BIEN 3 - Task #561

make VegBIEN ID fields plain-text instead of numeric

02/14/2013 04:47 AM - Aaron Marcuse-Kubitza

Status:	New	**Start date:**	02/14/2013
Priority:	Normal	**Due date:**	
Assignee:	Aaron Marcuse-Kubitza	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:			
Activity type:			

Description

Rationale

- this makes it possible to append data from multiple sources without having pkey collisions

Forming the ID value

- for datasource-specific data, use the datasource name + *intrinsic* identifiers such as a barcode or author plot code
    - *avoid* using the datasource's pkey if it's autogenerated, because this is random and not related to the data itself
    - use a standard path syntax:
        - LDAP dn
        - URN
        - XPath
- for hierarchical names, create a path from the names
    - include only as many parent nodes as are needed to guarantee global uniqueness
        - e.g. for taxonomic names, the family is globally unique, so higher taxa do not need to be included
- note that the ID field does not need to be the *only* way to uniquely identify a record, it just needs to be a way that is a *single text value*
- the value should be a human-readable text string rather than the cryptographic hash of such a string, to make it visually obvious what the ID refers to

Performance considerations

- to avoid affecting row lookup performance when using arbitrary-length strings as pkeys, a hash index should be added on the ID field
    - this is in addition to the usual btree index, which is still needed for merge joins and retrieving the data in sorted order
    - a hashed string is roughly equivalent to using just the hash value because the hash is usually cached
    - a hash value is roughly equivalent to an autogenerated integer, because it is a similar number of bytes long

History

#1 - 02/14/2013 05:45 AM - Aaron Marcuse-Kubitza

- Description updated

#2 - 02/14/2013 05:58 AM - Aaron Marcuse-Kubitza

- Description updated

#3 - 02/14/2013 06:21 AM - Aaron Marcuse-Kubitza

- Description updated

#4 - 02/14/2013 04:40 PM - Aaron Marcuse-Kubitza

- Description updated

#5 - 12/03/2013 05:54 AM - Aaron Marcuse-Kubitza

- Subject changed from make VegBIEN ID fields globally unique instead of autogenerated to make VegBIEN ID fields plain-text instead of numeric