Towards a technology for dictionary intermediated dynamic alignment of multilingual corpora.

A vision and some concrete steps

Dan Cristea[1,2], Eveline Wandl-Vogt[3], Mihaela Onofrei[2], Andrei Scutelnicu[1,2]

[1] "Alexandru Ioan Cuza" University of Iași, Faculty of Computer Science, 16 Berthelot St. Iași
[2] Institute of Computer Science, the Iasi branch of the Romanian Academy, 2 Codrescu St. Iași
[3] Austrian Centre for Digital Humanities (ACDH), Austrian Academy of Sciences Sonnenfelsgasse, 19 – A-1010 Wien

E-mail: dcristea@info.uaic.ro, eveline.wandl-vogt@oeaw.ac.at, mihaela.plamada.onofrei@gmail.com, andreiscutelnicu@gmail.com

Motivations and goals

- Tremendous advances acquired recently in linking linguistic open data (LLOD)

  (Chiarcos *et al.*, 2013)

  – YET: a coherent initiative of building a technology able to automatically keep updated a huge collection of multilingual language data is still a task for the future

Motivations and goals

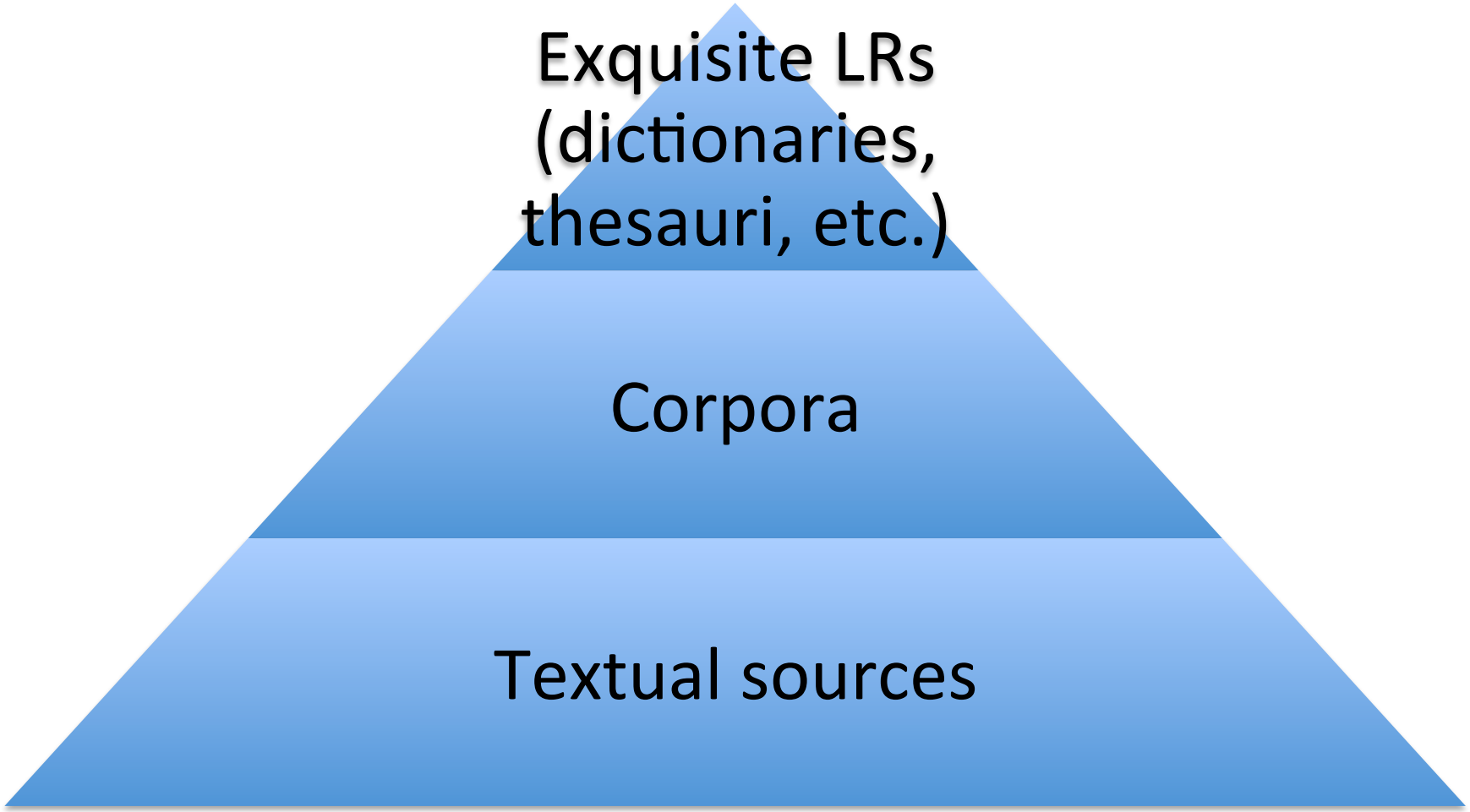
- Long running goals:
  1. Acquire linguistic corpora, on a continuous basis and in more languages
  2. Align them with dictionaries and other LR
  3. Align them among languages
     - technologically
     - content-based

