LINDAT/CLARIAH-CZ
Large Research Infrastructure
Digital Research Infrastructure for Language Technology, Arts and Humanities

„Digitální výzkumná infrastruktura pro jazykové technologie, umění a humanitní vědy“

Funded by:

Jan Hajíč
LINDAT/CLARIAH-CZ coordinator
Institute of Formal and Applied Linguistics
Computer Science School
Faculty of Mathematics and Physics
Charles University, Prague, Czechia
Christopher Cieri 1963-2023

Chris Cieri was a native Philadelphian who grew up among the diverse population in the neighborhoods of South Philadelphia. He had a lifelong affinity for the Italian language, and it was no surprise that when he entered the University of Pennsylvania as an undergraduate, he chose Linguistics for his major. In fact, he concluded his Masters’ degree concurrently with his Bachelor studies in 1985-1986; his thesis topic was *Italian Lexical Items in the English Speech of Italo-Americans*. Under the direction of Bill Labov, he received his PhD at Penn as well. His dissertation, *Modeling Phonological Variation in Multidialectal Italy*, was based on fieldwork in L’Aquila, Italy, of which he had fond memories and spoke about often.

Linguistic Data Consortium, director

Long-term member of LINDAT IAB
In memory of the many victims.

@Faculty of Arts, Dec. 21, 2023
LINDAT/CLARIAH-CZ

- **LINDAT = CLARIN & DARIAH & EHRI** in Czechia
  - European networks for **supporting research (in SSH)**
    - CLARIN: Language resources and technology
    - DARIAH: Digital humanities and arts
    - EHRI: European Holocaust Research Infrastructure
  - **LINDAT/CLARIAH-CZ: 2023-2026**
    - Combines membership in EU CLARIN, DARIAH and **EHRI** networks
    - 15 institutions in the Czech combined network (was: 4 in 2010, 11 in 2022)
    - UK, MU, Academy of Sciences, National and Moravian Libraries, NG, NFA
      - New in 2023:
        - National Archives
        - Masaryk Institute and Archives CAS
        - Terezín Memorial
        - Terezín Initiative Institute
        - complements the already included Center for Visual History Malach
LINDAT/CLARIAH-CZ

- LINDAT = CLARIN & DARIAH & EHRI in Czechia
  - European networks for research (in SSH)
    - CLARIN: Language resources and technology
    - DARIAH: Digital humanities and arts
    - EHRI: European Holocaust Research Infrastructure
  - LINDAT/CLARIAH-CZ: 2023-2026
    - Combines membership in EU CLARIN, DARIAH and EHRI networks
    - 15 institutions in the Czech combined network (was: 4 in 2010, 11 in 2022)
    - UK, MU, Academy of Sciences, National and Moravian Libraries, NG, NFA
      - New in 2023:
        » National Archives
        » Masaryk Institute and Archives CAS
        » Terezín Memorial
        » ITI
        » Complements the already included Center for Visual History Malach

EHRI Czech National Node

Top-quality research on the Holocaust is a prerequisite for informed discussion about Czech, European and world modern history and for understanding the risks and mechanisms of racism and genocide in their various forms. The European Holocaust Research Infrastructure (EHRI) connects collections and sources divided by borders and languages, promotes digital methods and supports researchers. The Czech EHRI national node is a gateway to EHRI services and community and a signpost for Holocaust research in the Czech Republic.

Research infrastructure

EHRI is funded by the European Commission under FP7, Horizon 2020 and Horizon Europe. Since 2014, EHRI has been a part of the research infrastructure LINDAT/CLARIAH-CZ, supported by the Ministry of Education, Youth and Sports since 2013.

Main goals and services

EHRI CZ provides the following types of services:

- Makes EHRI data and services available, including EHRI Portal, EHRI Document Blog, EHRI Editions, EHRI Geospatial Repository, Comenius Scholarship, and others.
- Creates data on sources on the history of the Holocaust in the Czech lands and uploads them to the EHRI Portal.
- Links and enhances the victim databases and methodologically supports their further development.
- Applies digital methods to the digitised sources, including automatic text and speech recognition, identification of places, historical actors, keywords, etc.
- Supports research using spatial methods and interactive maps, including the MemosMAP application.
LINDAT/CLARIAH-CZ
2010-2026

• LINDAT timeline + Czechia in ERICs
LINDAT/CLARIAH-CZ in European Research Infrastructures

- **ERIC = European Research Infrastructure Consortium**
  - Supervising body
    - ESFRI (European Strategy Forum on Research Infrastructures)
  - Members are countries (EU member or associated states)
    - ... or international organizations (NTU); associated countries possible (U.S.)

