Multilingual Language Technology in the age of Artificial Intelligence and Deep Neural Networks

Jan Hajič

Institute of Formal and Applied Linguistics
Computer Science School
Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic

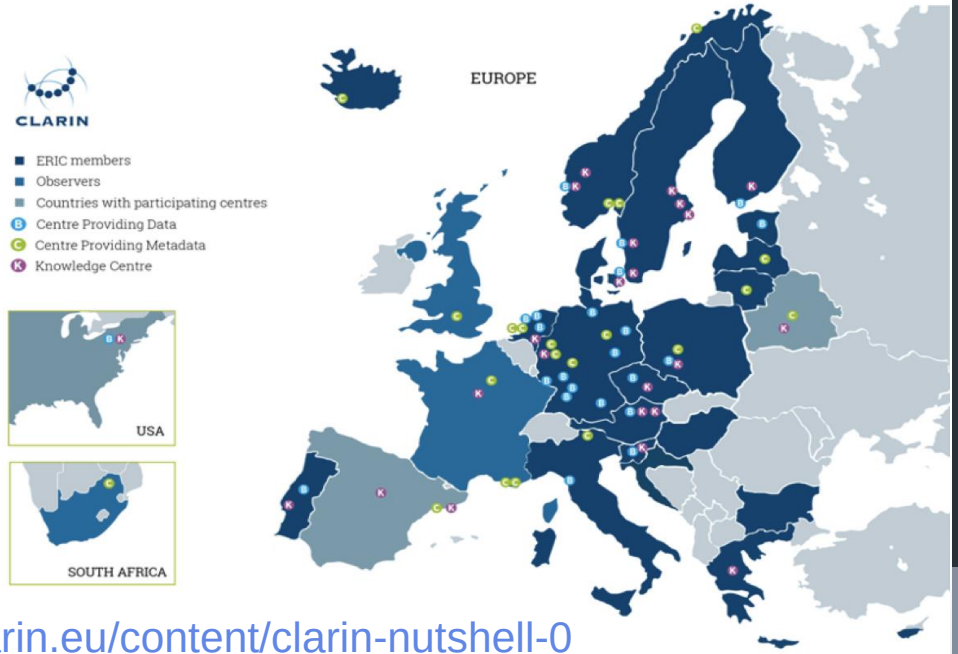
Using pictures created by Francesca Frontini, Fnaciska de Jong, Dieter van Uytvanck, Georg Rehm, Andy Way, Pavel Straňák

Outline

- European Networks
  - Language Resources & Technology
  - Fostering Language Equality
- Example of a complex Research Infrastructure
  - LINDAT/CLARIAH-CZ: resources, tools, services
- Future of Language Technology
  - AI methods / Large Language and Translation Models
  - Natural Language Understanding
- Conclusions

European Language Technology Networks: CLARIN

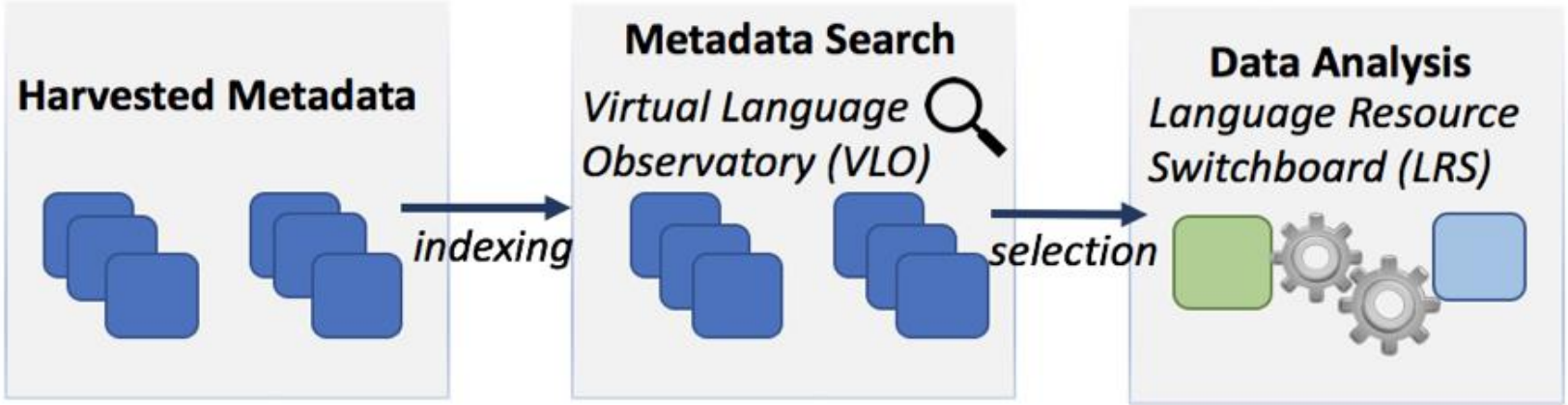
- CLARIN: since 2008, CLARIN ERIC since 2012
  - >50 centres
  - 21 members + 3 observers
- Provides
  - Resources
  - Tools
  - Services
- Standardized, open, virtual, distributed research infrastructure



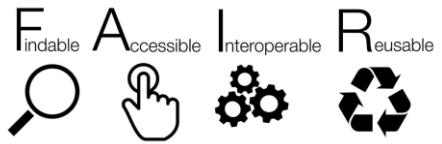
https://www.clarin.eu/content/clarin-nutshell-0

