Center for Content Extraction

Content Extraction Analytics
SIGDEV End-to-End Demo

21 May 2009
Introduction to Content Extraction

- New technologies can find Essential Elements of Information in documents

- The Center for Content Extraction provides “one stop shopping” for these technologies at NSA
Extraction can benefit SIGDEV from end to end

- Selection
- Translation & Transliteration
- Analysis
- Interpretation/Enrichment
- Retrieval
- Storage & Distribution
STAIRS Partners

S (Marina, CEA)

T (Cybertrans)

A (SNA/Paintball, Synapse)

I (Nymrod, Thundercloud)

R (Journeyman/CPE)

S (GoldenRetriever, SocioPath)
Implementation: CCE Extraction Architecture (LexHound)

**Subscription Based Customers - extracted report/transcript content**
- Marina (comms tracking)
- Synapse/EKS (link analysis)
- Nymrod (Name Matching)

**Web Service On Demand Customers**
- LexHound Web Demo
- CAMT (translation)
- TKB (target knowledge base)
- SNA (social network analysis)
- GIS (geo mapping)
- NTDC (terror cell tracking)
- Heryitch (UC collateral)
- GoldenRetriever (record building)
Elaboration: *The Central Importance of Storage*

- Each of the STAIRS Steps exploits stored information
  - Selection Dictionaries ("get it")
  - Linguistic Glossaries for Translation
  - Wikis etc for enrichment ("know it")

- Manual record-formation is slow, prone to omissions and inconsistencies
  - <200K Person Targets in TKB
  - Growth $\approx 20K$/year

- Automatic extraction accelerates storage
  - >3000K Citation Records in *Nymrod* Entity DB
  - Growth $\approx 1000K$/year
# Machine vs. Manual Chief-of-State Citations

<table>
<thead>
<tr>
<th>Nymrod (machine-extracted) Citations</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>Role</strong></td>
<td><strong>Code</strong></td>
<td><strong>Cites</strong></td>
</tr>
<tr>
<td>Abdullah Badawi</td>
<td>Malaysian Prime Minister</td>
<td>cos</td>
<td>&gt;100</td>
</tr>
<tr>
<td>Abdullahi Yusuf</td>
<td>Somali President (Mahmud 'Abbas) PA</td>
<td>cos</td>
<td>&gt;300</td>
</tr>
<tr>
<td>Abu Mazin</td>
<td>President</td>
<td>cos</td>
<td>&gt;200</td>
</tr>
<tr>
<td>Alan Garcia</td>
<td>Peruvian President</td>
<td>cos</td>
<td>&gt;100</td>
</tr>
<tr>
<td>Aleksandr Lukashenko</td>
<td>Belarusian President</td>
<td>cos</td>
<td>&gt;50</td>
</tr>
<tr>
<td>Alvaro Colom</td>
<td>Guatemalan President</td>
<td>cos</td>
<td>&gt;200</td>
</tr>
<tr>
<td>Alvaro Uribe</td>
<td>Colombian President</td>
<td>cos</td>
<td>&gt;700</td>
</tr>
<tr>
<td>Amadou Toumani Toure</td>
<td>Malian President</td>
<td>cos</td>
<td>&gt;50</td>
</tr>
<tr>
<td>Angela Merkel</td>
<td>German Chancellor</td>
<td>cos</td>
<td>&gt;300</td>
</tr>
<tr>
<td>Bashar al-Asad</td>
<td>Syrian President</td>
<td>cos</td>
<td>&gt;800</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Yuliya Tymoshenko</td>
<td>Ukrainian Prime</td>
<td>cos</td>
<td>&gt;200</td>
</tr>
</tbody>
</table>