Access to Memory (AtoM) - Feature #5173

Improve default ElasticSearch alphabetic sort to better reflect natural sort expectations.

06/03/2013 01:01 PM - Dan Gillean

Status:	Verified	**Start date:**	06/03/2013
Priority:	Medium	**Due date:**	
Assignee:		**% Done:**	100%
Category:	Search / Browse	**Estimated time:**	5.00 hours
Target version:	Release 2.6.0		
Google Code Legacy ID:		**Tested version:**	
Sponsored:	Yes	**Requires documentation:**	

Description

To recreate
Navigate to Browse > Archival descriptions
In the top right hand corner of the browse results, select Sort by: Alphabetic
View results

Error encountered
Results returned are sorted in alphabetical order, but different cases (upper/lower), accents, special characters (such as starting with quotation marks), leading white spaces, and other subtle changes will affect sort order, leading sometimes to unexpected results.

Expected result
All results displayed are re-sorted to display alphabetically when a user selects Sort by: Alphabetic.

Related issues:

Related to Access to Memory (AtoM) - Bug # 5206: Alphabetic sort produces une...	**Duplicate**	**06/08/2013**
Related to Access to Memory (AtoM) - Bug # 5643: Something wonky with sort or...	**Verified**	**09/21/2013**
Related to Access to Memory (AtoM) - Bug # 5599: Changing sort order on Autho...	**Verified**	**09/17/2013**
Related to AtoM Wishlist - Feature # 3095: Natural sort for mixed alphanumeri...	**New**	
Related to Access to Memory (AtoM) - Bug # 13071: Browse hierarchy is not dis...	**Feedback**	**05/29/2019**

History

#1 - 06/04/2013 04:05 PM - Jesús García Crespo

- Status changed from New to QA/Review

- % Done changed from 0 to 100

Applied in changeset atom|commit:57a8ab717f2d1f2d7ebd4ff664f5d8623b51e5c9.

#2 - 06/04/2013 04:34 PM - Dan Gillean

- File alphabeticSort.png added

- Status changed from QA/Review to Feedback

I am still not seeing any alphabetic sort, in an order that makes any sense to me. I am not sure if it has to do with the nature of the data (unseen white spaces, etc.) or if this feature is simply not working yet, but see the sample screen shot attached - the titles go from D to J to M to D to T to W to N, etc. without apparent order.

#3 - 06/04/2013 05:07 PM - Jesús García Crespo

Sorry, I had not updated the testing site. It should be working now.

#4 - 06/04/2013 05:08 PM - Jesús García Crespo

- Status changed from Feedback to QA/Review

#5 - 06/04/2013 06:31 PM - Jessica Bushey

- Status changed from QA/Review to Feedback

It appears to be working... but...
the problem is that the titles in a lot of the data have " " or é or () or they start with a number.

Could someone explain the logic to me. For example: symbols first, numbers second, then letters?
Because we will be asked to explain this to the users.
It will drive them nuts that é comes after w.

#6 - 06/04/2013 07:17 PM - Jesús García Crespo

- Target version changed from Release 2.0 - interim 1 to Release 2.0.0

That's how it works, and we have to do some research and see if we have other options by tuning ElasticSearch/Lucene.
But I'm moving this to 2.0 for now, we can improve it later. Thanks.

#7 - 08/05/2013 05:45 PM - Jesús García Crespo

- Estimated time set to 5.00

#8 - 09/21/2013 06:21 PM - Dan Gillean

This issue was more or less duplicated in #5206 (which I've marked as such) - but the testing notes are slightly different and useful for context, so copying them here:

To Reproduce

1) Archival descriptions

- Navigate to Browse > Archival descriptions and sort Alphabetically
- Look at first page of results. Jump to last page of results

Resulting error:

- first page results: Go from A___, to R___ to "A___ ... then later in page count, back to A___
- last page results: Include letters with accents, such as É___

2) Institutions

- Go to Browse > Institutions and sort Alphabetically. Look at results.

Resulting error: First page displays 2 results starting with C__ mixed in with the A__ results

Expected Result
This is difficult, since the error may be a result of some kind of data import issue. However, one would expect at least that the R__ results in the first page of the Browse archival description page, and the C__ results on the first page of Browse institutions, would appear in the right place.

