Access to Memory (AtoM) - Feature #13386

Remove unnecessary data from Elasticsearch index

07/16/2020 10:38 AM - Dan Gillean

Status:	QA/Review	**Start date:**	06/21/2019
Priority:	Medium	**Due date:**	
Assignee:		**% Done:**	0%
Category:	Search / Browse	**Estimated time:**	0.00 hour
Target version:	Release 2.7.0		
Google Code Legacy ID:		**Tested version:**	2.7
Sponsored:	No	**Requires documentation:**	

Description

This issue ticket is to continue work started in #13096, and aims to remove unnecessary noise from the search index, for better results and quicker indexing. In the archival description Elasticsearch documents, the data from related objects should be limited to the following to limit the data stored to relevant information:

Information objects

Repository:
Only keep the authorized form of name of the repository (addressed in #13096)

Creators:
- Keep authorized form of name, parallel names, other forms of name, etc.
- Keep admin/bio history
- Remove everything else

Name access points:
Keep authorized form of name, parallel names, other forms of name, etc. Remove everything else. (addressed in #13096)

Subjects, places, genres:
Keep only name, remove other fields (descriptions, etc)

Physical storage/objects
Remove all fields

Part of
Remove all fields

Other entities

Further analysis needs to be done to determine if individual fields should be removed from other entity type indices in AtoM.

However, one known factor - when editing a repository or actor, a job will be triggered to reindex all related descriptions. However, this job is often triggered unnecessarily, as only the authorized form of name (and actor history, entity type, dates) are displayed on related records. Ideally, we would limit the execution of this job so that it only runs when related fields are affected, and not every time a related record is edited.

Related issues:

Related to Access to Memory (AtoM) - Task # 13273: Use Elasticsearch's "updat...	**New**	**03/13/2020**
Related to Access to Memory (AtoM) - Bug # 13581: ES index update jobs freque...	**New**	**11/08/2021**
Copied from Access to Memory (AtoM) - Feature # 13096: Remove unnecessary rep...	**Verified**	**06/21/2019**

History

#1 - 07/16/2020 10:38 AM - Dan Gillean

- Copied from Feature #13096: Remove unnecessary repository and actor data from information object Elasticsearch index added

#2 - 08/10/2021 01:11 PM - Dan Gillean

Further notes for implementation:

Index updates we should fully remove

- IO index update when a repository theme is edited

(possibly more to come, this is the only obvious case I have found so far)

Updates on related entities that should trigger partial updates on linked information objects:

We don't need to necessarily listen for and review what fields are being updated in a related entity, and only trigger updates when the related fields have new save data. Instead, we can assume that any time these related entities are edited, only the following fields require index updates to the related descriptions, as they are the only ones that capture searchable strings:

Actors

- Authorized form of name
- Parallel name(s)
- Standardized name(s)
- Other name form(s)
- History

Entity type does not seem to be indexed as a string - only IDs are used for filters and facets. Consequently, a change in an actor entity type should not require a reindex of related descriptions

Repositories

- Authorized form of name
- Identifier

Information objects

Descriptions can also be related to each other.

- slug (we have the Rename module that can potentially edit the slug of a linked description)
- title
- ID
- identifier (for inherited reference codes)

All other entities that can be linked, have either already been addressed for descriptions (e.g. terms - see #11906), or else are only linked via ID, and searching for the actual strings (e.g. the authorized form of name of a linked function, etc) returns no results. Consequently, a reindex should not be needed for an update to the related term.

Note that **additions** and **deletions** require a different approach - a small update to remove the related id (and any other fields, e.g. the authorized form of name of an access point like a subject) from the index for the related entities would still be needed.

Updates on related entities that should trigger partial updates on linked actors:

Terms

- Subjects
- Places
- Occupations

For these taxonomies and terms, updates should trigger a partial update in the related actors, not a full reindex. Essentially any form of name.

See the related issue, #11906. This was addressed for descriptions, but at the time of the solution, it might not have been possible to add these terms to actors and browse them in related term pages. We should review the current implementation with this in mind.

Looking at the index, it appears we're fully adding all term fields for Subjects and Places currently:

- https://github.com/artefactual/atom/blob/qa/2.x/plugins/arElasticSearchPlugin/config/mapping.yml#L370-L371

I think this could be changed to just be a partial foreign type that indexes the ID and authorized form of name of the related term. Note the direct_subjects and direct_places on lines 390-391 as well - not sure what, if anything, needs to change there.

Updates to either the authorized form of name of these terms should trigger a search index update in any actors to which they are linked. Note that Occupation notes are not stored as part of the term, so index updates are handled separately, via the actor update where the occupation note can be updated.

All other terms (e.g. Actor entity type) seem to be linked only by ID. As such, updates to the term metadata itself should not need search index

updates for related actors.

Other entities

Similarly other entity relations (repositories, descriptions, functions, accessions, etc) either cannot be related, or are related only by ID. As such, updates to the related entity metadata itself should not need search index updates for related actors.