European Language Technology Networks: CLARIN

- The technical infrastructure



https://vlo.clarin.eu/
https://switchboard.clarin.eu/



European Language Technology Networks: CLARIN

https://www.clarin.eu/content/clarin-for-researchers

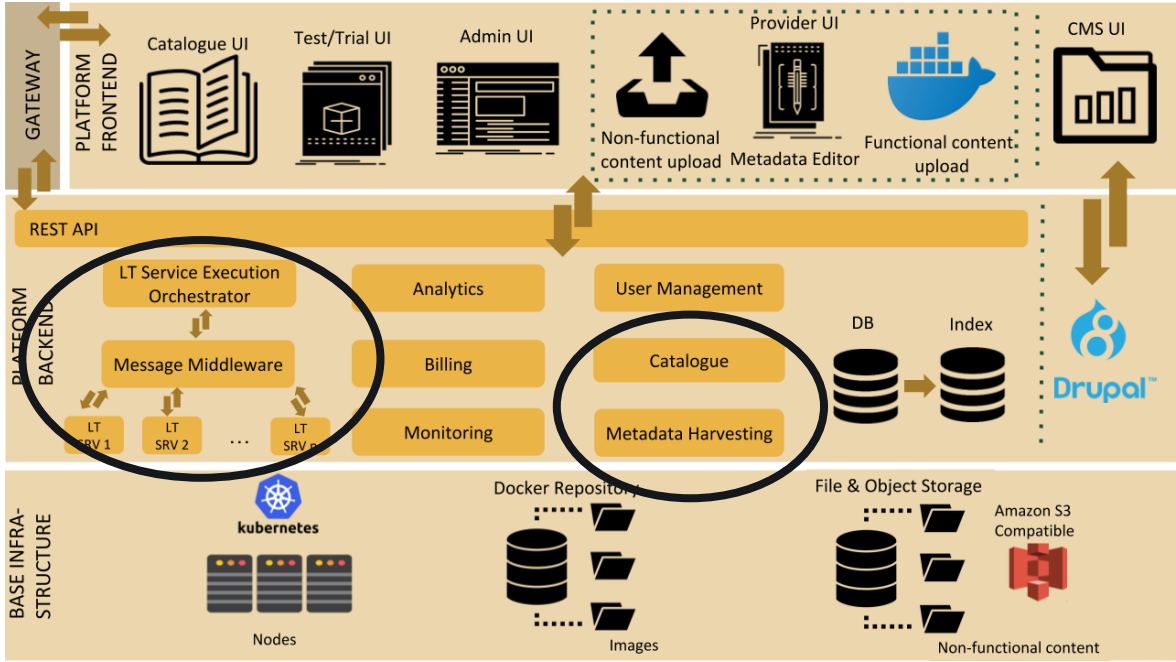
https://www.clarin.eu/content/knowledge-sharing

European Language Technology Networks: ELG

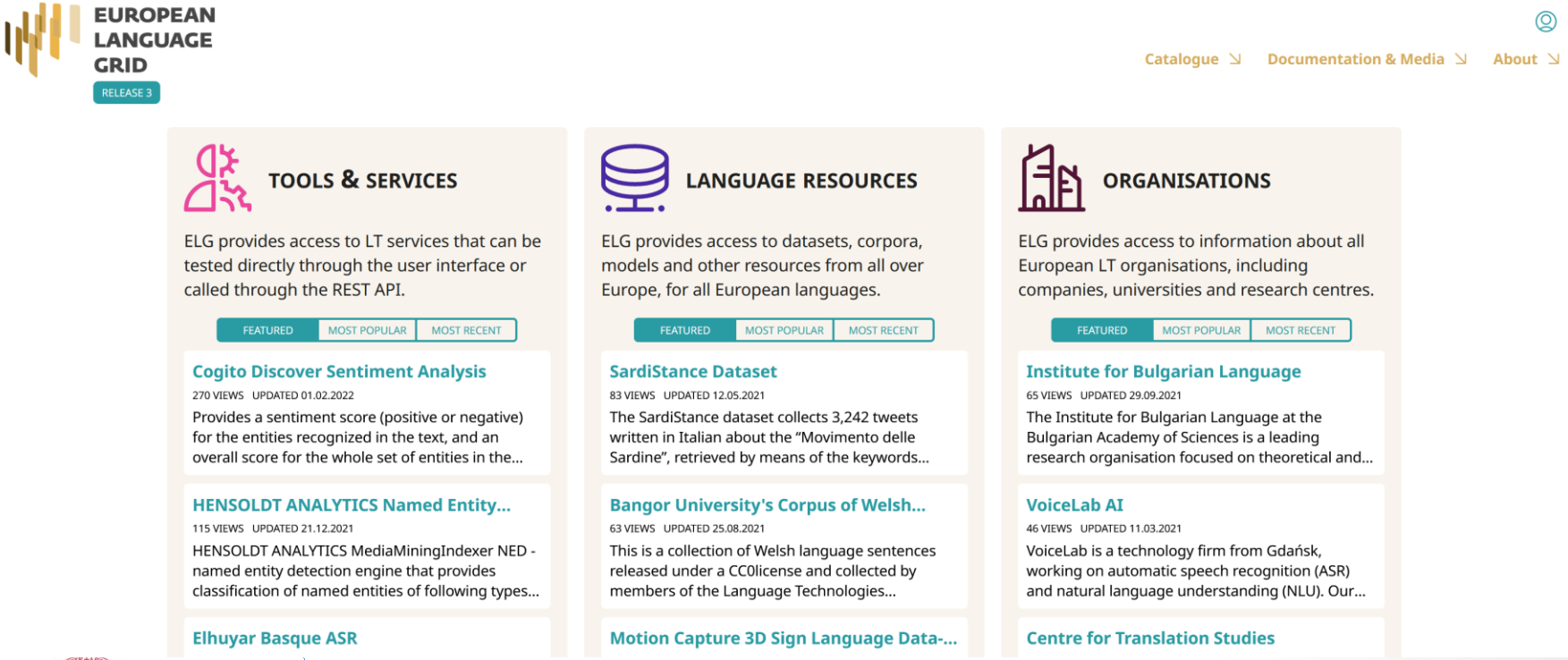
- European Language Grid
  - Centralized, all-in-one:
    - Language Resources (data, metadata, links to existing repositories)
    - Software tools, packages
    - (Running) services
    - Demos, catalogues (resources, institutions, stakeholders)
  - Marketplace for the European LT business space
  - Contributed from 15 external projects (FSTP funding)
    - And previous and concurrently running H2020 and infra projects
  - https://www.european-language-grid.eu

European Language Technology Networks: ELG

- Architecture
  - Services
    - Dockers
  - Data
    - Copied
    - Harvested
      - (CLARIN nodes)
    - Created
    - Curated

