DAM-LR at the INL
Archive Formation and Local INL

Remco van Veenendaal
veenendaal@inl.nl
http://imdi.inl.nl

Introducing…

Remco van Veenendaal
- Project manager DAM-LR
- Acting project manager Dutch HLT Agency (TST-centrale)
- Computational Linguist

Vincent Wagelaar
- Software engineer and system administrator DAM-LR

Overview

- The INL
- Archive Formation at the INL
- Local Solution at the INL
- Summary

The INL

The INL (1/2)

- INL studies the Dutch language
- Focus on lexicology: "Dutch language database"
  - Word lists, dictionaries, corpora and software tools
  - Jan 07: http://wnt.inl.nl
    - **The world's largest dictionary – free access**
- Track record of 40 yrs (1967-2007), about 40 employees
- Previous international projects
  - Parole (creation of electronic LRs)
  - Telri (network with Eastern Eu. LR creators)
  - Elan (infrastructure for Parole LRs)
  - Simple (semantics for Parole LRs)
  - Enabler (infrastructure for LRs)

The INL (2/2)

- Home of the Dutch HLT Agency (TST-centrale)
  - Initiative of the Dutch Language Union (NTU)
  - Single point of access for Dutch language resources (STEVIN)
    - Strengthen position of Dutch (in language technology)
    - Stimulation of (re)use of Dutch LRs
      - Acquisition, Maintenance, Support, Distribution (ITIL)
      - **IPR and service**
        - IPR, linguistic and technical consultancy (free)

- DAM-LR logical next step in international projects and DAM-LR and TST-centrale's services mutually beneficial
  - Management buy-in
  - Co-operation between DAM-LR, TST-centrale and IT department

Archive Formation at the INL

Archive Formation at the INL (1/2)

- 2004/5
  - Training at MPI: IMDI, IMDI corpora (Spoken Dutch Corpus, CGN 1.0) and IMDI software (Corex, LAMUS)
- 2005
  - Inventory of LRs (INL/TST-centrale)
  - Inventory of access rights for LRs
    - IPR!
  - Pilot version of archive with CGN 1.0
  - Hardware (dedicated DAM-LR server)
    - 3.4 GHz single core Intel Pentium 4
    - 1 Gb RAM and redundant 250 Gb disk storage
    - OS: FreeBSD 5.4 (Java 1.4.2)
  - Start of conversion of IMDI LRs to IMDI 3.0
  - 75 person months (Archive + Local INL)

Archive Formation at the INL (2/2)

- 2006
  - "**Production Line**" system for LRs
    - One system for storage, back-up, versioning, conversion/integration and distribution/exploration of LRs
      - Boekestein, M. et al. (2006). The functioning of the Centre for Dutch Language and Speech Technology. *Proceedings of the 5th International Conference of Language Resources.* Genoa : pp. 2303-2306.
  - CGN 2.0 and IFA corpus in archive
    - Many LRs in pipeline (next)
  - Handles generated and integrated in IMDI
  - 52 person months (Archive + Local INL)

CGN corpus 2.0

Spoken Dutch Corpus

- 900 hours spoken Dutch (Netherlands and Flanders)
- 12,780 audio fragments * annotation types
  - orthography, phonetics, word alignment, part-of-speech, …

- 12,780 IMDI session metadata files
- 127 IMDI (sub)corpus metadata files

- Frequency lists, tools, documentation

- About 120 GB

- Version 2.0 contains a lot of (meta)data bugfixes, annotations for 13 Flemish audio files, updated software and documentation