- **CLARIN ERIC**
  - Original network (since 2010)
  - Consortium members in
    - NCF, SCCTC, CLIC, Annual Conference committee, award committees

- **DARIAH ERIC**
  - Since 2014
  - Consortium members in
    - NCC, legal VCC

- **EHRI network (soon to be ERIC, January 2025)**
  - EHRI: 3 preparatory projects by EU
  - EHRI CZ: partner in all 3 preparatory projects
LINDAT/CLARIAH-CZ

- **Data, Tools, Services, Expertise**
  - Language Resources Data Repository
  - Software tools repository
  - Web services with both API and User Interactive Interface
- **Computing infrastructure**
  - New hardware for deep learning computations due in 2024-2026
    - Coming as we speak – for 7 partner institutions
- **Cooperation with other Czech RI from SSH**
  - Czech National Corpus (CNC)
  - Czech Literary Bibliography (CLB)
  - Archaeological Information System of the Czech Republic (AIS CR)
LINDAT/CLARIAH-CZ in 2023

- Web at https://lindat.cz
LINDAT/CLARIAH-CZ in 2023 (1)

• Repository development (CLARIN DSpace)
  – Upgrade to version 7.x of DSpace – major undertaking, w/external partners
  – Will be used by Charles University Library for Open Data Library

• Universal Dependencies project
  – General support, official editions (every 6 months), now at v2.13 (Nov. 2023)
  – Tools: UDPipe (now UDPipe 2 -> LinPipe soon)

• Uniform Meaning Representation project
  – With partners in the U.S. (Brandeis, Colorado, New Mexico)
  – Official release repository for UMR 1.0
LINDAT/CLARIAH-CZ in 2023 (2)

- Prague Dependency Treebank – Consolidated ed. (PDT-C 1.0 -> 2.0)
  - Approx. 4MW, manual dependency syntax for all parts underway
    - PDT, PCEDT, PDTSC, Faust; TR annotation same, a-layer: full manual, corrections
- Prague Discourse Treebank – on top of PDT
  - Adding cross-sentence annotation
- SynSemClass
  - Multilingual (cs, en, de, es, [kr in the works]) event type ontology, now v5.0
    - A.k.a synonym dictionary of verbs, with valency, linked to corpora and other lexical resources
    - New search tool in 2023
- TEI:TOK
  - Platform for rich annotation and multimedia content
  - Many new datasets, continued ParlaMint project (led by CLARIN-SI)
- New ASR service (UWB Pilsen)
- Consolidation of MT interface underway
LINDAT/CLARIAH-CZ in 2023 (2)

- Prague Dependency Treebank – Consolidated ed. (PDT-C 1.0 -> 2.0)
  - Approx. 4MW, manual dependency syntax for all parts underway
- PDT, PCEDT, PDTSC, Faust; TR annotation same, a-layer: full manual, corrections
- Prague Discourse Treebank – on top of PDT
  - Adding cross-sentence annotation
- SynSemClass
  - Multilingual (cs, en, de, es, [kr in the works]) event type ontology, now v5.0
  - A.k.a synonym dictionary of verbs, with valency, linked to corpora and other lexical resources
- New search tool in 2023
- TEI:TOK
  - Platform for rich annotation and multimedia content
  - Many new datasets, continued Parliament project (led by CLARIN-SI)
- New ASR service (UWB Pilsen)
- Consolidation of MT interface underway

January 5, 2024 LINDAT/CLARIAH-CZ International Advisory Board Meeting
• New ASR service (UWB Pilsen)
• Consolidation of MT interface underway
LINDAT/CLARIAH-CZ services

• Running services (https://lindat.cz/en/services)