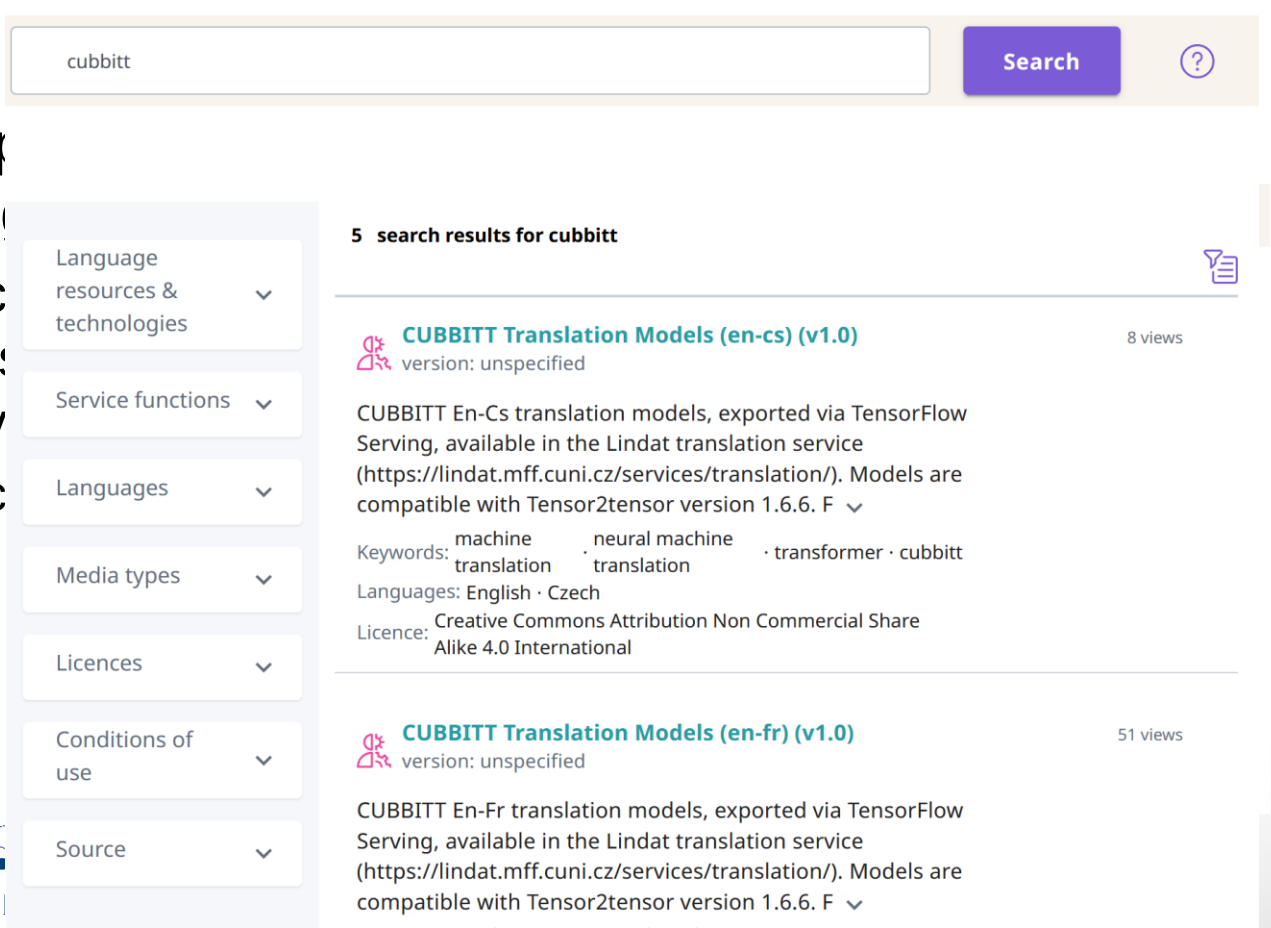


European Language Technology Networks: ELG



CLARIN
EUROPEAN LANGUAGE GRID
EUROPEAN LANGUAGE EQUALITY
Hippolyta

- Examp
  catalo
  - Mac
    trans
    serv
    - c

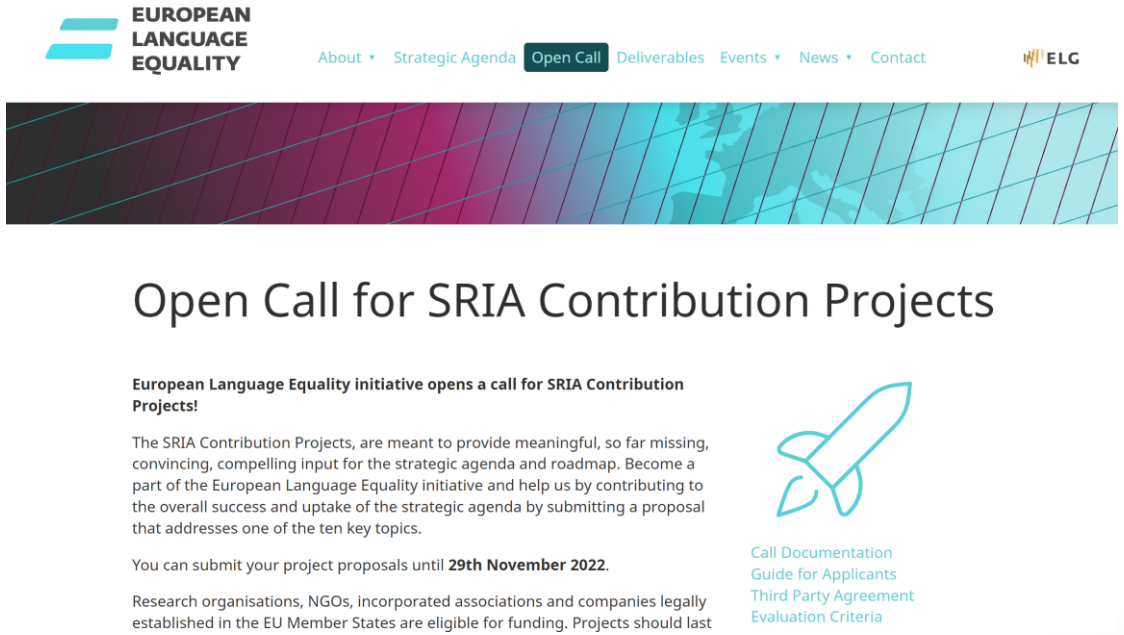


European Language Technology Networks: ELE/ELE2

- European Language Equality
  - https://european-language-equality.eu/
  - Foster and promote multilinguality and equal (digital technology) treatment of European languages
  - Draw up **strategic research agenda and roadmap**
    - From current state of Language Technology to **Digital Language Equality**
    - Provide path for implementing the agenda by 2030
- Now in second phase (2022-2023)
  - Open Call for contributions – everybody welcome

