EDIT - bug #9905

Parser is slow for names with authorteams

01/03/2022 09:58 PM - Andreas Müller

Status:	Closed	**Start date:**	
Priority:	Highest	**Due date:**	
Assignee:	Andreas Müller	**% Done:**	100%
Category:	cdmlib	**Estimated time:**	0:00 hour
Target version:	Release 5.30		
Severity:	critical	**Found in Version:**	

Description

Probably the reason is that potential deduplication candidates are not found correctly and the list of candidates is too long.

=====

ERS:

in letzter Zeit habe ich den Eindruck, dass die Performance des Parsers im Freitext-Editor stark nachgelassen hat. Es scheint mir ewig zu dauern, bis der Parser Namen, Autoren und Nomenklatorische Referenz atomisiert hat und das Taxon zum weiteren Editieren frei wird, die Sanduhr läuft und läuft…

KL:

danke für die Rückmeldung! Tritt dieses Phänomen im Zusammenhang mit dem neuen Release auf oder gab es das Problem schon länger bzw tritt das erst seit kurzem auf?
Ich frage, weil es ja an der Performance des CdmServers oder wirklich am Parser liegen kann. Ich werde auf jeden Fall aber auch nochmal testen und debuggen.

ERS:

ich habe den Eindruck, dass es mit dem neuen Release zusammenhängt. Um das Problem zu konkretisieren:

ich habe ein neues Taxon eingegeben (Rechtsklick auf den Elternknoten -> Neu-> Taxon-> Neues Taxon.
Hier gebe ich den kompletten Namen mit Autoren und Referenz ein, danach auf „finish and open“. Es dauert etwa 40-50 Sekunden, bis die Sanduhr abläuft und der Freitext-Editor sich öffnet. Daher vermute ich, dass die Verzögerung an der Stelle durch den Parser kommt, oder?

KL:

ich habe gestern schon mal ein bisschen getestet und mir scheint, dass tatsächlich beim Suchen der möglichen Kandidaten für Autoren irgendwie zu viele Ergebnisse zurück kommen, aber das muss ich nochmal verifizieren und dann an Andreas M. weiter leiten. Wobei an dieser Stelle eigentlich schon im Sommer gearbeitet wurde. Aber manchmal fällt es eben erst später auf.

ERS:

das klingt für mich insofern plausibel, als dass es deutlich schneller geht, wenn nur ein Autor vorhanden ist und nicht mehrere.

KL:

Hallo Andreas, Eckhard hat ja ... angemerkt, dass der Parser bei längeren Autorenlisten sehr langsam geworden ist. Ich habe mir das oberflächlich mal angesehen und bei den „potentiellen“ Autoren erscheinen sehr viele, bei denen mir nicht klar ist, wie sie in die Liste der potentiell matchenden Personen kommen, vielleicht kannst Du da selber nochmal drauf gucken?

AM:

Kannst du mir noch sagen, an welcher Stelle im Code das genau ist?
Und mit welchen Daten hast du getestet?

KL:

ich habe in meiner caryophyllales_ssp Instanz getestet und zum Beispiel, wenn man Freitag & G.Kadereit als Autorenteam angibt, dann bekommt man eine sehr lange Liste von candidates (s. u. und diese ist noch nicht vollständig…)

der code ist in CdmGenericDaoImpl ungefähr Zeile 658

bei einzelnen Autoren scheint es gut zu funktionieren, da kommen soweit ich das gerade eben nochmal getestet habe, passende Kandidaten (ich habe nur L. getestet)

```
Match candidate did not match: Eckl. & Zeyher, C.L.P.
[local-test] 2022-01-03 10:59:19,095 INFO [eu.etaxonomy.cdm.persistence.dao.hibernate.common.CdmGe
nericDaoImpl] - Match candidate did not match: Rose, J.N. & Standley, P.C.
[local-test] 2022-01-03 10:59:19,095 INFO [eu.etaxonomy.cdm.persistence.dao.hibernate.common.CdmGe
nericDaoImpl] - Match candidate did not match: Alvarado-Sizzo, H., Casas, A., Parra, F., Arreola-N
ava, H. J., Terrazas, T. & Sánchez, C.
[local-test] 2022-01-03 10:59:19,095 INFO [eu.etaxonomy.cdm.persistence.dao.hibernate.common.CdmGe
nericDaoImpl] - Match candidate did not match: Barrios, D., González-Torres, L.R., Arias, S. & Maj
ure, L.C.
[local-test] 2022-01-03 10:59:19,095 INFO [eu.etaxonomy.cdm.persistence.dao.hibernate.common.CdmGe
nericDaoImpl] - Match candidate did not match: Fuertes, J. F. & Nieto Feliner, G.
[local-test] 2022-01-03 10:59:19,095 INFO [eu.etaxonomy.cdm.persistence.dao.hibernate.common.CdmGe
nericDaoImpl] - Match candidate did not match: Dinter, M.K. & G.Schellenb.
[local-test] 2022-01-03 10:59:19,095 INFO [eu.etaxonomy.cdm.persistence.dao.hibernate.common.CdmGe
nericDaoImpl] - Match candidate did not match: Eckl. & Zeyher, C.L.P.
[local-test] 2022-01-03 10:59:19,095 INFO [eu.etaxonomy.cdm.persistence.dao.hibernate.common.CdmGe
nericDaoImpl] - Match candidate did not match: Eckl. & Zeyher, C.L.P.
...
```

AM:

ok, ich schau mir das an. Kannst du evtl. noch mal kurz testen, ob es ein Problem ist, was nur bei Teams auftritt?