IFA corpus

IFA Corpus of Spoken Dutch

- Open source (GPL) database of hand-segmented Dutch speech

- 8 speakers and 8 speaking styles
- 50,000 words / 5½ hours

- IMDI 3.0 session and (sub)corpus metadata
- Documentation

- 400+ MB

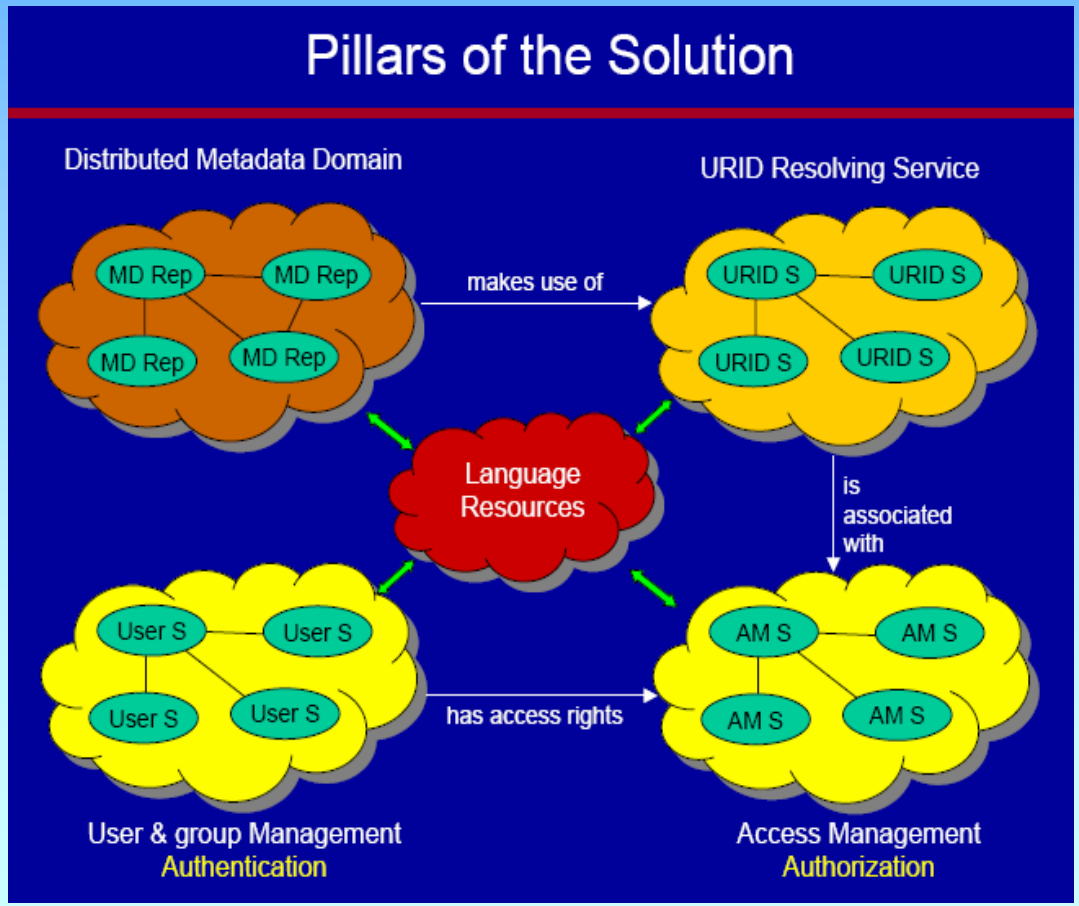
Resources to add in 2007

RBN (reference list of Dutch); RBBN (reference list of Flemish Dutch); OMBI (dictionary editor); several bilingual lexicons; Official Dutch word lists of 1995 and 2005 (lexicons); dictionary of Old-Dutch; E-LEX (large Dutch electronic single word and multi-word lexicon; 5, 27 and 38 million words written text corpora and exploration software; 20 million words Parole written text corpus, lexicon and exploration software; Neologisms list for Dutch; ANW corpus (100 million word text corpus); Nl-En, Nl-Fr, Fr-Nl and En-Nl translation dictionaries, corpora and software; Terminology lexicons and software (TermExtractor; E-ANS (Dutch grammar rules); Regional (dialect) dictionaries; STEVIN D-Coi (pilot for Written Dutch Corpus); STEVIN WDC (500 million word text corpus); STEVIN Jasmin-CGN (extension of CGN); STEVIN COREA (co-reference tools and annotated text corpus); STEVIN IRME (multi-word database and software); STEVIN Autonomata (spoken name database and software); STEVIN Daeso (corpus and software for semantics); STEVIN DPC (Nl-En and Nl-Fr parallel corpora); STEVIN Lassy (syntactically annotated corpus); STEVIN Midas (software for robustness in ASR); STEVIN N-best (benchmark for Dutch ASRs); STEVIN can Praat (software for speech research); STEVIN Spraak (ASR toolkit for Dutch); STEVIN Cornetto (software toolkit for lexical semantics); 14th century written text corpus; Eindhoven writtten text corpus (the first written Dutch text corpus)

