Kutatási adatok az MTA Nyelvtudományi Intézetében

linginst@nytud.mta.hu

Oravecz Csaba
Nyelvtechnológiai Kutatócsoport
www.nytud.hu

Kutatási adatok ws, 2015. május 14.

Kutatók és adatok

Társadalomtudomány természettudományos módszerekkel

- több mint 100 kutató
- változatos kutatási területek
  (nyelvtörténet [...] $\rightarrow$ fonetika [...] $\rightarrow$ számítógépes nyelvészet)
- változatos adatok
  (kézirat [...] $\rightarrow$ hangfelvétel [...] $\rightarrow$ nemzetközi szabványokhoz illeszkedő XML annotált adatbázis)

Kutatási adatok, metaadatok

- közel 30 féle gyűjtemény, adatbázis
- nyelvi adatok:
  - több százezer kéziratos cédula és gyűjteményes füzetek
  - több száz órányi hang- és videófelvétel
  - több millió lexikális tételt és
  - több milliárd szövegszót tartalmazó adatbázisok
- metaadatok:
  - kéziratos fejléc
  - külön táblázat
  - SQL adatbázis
  - T(ext) E(ncoding) I(nitiative) XML fejléc
  - szabványos leíró séma (ISLE Meta Data Initiative – IMDI)

Gyűjteménykezelés I.

Tárolás

- papírdobozban
- CD-n, saját számítógépen (fájlonként)
- szerveren (megosztással, könyvtárstruktúrában)
- szerveren adatbázisban
- repozitóriumban

Gyűjteménykezelés II.