In this talk…

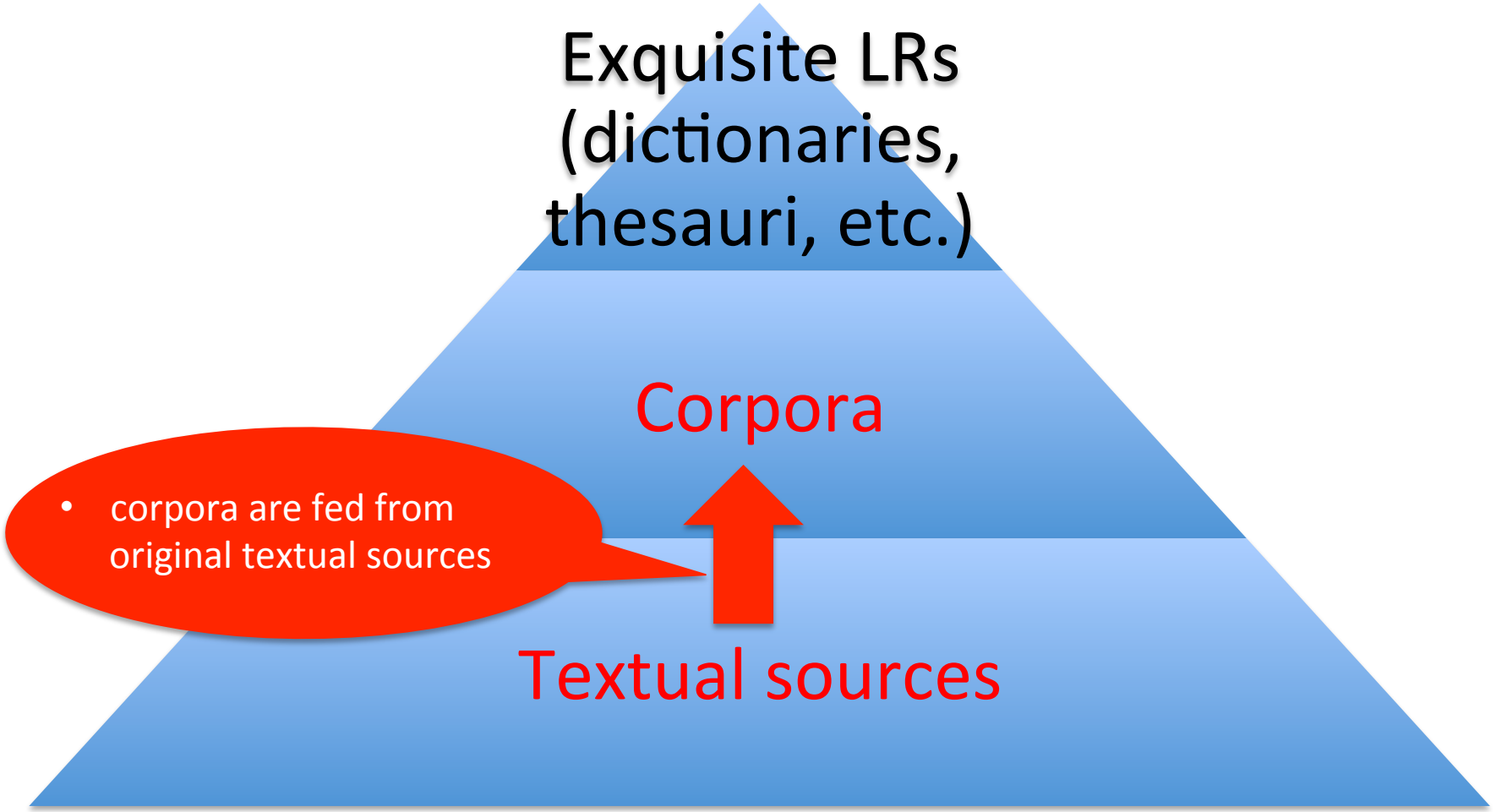
- A methodology of continuously acquiring a language corpus
- Show a pair of languages whose corpora are in the process of being technologically synchronised <= the first stage of alignment
- Show immediate applications and think for the future

Notice of a miss in this meeting: the expression of need to align dictionaries with corpora.

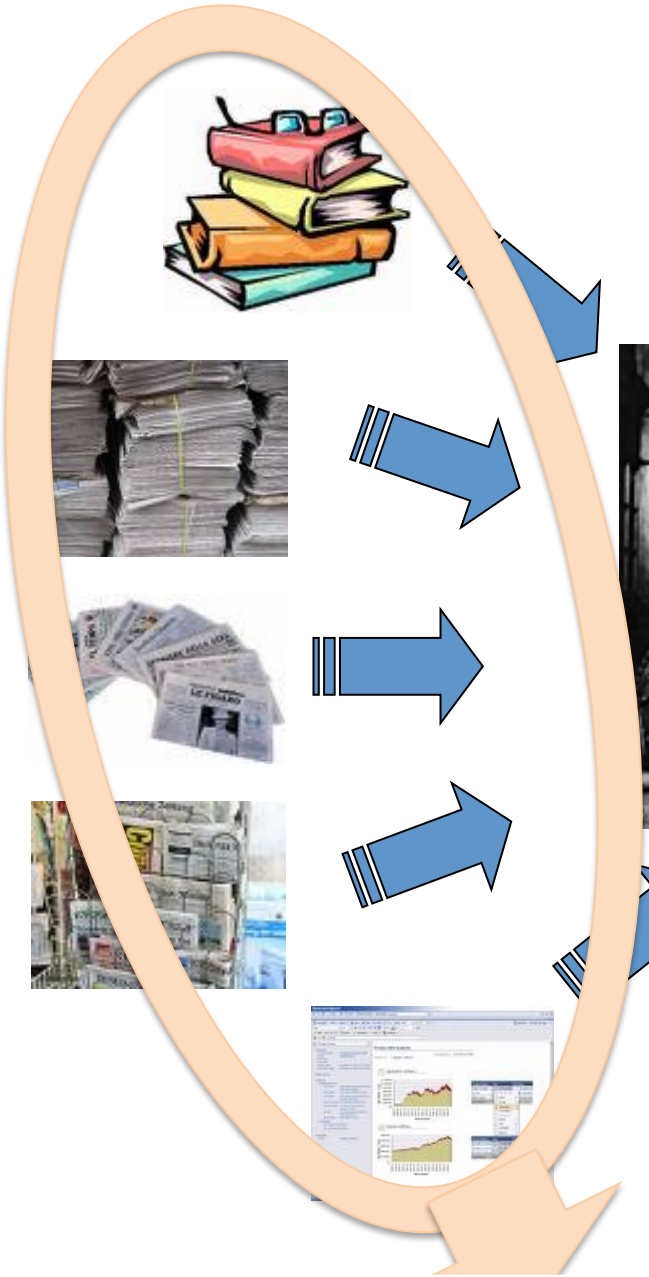
Pyramid of sublimation of LRs for a language

