Xref sources and parsing

  - Refseq_peptide
  - Refseq_dna
  - IPI (International Protein Index)
  - UniGene
  - EMBL
  - PDB
  - protein_id
  - PUBMED + Medline
  - GO
  - EntrezGene
  - InterPro
- SPECIES SPECIFIC ENTRIES
  - Human
    - MIM - Online Mendelian Inheritance in Man
    - HGNC
    - CCDS
  - Mouse
    - MGI
  - Rat
    - RGD
  - Zebra fish
    - ZFIN_ID
  - C Elegans
    - wormpep_id, wormbase_locus, wormbase_gene, wormbase_transcript

UniProt/Swissprot - UniProt/Trembl (UNIversal PROTein resource)

The files can come in two types:

1) Contains data for all species

```
ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/uniprot_sprot.dat.gz
```

or

```
ftp://ftp.ebi.ac.uk/pub/databases/uniprot/knowledgebase/uniprot_trembl.dat.gz
```

This is the normal case.

2) Contains data for one species only

```
ftp://ftp.ebi.ac.uk/pub/databases/integr8/uniprot/proteomes/17.D_melanogaster.dat.gz
```

These are primary Xrefs in that they contain sequence and hence can be
mapped to the Ensembl entities via normal alignment methods (we use
Exonerate).

This is a list of dependent Xrefs that might be added:

```
EMBL
    PDB
    protein_id
```

Note: For human, mouse and rat we also take the direct mappings from uniprot for the SWISSPROT entries.
Those not mapped by uniprot are then processed in the normal way.

Refseq_peptide

The files come in two types those for specific species i.e.

```
ftp://ftp.ncbi.nih.gov/genomes/Canis_familiaris/protein/protein.gbk.gz
```

or as a series of numbered none specific species files i.e.

```
ftp://ftp.ncbi.nih.gov/refseq/release/vertebrate_other/vertebrate_other3.protein.gpff.gz
```

These files are parsed by the parser RefSeqGPFFParser.pm

These are primary Xrefs in that they contain sequence and hence can be mapped to the Ensembl entities via normal alignment methods (we use Exonerate).

Below is a list of dependent Xrefs that might be added:

EntrezGene

Refseq_dna

The files come in two types those for specific species i.e.

```
ftp://ftp.ncbi.nih.gov/genomes/Gallus_gallus/RNA/rna.gbk.gz
```

or as a series of numbered none specific species files i.e.

```
ftp://ftp.ncbi.nih.gov/refseq/release/vertebrate_mammalian/vertebrate_mammalian46.rna.fna.gz
```

These files are parsed by the parser RefSeqParser.pm

These are primary Xrefs in that they contain sequence and hence can be mapped to the Ensembl entities via normal alignment methods (we use Exonerate).

IPI (International Protein Index)

Comes as species specific file i.e.

```
ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.fasta.gz
```

The files have something like

```
>IPI:IPI00000005.1|SWISS-PROT:P01111|TREMBL:Q5U091|ENSEMBL:ENSP00000261444;
                       ENSP00000358548|REFSEQ:NP_002515|VEGA:OTTHUMP00000013879 Tax_Id=9606 GTPase
NRas precursor
sequence..................
```

But most of the header information is ignored except for the description and the IPI value. The sequence is used to position the IPI Xref.

These are primary Xrefs in that they contain sequence and hence can be mapped to the Ensembl entities via normal alignment methods (we use Exonerate).

Has no dependent Xrefs.

UniGene

Comes as species specific file i.e.

```
ftp://ftp.ncbi.nih.gov/repository/UniGene/Bos_taurus/Bt.seq.uniq.gz
    ftp://ftp.ncbi.nih.gov/repository/UniGene/Bos_taurus/Bt.data.gz
```

These are primary Xrefs in that they contain sequence and hence can be mapped to the Ensembl entities via normal alignment methods (we use Exonerate). No longer loaded via UniProt.

Has no dependent Xrefs.

EMBL

These are dependent Xrefs and are linked to Ensembl via the UniProt entries.

PDB

Protein Data Bank entries are dependent Xrefs and are linked to Ensembl via the UniProt entries.

protein_id

These are dependent Xrefs and are linked to Ensembl via the UniProt entries.