Formátum

- papírcédula
- doc, xls, txt, xml
- wav, mp3, mp4

Kezelés

- archiválási protokoll nem jellemző, szabályozás nincs

Hozzáférhetőség

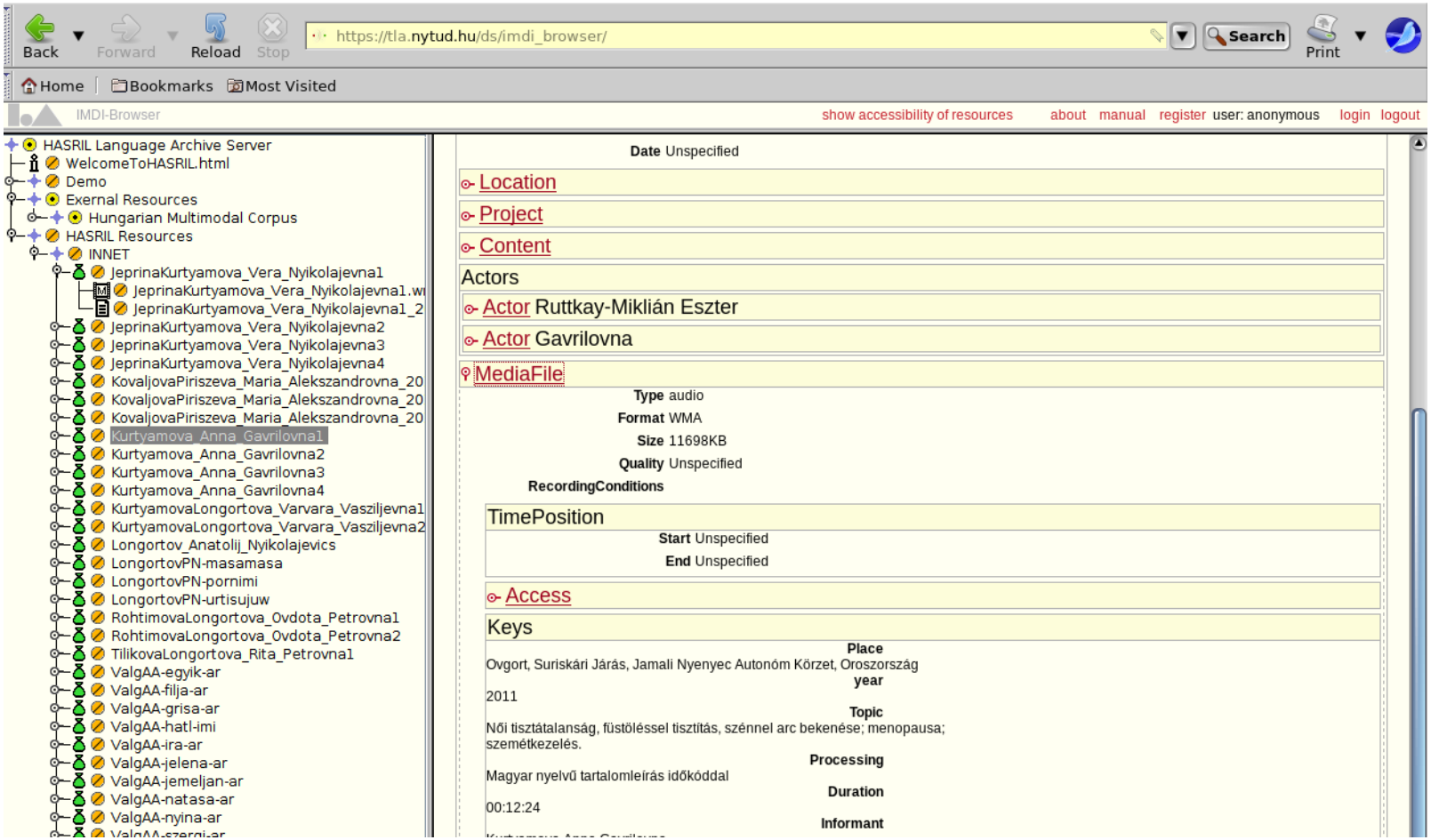
Elérés

- saját felhasználás
- kutatócsoporton belül
- külön megállapodás alapján
- dedikált felhasználói felületen regisztráció után
- szabadon elérhető

