A HuComTech adatbázis mint kutatási infrastruktúra

Váradi Tamás

varadi.tamas@nytud.mta.hu

MTA Nyelvtudományi Intézet

Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály

Háttér

Nyers adat → nyelvi erőforrás

- annotáció
- metaadatok
- szabványok

Erőforrások → infrastruktúra

- CLARIN (2008 -)
- CESAR@META-SHARE (2011 - 2013)
- HUNCLARIN - NEKIFUT (2009 - 2014)

A nyelvi erőforrások fogalma

Nem pusztán nyers adatok halmaza

- reprezentatív minta
- annotáció = hozzáadott érték
  - XML technológia
  - szabványos kódolási útmutató TEI (Text Encoding Initiative)
- metaadatok

A nyelvi infrastruktúra fogalma

Nem pusztán erőforrások halmaza

- központi gondolat: megosztás
- metaadatok és szabványos annotáció
- az erőforrásokmetaadat megfelelő gondozása
  - lehetőleg szabványos és egységes annotáció
  - egyértelmű és átlátható jogi helyzet
- interoperabilitás
- fenntarthatóság

CLARIN

About ▾ Services ▾ Centres ▾ Technology ▾ Events Media ▾ News Contact

- CLARIN Portal (new users start here!)
- Search language resources
- Depositing Services
- Easy access to protected resources
- Virtual Language Observatory
- Web Services

Search

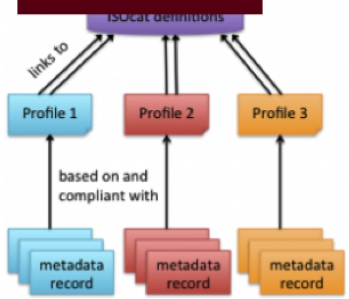
Home › Technology

Technology

- Component Metadata
  - CMDI examples
  - CMDI tutorial September 2012
  - Component Registry Documentation
  - Component Registry REST interface documentation
  - OAI-PMH for CMDI
  - Feedback
- Content Search
- Federated Identity
- Persistent Identifiers
- Repositories
- Specification Documents

C…data

I…nt Metadata



Metadata for language resources and tools exists in a multitude of formats. Often these descriptions contain specialized information for a specific research community (e.g. TEI headers for text, IMDI for multimedia collections).

To overcome this dispersion CLARIN has initiated the Component MetaData Infrastructure (CMDI). It provides a framework to describe and reuse metadata blueprints. Description building blocks ("components", which include field definitions) can be grouped into a ready-made description format (a "profile"). Both are stored and shared with other users in the Component Registry to promote reuse. Each metadata record is then expressed as an XML file, including a link to the profile on which it is based.

The CMDI approach combines architectural freedom when modeling the metadata with powerful exploration and search possibilities over a broad range of language resources.

Further information and examples

- For a detailed explanation, take a look at the Frequently Asked Questions about CMDI
- For a general introduction, read the CMDI section in the CLARIN-D User Guide
- Examples and data sets
- Document about granularity and modelling issues with CMDI (draft)

