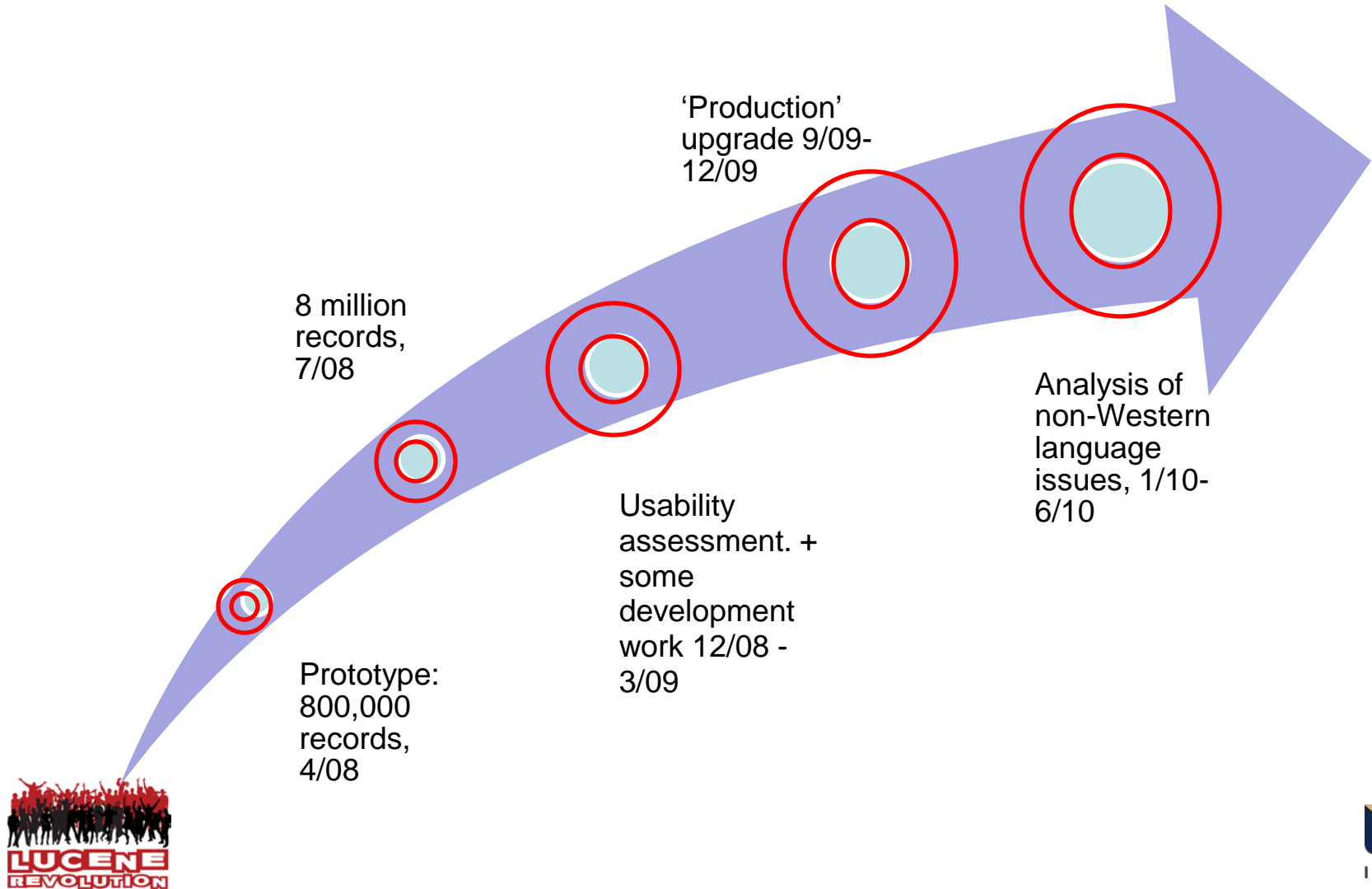


Scaling the Stacks of Babel: Challenges and Opportunities for Solr-Powered Multilingual Multi-script Resource Discovery and Access in Research Libraries