PUBMED + Medline

These are no longer stored due to the large numbers of these. If you want to add these then see the UniProtParser and RefseqPArser for more details.

GO

Can come in a species specific file or can contain all species.

```
ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz
    ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association.goa_human.gz
```

GO information in the UniProt and RefSeq files are ignored and just the information from the above files are used. The files have references to UniProt and RefSeq entries and so the GO entries are set to be dependent Xref on these.

EntrezGene

Gene-centred information at NCBI is stored as a dependent Xref and is obtained from the RefSeq entries.

InterPro

InterPro is a database of protein families, domains and functional sites and gets it data from the file

```
ftp://ftp.ebi.ac.uk/pub/databases/interpro/interpro.xml.gz
```

NOTE: InterPro has its own table and hence the Xrefs are stored but are not linked to the Ensembl entities directly but a list of InterPro and identifiers are stored. The identifiers stored are of the type

```
PROSITE, PFAM, PRINTS, PREFILE, PROFILE, TIGRFAMs
```

SPECIES SPECIFIC ENTRIES

Human

MIM - Online Mendelian Inheritance in Man

Descriptions and types are obtained from the file

```
ftp://grcf.jhmi.edu/OMIM/omim.txt.Z
```

This creates two set of Xrefs:

1) MIM_GENE (disease genes and other expressed gene)
2) MIM_MORBID (the disease genes)

Note those in set 2 will also be in set 1.

These MIM Xrefs are linked to UniProt/SwissProt entries using the UniProtParser.pm creating dependent Xrefs. Note if the Swissprot entry does not specify whether the MIM entry is a phenotype or a gene then it is ignored. For this same reason MIM dependent Xrefs are NOT obtained from the RefSeq entries.

So when the Swissprot entries are matched to Ensembl the MIM entries will also be matched.

HGNC

The Human Genome Organisation Xrefs are obtained from various sources:-

1) HGNC (ensembl_mapped)
HGNC has direct mapping to ensembl which have been manually curated.
So information is obtained from the script http://www.genenames.org/cgi-bin/hgnc_downloads.cgi

2) CCDS
The HGNC's are connected to the same ensembl object that the CCDS are linked to. We connec to the ccds database to get this information.

3) Vega
This is made from the Havana manually curated database.

4) HGNC
HGNC has links to other databases like uniprot,refseq etc and these can be used to link to ensembl

Which of these is chosen at the mapping stage is based on the prioritys of the sources. Here they are listed in order above.
This is known as a priority xref as the mapping with the best priority is chosen.

CCDS

The CCDS database identifies a core set of human protein coding regions that are consistently annotated by multiple public resources and pass quality tests.

A local file is used here:

```
file:CCDS/CCDS.txt
```

The file contains a list of CCDS identifiers and the Ensembl entities they match to. So direct Xrefs are created for these.

Mouse

MGI

Previously known as 'MarkerSymbol'.

```
ftp://ftp.informatics.jax.org/pub/reports/MRK_SwissProt_TrEMBL.rpt
    ftp://ftp.informatics.jax.org/pub/reports/MRK_Synonym.sql.rpt
```

This is mouse specific Xref being the Mouse Genome Informatics data.
The files have references to UniProt entries and so the GO entries are
set to be dependent Xrefs on these.

Rat

RGD

Rat Genome Database entries are populated by using the file

```
ftp://rgd.mcw.edu/pub/data_release/GENES
```

The RGD Xrefs are dependent Xrefs on the Refseq entries.

Zebra fish

ZFIN_ID

The two files

```
http://zfin.org/data_transfer/Downloads/refseq.txt
    http://zfin.org/data_transfer/Downloads/swissprot.txt
```

contains list of ZFIN identifiers and RefSeq or Swissprot identifiers
depending on the file.

This creates a set of dependent Xrefs on RefSeq and UniProt entries.

C Elegans

wormpep_id, wormbase_locus, wormbase_gene, wormbase_transcript

Uses the file

```
ftp://ftp.sanger.ac.uk/pub/databases/wormpep/wormpep180/wormpep.table180
```

and the database (last release should do)

```
mysql:ensembldb.ensembl.org:3306:caenorhabditis_elegans_core_46_170b:anonymous
```

This creates direct Xrefs for all these.