Hi Everyone:

As one of our routine tasks, we analyze the results of LUI assignment based on OCCS reports on Merge, Split, and Split-Merge (from Soma) to monitor the behavior of LuiNorm, enhance LuiNorm and its associated LVG flow components, and fix Lexicon data for the next release. Based on the report of last year, there is no software change suggested on the LuiNorm program. The only change for LuiNorm is the data from updated UMLS (atoms) and LEXICON. Accordingly, the results of LuiNorm should be the same between 2009 and 2010 releases except for terms contain new words from LEXICON or UMLS. Theoretically, terms have same LuiNorm results should have same LUI assignment and no merge, split, or split-merge should be introduced between releases.

In addition, there are two possible issues observed on the last year OCCS LUI assignment Java programs and scripts from the report of last year:

- Ambiguity tags, such as <1>, <2>, etc.
- stopwords, such as “with”, “NOS”, “in”, “by”, “and”, etc.

As Soma indicated, the above two issues were addressed in the 2010 release by enhanced algorithm on the OCCS LUI assignment Java programs and scripts. As expected, these changes introduce in different LUI assignment for merge, split, and split-merge cases.

In our study, first, we observe the total number of merge cases (SUIs) increased from 2618 to 6382 from 2009 to 2010 release. The main reason of this increased amount of merge strings is the enhanced algorithm on the OCCS LUI Assignment Java programs and scripts. There are only 5.36% (343/6382) of merge strings are caused by the different LuiNorm results. Second, we observe the total numbers of all split strings (100) and split-merge strings (320) stay in the same level of small amount as the last release. There are 35 and 33 strings have different LuiNorm results for split and split-merge cases, respectively. All these three cases with different LuiNorm results are caused by the different base forms or Canonical forms resulting from new words in UMLS, new words in LEXICON, new inflectional rules in LEXICON, and new spelling variants in the LEXICON. This small amount of changes is expected to happen every year and considered as system enhancement.

In conclusion, LuiNorm.2010 behaves very well and no software change requests or lexical records updates are found from this study. We also assume the OCCS enhanced algorithm on LUI-assignment programs fix issues from the report of last year. Accordingly, we expect low magnitude of merge, split, and split-merge for the next release (2011). We are pleased with the results and considering the algorithm of LuiNorm and OCCS LUI assignment programs are stable. Please refer to the web site for more details: http://lexlx1.nlm.nih.gov/LexSysGroup/Projects/lvg/2010/docs/designDoc/LifeCycle/test/luiAssign/index.html.

==========

- **Merge cases:**
  - Summary:
    There are 6382 SUIs with 3212 Luis merged into 1584 Luis. As discussed above, most of these merge cases (94.63%) are caused by the enhanced algorithm from OCCS LUI-assignment programs and scripts. Only 5.37% (343/6382) strings merge caused by the different LuiNorm results. Table 1 shows the percentage distribution of all causes of merge cases. There are 1.64% and 3.73% of merge cases caused by the changes of Canonical forms and base forms, respectively. Please refer the analysis section below for details.

Causes of merge cases	**Merge Strings**	**Percentage**	**Examples**
Enhanced Software on OCCS LUI assignment programs	6039	94.63%	▪ cancer <1> ▪ CANCER <NOS> ▪ Back pain <3> ▪ Pain in back <1>
Different Canonical form (new data in LEXICON & UMLS)	105	1.64%	▪ Salazodine ▪ Quaternifolia ▪ T11
Different Base form (new data in Lexicon)	238	3.73%	▪ itching ▪ Wilms

Table 1. Percentage distribution of merge causes

  - Merge analysis:
    1). By the enhanced software on OCCS LUI assignment programs:
    As suggested from the report of last year, issues of handling ambiguity tags and stopwords were addressed on the LUI-Assignment programs and scripts. This software change introduce major amount (94.63%) of merge strings for this release. For example, “cancer <10>” is merged into cancer class; “Other <106>” is merged into “other” class; and “Pain in back <1>” is merged into “back pain” class. The total number of merge cases is expected to be dropped down to low magnitude for the next release assuming no algorithm change for the next release.

    2). By Canonical form (caused by new data in UMLS - atoms or LEXICON):
    There is no software change in canonical forms and LuiNorm in Lexical Tools. In other words, the results of canonical forms should be the same as the previous release except for strings contain new data from UMLS – atoms or LEXICON. Below, we illustrate two examples in merge cases from these causes:

    Example 1 - New words in UMLS

"Salazodine" is a new word in UMLS. "salazodine" belongs to same canonical class of "salazodin" in 2010 while it was not in 2009. Accordingly, "salazodine" is merged with "salazodin" into one LUI in 2010.