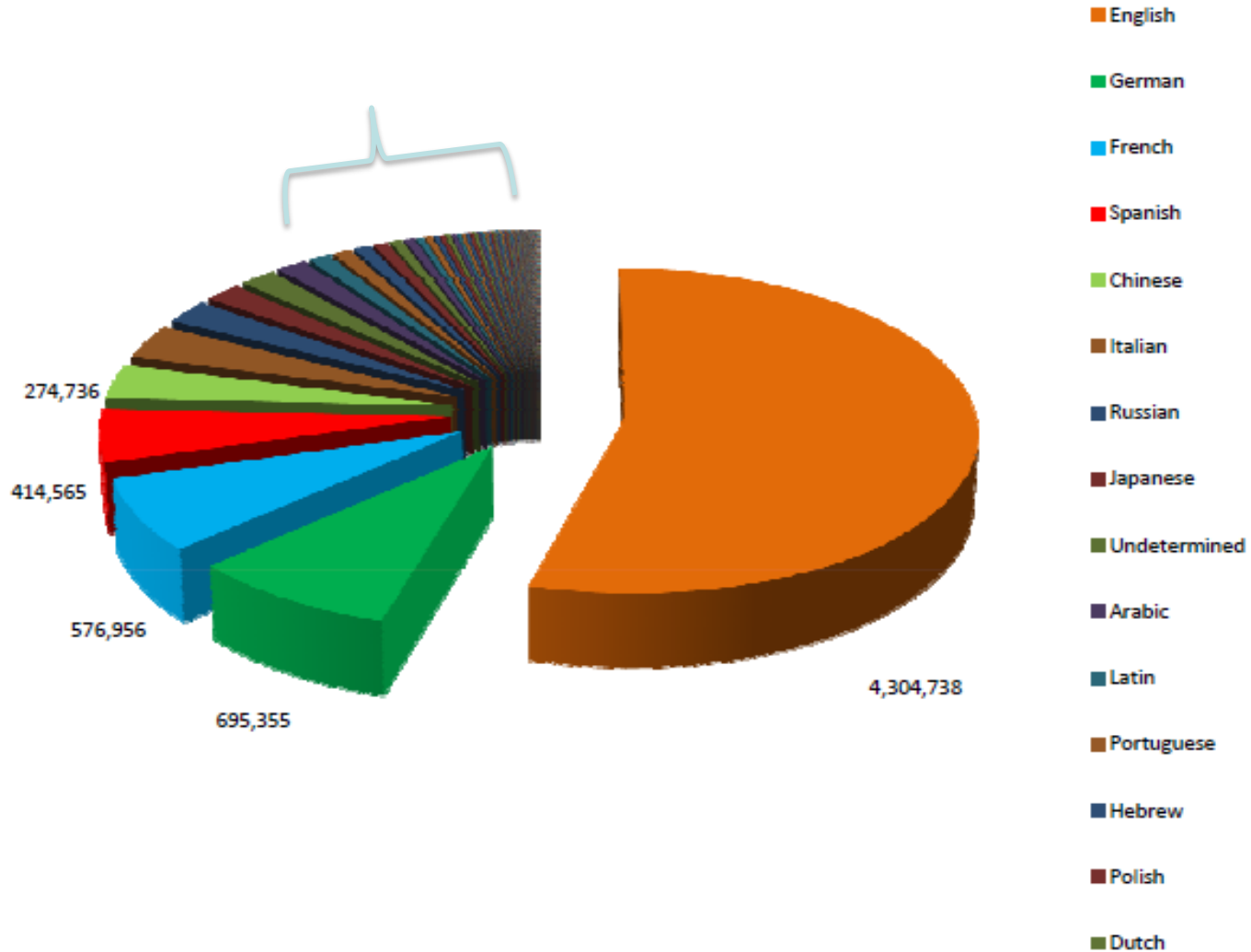
**Jeffrey Barnett, Daniel Lovins, Yale University Library, Oct. 8,
2010**