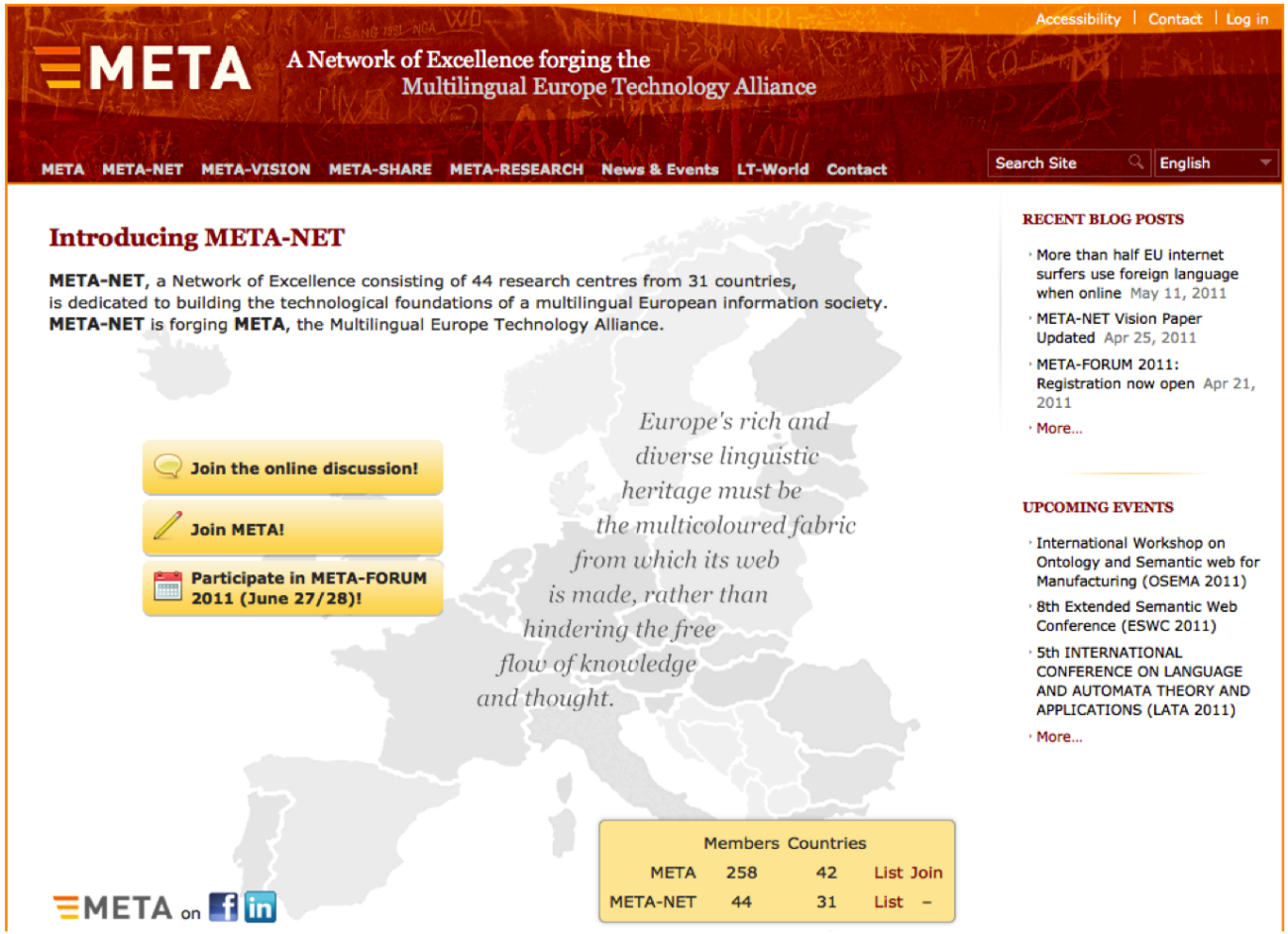
Virtual Language Observatory

NYELVTUDOMÁNYI INTÉZET
MAGYAR TUDOMÁNYOS AKADÉMIA

A MAGYAR TUDOMÁNY ÜNNEPE

META-NET

NYELVTUDOMÁNYI INTÉZET
MAGYAR TUDOMÁNYOS AKADÉMIA



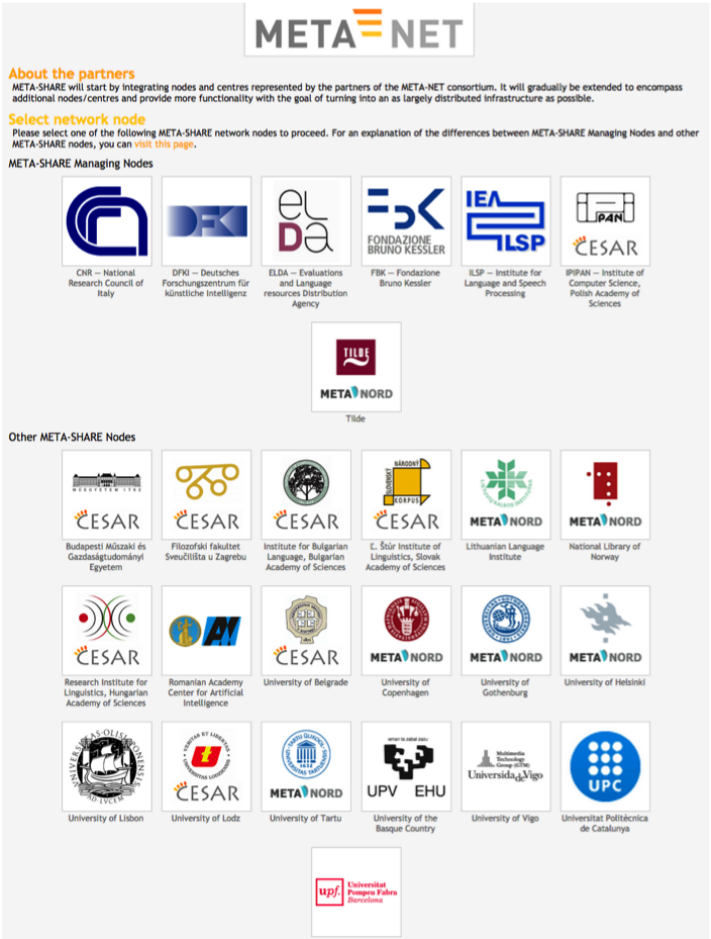
A MAGYAR TUDOMÁNY ÜNNEPE

CESAR

NYELVTUDOMÁNYI INTÉZET
MAGYAR TUDOMÁNYOS AKADÉMIA

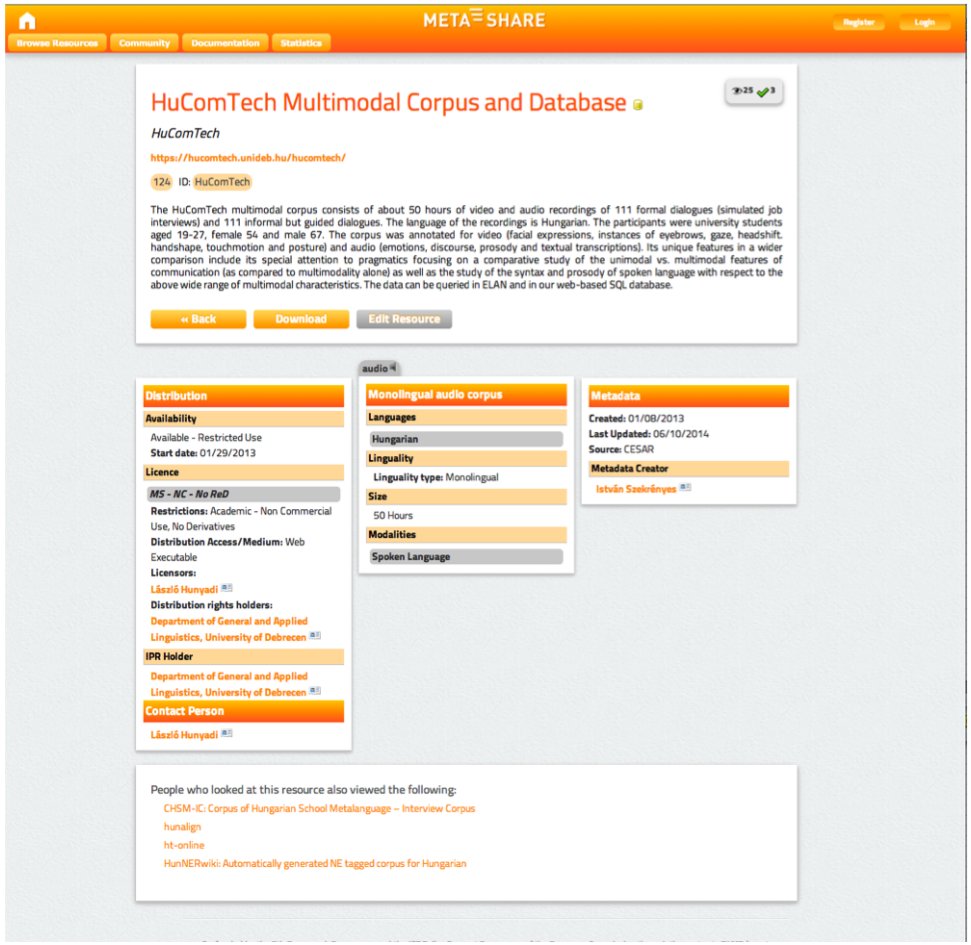
A MAGYAR TUDOMÁNY ÜNNEPE

META-SHARE


META-NET
About the partners
META-SHARE will start by integrating nodes and centres represented by the partners of the META-NET consortium. It will gradually be extended to encompass additional nodes/centres and provide more functionality with the goal of turning into an as largely distributed infrastructure as possible.
Select network node
Please select one of the following META-SHARE network nodes to proceed. For an explanation of the differences between META-SHARE Managing Nodes and other META-SHARE nodes, you can visit this page.
META-SHARE Managing Nodes
CNR – National Research Council of Italy
DFKI – Deutsches Forschungszentrum für künstliche Intelligenz
ELDA – Evaluations and Language resources Distribution Agency
FBK – Fondazione Bruno Kessler
ILSP – Institute for Language and Speech Processing
IPIPAN – Institute of Computer Science, Polish Academy of Sciences
Tilde
Other META-SHARE Nodes
Budapesti Műszaki és Gazdaságtudományi Egyetem
Filozofski fakultet Sveučilišta u Zagrebu
Institute for Bulgarian Language, Bulgarian Academy of Sciences
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
Lithuanian Language Institute
National Library of Norway
Research Institute for Linguistics, Hungarian Academy of Sciences
Romanian Academy Center for Artificial Intelligence
University of Belgrade
University of Copenhagen
University of Gothenburg
University of Helsinki
University of Lisbon
University of Lodz
University of Tartu
University of the Basque Country
University of Vigo
Universitat Politècnica de Catalunya