KL:

bei einzelnen Personen scheint das Problem nicht zu bestehen, da kommt nur eine Liste, wenn es wirklich mehrere Einträge mit dem selben Namen gibt, wie z.B. bei L.

ERS:

Im Nachgang zu meinen bisherigen Nachrichten:

Testet doch einmal die Eingabe folgenden Namens in den Freitext-Editor und seht, wie sich der Parser verhält.

Berberocarum Zakharova & Pimenov in Nordic J. Bot. 2021(e03206): 10. 2021

Es gibt auch zwei Problem-Meldungen, die ich eigentlich nicht verstehe:

„check rank“ (obwohl der Rang korrekt als Genus erkannt wird)
„reference title not parsable“ (liegt das an der ungewöhnlichen Bandangabe für electronic journals? Damit müsste der Parser aber auch umgehen können, d.h., dort sollte man alle Möglichkeiten offenhalten und die Eingabe nicht irgendwie beschränken)

Eine generelle Frage, die ich beobachtet habe: der Parser fängt ja anscheinend sofort an zu arbeiten, sobald ich etwas eintippe. Wenn ich also nicht mit copy&paste arbeite, sondern wirklich Freitext eintippe, wird sofort versucht zu parsen, auch wenn ich noch nicht fertig bin und z.B. die Seitenzahl oder etwas anderes noch kurz nachschlagen muss. Wenn meine Eingabe also noch nicht abgeschlossen ist, kommen dann Fehlermeldungen.
Bisher war mir das nicht so aufgefallen, da der Parser sehr schnell war. Aber wie gesagt, jetzt läuft es in dem Bereich (gefühlt seit einem der letzten Releases) sehr langsam.

Vielleicht hilft das Beispiel oben ja weiter, das Problem einzugrenzen. Schaut doch mal, wie lange das Parsen dieses Namens bei euch dauert und welche Schwierigkeiten dabei entstehen.

parallel AM:

Katja hat da drauf geschaut und bereits ein mögliches Problem gefunden. Kann es sein, dass das Problem nur bei Namen auftaucht, die ein Autorenteam und nicht einen Einzelautor haben?
Ich schaue mir das noch im Detail an und hoffe, dass wir es zum nächsten Release gelöst bekommen.

ERS:

da haben sich unsere mails gerade überschnitten. Ja, auf jeden Fall ist es bei Autorenteams besonders langsam.

Related issues:	
Copied to EDIT - task #9964: Improve matching of collections	**New**

Associated revisions

Revision 23ac1700 - 03/02/2022 12:35 PM - Andreas Müller

ref ##9943 add (commented) test for #9943 (too many matching candidates for teams)

Revision bbc29125 - 03/02/2022 12:41 PM - Andreas Müller

ref #9905 fix ticket number and add javadoc

Revision a253e382 - 03/03/2022 08:42 AM - Andreas Müller

ref #9905 preliminary fix handling of collections with match mode "match" using count and Person.nomenclaturalTitle

Revision 360e1668 - 03/03/2022 11:51 AM - Andreas Müller

ref #9905 fix exception when matching TeamOrPersonBase

History

#1 - 01/03/2022 10:08 PM - Andreas Müller

- Status changed from New to In Progress

- Priority changed from New to Highest

- Severity changed from normal to critical

#2 - 02/24/2022 06:49 PM - Andreas Müller

- Tags changed from parser, deduplication, euro+med to parser, deduplication, euro+med, 5.30

#3 - 02/25/2022 09:06 AM - Andreas Müller

- Target version changed from Unassigned CDM tickets to Release 5.45

#4 - 03/02/2022 10:49 AM - Andreas Müller

The problem is that in CdmGenericDaoImpl.matchNonComponentType() - currently line 775 - the MatchMode Match is not yet implemented for collections:

```
if (isMatch(matchModes)){
    if (propertyType.isCollectionType()){
        //TODO collection not yet handled for match
    }
```

#5 - 03/02/2022 11:05 PM - Andreas Müller

- % Done changed from 0 to 30

JPA allows addressing a certain list entry by index. Example here:
https://www.logicbig.com/tutorials/java-ee-tutorial/jpa/criteria-api-collection-operations.html

#6 - 03/02/2022 11:17 PM - Andreas Müller

- Copied to task #9964: Improve matching of collections added

#7 - 03/03/2022 12:13 PM - Andreas Müller

- Status changed from In Progress to Resolved

- % Done changed from 30 to 50

#8 - 03/03/2022 02:05 PM - Andreas Müller

- Status changed from Resolved to Closed

- % Done changed from 50 to 100

The work around works for me and KL also tested it and finally did not find any errors anymore. So I close this ticket.

#9 - 03/03/2022 02:28 PM - Andreas Müller

- Target version changed from Release 5.45 to Release 5.30

#10 - 03/04/2022 03:43 PM - Andreas Müller

- Tags changed from parser, deduplication, euro+med, 5.30 to parser, deduplication, euro+med

#11 - 05/12/2022 10:25 PM - Andreas Müller

- Related to bug #10061: Current in-TaxEditor documentation needs to be removed added

#12 - 05/12/2022 10:30 PM - Andreas Müller

- Related to deleted (bug #10061: Current in-TaxEditor documentation needs to be removed)