• Each service available as
  – Software to download – for those able to run it locally, modify etc (Open Source)
  – Models to download – e.g. for machine translation, for locally run software
  – API, for running programmatically but remotely
  – User interface (running on small samples, cut & paste, etc.) – open to everybody

• Examples
  – UDPipe 2.0
    • Linguistic analysis: segmentation, morphology, lemmatization, POS, parsing, NER
    • 100+ languages
  – Machine translation
    • en-cs and en-fr at human level of accuracy
    • Includes uk-cs
    • Also on Android (Charles Translator)
  – Dictionaries
    • Valency, semantics
LINDAT/CLARIAH-CZ services

• Running services (https://lindat.cz/en/services)
• Each service available as
  – Software to download – for those able to run it locally, modify etc (Open Source)
  – Models to download – e.g. for machine translation, for locally run software
  – API, for running programmatically but remotely – User interface (running on small samples, cut & paste, etc.) – open to everybody

Examples

– UDPipe 2.0
  • Linguistic analysis: segmentation, morphology, lemmatization, POS, parsing, NER
  • 100+ languages

– Machine translation
  • en-cs and en-fr at human level of accuracy
  • Includes uk-cs
  • Also on Android (Charles Translator)

– Dictionaries
  • Valency, semantics

January 5, 2024 LINDAT/CLARIAH-CZ International Advisory Board Meeting
LINDAT/CLARIAH-CZ services

- Running services (https://lindat.cz/en/services)
- Each service available as:
  - Software to download – for those able to run it locally, modify etc. (Open Source)
  - Models to download – e.g. for machine translation, for locally run software
  - API, for running programmatically but remotely – User interface (running on small samples, cut & paste, etc.) – open to everybody

Examples
- UDPipe 2.0
  - Linguistic analysis: segmentation, morphology, lemmatization, POS, parsing, NER
  - 100+ languages
- Machine translation
  - en-cs and en-fr at human level of accuracy
  - Includes uk-cs
  - Also on Android (Charles Translator)
- Dictionaries
  - Valency, semantics

January 5, 2024 LINDAT/CLARIAH-CZ International Advisory Board Meeting
LINDAT/CLARIAH-CZ services

- Running services (https://lindat.cz/en/services)
  - Software to download – for those able to run it locally, modify etc (Open Source)
  - Models to download – e.g. for machine translation, for locally run software
  - API, for running programmatically but remotely – User interface (running on small samples, cut & paste, etc.) – open to everybody

Example

- UDPipe 2.0
  - Linguistic analysis: segmentation, morphology, lemmatization, POS, parsing, NER
  - 100+ languages

- Machine translation
  - en-cs and en-fr at human level of accuracy
  - Includes uk-cs
  - Also on Android (Charles Translator)

- Dictionaries
  - Valency, semantics

January 5, 2024 LINDAT/CLARIAH-CZ International Advisory Board Meeting
LINDAT/CLARIAH-CZ partner services (1)

- **Institute of the Czech Language, Czech Academy of Sciences (CAS)**
  - Internet Language Reference Book (“Internetová jazyková příručka”) – 40 mil. acc/yr
  - In cooperation with Masaryk Univ., Faculty of Informatics

- **Faculty of Arts, Charles University (FF UK)**
  - LINDAT team participates at building the Center for Digital Humanities at FF UK
  - Corpora (diachronic, educational), Exhibitions (manuscripts)
  - Didaktikon project – educational project with National Library, other universities

- **Masaryk University Brno – Faculty of Arts**

- **University of West Bohemia, Faculty of Applied Sciences**
  - Technology – speech recognition services

- **Libraries (National Library, Moravian Library, Library of CAS)**
  - Upgrade to Kramerius 7 (underlying library management system)
  - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)
• Institute of the Czech Language
  – Internet Language Reference Book ("Internetová jazyková příručka") – 40 mil. acc/yr
  – In cooperation with Masaryk University, Faculty of Informatics

• Faculty of Arts, Charles University (FF UK)
  – LINDAT team participates in building the Center for Digital Humanities at FF UK
  – Corpora (diachronic, educational), Exhibitions (manuscripts)
  – Didaktikon project – educational project with National Library, other universities