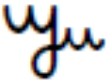









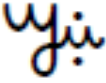









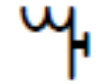









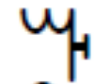









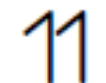









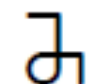








Evolution of Solr-Powered Catalog at Yale ('Yufind')



Language Challenge



Example Script: Vai

| | | | | | | | | | |
|---------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
|  A500 |  A510 |  A520 |  A530 |  A540 |  A550 |  A560 |  A570 |  A580 |  A590 |
|  A501 |  A511 |  A521 |  A531 |  A541 |  A551 |  A561 |  A571 |  A581 |  A591 |
|  A502 |  A512 |  A522 |  A532 |  A542 |  A552 |  A562 |  A572 |  A582 |  A592 |
|  A503 |  A513 |  A523 |  A533 |  A543 |  A553 |  A563 |  A573 |  A583 |  A593 |
|  A504 |  A514 |  A524 |  A534 |  A544 |  A554 |  A564 |  A574 |  A584 |  A594 |
|  A505 |  A515 |  A525 |  A535 |  A545 |  A555 |  A565 |  A575 |  A585 |  A595 |



Character Mapping: CJK

“Mao Zedong”

毛泽东

Simplified

毛澤東

Traditional

毛沢東

Kanji

Configuring FieldType in schema.xml

```
<!-- Text Field for original texts (multilingual) -->
- <fieldType name="text_880" class="solr.TextField" positionIncrementGap="100">
  - <analyzer type="index">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="schema.UnicodeNormalizationFilterFactory" version="icu4j" composed="false"
remove_diacritics="true" remove_modifiers="true" fold="true"/>
    <filter class="solr.ISOLatin1AccentFilterFactory"/>
    <filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1"
catenateWords="1" catenateNumbers="1" catenateAll="0"/>
    <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.RemoveDuplicatesTokenFilterFactory"/>
  </analyzer>
  - <analyzer type="query">
    <tokenizer class="solr.StandardTokenizerFactory"/>
    <filter class="schema.UnicodeNormalizationFilterFactory" version="icu4j" composed="false"
remove_diacritics="true" remove_modifiers="true" fold="true"/>
    <filter class="solr.ISOLatin1AccentFilterFactory"/>
    <filter class="solr.WordDelimiterFilterFactory" generateWordParts="1" generateNumberParts="1"
catenateWords="0" catenateNumbers="0" catenateAll="0"/>
    <filter class="solr.StopFilterFactory" ignoreCase="true" words="stopwords.txt"/>
    <filter class="solr.LowerCaseFilterFactory"/>
    <filter class="solr.RemoveDuplicatesTokenFilterFactory"/>
  </analyzer>
```