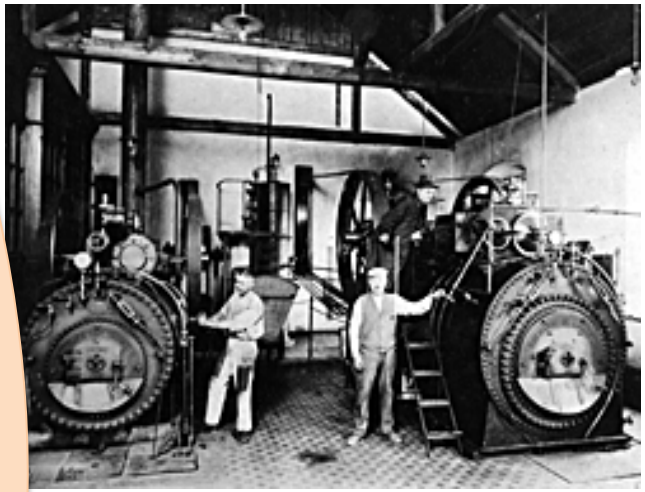


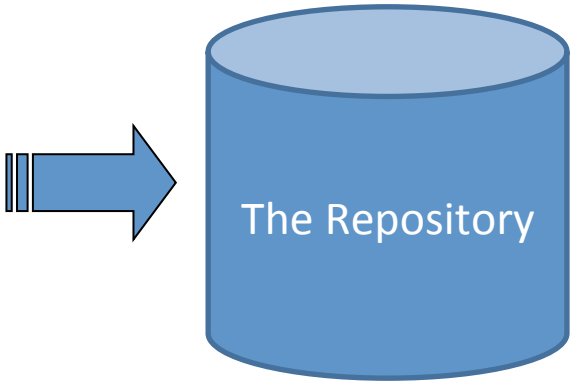
The basic acquisition link

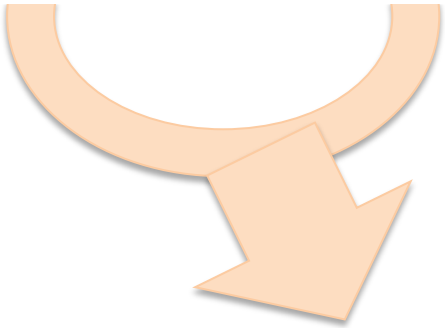


Collecting and processing textual sources







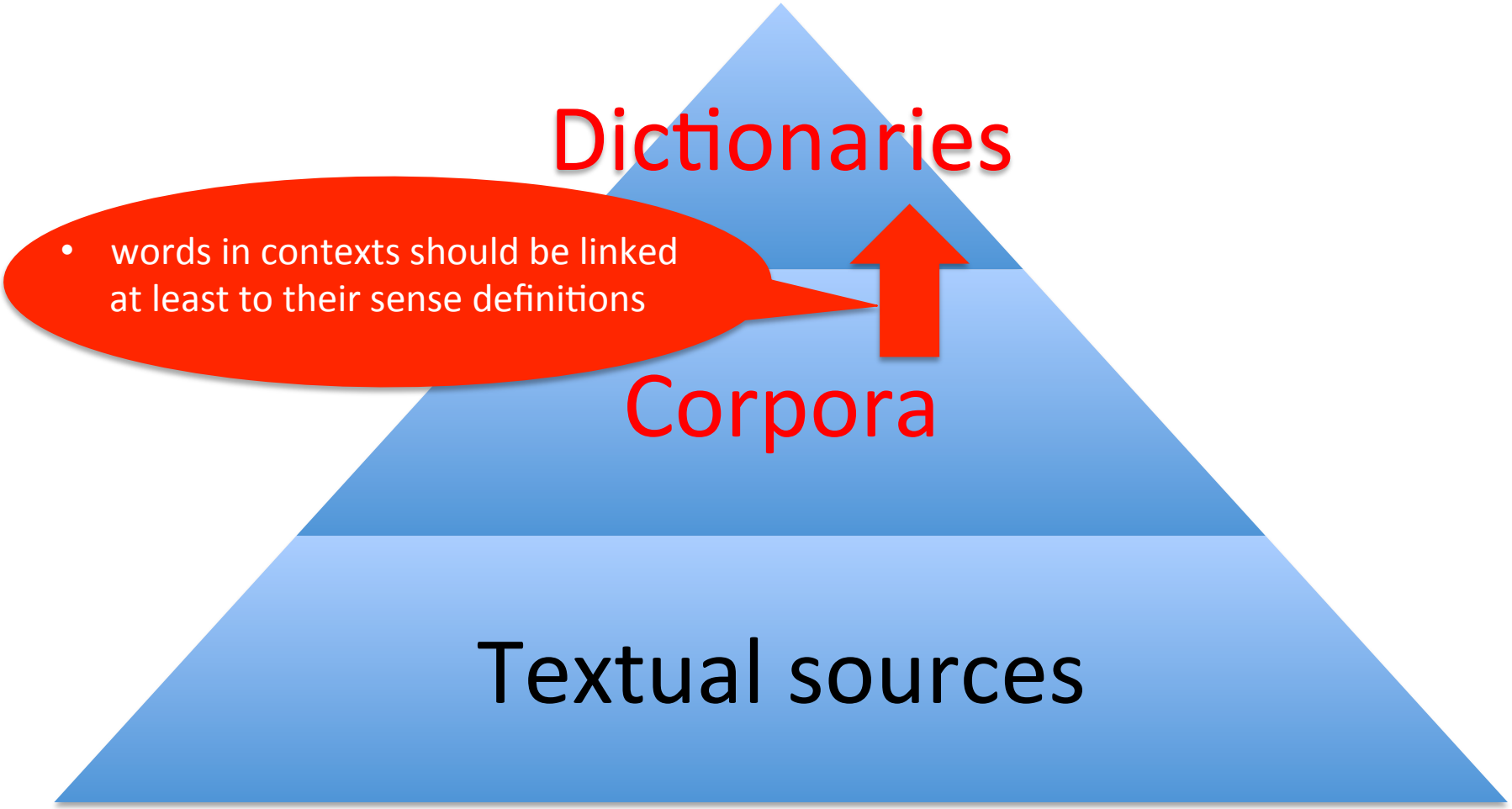