• Masaryk University Brno – Faculty of Arts
  – Own digital repository at https://digitalia.phil.muni.cz/en (Digitalia MUNI ARTS | PHIL)

• University of West Bohemia, Faculty of Applied Sciences
  – Technology – speech recognition services

• Libraries (National Library, Moravian Library, Library of CAS)
  – Upgrade to Kramerius 7 (underlying library management system)
  – Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)

---

January 5, 2024 LINDAT/CLARIAH-CZ International Advisory Board Meeting
LINDAT/CLARIAH-CZ partner services (1)

• Institute of the Czech Language
  – Internet Language Reference Book ("Internetová jazyková příručka") – 40 mil. acc/yr
  – In cooperation with Masaryk Univ., Faculty of Informatics

• Faculty of Arts, Charles University (FF UK)
  – LINDAT team participates at building the Center for Digital Humanities at FF UK
  – Corpora (diachronic, educational), Exhibitions (manuscripts)
  – Didaktikon project – educational project with National Library, other universities

• Masaryk University Brno – Faculty of Arts
  – Own digital repository at https://digitalia.phil.muni.cz/en (Digitalia MUNI ARTS | PHIL)

• University of West Bohemia, Faculty of Applied Sciences
  – Technology – speech recognition services

• Libraries (National Library, Moravian Library, Library of CAS)
  – Upgrade to Kramerius 7 (underlying library management system)
  – Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)
• Institute of the Czech Language
  – Internet Language Reference Book (“Internetová jazyková příručka”) – 40 mil. acc/yr
  – In cooperation with Masaryk University, Faculty of Informatics
• Faculty of Arts, Charles University (FF UK)
  – LINDAT team participates at building the Center for Digital Humanities at FF UK
  – Corpora (diachronic, educational), Exhibitions (manuscripts)
  – Didaktikon project – educational project with National Library, other universities
• Masaryk University Brno – Faculty of Arts
  – Own digital repository at https://digitalia.phil.muni.cz/en (Digitalia MUNI ARTS | PHIL)
• University of West Bohemia, Faculty of Applied Sciences
  – Technology – speech recognition services
• Libraries (National Library, Moravian Library, Library of CAS)
  – Upgrade to Kramerius 7 (underlying library management system)
  – Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)
LINDAT/CLARIAH-CZ partner services (1)

- Institute of the Czech Language
  - Internet Language Reference Book ("Internetová jazyková příručka") – 40 mil. acc/yr
  - In cooperation with Masaryk Univ., Faculty of Informatics

- Faculty of Arts, Charles University (FF UK)
  - LINDAT team participates at building the Center for Digital Humanities at FF UK
  - Corpora (diachronic, educational), Exhibitions (manuscripts)
  - Didaktikon project – educational project with National Library, other universities

- Masaryk University Brno – Faculty of Arts
  - Own digital repository at https://digitalia.phil.muni.cz/en (Digitalia MUNI ARTS | PHIL)

- University of West Bohemia, Faculty of Applied Sciences
  - Technology – speech recognition services

- Libraries (National Library, Moravian Library, Library of CAS)
  - Upgrade to Kramerius 7 (underlying library management system)
  - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)
• Libraries (National Library, Moravian Library, Library of CAS)
  – Upgrade to Kramerius 7 (underlying library management system)
  – Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)
Involved platforms

We present a list of individual platforms involved in the Digitalia MUNI ARTS infrastructure. By joining the infrastructure, long-term protection and quality care of the contained data is ensured. Links to other faculty platforms can be found in a separate catalogue.

Digital Library of Arne Novák
The aim of the project is to make the extensive work of Arne Novák accessible. The collection includes digitized materials from the Central Library of the FF MU.

Projectiles
The factographic database provides access to archaeological data and visualisations of Early Bronze Age arrowheads from Moravian and Slovak sites.

Cinematic Brno
The database contains information about individual cinemas, their programmes and films shown in Brno in 1918–1945.

Digital Library of the FF MU
The digital library contains freely available full texts of journals, monographs and proceedings published at the Faculty of Arts of Masaryk University.

