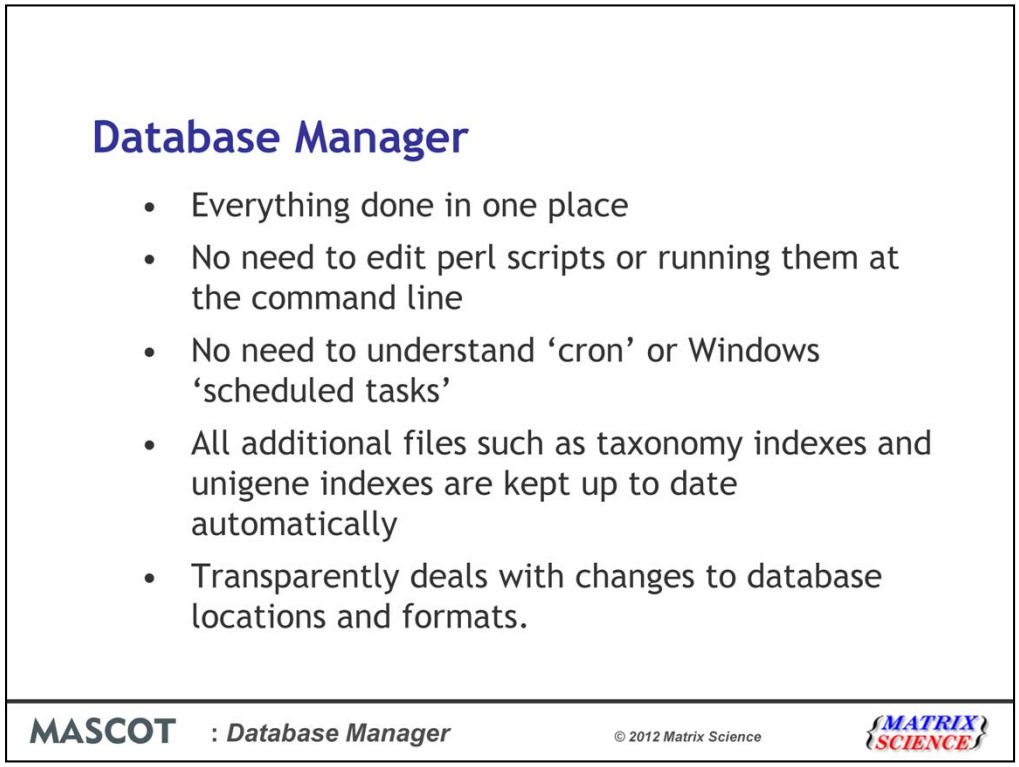


From the databases screen, click on SwissProt and then click on the Edit Schedule button



Here, I've decided to update it weekly on Sunday at 2:25 am.

In practice, SwissProt only gets updated monthly. However, database manager is pretty smart and will only download files if they have changed, so it doesn't matter that it's checking more often than it needs to.



I hope I've shown that it really is easy to use database manager and that this is a great improvement on the different utilities that we provided in previous versions. As you've seen, there's just one utility that does it all.

Also, there's no requirement to edit perl scripts, like the db_update.pl script, and there's no need to run anything at a command line any more.

There's also no need to understand cron, or Windows scheduled tasks.

All additional files such as taxonomy indexes and unigene indexes are kept up to date automatically, so you don't need to worry about downloading them and making sure that they are unpacked and put into the correct directory.

And finally, when the NCBI or EBI change a database format or location, this should all be dealt with transparently.