Providers

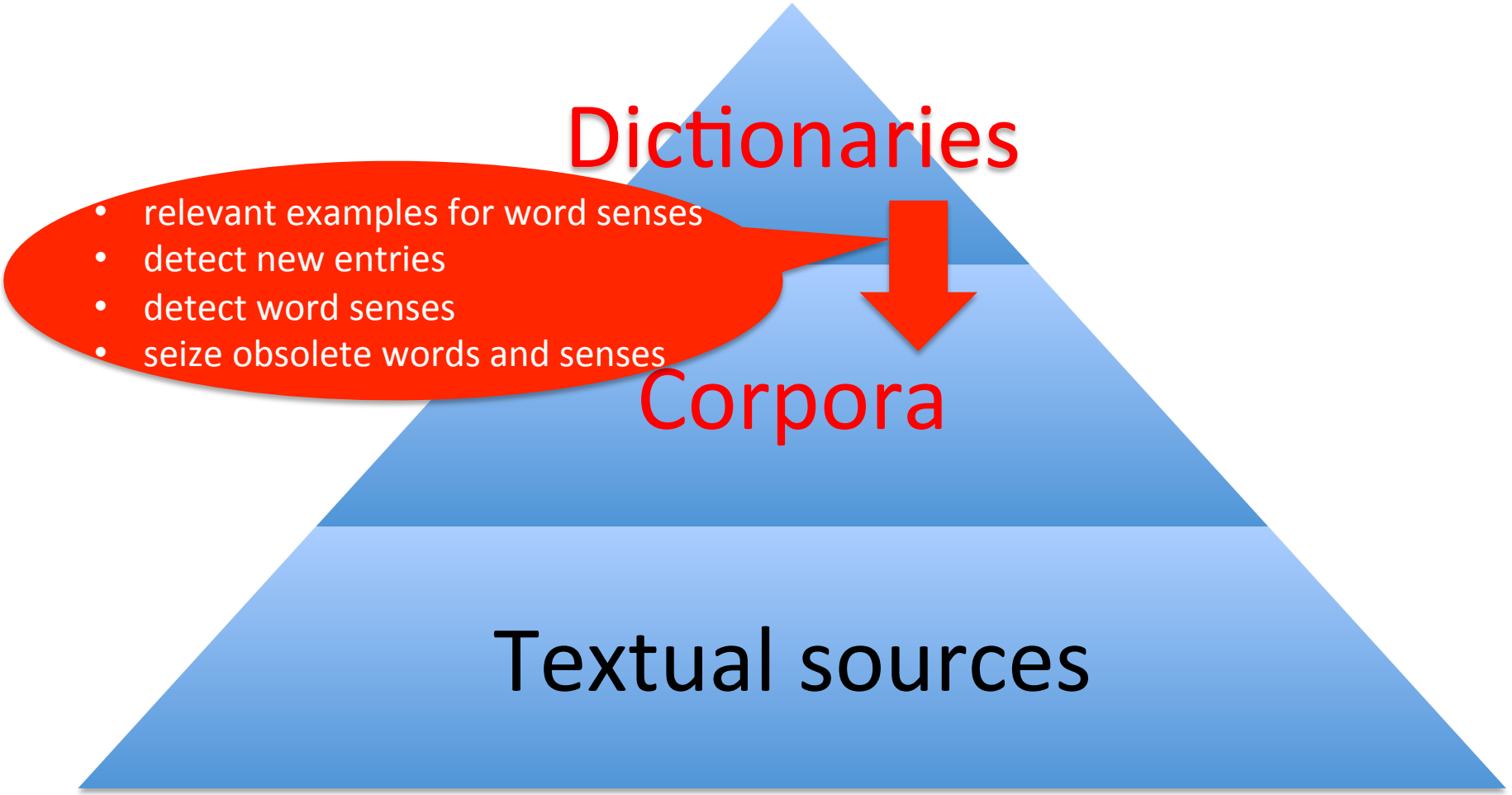
- Editing houses
- Public and particular persons producing printings for public consume
- Recording houses and studios
- Universities, research institutes
- etc.