Beauty Patterns
The image database offers several hundred samples of liturgical vestments and other ceremonial fabrics, mostly from the 18th century.
LINDAT/CLARIAH-CZ partner services (1)

- Institute of the Czech Language
  - Internet Language Reference Book ("Internetová jazyková příručka") – 40 mil. acc/yr
  - In cooperation with Masaryk Univ., Faculty of Informatics

- Faculty of Arts, Charles University (FF UK)
  - LINDAT team participates at building the Center for Digital Humanities at FF UK
  - Corpora (diachronic, educational), Exhibitions (manuscripts)
  - Didaktikon project – educational project with National Library, other universities

- Masaryk University Brno – Faculty of Arts
  - Own digital repository at https://digitalia.phil.muni.cz/en (Digitalia MUNI ARTS | PHIL)

- University of West Bohemia, Faculty of Applied Sciences
  - Technology – speech recognition services

- Libraries (National Library, Moravian Library, Library of CAS)
  - Upgrade to Kramerius 7 (underlying library management system)
  - Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)
UWebASR HTTP API

UWebASR HTTP API is a simple interface to speech recognition engine. The input data can be passed directly within the HTTP request (POST method) or as a link to a file in the form of a URL (GET method). Live audio stream recognition from a given URL is supported, as well. The output format includes plain text, machine-readable XML and JSON formats, and the WebVTT format for web captions. Recognition results (except TRS format) are streamed continuously.

For the UWebASR use the following values of variables:

- **app_id** - use one of the following values to use the models provided by the UWebASR service:
  - malach/en - English Wav2Vec 2.0 model
  - malach/de - German Wav2Vec 2.0 model
  - malach/cs - Czech Wav2Vec 2.0 model
  - malach/sk - Slovak Wav2Vec 2.0 model

In other words, the full HTTP API endpoint for Czech is https://uwebasr.zcu.cz/api/v2/lindat/malach/cs.

The UWebASR HTTP API uses the underlying SpeechCloud platform. The platform has an architecture employing a set of real-time workers. Therefore it is possible that all of the workers are allocated and no further requests could be processed at the moment. This situation is indicated by the 503 HTTP status.

Each session employing the worker is limited to at most 3600 seconds. Longer sessions are automatically terminated. Since the speech recognizer runs faster than real-time in most situations, it means that the maximum length of the audio processed in the API is actually longer than 3600 seconds.
LINDAT/CLARIAH-CZ partner services (1)

• Institute of the Czech Language
  – Internet Language Reference Book (“Internetová jazyková příručka”) – 40 mil. acc/yr
  – In cooperation with Masaryk Univ., Faculty of Informatics

• Faculty of Arts, Charles University (FF UK)
  – LINDAT team participates at building the Center for Digital Humanities at FF UK
  – Corpora (diachronic, educational), Exhibitions (manuscripts)
  – Didaktikon project – educational project with National Library, other universities

• Masaryk University Brno – Faculty of Arts
  – Own digital repository at https://digitalia.phil.muni.cz/en (Digitalia MUNI ARTS | PHIL)

• University of West Bohemia, Faculty of Applied Sciences
  – Technology – speech recognition services

• Libraries (National Library, Moravian Library, Library of CAS)
  – Upgrade to Kramerius 7 (underlying library management system)
  – Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)
• Institute of the Czech Language
  – Internet Language Reference Book ("Internetová jazyková příručka") – 40 mil. acc/yr
  – In cooperation with Masaryk University, Faculty of Informatics

• Faculty of Arts, Charles University (FF UK)
  – LINDAT team participates at building the Center for Digital Humanities at FF UK
  – Corpora (diachronic, educational), Exhibitions (manuscripts)
  – Didaktikon project – educational project with National Library, other universities

• Masaryk University Brno – Faculty of Arts
  – Own digital repository at https://digitalia.phil.muni.cz/en (Digitalia MUNI ARTS | PHIL)

• University of West Bohemia, Faculty of Applied Sciences
  – Technology – speech recognition services

• Libraries (National Library, Moravian Library, Library of CAS)
  – Upgrade to Kramerius 7 (underlying library management system)
  – Czech Digital Library in all 3 institutions (137 mil. pages, 300,000+ books)
LINDAT/CLARIAH-CZ
partner services (2)