European Language Technology Networks: ELE/ELE2

- Open Call



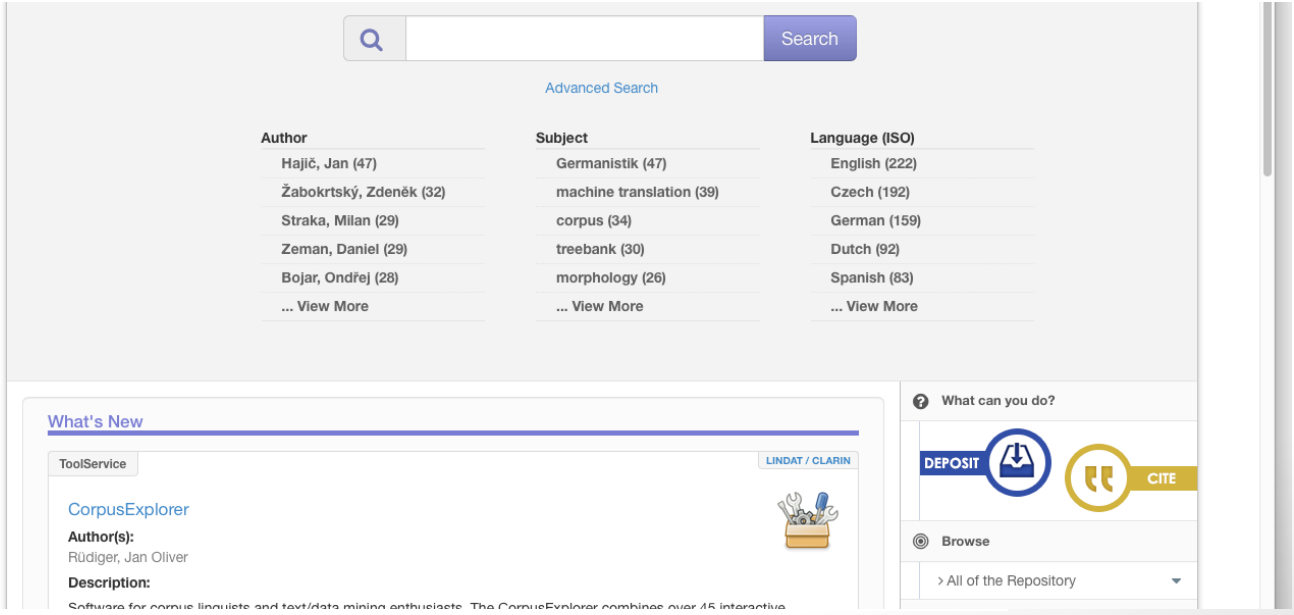
- https://european-language-equality.eu/open-call/

LINDAT/CLARIAH-CZ

- CLARIN “node” at https://lindat.cz
- Certified technical centre
  - Repository
    - Data, tools, metadata, licensing information
    - Single sign-on, persistent IDs
  - Services
  - Distributed network for Digital Humanities and Arts
    - Libraries, gallery, film archive, historical and other academy partners
    - Also part of DARIAH ERIC and (soon) of EHRI ERIC

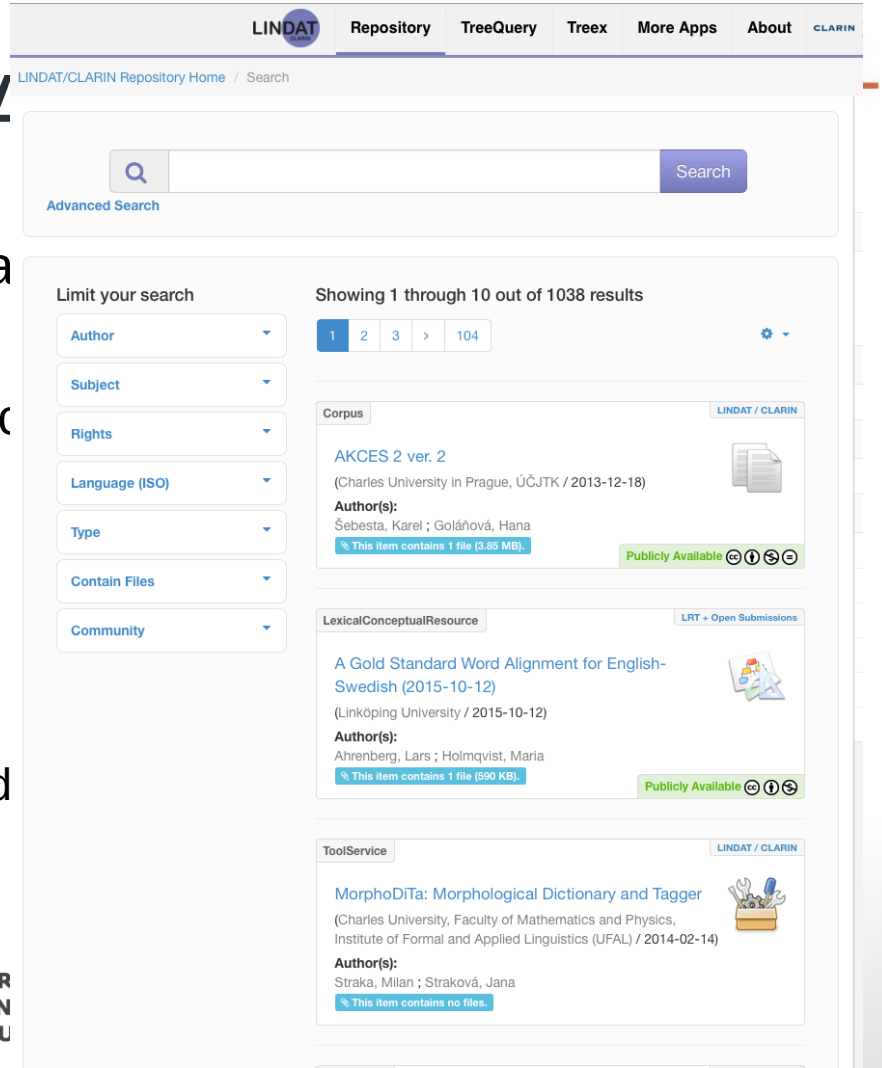
LINDAT/CLARIAH-CZ

- Core point: data repository – standardized, certified
  - Preserve and find language data and NLP tools