Ideally, the parameters for sorting alphabetically could be tweaked so that:

- Accents do not push results to the end of the sort order, but appear in order, so that "E__" and "É__" results would appear together, for example
- When the first character is a symbol (such as: [, (, ", ', etc.) they are excluded from consideration in the sort order.

This has been filed as an issue for consideration in 2.0, since the primary errors may be a result of some kind of data issue and not the application itself, and because the intelligent sort options (accents, special characters) border on the inclusion of a new feature.

#9 - 09/21/2013 06:23 PM - Dan Gillean

See also: #5643, and #5599 - leaving those two for now, but they are all clearly related.

#10 - 09/21/2013 06:36 PM - Tim Hutchinson

To add to this, I also noticed that titles starting with lower case letters show up at end (A to Z then a to z). You wouldn't think anyone would create titles with lower case, but we seem to have a lot particularly at item level!

#11 - 09/23/2013 10:28 AM - Dan Gillean

- Tracker changed from Bug to Feature

- Subject changed from Alphabetic sort in Browse archival descriptions does not work as expected to Improve default ElasticSearch alphabetic sort to better reflect natural sort expectations.

- Description updated

- Status changed from Feedback to New

- Target version changed from Release 2.0.0 to Release 2.1.0

Issues in AtoM with the sort order not working have been resolved. Remaining issues have to do with the default order of search results in ElasticSearch. Changing this issue ticket to reflect this - we will want, in a future release, to review and revise the sort order in ElasticSearch to optimize it and deal with better "natural" sorting, handling different cases, accents, special characters, leading white space, etc.

#12 - 09/06/2014 05:55 PM - Jesús García Crespo

- Target version changed from Release 2.1.0 to Release 2.2.0

#13 - 03/17/2015 11:27 AM - Sarah Romkey

- Target version deleted (Release 2.2.0)

#14 - 06/05/2015 07:31 PM - Dan Gillean

- Project changed from Access to Memory (AtoM) to AtoM Wishlist

- Category deleted (Search / Browse)

Moved to AtoM wishlist until sponsored for inclusion.

#15 - 06/05/2015 07:38 PM - Dan Gillean

- Related to Feature #3095: Natural sort for mixed alphanumeric identifiers added

#16 - 06/28/2015 12:20 AM - Jesús García Crespo

- Assignee deleted (Jesús García Crespo)

#17 - 09/23/2019 12:59 PM - David Juhasz

- Status changed from New to QA/Review

- Assignee set to Dan Gillean

- Target version set to Release 2.6.0

I have merged PR#971 to qa/2.6.x which improves alphabetic sorting by creating an "alphasort" ES filed that is lower cased (so case doesn't influence sort order), removes some punctuation*, and does asciifolding.

(*) The current list of removed punctation is:

'"_-?!.()[]#*`:;

I've updated the sort logic to use the new alphasort field for:

- Accession record title (accession_i18n.title)
- Archival description title(information_object_i18n.title)
- Archival institution name (repository_i18n.authorized_form_of_name)
- Authority record name (actor_i18n.authorized_form_of_name)
- Term name (term_i18n.name)

I **think** I updated all sorts using the above fields, but testing may turn up cases I have missed.

#18 - 09/23/2019 01:05 PM - David Juhasz

N.B. A full search:populate is required to add the "alphasort" field for the relevant resource types.

#19 - 09/23/2019 06:01 PM - Dan Gillean

- Project changed from AtoM Wishlist to Access to Memory (AtoM)

- Category set to Search / Browse

- Status changed from QA/Review to Verified

- Assignee deleted (Dan Gillean)

- Sponsored changed from No to Yes

There are still some results that aren't perfectly sorted the way one might expect in a natural sort (generally having to do with punctuation or spaces in the title), but this is a VAST improvement! Seems to work in other latin-character alphabets fine as well.

Note this does not change the asciibetical sorting of identifiers/reference codes/numbers. sequences will still follow ascii sort order - e.g. 1, 10, 100, 11, 2, 20, 21, 3, etc.

#20 - 10/28/2019 11:17 AM - Dan Gillean

- Related to Bug #13071: Browse hierarchy is not displaying alphabetically added

Files

alphabeticSort.png	154 KB	06/04/2013	Dan Gillean