Actor to actor relations also don't seem to need much, if anything. Right now we are indexing the entire relation as a foreign type:

- https://github.com/artefactual/atom/blob/qa/2.x/plugins/arElasticSearchPlugin/config/mapping.yml#L372

However, the authorized form of name, or any other forms of name, are not actually indexed as strings on related authority records.

Note that **additions** and **deletions** require a different approach - a small update to remove the related id (and any other fields, e.g. the authorized form of name of an access point like a subject) from the index for the related entities would still be needed.

#3 - 08/10/2021 01:38 PM - Dan Gillean

Accessions

Right now we are indexing the entire actor record on an accession when a creator is linked. However, the only information shown is the authorized form of name. See:

- https://github.com/artefactual/atom/blob/qa/2.x/plugins/arElasticSearchPlugin/config/mapping.yml#L413

This should be changed to a partial foreign type that only includes:

- Authorized form of name
- Parallel name(s)
- Standardized name(s)
- Other name form(s)

I don't think we need to index actor histories on accessions, since this is never shown and could lead to a lot of unnecessary noise in the results.

#4 - 08/10/2021 01:46 PM - Dan Gillean

- Related to Task #13273: Use Elasticsearch's "update by query API" to update related resources added

#5 - 11/05/2021 01:41 PM - Steve Breker

- File AtoM_index_dump_pre_changes.json added

Updated creators and inherited creators linked to information objects.

See attached file for example of old ES creator linkages: AtoM_index_dump_pre_changes.json

New creator ES index tree:


```
    "creators": [
     {
      "id": "487",
      "i18n": {
       "languages": ["en"],
       "en": {
        "authorizedFormOfName": "Steve",
        "history": "SB History"
       }
      },
      "otherNames": [
       {
        "sourceCulture": "en",
        "i18n": { "languages": ["en"], "en": { "name": "Grover" } }
       }
      ],
      "parallelNames": [
       {
        "sourceCulture": "en",
        "i18n": { "languages": ["en"], "en": { "name": "Steeeeeve" } }
       }
      ],
      "standardizedNames": [
       {
        "sourceCulture": "en",
        "i18n": {
         "languages": ["en"],
         "en": { "name": "Steven P Stevenson" }
        }
       }
      ]
     }
    ],
```

#6 - 11/05/2021 02:59 PM - Steve Breker

Subject, place, genre terms updated to only index the i18n name (and id) with each IO.

Example post-modification:


```
    "places": [
      {
        "id": "485",
        "i18n": {
          "languages": ["en"],
          "en": { "name": "Victoria" }
        }
      }
    ],
    "directSubjects": ["497"],
    "subjects": [
      {
        "id": "497",
        "i18n": { "languages": ["en"], "en": { "name": "Beer" } }
      },
      {
        "id": "483",
        "i18n": { "languages": ["en"], "en": { "name": "Wine" } }
      }
    ],
    "directGenres": ["391"],
    "genres": [
      {
        "id": "391",
        "i18n": {
          "languages": ["en", "es", "fr", "pt"],
          "en": { "name": "Books" },
          "es": { "name": "Libros" },
          "fr": { "name": "Livres" },
          "pt": { "name": "Livros" }
        }
      }
    ],
```

#7 - 11/05/2021 05:57 PM - Steve Breker

- File deleted (AtoM_index_dump_pre_changes.json)

#8 - 11/05/2021 05:58 PM - Steve Breker

- File AtoM_index_dump_pre_changes.json added

#9 - 11/05/2021 06:25 PM - Steve Breker

Updated creators linked to accessions to only index:

- authorizedFormOfName
- otherNames
- parallelNames
- standardizedNames

#10 - 11/08/2021 03:28 PM - Dan Gillean

- Subject changed from Remove unnecessary data from Elasticsearch index and reduce unnecessary re-index operations to Remove unnecessary data from Elasticsearch index

#11 - 11/08/2021 03:38 PM - Dan Gillean

- Related to Bug #13581: ES index update jobs frequently run when not needed. added

#12 - 11/08/2021 03:40 PM - Dan Gillean

Quick update:

Originally this ticket covered both unnecessary fields in the ES index **and** unnecessary search index jobs being run. These are in fact two separate problems, and correspondingly, this ticket has been updated only to cover the former, while a new ticket - #13581 - has been created to cover the latter.

#13 - 11/08/2021 05:23 PM - Steve Breker

PR containing work to remove unnecessary data being indexed in Elasticsearch:

https://github.com/artefactual/atom/pull/1472

#14 - 11/08/2021 05:23 PM - Steve Breker

- Status changed from New to QA/Review

#15 - 11/08/2021 05:23 PM - Steve Breker

- Target version set to Release 2.7.0

- Tested version 2.7 added

Files

AtoM_index_dump_pre_changes.json	28.6 KB	11/05/2021	Steve Breker