Description of seqenv's main internal data structures

Lucas Sinclair, lucas.sinclair@me.com

Last modified:

Tue Mar 01 2016

All numbers are fake
This document just describes a plan for processing data.

seq_to_gis	*One to many*
C0 -> [343206452, 343204350, 199999999]	
C1 -> [234523451, 343204354]	
C2 -> []	

unique_gis	*Set*
343206452, 343204350, 199999999, 234523451	

source_db (previously gi_to_text)	*One to two*
343206452 -> "the deepest sea mud of the Mariana Trench"	pb421341233
343204354 -> "isolated from terrestorial samples near ..."	pb421341233
343204350 -> "the deepest sea mud of the Mariana Trench"	pb534533443
... etc etc	

gis_with_text	*Set*
343206452, 199999999	

unique_texts	*Set*
"the deepest sea mud of the Mariana Trench", ...	

text_to_matches	*One to many to many*
"the deepest sea mud of the Mariana Trench"	-> [ENVO:123123, ENVO:123777, ENVO:123777]
"isolated from terrestorial samples near ..."	-> [ENVO:23452345]
"Everglades wetlands"	-> []

text_to_counts	*One to many to one*
"the deepest sea mud of the Mariana Trench" -> (ENVO:00002007 -> 1, ENVO:00002041 -> 2)	
"isolated from terrestorial samples near ..." -> (ENVO:00001993 -> 1, ENVO:00003452 -> 1)	

seq_to_counts	*One to many to one*
C0 -> (ENVO:00002007 -> 0.667, ENVO:00002041 -> 0.333)	
C1 -> (ENVO:00002007 -> 0.123, ENVO:00002041 -> 0.821)	
C2 -> ()	

TODO:

serial_to_concept
lorem

child_to_parents
lorem

concept_to_name
lorem

df_abundances
lorem