• Institute of Philosophy CAS
  – MEMORI project – early middle ages, Czech Medieval Sources Online
  – Jan Patočka Archive – Czech philosopher, dissident 1968-89

• Institute of History CAS
  – Czech historical bibliography

• National Gallery
  – Complete overhaul of digital collection hardware and software environment

• National Film Archive
  – Uses LINDAT repository for all open film documents (historical so far)
  – Direct play from repository

• EHRI group
  – Starting to integrate with LINDAT and Center for Visual History Malach
  – Documentation of documents digitized earlier (pre-LINDAT) (NA, ITI, PT)
  – EHRI CZ presentation and organization
  – Preparation of CLARIN-EHRI workshop, Prague, March 27-28, 2024
• Institute of Philosophy CAS
  – MEMORI project – early middle ages, Czech Medieval Sources Online
  – Jan Patočka Archive – Czech philosopher, dissident 1968-89

• Institute of History CAS
  – Czech historical bibliography

• National Gallery
  – Complete overhaul of digital collection hardware and software environment

• National Film Archive
  – Uses LINDAT repository for all open film documents (historical so far)
  – Direct play from repository

• EHRI group
  – Starting to integrate with LINDAT and CVHM
  – Documentation of documents digitized earlier (pre-LINDAT) (NA, ITI, PT)
  – EHRI CZ presentation and organization
  – Preparation of CLARIN-EHRI workshop, Prague, March 27-28, 2024

January 5, 2024
LINDAT/CLARIAH-CZ

- Institute of Philosophy CAS
  - MEMORI project – early middle ages, Czech Medieval Sources Online
  - Jan Patočka Archive – Czech philosopher, dissident 1968-89
- Institute of History CAS
  - Czech historical bibliography
- National Gallery
  - Complete overhaul of digital collection hardware and software environment
- National Film Archive
  - Uses LINDAT repository for all open film documents (historical so far)
  - Direct play from repository
- EHRI group
  - Starting to integrate with LINDAT and CVHM
  - Documentation of documents digitized earlier (pre-LINDAT) (NA, ITI, PT)
  - EHRI CZ presentation and organization
  - Preparation of CLARIN-EHRI workshop, Prague, March 27-28, 2024

January 5, 2024 LINDAT/CLARIAH-CZ International Advisory Board Meeting
• Institute of Philosophy CAS
  – MEMORI project – early middle ages, Czech Medieval Sources Online
  – Jan Patočka Archive – Czech philosopher, dissident 1968-89

• Institute of History CAS
  – Czech historical bibliography

• National Gallery
  – Complete overhaul of digital collection hardware and software environment

• National Film Archive
  – Uses LINDAT repository for all open film documents (historical so far)
  – Direct play from repository

• EHRI group
  – Starting to integrate with LINDAT and CVHM
  – Documentation of documents digitized earlier (pre-LINDAT) (NA, ITI, PT)
  – EHRI CZ presentation and organization
  – Preparation of CLARIN-EHRI workshop, Prague, March 27-28, 2024
Internal & User Publications in 2023

- Research Infrastructure made for (research) **users**
- LINDAT/CLARIAH-CZ
  - Distributed, virtual, open
- RI-Internal publications
  - About building the RI
    - Hardware, repository software, organizational matters, legal, networking (ELE)
  - Creating datasets, software, services
    - Papers typically at LREC, ACL, …
- User publications (NB: hard to find…) – using Zotero to collect
- 2023: [incomplete yet]
  - RI-Internal publications: 47
  - User / LINDAT partners: 339
  - User / rest of the world: **725**
Educational support

- CLARIN K-Centre for treebanking (with CLARINO Norway)
- Didaktikon (UK & City of Prague, @Campus Hybernská)
  - Educational activities for primary and secondary school level
- Courses
  - Regular (FF UK, FF MU)
    - E.g., Digital Humanities (fall 2023) at FF MU,
  - Continued education (FF UK)
  - Specialized open seminars
    - CVHM MFF UK (Holocaust topics) for teachers
  - Support for teaching in general (tools, data, services/UI)
- Students
  - Employed as data curators, annotators (universities)
  - Technological support (HW cluster, software, data)
Educational support