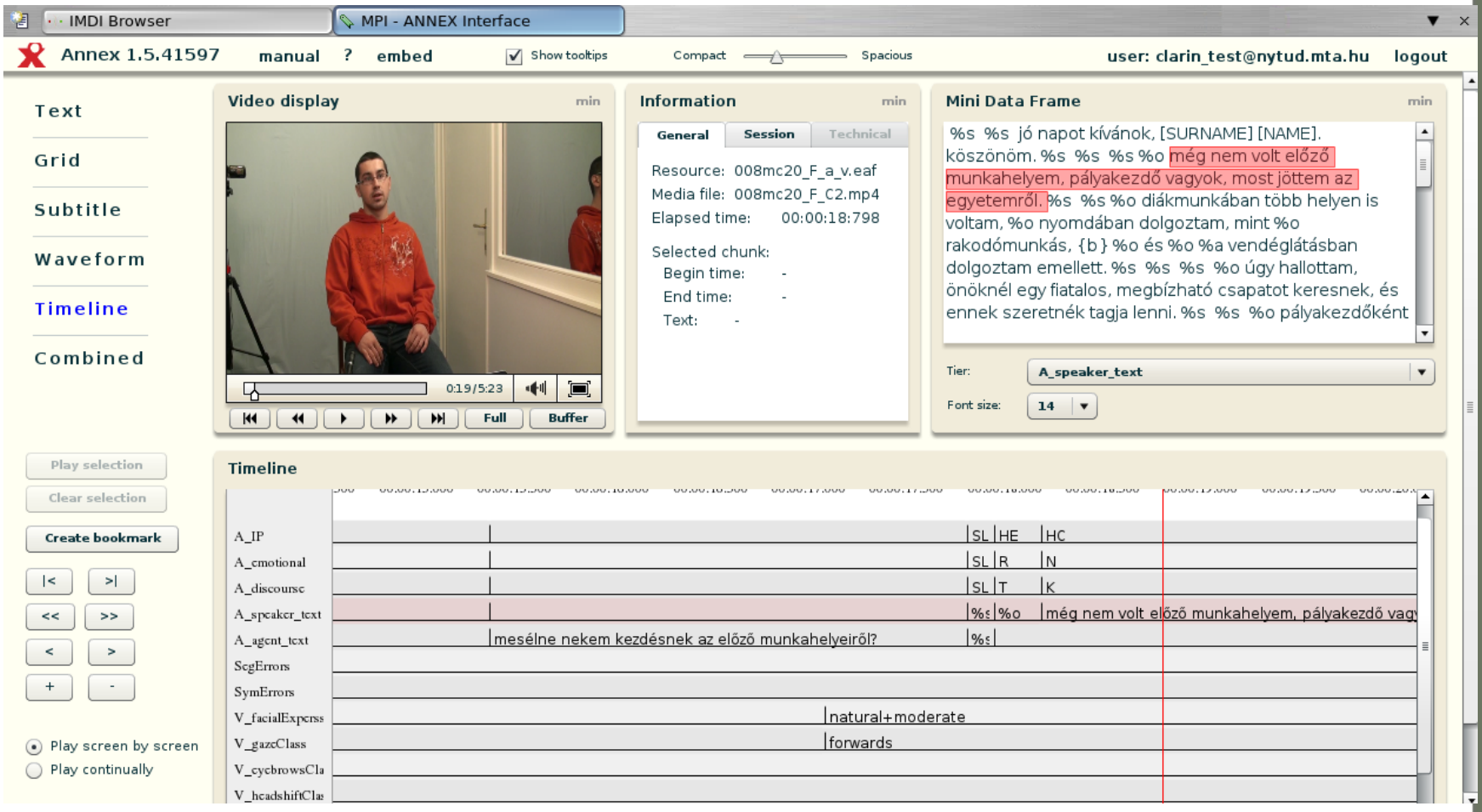
Megosztás

- nincs általános szabályozás
- adatvédelem, kutatói hozzáállás

NYTI archívum



Többrétegű adatbázis



Nyelvtechnológiai Kutatócsoport

Nemzetközi kapcsolatok

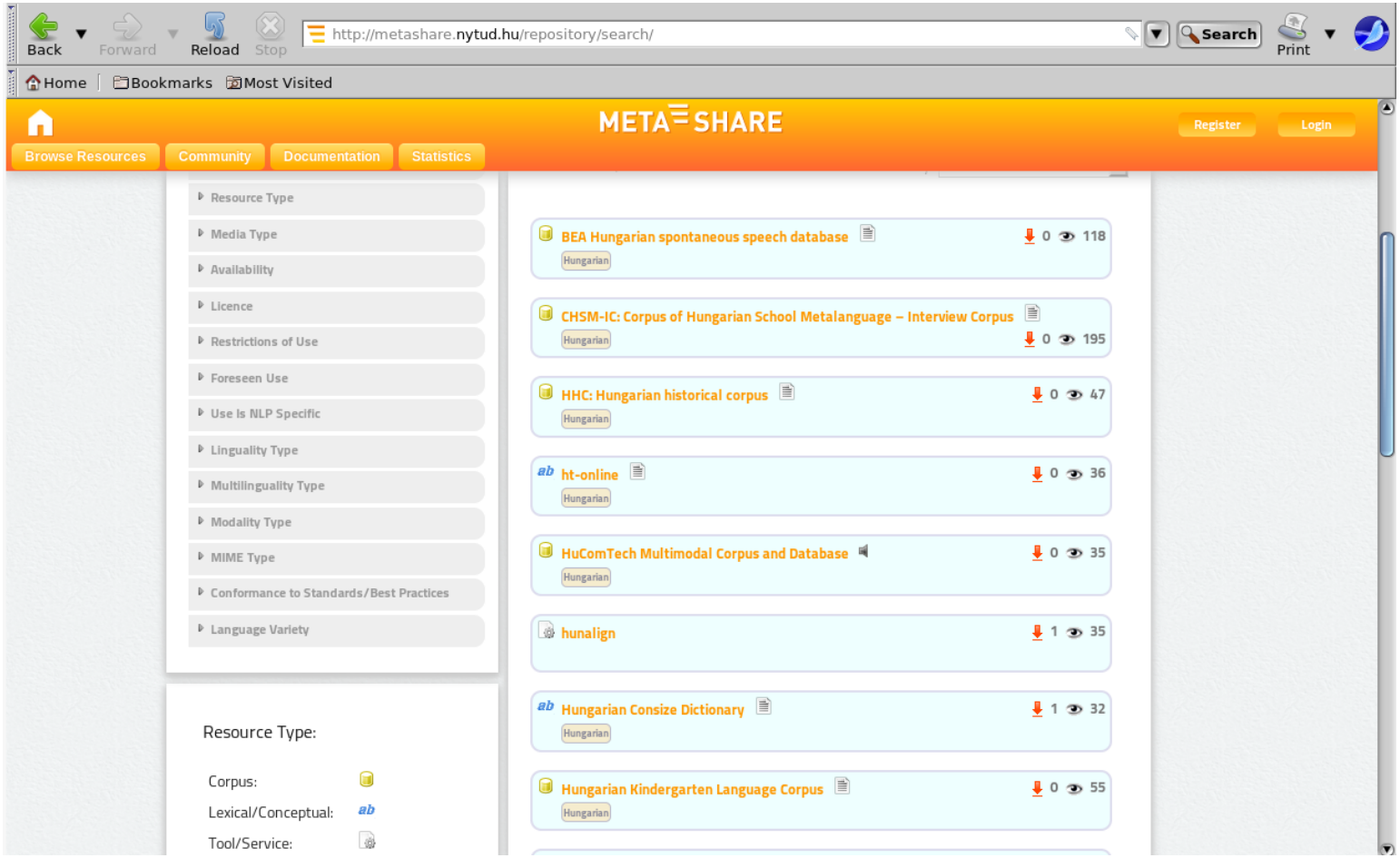
Európai elosztott kutatási infrastruktúrák

- CLARIN (www.clarin.eu)
- META-NET kiválósági hálózat META-SHARE disztribúciós rendszer magyar csomópont

Együttműködés

- Max Planck Institute, Nijmegen

META-SHARE magyar csomópont



Összefoglalás

✘

- nagyon változatos, eltérő kezelést igénylő adatok
- ad-hoc, kialakított szabályozás és infrastruktúra nélküli tárolás és archiválás
- szabványos metaadatolás ritka

✔

- nemzetközi kapcsolatok, részvétel európai hálózatokban
- helyben üzemelő gyűjteménykezelő keretrendszer

Köszönöm a figyelmet