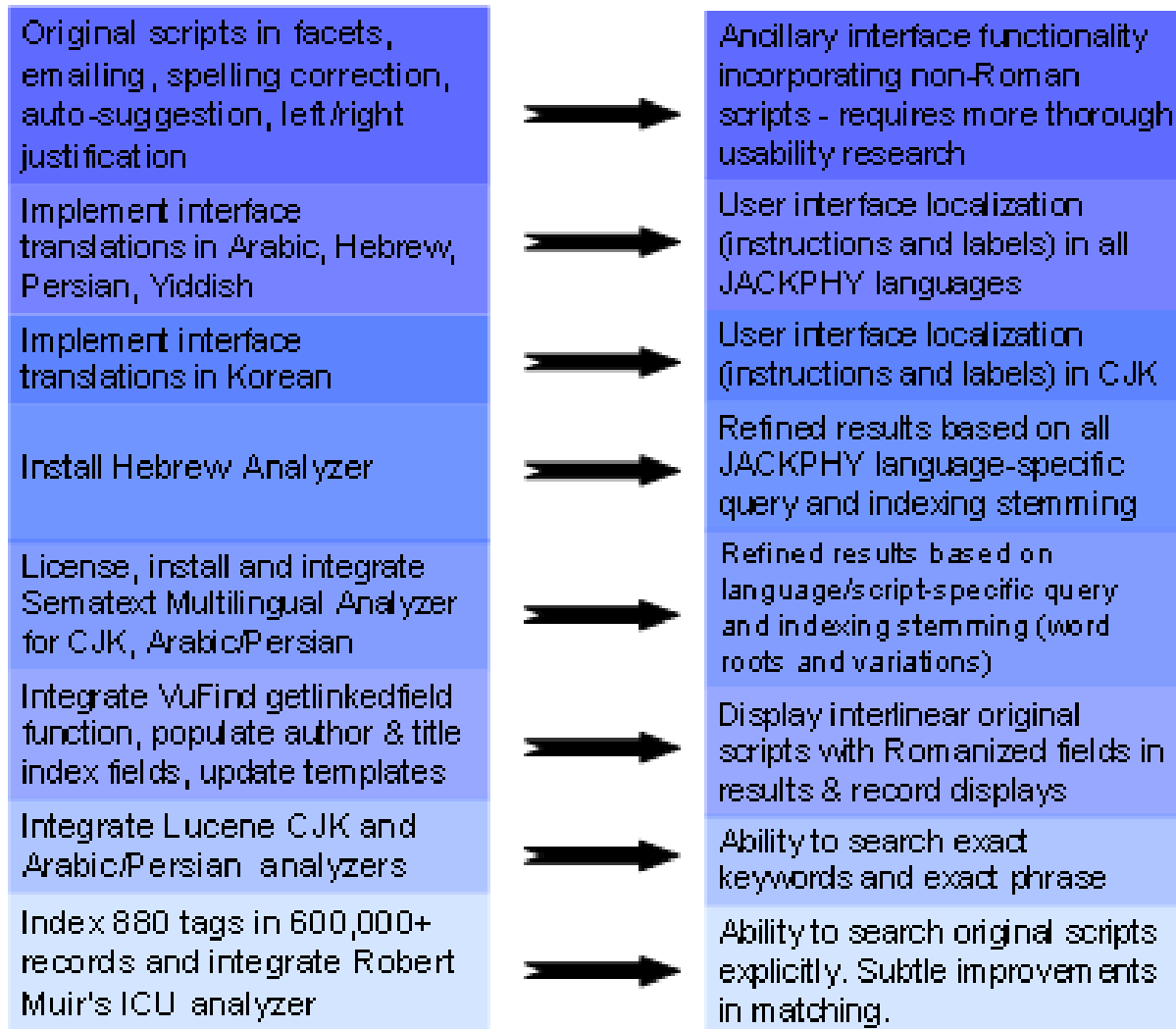


Configuring Field Content in marc_local.properties

```
# Uncomment the following settings to use the .bsh scripts in
import/scripts/
# instead of the built-in SolrMarc functionality found in the .jar file.
# (by default, the scripts have the same behavior as the built-in
functions,
# but the external scripts are easier to customize to your needs).
language = 008[35-37]:041a:041d:041j, language_map.properties
author2-role = 700e:710e
author additional = 505r
author_880 = custom, getLinkedField(100ab)
title_880 = custom, getLinkedField(245ab)
title_alt = 130adfgklnpst:240a:246a:730adfgklnpst:740a
topic_880 = custom, getLinkedField(600a)
```



Report Recommendations



Contact

- jeffrey.barnett@yale.edu
- daniel.lovins@yale.edu



More Information

- **“Investigating Multilingual, Multi-script Support in Lucene/Solr Library Applications” (June, 2010). Unpublished report for the Yale University Library. Jeffrey Barnett, Daniel Lovins, Audrey Novak, Charles Riley, Keiko Suzuki.**

<https://collaborate.library.yale.edu/yufind/public/FinalReportPublic.pdf>

