INERA
INC

eXtyles®

User Documentation

NCBI Matching/Linking

Last updated January 2014

Inera Incorporated

19 Flett Road
Belmont, MA 02478
+1 617 932 1932
eXtyles-support@inera.com
www.inera.com

NCBI Matching/Linking

Matching single accession numbers

The eXtyles NCBI Matching/Linking module looks for accession numbers to entries in the various NCBI databases and some EBI databases. If it finds a match, it applies a character style to the accession number specific to the appropriate database, and it also creates a hyperlink to the record for that accession number in the database.

For example, when the module sees this text:

> More recently, closely related strains were also isolated … from the ciliate Collinia sp. endoparasitic in euphausiids from the Gulf of California (unpublished GenBank record EU090136), and in a culture-independent analysis of the microbial burden and diversity in commercial airline cabins [7].

it applies the character style db_ncbi_genbank and this link to the text "EU090136":

> http://www.ncbi.nlm.nih.gov/nuccore/156744481

Clicking on this link brings up the NCBI GenBank record for this sequence. The linked text is displayed:

> More recently, closely related strains were also isolated … from the ciliate Collinia sp. endoparasitic in euphausiids from the Gulf of California (unpublished GenBank record EU090136), and in a culture-independent analysis of the microbial burden and diversity in commercial airline cabins [7].

Matching ranges of accession numbers

If the module finds a range of accession numbers, it embeds a link that includes each accession number in the range (up to a user-configurable limit, set to 100 by default).

For example, when the module sees this text:

> The GenBank/EMBL/DDBJ accession numbers for the bovine RVC sequences determined in this study are AB738402–AB738417, as detailed in Fig. 1.

it applies the character style to the entire range and generates this link:

> http://www.ncbi.nlm.nih.gov/nuccore/430726479%20430726477%20430726475%20430726473%20430726471%20430726469%20430726467%20430726465%20430726463%20430726461%20430726459%20430726457%20430726455%20430726453%20430726451%20430726449

which brings up a list of all 16 accession numbers in the range.

> The GenBank/EMBL/DDBJ accession numbers for the bovine RVC sequences determined in this study are AB738402–AB738417, as detailed in Fig. 1.

If a range of accession numbers greater than the maximum limit is encountered (set to 100 by default), the NCBI Matching/Linking module will style the range as shown in the previous example, but the link will point only to the first and last accession numbers in the range, not to all of the intermediate values.

Databases Linked

NCBI hosts a number of sequence and structure databases for nucleotide and amino acid sequence data. It also hosts a number of literature databases, most obviously PubMed and PubMed Central, and databases that contain other original data, such as the Gene Expression Omnibus (GEO) database, and sources of secondary data such as the HIV-1, Human Protein Interaction database and MedGen. A full list can be found at http://www.ncbi.nlm.nih.gov/guide/all/#databases_ . The NCBI Matching/Linking module also links to a few EBI databases; more information about EBI can be found at http://www.ebi.ac.uk.

The NCBI Matching/Linking module uses 17 different character styles to indicate the database that has been matched to. These are shown in the following table:

Style	Database
EBI databases	
db_ebi_ArrayExpress_Array	A database of genome arrays used in ArrayExpress experiments
db_ebi_ArrayExpress_Experiment	A database of genomics experiments, results of which are included in ArrayExpress
db_ebi_ArrayExpress_GEO	Functional genomics data, either submitted directly to ArrayExpress or imported from the NCBI GEO database
NCBI databases	
db_ncbi_ccds	CCDS, the Consensus CDS Project
db_ncbi_dbGap	dbGaP, the Database of Genotypes and Phenotypes
db_ncbi_dbnsp	dbSNP, the Database of Short Genetic Variations or single nucleotide polymorphisms (SNPs)
db_ncbi_entrezgene	Entrez Gene
db_ncbi_genbank	GenBank – both accession numbers and GI numbers are recognized
db_ncbi_genpept	Translated protein sequences from GenBank (see Entrez Protein)
db_ncbi_geo	GEO, Gene Expression Omnibus database
db_ncbi_omim	OMIM, Online Mendelian Inheritance in Man
db_ncbi_pdb	Protein Data Bank; these records form part of Entrez Protein
db_ncbi_refseq	RefSeq, the NCBI Reference Sequence database

db_ncbi_SRA	SRA, the Sequence Read Archive, containing raw sequence data from next-generation sequencing platforms
db_ncbi_swissprot	Swiss-Prot protein sequence database; these records form part of Entrez Protein
db_ncbi_unigene	UniGene
db_ncbi_other	Records that match an NCBI database covered by the module but not listed above

Databases Not Linked

Not all of the NCBI databases are covered by NCBI Matching/Linking. Specifically, the module does not attempt to link to the following databases:

- PubMed (covered by the eXtyles PubMed Reference Linking module)
- PubMed Central
- GEO Profiles (a distinct database from the GEO database)
- MeSH
- NCBI Bookshelf
- NLM Catalog

How It Works

Each of the NCBI databases has certain constraints on the form that its accession numbers can take. For example, sequences in GenBank that were originally direct submissions to DDBJ (the DNA Data Bank of Japan) can be in the form of the letter “D” followed by five digits or the letters “AB” followed by six digits. Direct submissions to EMBL can take the form of five digits preceded by “V”, “X”, “Y”, or “Z” or six digits preceded by “AJ”, “AM”, “FM”, “FN”, “FO”, “HE”, “HF”, or “HG”. EST (expressed sequence tag) sequence accession numbers follow different rules. The eXtyles NCBI Matching/Linking module uses these rules to look for strings in the text that might match one of the databases.

If a string that matches the rules for one of the databases is found, the module queries that database to see whether it contains an accession that matches the found string and, if it does, it links that string in the text and applies the appropriate character style.

If at least one match is found during this “first pass”, the NCBI Matching/Linking module then runs a second, less strict pass, under the assumption that, if the document contains at least one accession number, it’s worth looking for other patterns that might be accession numbers that contain errors or that have not yet been released by the database.

During this second pass, the module looks for strings in the text that loosely match the database rules but don’t correspond to an entry in the database. The module then attaches a Word comment to those pieces of text to alert the editor to the possibility that they correspond to a database record.

The second pass only takes place if the module has already found at least one match to a database in the document; this is designed to reduce the likelihood of false-positives in content of a subject matter where accession numbers don't appear.

The module also uses some semantic clues to tell it that a potential match is either more likely or less likely to match one of the databases. For example, if text such as "accession number" or "GenBank" appears in the proximity of the match, it is more likely to be an accession number. In contrast, if terms such as "grant", "award", or "funding" appear, the module assumes that the possible match is actually a grant number of some kind, which not infrequently have a structure that matches one of the NCBI databases.

Excluded Document Parts

The module excludes the reference section of the document from the matching process by default.

It is also possible to set the module up to omit other paragraph styles on a customer-specific basis. For example, if accession numbers would never appear in the acknowledgments section of your content, this paragraph style could be excluded, and this might help to avoid false-positive matches against grant numbers. You could exclude the affiliations, if complex room numbers or postal codes throw up warnings.