Estimating indexing time for different combinations of tab00 / tab11 entries

- **Article Type:** General
- **Product:** Aleph
- **Product Version:** 18.01

Description:
We are in the middle of a project to configure v18 authority control functionality, refresh the auth records in abc10, and reindex all the bib headings. In the course of this project, Tech Services has asked for a number of new indexes or changes to existing indexes, and we'd like to be able to present some data about the trade-offs of adding more ACC indexes in tab00.eng - or adding 100 new lines to tab11_acc, for example. We were planning to do some performance testing here on indexing based on two versions - one with everything they want, and the other with a "streamlined" version of the indexes. The question is what configuration changes to make for the "streamlined" version. The two that immediately came to my mind were the ones I mentioned above: overall number of ACC indexes and all the tags (and permutations of the tags like for tab_xyz functionality) included in the index. Do you have a good sense of what factors we should really be considering here?

Resolution:
For p_manage_01 (Word):

The factors influencing the time for p_manage_01 are the number of documents and the number of word-index entries generated for the documents.

UTIL-F-1-28 ("Display Word Indexing for a Single Record") is useful in seeing this. Take the util f/1/28 for USM01 doc 000012345 shown below. I count roughly 62 000012345 lines, each representing a word being sent to an index. (The third column is the index number, "001" = "W-001" = WORDS, "009" = "W-009" = "W-publishers" , etc.)

I think the main factor in the indexing time is not the number of indexes but the number of word-index entries as shown below. If one scenario results in significantly more word-index entries, it will require more p_manage_01 time.

I do not think that the relationship is linear ... less than linear.

Other factors can be the number of documents (1,000 docs with 100,000 word-index entries will require less time than 5,000 docs with 20,000 word-index entries) and the number of different words in the documents (the more unique words, the more time).

But I think the main factor is the number of word-index entries.

For p_manage_02 (Browse):

The UTIL-F-1-29 ("Display Headings Indexing for a Single Record") can be used in a similar fashion to determine the number of Browse/ACC index entries for a particular tab00 / tab11_acc combination.

ue_01: The effect of additional indexes upon the ue_01 indexing daemon is exactly the same as that described above for

p_manage_01 and p_manage_02 and the same principles can be used in estimating the effect.

util f/1/28 for usm01 doc 000012345:

FMT BK
Load: /tmp/utf_files/exlibris/aleph/a18_1/usm01/tab/tab_word_breaking
Load: /exlibris/aleph/a18_1/alephe/unicode/unicode_to_word_gen
000012345 0001 0011 bk

LDR 00385na9^^22001455|^4500

001 000012345-5

005 20021206084957.0

008 881209|||||||||^^^|||||||||||||||^^^||

24510 $$aJourn?©es Anthropologiques de Valbonne.
000012345 0152 0001 journees
000012345 0153 0001 anthropologiques
000012345 0154 0001 de
000012345 0155 0001 valbonne
000012345 0156 0001 journeesanthropologiques
000012345 0157 0001 anthropologiquesde
000012345 0158 0001 devalbonne
000012345 0209 0002 journees
000012345 0210 0002 anthropologiques
000012345 0211 0002 de
000012345 0212 0002 valbonne
000012345 0213 0002 journeesanthropologiques
000012345 0214 0002 anthropologiquesde
000012345 0215 0002 devalbonne

260 $$aParis :$$bCentre National de la Recherche Scientifique,$$c1983-

000012345 0266 0001 paris
000012345 0267 0001 centre
000012345 0268 0001 national
000012345 0269 0001 de
000012345 0270 0001 la
000012345 0271 0001 recherche
000012345 0272 0001 scientifique
000012345 0273 0001 1983-
000012345 0274 0001 scientifique
000012345 0275 0001 1983
000012345 0276 0001 scientifique
000012345 0277 0001 1983
000012345 0278 0001 pariscentre
000012345 0279 0001 centrenational
000012345 0280 0001 nationalde
000012345 0281 0001 dela

000012345 0282 0001 larecherche
000012345 0283 0001 recherchescientifique
000012345 0284 0001 scientifique1983-
000012345 0285 0001 scientifique1983
000012345 0286 0001 scientifique1983
000012345 0337 0016 paris
000012345 0388 0009 centre
000012345 0389 0009 national
000012345 0390 0009 de
000012345 0391 0009 la
000012345 0392 0009 recherche
000012345 0393 0009 scientifique
000012345 0394 0009 centrenational
000012345 0395 0009 nationalde
000012345 0396 0009 dela
000012345 0397 0009 larecherche
000012345 0398 0009 recherchescientifique
000012345 0449 0010 1983-
000012345 0450 0010 1983
000012345 0451 0010 1983

300 $$av.
000012345 0502 0001 v
000012345 0502 0007 v

STA $$aSUPPRESSED
000012345 0553 0012 suppressed

PST $$0Z30$$1000012345000010$$bWID$$cGEN$$oBOOK$$eOI$$fN$$rUSM60-000000000$$3B
ook$$4Main Library$$5General$$7Order initiated

SBL $$aWID
000012345 0604 0014 wid

LOC $$bWID$$cGEN$$oBOOK
000012345 0655 0001 wid
000012345 0655 0013 wid
000012345 0656 0001 gen
000012345 0656 0013 gen
000012345 0657 0001 widgen
000012345 0657 0013 widgen

NTL $$ajourneesanthropologiquesd
000012345 0708 0023 journeesanthropologiquesd

(keywords: p-manage-01 p-manage-02 manage_01 manage_02)

- **Article last edited:** 10/8/2013