LINDAT/CLA

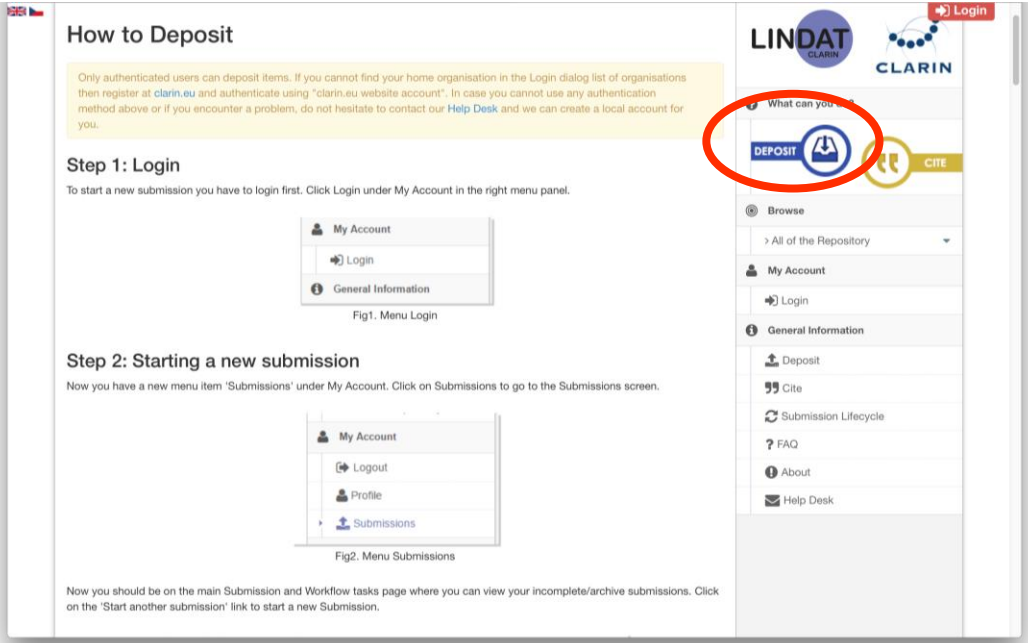
- OPEN ACCESS (as much as it ca
- > 500 registered users
  - submitters & users signing lic
- 200+ Data Records
  - > 1000 Metadata Records
- 1001 languages
  - 200 TB+ Data in Repository
    - + 1PB of UCS Shoah Found
- 40,000 accesses per month



CLARIN
EUROPEAN LANGUAGE GRID

LINDAT/CLARIAH-CZ

- Safe preservation
  - upload and don’t worry
- Discovery & Reuse
- Direct data citation
  - Google Scholar, Datasets
- Licensing
  - Open Access
  - … more options
- Versioning
- Worldwide (for everyone)

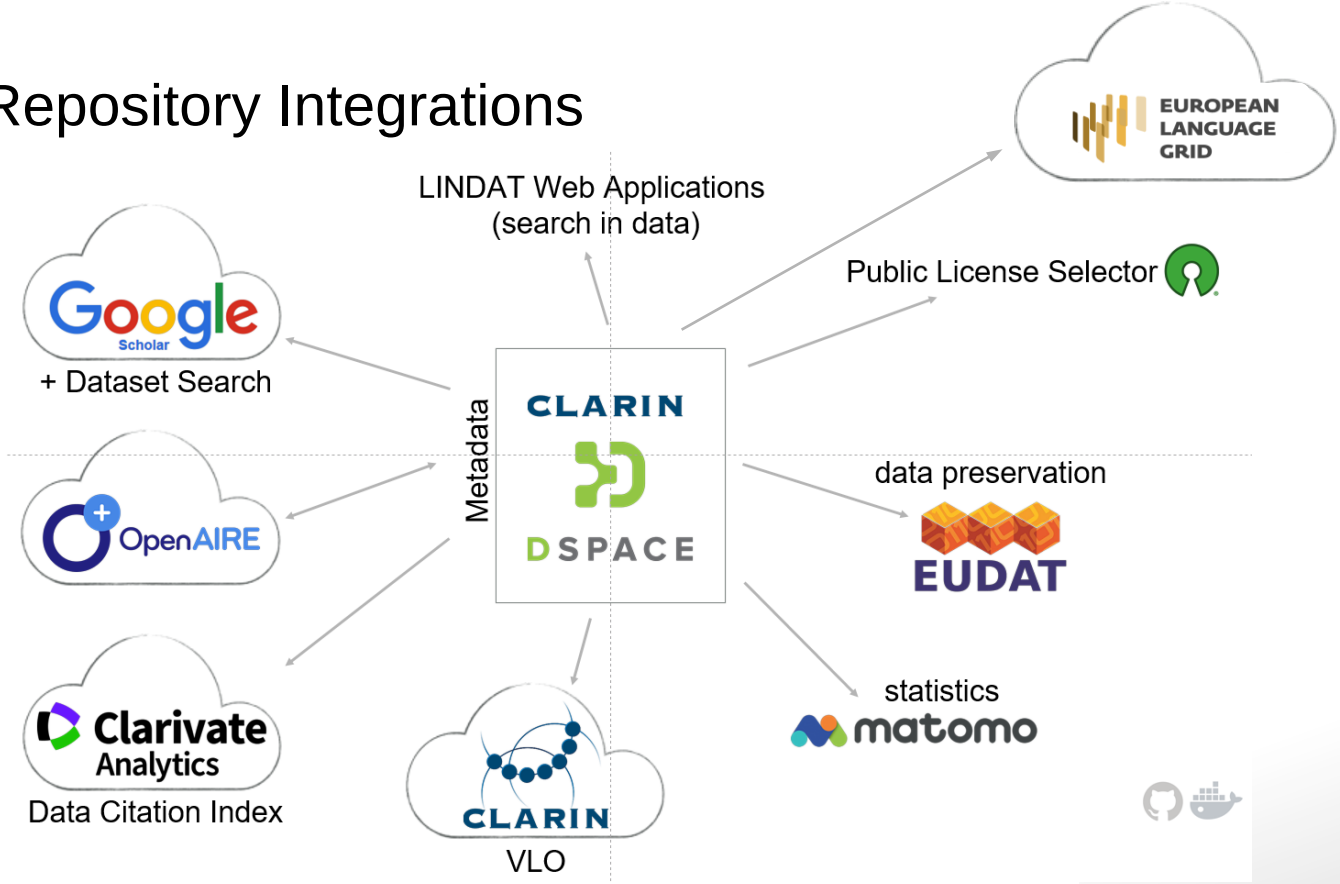


OPEN ACCESS FORCE11 The Future of Research Communications and e-Shcolarship GO FAIR OpenAIRE