Willing to donate their textual data!

Linking corpora onto dictionaries



Linking dictionaries onto corpora

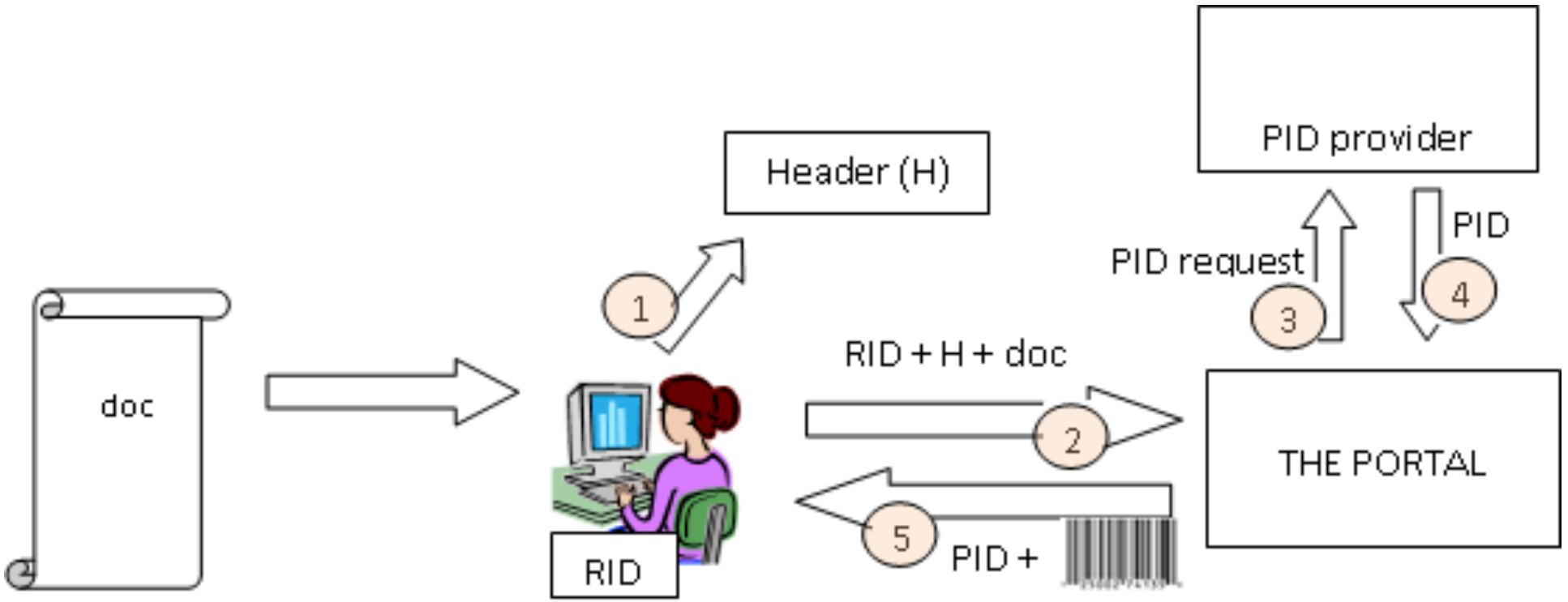


In a perfect world…

- Language corpora are being accumulated daily, to keep with language evolution
- Links between words and dictionaries should be permanently updated =>
  - new entries and/or new senses in dictionaries, with corresponding definitions, are being generated when necessary
  - new words and their senses are detected
- Languages are aligned:
  - a dedicated environment: the LOD version of Wiktionary, consistently connected onto each specific language dictionary – an universal inventory of linguistic concepts

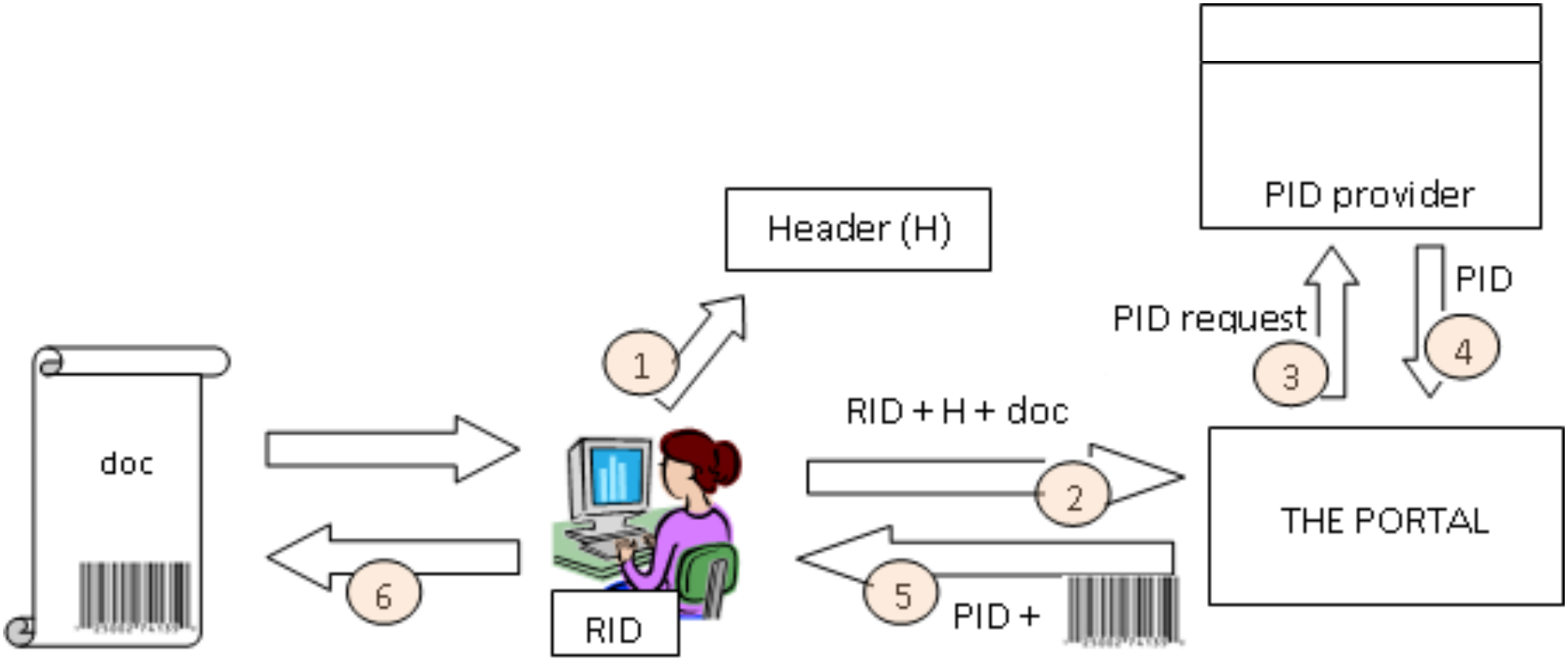
(Declerck *et al.*, 2014)