• Didaktikon (UK & City of Prague, @Campus Hybernská)
  – Educational activities for primary and secondary school level
  – Courses
    – Regular (FF UK, FF MU)
    – E.g., Digital Humanities (fall 2023) at FF MU,
    – Continued education (FF UK)
    – Specialized open seminars

• CVHM MFF UK (Holocaust topics) for teachers
  – Support for teaching in general (tools, data, services/UI)

• Students
  – Employed as data curators, annotators (universities)
  – Technological support (HW cluster, software, data)
Educational support

- Didaktikon (UK & City of Prague, @Campus Hybernská)
  - Educational activities for primary and secondary school level

- Courses
  - Regular (FF UK, FF MU)
    - E.g., Digital Humanities (fall 2023) at FF MU
  - Continued education (FF UK)
  - Specialized open seminars
    - CVHM MFF UK (Holocaust topics) for teachers
  - Support for teaching in general (tools, data, services/UI)

- Students
  - Employed as data curators, annotators (universities)
  - Technological support (HW cluster, software, data)
Educational support

- Didaktikon (UK & City of Prague, @Campus Hybernská)
  - Educational activities for primary and secondary school level

- Courses
  - Regular (FF UK, FF MU)
    - E.g., Digital Humanities (fall 2023) at FF MU,
  - Continued education (FF UK)
  - Specialized open seminars

- CVHM MFF UK (Holocaust topics) for teachers
  - Support for teaching in general (tools, data, services/UI)

- Students
  - Employed as data curators, annotators (universities)
  - Technological support (HW cluster, software, data)

January 5, 2024 LINDAT/CLARIAH-CZ International Advisory Board Meeting
Specific datasets: Shared Tasks Support

- Supporting Shared Tasks
  - by data, manpower, participation
  - “Shared Task” = worldwide open competition (~kaggle etc.)
    - on a defined task
    - data and metrics provided by organizers
- CoNLL Shared Task on dependency parsing (2017, 2018)
  - Using the Universal Dependencies collection, additional data
- CoNLL Shared Task on Meaning Representation Parsing (MRP)
  - Run in 2019 and 2020, Stephan Oepen & colleagues, co-organization
- CRAC Shared Task on Multilingual Coreference Resolution
  - 2022, 2023 co-organized, datasets in repository (EMNLP 2023)
- WMT Shared Tasks and Subtasks
  - Data/repository support for WMT
Specific datasets:

Shared Tasks Support

- Supporting Shared Tasks
  - by data, manpower, participation
  - "Shared Task" = worldwide open competition (~Kaggle etc.)
- on a defined task
- data and metrics provided by organizers

CoNLL Shared Task on dependency parsing (2017, 2018)
- Using the Universal Dependencies collection, additional data

CoNLL Shared Task on Meaning Representation Parsing (MRP)
- Run in 2019 and 2020, Stephan Oepen & colleagues, co-organization

CRAC Shared Task on Multilingual Coreference Resolution

CRAC 2023 Shared Task on Multilingual Coreference Resolution

Overview
Coreference resolution is the task of clustering together multiple mentions of the same entity appearing in a textual document (e.g., Joe Biden, the U.S. President and he). This CoLab-powered shared task deals with multilingual coreference resolution and is associated with the CRAC 2023 Workshop (the Sixth Workshop on Computational Models of Reference, Anaphora and Coreference) held at EMNLP 2023.

Official Results
The following table shows four versions of the CoNLL metric macro-averaged over all datasets:

- head-match excluding singletons (the primary metric, see below),
- partial-match excluding singletons,
- exact-match excluding singletons and
- head-match with singletons.

A more detailed evaluation will be provided in the shared task overview paper.