CLARIN EUROPEAN LANGUAGE GRID EUROPEAN LANGUAGE EQUALITY Hippolyta

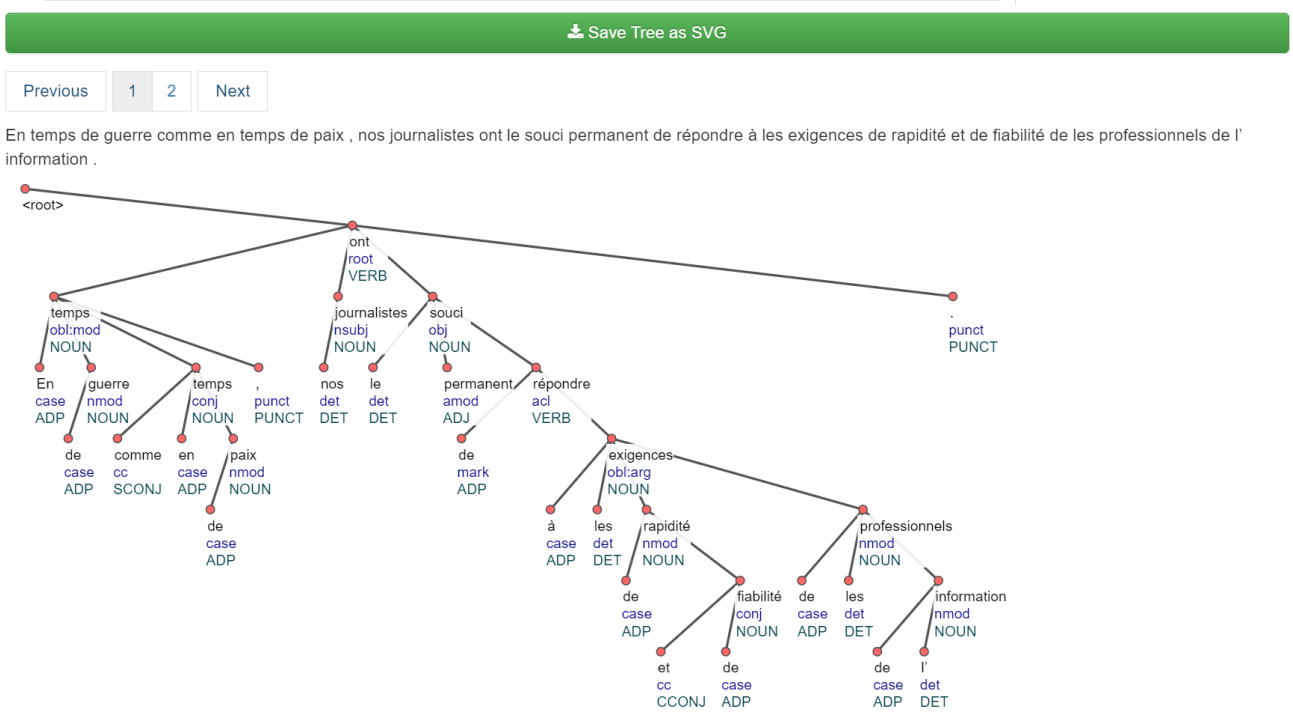
LINDAT/CLARIAH-CZ

- Repository Integrations



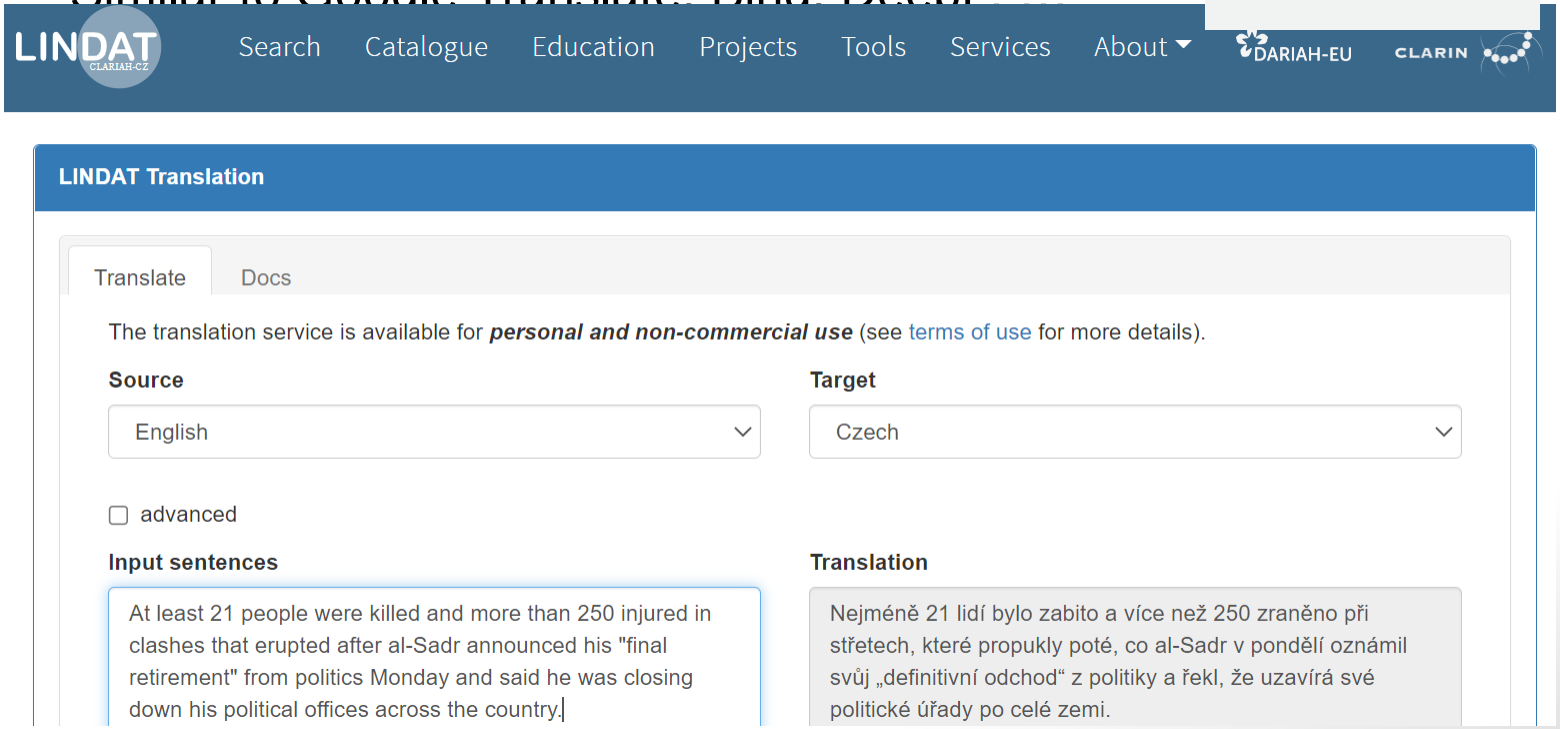
LINDAT/CLARIAH-CZ

- Text analysis services – UDPipe 2.0 (and other)



LINDAT/CLARIAH-CZ

- Machine Translation service
  - Similar to Google Translate, Bing, DeepL, ...



Charles Translator

Language Technology Future

- In past few years…
  - Huge success in applying Deep Learning
    - End2end systems, large language models, Multilanguage models
  - Need for large data (and computing resources)
  - Advances in applications
    - Speech
    - Language
    - Multimodal (vision, image, video, …)
    - Integrated structured and unstructured data

Language Technology Future

- In past few years…
  - Huge success in applying Deep Learning
    - End2end systems, ***large language models, Multilanguage models***
  - ***Need for large data (and computing resources)***
  - Advances in applications
    - Speech
    - Language
    - Multimodal (vision, image, video, …)
    - Integrated structured and unstructured data

The HPLT (Hippolyta) project

- High-Performance Language Technology
- Horizon Europe DATA call, 2022-2025
- Goals
  - Collect large data from Internet Archive (SF, CA, USA)
    - Approx. 12 PB
  - Extract text, clean, identify, deduplicate, pseudonymize, describe, …
  - Train language and translation models: 24 EU + min. 16 other
    - xBERTy, GPT-x, Transformer, future SoTA
    - make them openly available (OpusMT, Huggingface, possibly other repos)
  - Evaluate models – keep a dashboard
  - Demonstrate use of EU HPC Centres in a distributed manner
    - Huge compute demands: just for cleaning, 20 mil. CPU hours

The HPLT (Hippolyta) project

- Project partners
  - University of Edinburgh (Ken Heafield): scientific coordinator,
  - Charles University (UFAL/LINDAT, Jan Hajic, Dusan Varis, Jindrich Helcl, Martin Popel, Pavel Stranak, Barbor Hladka)
    - coordinator
  - University of Helsinki (Finland, Jorg Tiedemann, OpusMT)
  - University of Turku (Finland, Sampo Pyysalo, Filip Ginter)
  - University in Oslo (Norway, Stephan Oepen)
  - Prompsit (Spain, Gema Ramirez)
  - HPCs:
    - CESNET (Czechia, Ludek Matyska, David Antos)