Communication:
Provider – Portal – PID provider



(Cristea, 2011)

Communication:
Provider – Portal – PID provider



(Cristea, 2011)

Two big corpora into one: DRUKOLA (2016-2018)

Original title: *Sprachvergleich korpustechnologisch Deutsch - Rumänisch*

- A project funded by Alexander von Humboldt Foundation (Cosma *et al.*, 2016)
- Research Group Linkage Programme
  - University of Bucharest
  - Institute for the German Language in Mannheim
  - Romanian Academy as associate partner:
    - Institute for Artificial Intelligence Bucharest
    - Institute of Computer Science Iaşi

DRUKOLA

- Concrete tasks
  - construction and provision of comparable corpora, i.e. apply similar principles and accessing technologies to:
    - DeReKo, a German Reference Corpus, and
    - CoRoLA, a Romanian corpus
  - development of criteria for building comparable virtual sub-corpora, based on metadata and other text properties
  - exploration of quantitative differences wrt. to different variables and their distributional properties
  - conduction of corpus-based comparative case studies
  - development of a corpus technology to share the corpus and research results in a common platform
  - building a crystallization structure for an European Reference Corpus

Harmonization of DeReKo and CoRoLa

- **Syntactical interoperability**
  - **metadata** comply with CMDI (Component MetaData Infrastructure) and TEI-P5 standards
- **Semantic interoperability**
  - e.g. for the metadata categories that are used for the construction of virtual corpora

The general procedure for the harmonization of data categories and value sets:

- **define functions that map the original data to more coarse-grained taxonomies**

Additional harmonization on lower levels

- integrate CoRoLa into the KorAP **corpus query engine**
- adopt the GGS query mechanism developed for CoRoLa as an auxiliary search engine to express constraints that would exploit the multi-layered annotation of DeReKo

DeReKo – Deutsches Referenzkorpus

- at Institut für Deutsche Sprache, Mannheim, since 1964

(Kupietz *et al.*, 2010)

- the world's largest collection of German texts

  >25 billion tokens

- a broad variety of text types with a quantitative focus on newspaper texts and a rapidly growing portion of computer mediated communication