Example 2 - New words in UMLS
"T11e" is a new word in UMLS. "t11e" belongs to same canonical class of "T11" in 2010 while it was not in 2009. Accordingly, "t11e" is merged with "t11" into one LUI in 2010.

3). By base forms (new data in Lexicon):
The results of base forms from Lexical tools mainly depend on the data of LEXICON. The base form of a word might be different from last version if there are new lexical record(s), modified or new inflectional rules, or deleted lexical records associated with this word. These cases are expected to be observed every year and are considered as an enhancement between releases. Two examples are illustrated as follows:

Example 1 - New Lexical Records
"itch" is a new record (E0693125) with part of speech of verb in LEXICON 2010. According to the inflection rules (variants=reg) in the lexical record, "itch" has inflectional variant of "itching" and merged together to the same LUI.

Example 2 - New Lexical Records
"Wilm" is a new record (E0667830) with part of speech of noun in LEXICON 2010. According to the inflection rules (variants=reg) in the lexical record, "wilm" has inflectional variant of "wilms" and merged together to the same LUI.

==========

- **Split:**
  - Summary:
    There are 100 SUIs with 29 LUIs split to 58 LUIs. As expected, these numbers stay in the low magnitude as the previous release (2009). Most of these split cases (82.76%, 24/29) are caused by different base forms. As mentioned above in Merge section, this is caused by the data change in UMLS or LEXICON. In addition, new canonical forms introduced 17.24% (5/29) on the split cases. The total number of split cases is very small and we did not find any unexpected behavior of LuiNorm in the split cases. Table 2 shows the percentage distribution of each cause.

Causes	**Split No.**	**Percentage**	**Example**
Canonical form (new data in UMLS & Lexicon)	5	17.24%	▪ miurus ▪ indistinguendum
Base form (new data in Lexicon)	24	82.76%	▪ Peristalses ▪ viannias

Table 2. Percentage distribution of split causes

- Split analysis:
  1). By Canonical form:
  As mentioned in the merge cases, different canonical forms can be caused by the data change in UMLS or LEXICON. Bellows, we illustrate two examples of split cases caused by data change in resulting different Canonical forms. Both cases are caused by terms removed from UMLS-atoms.data. It may be worthy to trace down the cause for these records to be removed even this have very little impact on split cases.

  Example 1 – Deleted words in UMLS
  The old term, “Noturus miuris” was removed from atoms.data 2009 and does not exist in atoms.data.2010. Accordingly, “miuri” and “miuris” do not have the same canonical form of “miuru” and results in split case.

  Example 2 – Deleted words in UMLS
  The old term, “Nostoc indistinguenda” was removed from atoms.data 2009 and does not exist in atoms.data.2010. Accordingly, “indistinguendum” does not have the same canonical form of “indistinguenda” and results in split case.

  2). By base forms (new data in Lexicon):
  Split cases can be caused from different based forms by new lexical records to result in different LuiNorm results. Bellows we illustrate two examples of split cases caused by data change in LEXICON to have different base form and results in different Canonical forms.

  Example 1 - New Lexical Records
  “peristalse” is a new record (E0692542) with part of speech of verb in LEXICON 2010. According to the inflection rules (variants=reg) in the lexical record, “peristalse” has inflectional variant of “peristalses” and split from “peristalsis” to a different LUI.

  Example 2 - New Lexical Records
  “viannia” is a new record (E0657227) with part of speech of noun in LEXICON 2010. According to the heuristic rules in the inflections generation by rules in the lexical tools, “viannia” is removed as the base form of “viannias” because “viannia” is known to LEIXCON. This effect split “viannias” from “viannia” to a different LUI.

==========

- **Split_Merge:**
  - Summary:
  Split merge cases are the cases when some words split first, then merge (with others) again. There are 320 SUIs with 48 LUIs split-merge cases. The total number (48) of split merge cases in 2009 is relatively small of LUI assignment. Table 3 shows the percentage distribution of each cause. There are 93.75%

(45/48) and 6.25% (3/48) caused by new base forms and enhanced algorithm on OCCS LUI-Assignment programs, respectively.

Causes	**Split-Merge No.**	**Percentage**	**Example**
Base form (new data in Lexicon)	45	93.75%	▪ SAT ▪ IPS
Enhanced Software on OCCS LUI assignment programs	3	6.25%	▪ Penicillin VK <2>

Table 3. Percentage distribution of split-merge causes

- Split_Merge analysis:
  The cause of this case is the combination of above two (split and merge). Potentially, terms in the split-merge cases might belong to same canonical class. This is the data we use to enhance the algorithm of canonical form and make canonical class covers bigger range (more words). In this analysis (2010), we did not found anything to enhance LuiNorm from these 48 cases. Please refer to the discussion in split or merge session for details.