HUCOMTECH in META-SHARE

NYELVTUDOMÁNYI INTÉZET
MAGYAR TUDOMÁNYOS AKADÉMIA



A MAGYAR TUDOMÁNY ÜNNEPE

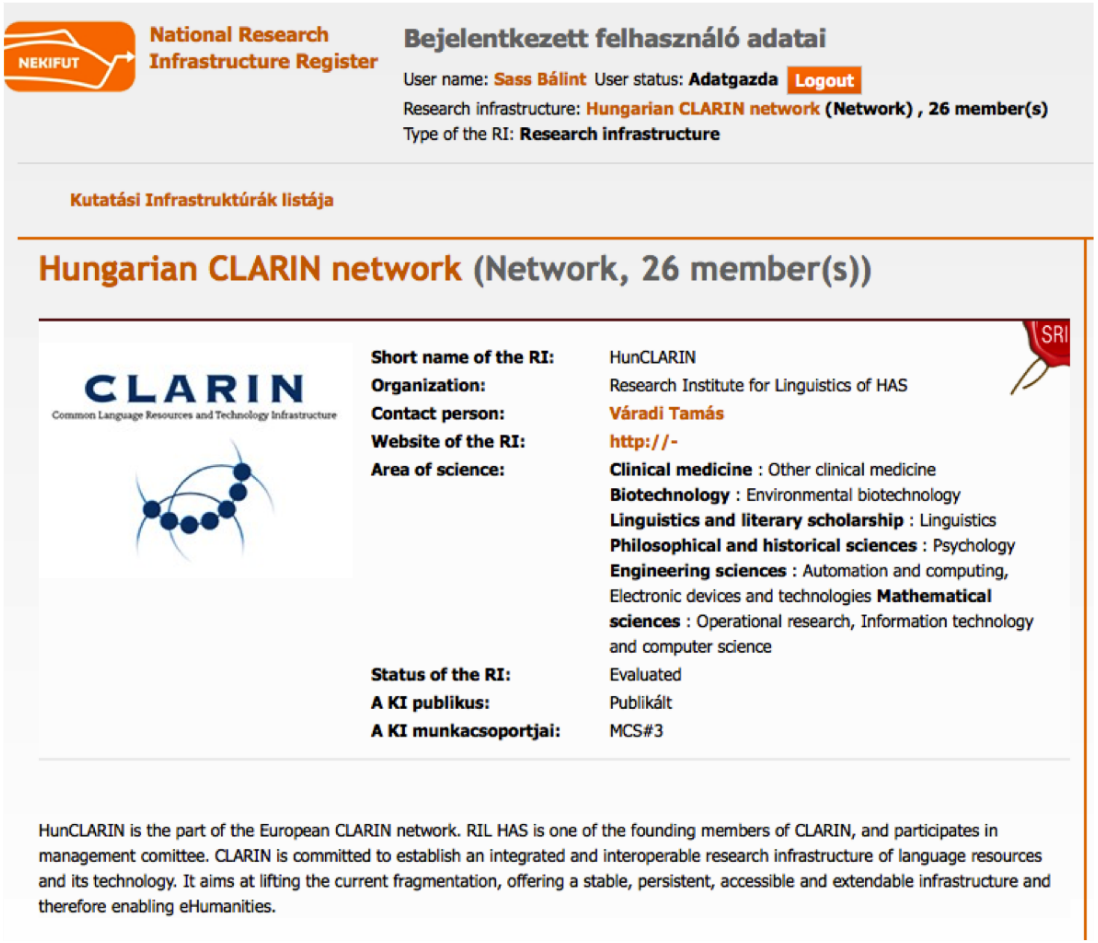
Interoperabilitás

NYELVTUDOMÁNYI INTÉZET
MAGYAR TUDOMÁNYOS AKADÉMIA

A MAGYAR TUDOMÁNY ÜNNEPE

Fenntarthatóság

NYELVTUDOMÁNYI INTÉZET
MAGYAR TUDOMÁNYOS AKADÉMIA



National Research Infrastructure Register

Bejelentkezett felhasználó adatai

User name: Sass Bálint User status: **Adatgazda** Logout
Research infrastructure: Hungarian CLARIN network **(Network) , 26 member(s)**
Type of the RI: **Research infrastructure**

Kutatási Infrastruktúrák listája

Hungarian CLARIN network (Network, 26 member(s))

SRI

CLARIN
Common Language Resources and Technology Infrastructure

Short name of the RI:	HunCLARIN
Organization:	Research Institute for Linguistics of HAS
Contact person:	Váradi Tamás
Website of the RI:	http://-
Area of science:	**Clinical medicine** : Other clinical medicine **Biotechnology** : Environmental biotechnology **Linguistics and literary scholarship** : Linguistics **Philosophical and historical sciences** : Psychology **Engineering sciences** : Automation and computing, Electronic devices and technologies **Mathematical sciences** : Operational research, Information technology and computer science
Status of the RI:	Evaluated
A KI publikus:	Publikált
A KI munkacsoportjai:	MCS#3

HunCLARIN is the part of the European CLARIN network. RIL HAS is one of the founding members of CLARIN, and participates in management comittee. CLARIN is committed to establish an integrated and interoperable research infrastructure of language resources and its technology. It aims at lifting the current fragmentation, offering a stable, persistent, accessible and extendable infrastructure and therefore enabling eHumanities.

A MAGYAR TUDOMÁNY ÜNNEPE

Köszönöm a figyelmet!