Local Solution at the INL

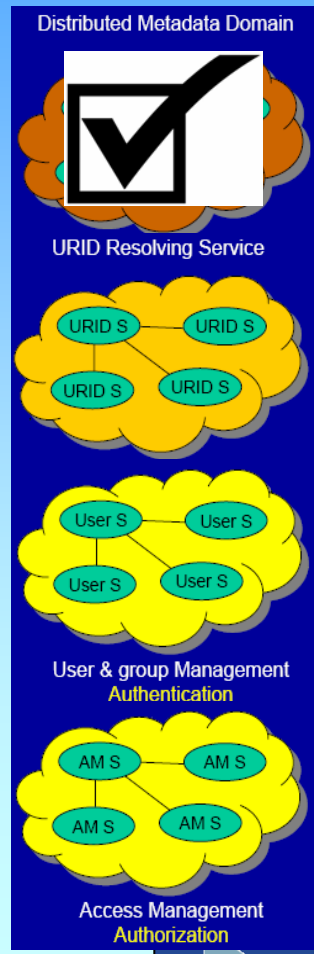
Local Solution at the INL (1/3)

- Finding and implementing solutions for the four access pillars



Local Solution at the INL (2/3)

- 2004/5
  - Training at MPI also included IMDI portal software
- 2005
  - Pilot implementation of IMDI portal
  - Test LR: CGN 1.0
  - Experiences/documentation on project wiki
  - **Solution for distributed metadata domain (IMDI)**



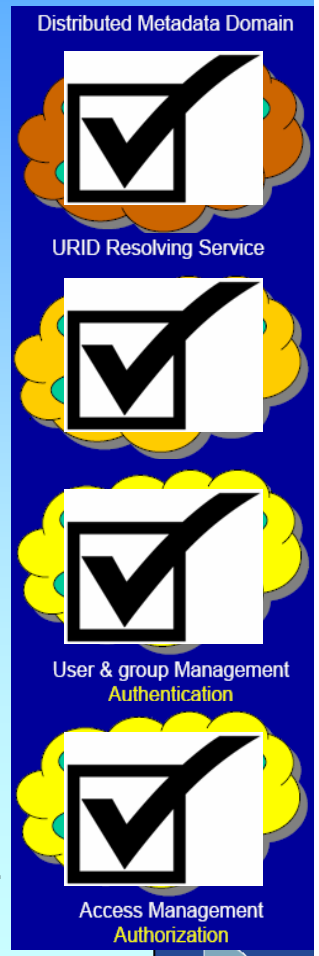
Local Solution at the INL (3/3)

- 2006
  - **Solution for URID service (Handle)**
  - **Solution for authentication (LDAP)**
  - **Solution for authorisation (Shibboleth)**

  - PKI: Leiden University as Certification Authority
  - Portal beta online (http://imdi.inl.nl)
  - CGN 2.0 and IFA corpus as test LRs

  - Solutions for all "pillars" up and running

  - Towards a distributed solution
    - INL portal links to MPI and Lund archives
    - Succesful Shibboleth test between MPI and INL (November 2006)



Summary

- DAM-LR at INL in close co-operation with TST-centrale and IT department
- Archive Formation + Local Solution
  - 75 pm in 2005 (total: 108 pm)
  - 52 pm in 2006 (total: 78 pm)

- Solution for distr. metadata domain: IMDI
- Solution for URID service: Handle (licensed)
- Solution for authentication: (open)LDAP
- Solution for authorisation: Shibboleth

- Two large LRs in beta portal: http://imdi.inl.nl
- Start of distributed solution: links MPI & Lund
  - **Successful Shibboleth test between MPI and INL**

End

Questions?