CoRoLa – the Digital Corpus of the Contemporary Romanian Language

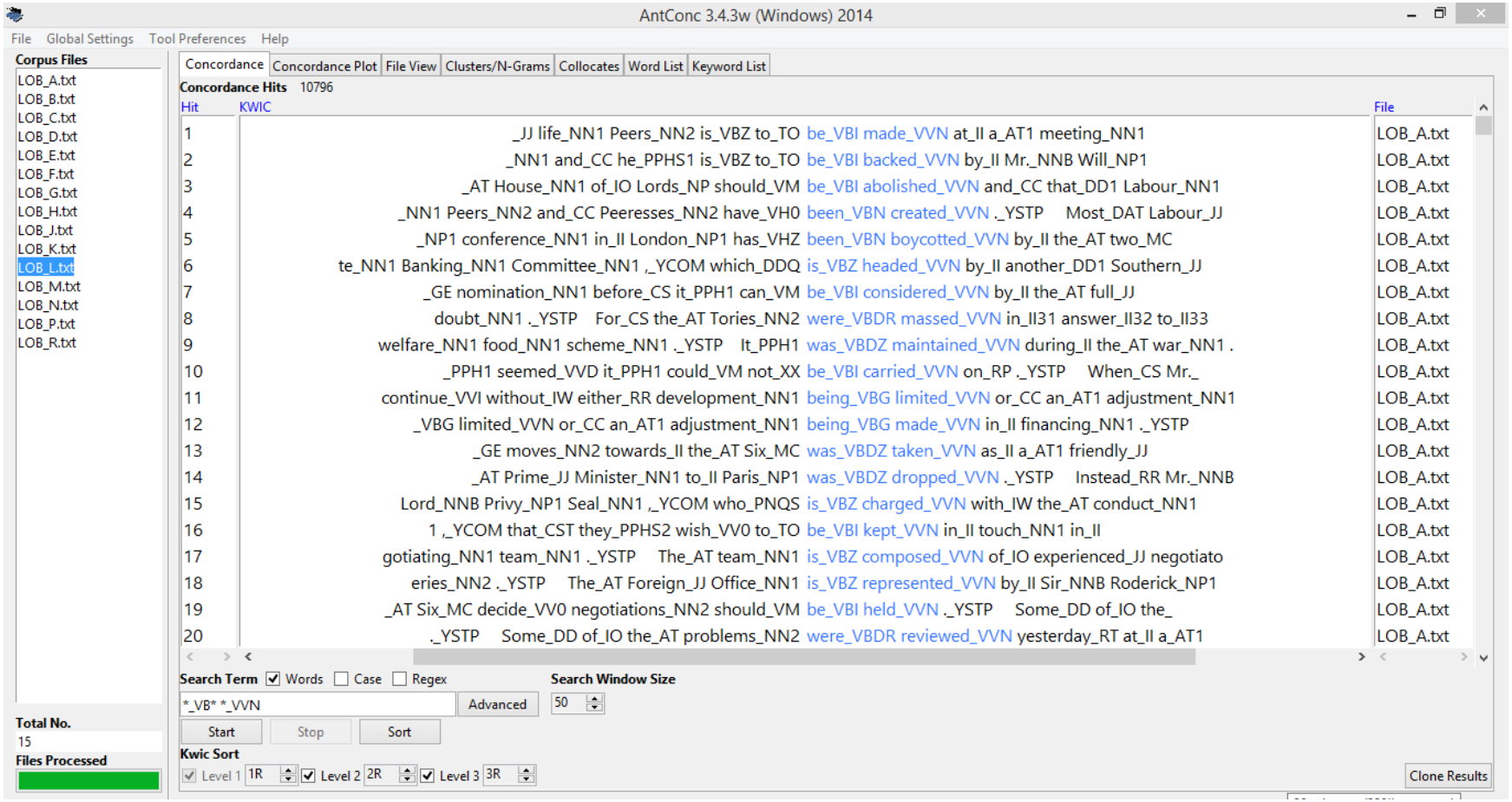
- **written** texts – 500 million tokens
- **speech** records – 300 hours
- language varieties: **standard literary language**
- time period **1945-1989&1990-today**
- type of metadata: **CMDI** standard http://www.meta-net.eu/meta-share/index_html)
- type of annotations: currently, **in-line**, but finally **stand-off**

(Barbu Mititelu *et al.*, 2014)
(Cristea *et al.*, 2015)
(Tufiș *et al.*, 2016)

Corpus interrogation platform: KORAP

- corpus data can be stored at different locations
- virtual corpora can easily be defined based on metadata properties
- unlimited maximum corpus size
- unlimited number of annotation layers
- support for multiple query languages
- open-source

Key Word In Context (KWIC)