    - Sigma2 (Norway, Hans Eide)
    - Cooperation with LUMI, EuroHPC, Karolina (IT4Innovations), possibly others

Natural Language Understanding

- Common theme of many roadmaps, agendas, etc.
- Are we there yet?
  - Large language models (esp. GPT-x) can generate fluent text (incl. theatre plays! See e.g. the TheAItre project)
  - Translation beats humans (some domains)
  - Dialog systems can pass Turing test (in the order of minutes or low tens of minutes)
- Questions remain:
  - Explainability, interpretability, post-output error detection, (de)bias(ing), ethical issues, …

Natural Language Understanding

- Possible direction of further research
  - Both large(r) language models and symbolic methods
    - Combination?
- Can knowledge representation be useful?
  - What “knowledge representation”?
- Goal of many research projects
  - Building ontologies: Cyc (and successors), Babelnet, but also Wikipedia, wikidata, databases, …
  - Building semantic representations
    - From the language point of view
    - From the logical/reasoning/inference side

Natural Language Understanding

- Human-like perception of read/heard language (text, voice)

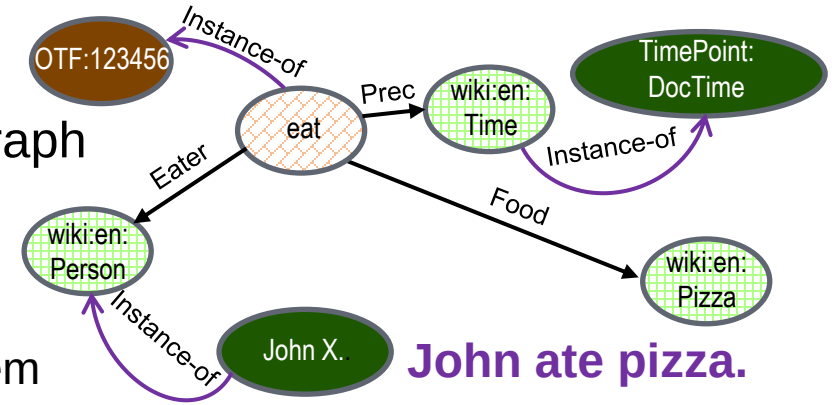
  - Key point: ***representation*** (of the knowledge/content acquired)
  - Unclear how people remember and structure knowledge and facts…
    - … Mathematics and Logic to the rescue
      - Graphs (with known properties and algorithms)
      - Graph search, graph transformations, graph matching
    - Ontologies to include contents
    - Logic: inferencing, contradiction detection, deduction, …

Natural Language Understanding

- LUSyD: Language Understanding: from Syntax to Discourse
  - Project 2020-2024, Czech Science Foundation
    - Linguistic & Computational (DNN, Language Models)
  - Relation between language form and knowledge representation
    - Morphology, syntax, semantics, discourse
  - Come up with ***Knowledge Representation***
    - Link to existing language (annotation/lexical) resources
    - Demonstrate on ***multiple*** (as many as possible) ***languages***
      - Plan: Czech, English, German, Spanish, Chinese, Korean (Arabic, …)
    - Create ***eventive grounding ontology***, populated with multiple languages (SynSemClass), linked to FrameNet, WordNet etc.