<table>
<thead>
<tr>
<th>system</th>
<th>head-match</th>
<th>partial-match</th>
<th>exact-match with singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>CorPipe</td>
<td>74.90</td>
<td>71.46</td>
<td>76.82</td>
</tr>
<tr>
<td>Anonymous</td>
<td>70.41</td>
<td>67.09</td>
<td>73.20</td>
</tr>
<tr>
<td>Ondta</td>
<td>69.19</td>
<td>63.01</td>
<td>68.37</td>
</tr>
<tr>
<td>McGill</td>
<td>65.43</td>
<td>63.13</td>
<td>68.23</td>
</tr>
<tr>
<td>DeepBlueAI</td>
<td>62.29</td>
<td>59.95</td>
<td>54.51</td>
</tr>
<tr>
<td>DFKI-Adapt</td>
<td>61.86</td>
<td>59.18</td>
<td>53.94</td>
</tr>
<tr>
<td>Morfbase</td>
<td>59.53</td>
<td>56.89</td>
<td>52.07</td>
</tr>
<tr>
<td>BASELINE</td>
<td>56.96</td>
<td>54.75</td>
<td>49.32</td>
</tr>
<tr>
<td>DFKI-MPrompt</td>
<td>53.76</td>
<td>50.42</td>
<td>46.83</td>
</tr>
</tbody>
</table>
LINDAT/CLARIAH-CZ

international cooperation

• International cooperation (support or development)
  – ELG European Language Grid, EU H2020 Call 29a
    • Automatic metadata harvesting continues

• SSHOC, Infrastructural H2020 project, followup
  – Part of CLARIN ERIC participation, EOSC Marketplace

• Humane-AI-Net: AI Center of Excellence, Call 48
  – Microprojects w/DFKI, INRIA, etc.

• ATRIUM 2024-2028, Horizon Europe INFRA
  – Archaeology, MT and object recognition
    • In cooperation with Czech RI: Archaeological Information System

• Everse 2024-2028, Horizon Europe INFRA
  – Long/term preservation of research software (SSH), w/CLARIN

• Support for other Horizon Europe projects
  – RESQ+ (IR), MEMORISE (Holocaust access)
  – HPLT (LLM data collection, LLM/TM building, 2022-2025)
• International cooperation
  – ELG European Language Grid
    • Automatic metadata harvesting
  – SSHOC, Infrastructural H2020 project
    • Part of CLARIN ERIC participation, EOSC Marketplace
  – Humane-AI-Net: AI Center of Excellence
    • Microprojects with DFKI, INRIA, etc.
  – ATRIUM 2024-2028
    • Archaeology, MT and object recognition
      • In cooperation with Czech RI: Archaeological Information System
  – Everse 2024-2028, Horizon Europe INFRA
    • Long-term preservation of research software (SSH), with CLARIN
  • Support for other Horizon Europe projects
    – RESQ+ (IR), MEMORISE (Holocaust access)
    – HPLT (LLM data collection, LLM/TM building, 2022-2025)
• Czech European Open Science Cloud Association and Secretariat (EOSC CZ)
  – National support for central tasks and organization

• EOSC CZ Open Science I (2024-2028)
  – Project: National Repository Platform
    • Lead by Czech eInfra “CESNET”
    • LINDAT/CLARIAH-CZ significant partner (0,8 mil. EUR)
      – CLARIN-DSpace chosen as one of four repositories recommended nationwide
    • Funds repository(ies) development, upgrades, adaptation for national use

• EOSC CZ Open Science II (2025-2028)
  – Project led by Charles University (HQ)
    • LINDAT/CLARIAH-CZ will lead SSH workpackage (4 mil. EUR total), 1 of 8
      – Czech SSH RIs as partners, FSTP-like funding for others
LINDAT/CLARIAH-CZ beyond 2023

• Future continuation
  – Contract (with cuts) until 2026 (originally: 1.8x more funding, -2029)
  – LINDAT/CLARIAH-CZ on Czech Roadmap
  – Czech Republic joins EHRI ERIC in early 2025
  – Durable equipment (hardware, computing) grant under review
    • 2024-2026, 2 mil. EUR (shared by 7 partners of the RI)

• Evaluation
  – International panel + quantitative scientometrics at national level
    • Crucial for funding 2027-???? (legal changes – special status?)
  – Starts late 2024
  – IAB may play a role

• Integration with / support to EOSC CZ
Thank you!

https://lindat.cz