Infrastructure - Task #3022

Story # 3023 (Closed): Review and where necessary alter, add, remove index fields

Update search index solr schema regarding indexed fields

2012-06-26 14:50 - Skye Roseboom

Status:	Closed	**Start date:**	2012-06-26
Priority:	Normal	**Due date:**	
Assignee:	Skye Roseboom	**% Done:**	100%
Category:	d1_indexer	**Estimated time:**	0.00 hour
Target version:	Sprint-2012.39-Block.5.4		
Milestone:	CCI-1.1	**Story Points:**	
Product Version:	*		

Description

Need to review which fields are 'indexed' and which are not in the search index's solr schema.

AuthoritativeMN was found to be not indexed which is likely wrong. Also need to review 'size' for use of index and a sortable long data type. These issues indicate the need for a wider review of the data types in the schema regarding sortable data type and indexed.

Indexed indicates solr will maintain a searchable index for the data field. It requires more disk space and index update time to maintain each indexed field.

History

#1 - 2012-06-26 14:55 - Skye Roseboom

Link to the search index schema:

https://repository.dataone.org/software/cicore/trunk/cn-buildout/dataone-cn-index/usr/share/dataone-cn-index/debian/index-solr-schema.xml

#2 - 2012-06-26 15:14 - Skye Roseboom

non-indexed fields:

size - updating type to slong and making indexed.
checksum
checksumAlgorithm
authoritativeMN
replicationAllowed
numberReplicas
preferredReplicationMN
blockedReplicationMN
replicaMN
replicaVerifiedDate

ogcUrl
dataUrl
webUrl
(all the URL are variations of the resolve endpoint to get the data or metadata. Used by mercury search.)

#3 - 2012-06-26 15:15 - Dave Vieglais

- Parent task set to #3023

#4 - 2012-06-26 16:38 - Skye Roseboom

- Status changed from New to In Progress

#5 - 2012-06-26 16:44 - Dave Vieglais

size - updating type to slong and making indexed. Also need to verify that slong renders as expected when returning search results.

checksum: It *may* be helpful to search on checksum as a proxy for identifier. This is a low priority however and can remain un-indexed for now.

checksumAlgorithm: It is useful for analysis to obtain a quick estimate of the algorithms in use. However this is an edge case that can easily be implemented, albeit less efficiently through a script that scans listObjects response. Remains unindexed.

authoritativeMN: Probably not useful to an end user, but could be useful internally for auditing processes. Remains unindexed.

replicationAllowed: These replication related entries may be helpful for auditing processes in the future, but can remain unindexed for now. Remains unindexed.

numberReplicas: Remains unindexed

preferredReplicationMN: Remains unindexed.

blockedReplicationMN: Remains unindexed.

replicaMN: Remains unindexed.

replicaVerifiedDate: Remains unindexed.

These fields are not relevant in DataONE and are really a hangover from the Mercury specific implementation. Unless there's good data to populate them, they can remain unindexed:
ogcUrl
dataUrl
webUrl

#6 - 2012-06-26 18:42 - Skye Roseboom

Great, thanks for feedback. Will just update 'size' field for now. Will test in cn-dev environment to ensure proper display of the 'slong' data type.

#7 - 2012-06-26 19:47 - Skye Roseboom

Updated solr schema has been placed in the 1.0.0 buildout for release in 1.0.2 patch of the index-processor. Index-tool contains the index-processor so it may also need to be updated (pom dependency to new index jar). Will handle updating pom.xml of index-processor and index-tool prior to creating new 1.0.2 patch release tags.

#8 - 2012-07-03 14:18 - Skye Roseboom

- Milestone changed from CCI-1.0.2 to CCI-1.2

Delaying this update. Trouble updating the schema in place without causing errors in search interfaces.

Going to accumulate schema changes into a single update and also investigate using a second solr core where a new index can be built and then swapped into the live index - to avoid disruptions in search.

#9 - 2012-10-03 02:47 - Skye Roseboom

- Milestone changed from CCI-1.2 to CCI-1.1

#10 - 2012-10-08 16:54 - Skye Roseboom

- Status changed from In Progress to Closed

- translation missing: en.field_remaining_hours set to 0.0

Changed 'size' field to be indexed and updated field data type to support ranged queries on this field.

Tested on cn-dev-unm-1.