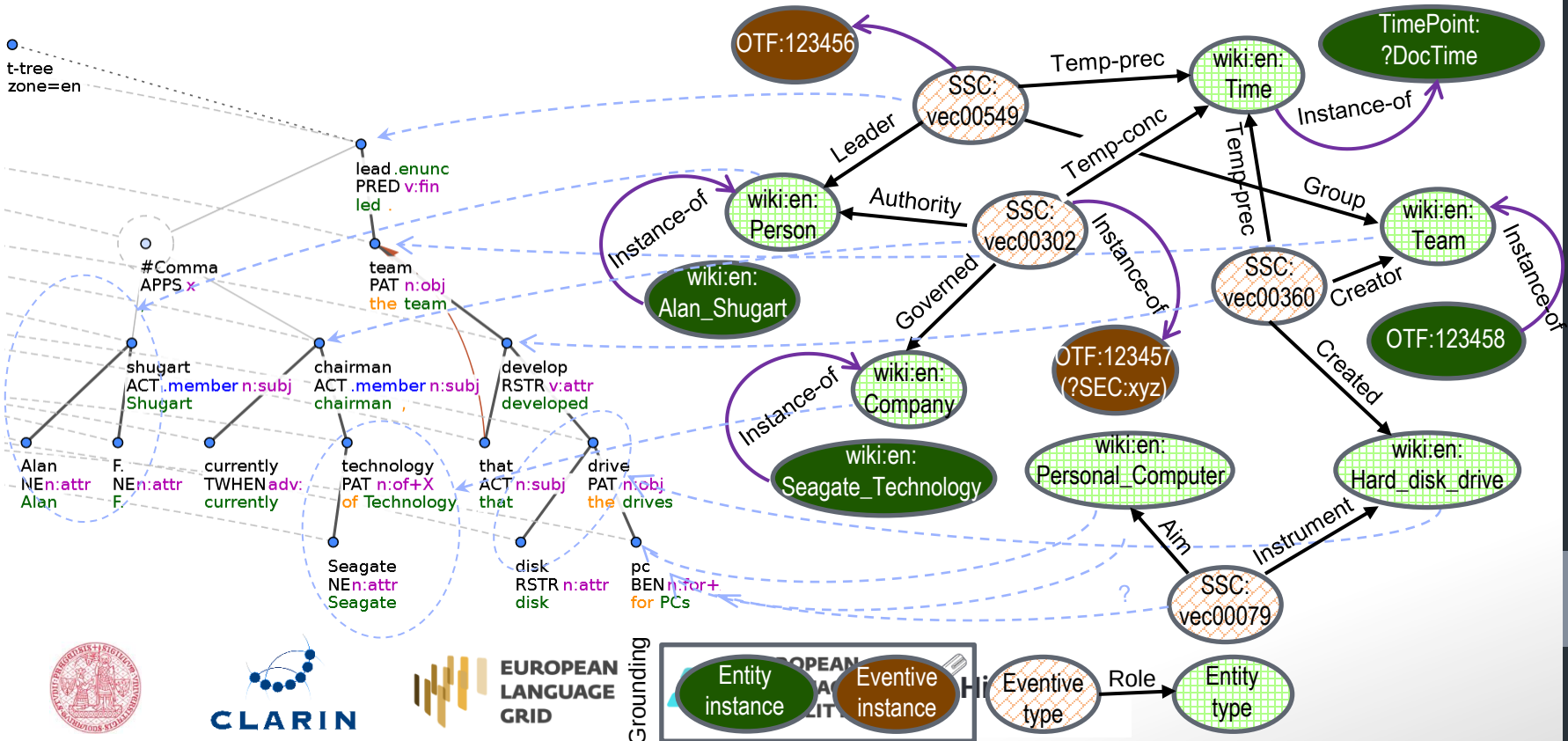
Natural Language Understanding

- The Principles
  - Knowledge Representation Graph
    - Nodes representing
      - ***Events***
      - ***Entities***
    - Edges: ***relations*** between them
  - All nodes ***grounded***
    - *Ontologies*
  - Relations: fixed set
  - Nodes linked to original text/speech/…
    - For Machine Learning in general, applications



NLU: real example (WSJ)

Alan F. Shugart, currently chairman of Seagate Technology, led the team that developed the disk drives for PCs.



Thank you!

https://ufal.mff.cuni.cz
https://lindat.cz
https://lindat.cz/services
Twitter: @LindatClariahCZ
Twitter: @ufal_cuni

References (NLU)

- Zdeňka Urešová, Karolina Zaczynska, Peter Bourgonje, Eva Fučíková, Georg Rehm, Jan Hajič. 2022. Making a Semnatic Event-type Ontology Multilingual. LREC 2022, Marseille, France (this presentation).
- Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel, Dordrecht.
- Zdeňka Urešová, Jan Štěpánek, Jan Hajič, Jarmila Panevová, and Marie Mikulová. 2014. *PDT-Vallex.* LINDAT/CLARIN digital library. http://hdl.handle.net/11858/00-097C-0000-0023-4338-F.
- Zdeňka Urešová, Eva Fučíková, and Jana Šindlerová. 2016. *CzEngVallex: a bilingual Czech-English valency lexicon.* The Prague Bulletin of Mathematical Linguistics, 105:17–50.
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018a. Creating a Verb Synonym Lexicon Based on a Parallel Corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18*), Miyazaki, Japan, May. European Language Resources Association (ELRA).
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018b. Defining verbal synonyms: between syntax and semantics. In Dag Haug, Stephan Oepen, Lilja Ovrelid, Marie Candito, and Jan Hajič, editors, *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)* (Pub. No. 155), pages 75–90, Linköping, Sweden. Universitetet i Oslo, Linköping University Electronic Press.
- Zdeňka Urešová, Eva Fučíková, Eva Hajičová, and Jan Hajič. 2018c. Synonymy in Bilingual Context: The SynSemClass Lexicon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, Santa Fe, New Mexico, USA, August 20-26, 2018, pages 2456–2469.