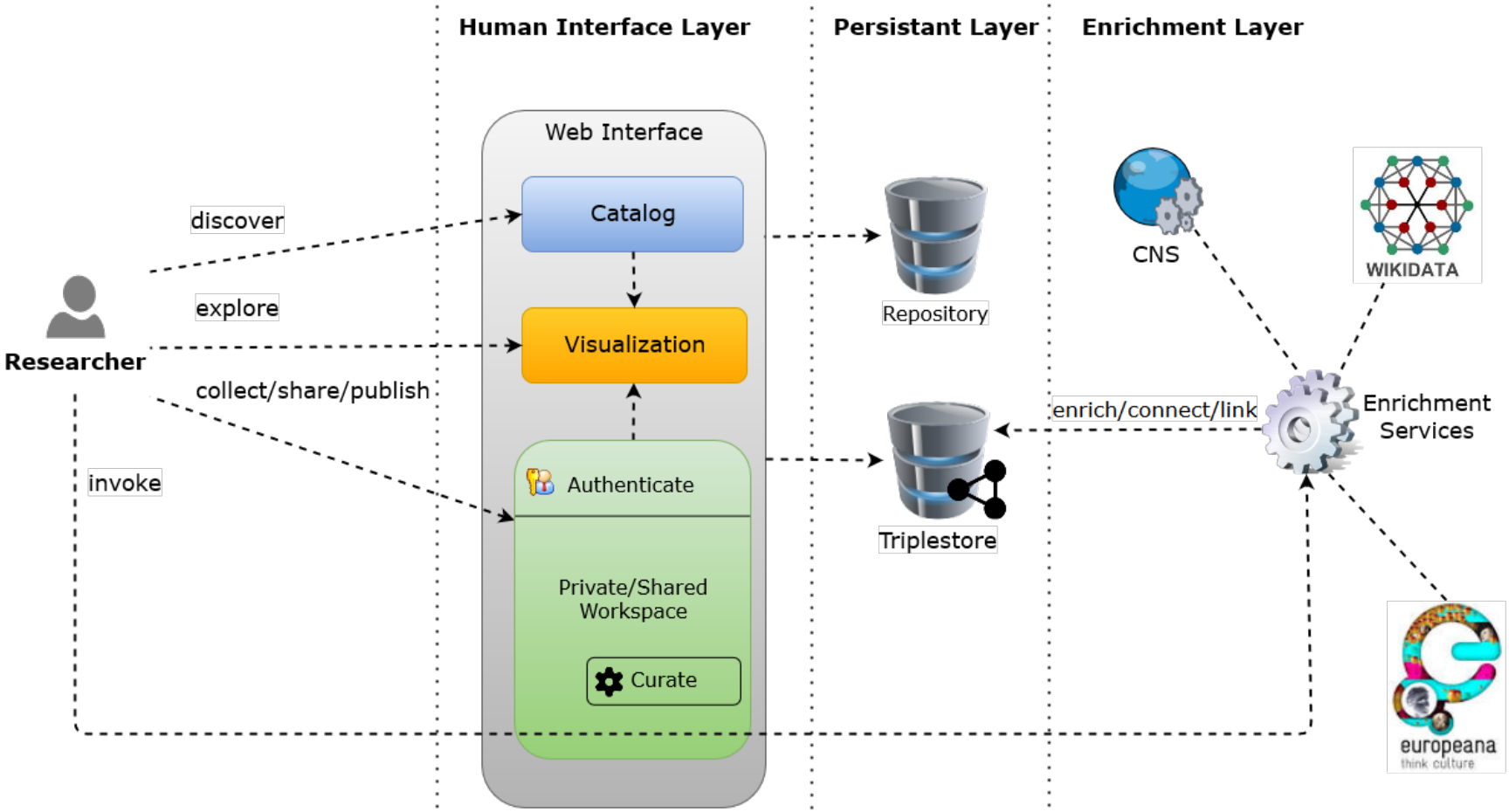
Applications of DRUKOLA

- In a wide range of areas:
  - corpus linguistics
  - computational linguistics
  - applied linguistics, cross-linguistic studies
  - applied computer science
  - research infrastructure development
  - research politics
  - (L1-specific L2-teaching and learning)
  - (machine translation)

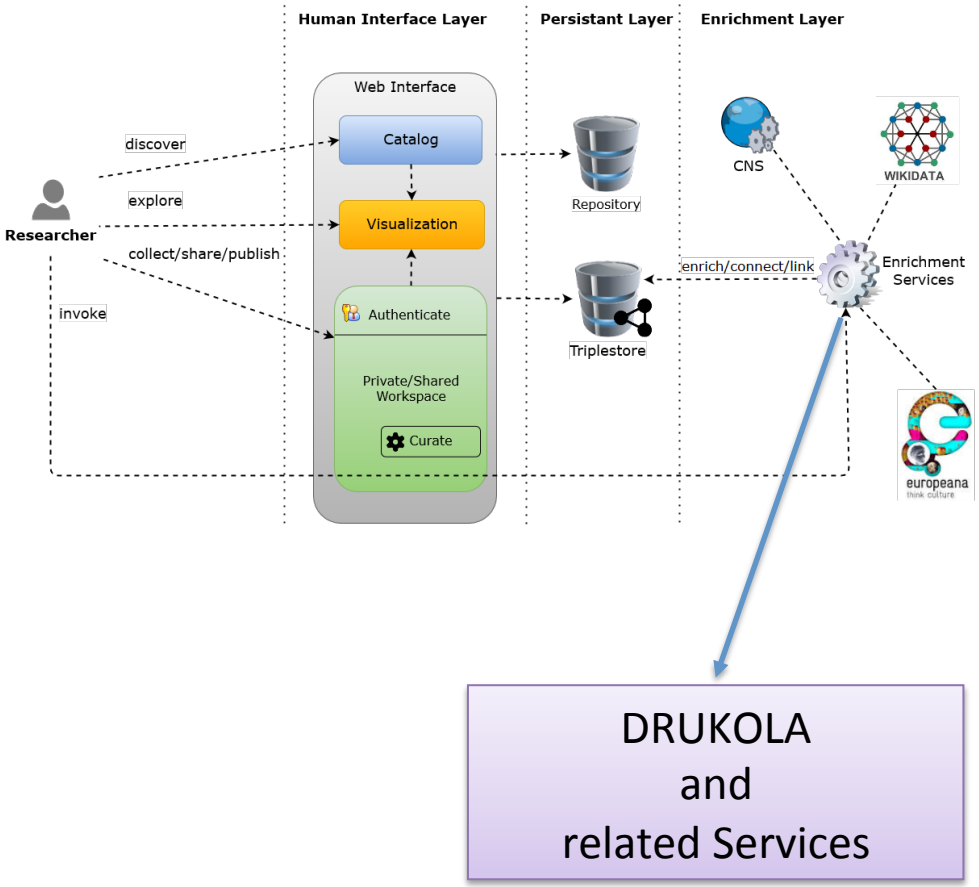
Case Study:
Biodiversity and Linguistic Diversity

- Examples:
  detect names of plants in the corpora
- Retrieve related entries from the following resources
  - Lexicographically curated databases, e.g.
    Database of Bavarian Dialects (DBÖ)
  - Botanically, taxonomically curated databases, e.g. some beyond
    Biodiversity and Linguistic Diversity Project
  - Aggregators, e.g.
    Wikidata and Babelnet

Case Study:
Biodiversity and Linguistic Diversity



Case Study:
Biodiversity and Linguistic Diversity



Case Study:
Biodiversity and Linguistic Diversity

- Train the recognizer based on results
- Detect new entries in other resources

- Add data to the portal
- Add an enrichment service to the portal
- Learn from data available within the portal
- Enable Pan-European cultural studies, e.g. semi-automatically help detecting naming concepts etc.
- Contribute to a multilingual Pan-European plant-names dictionary

Thank you!

- Acknowledgements:
  - COROLA – a project of the Romanian Academy
  - DRUKOLA – a project funded by the Alexander von Humboldt Foundation
  - exploreAT! – a project funded by the Austrian Academy of Sciences

Some bibliography

- Barbu Mititelu, V., Irimia, E., and Tufiş, D. 2014. CoRoLa – the reference corpus of contemporary Romanian language. In *Proceedings of the ninth International Conference on Language Resources and Evaluation – LREC* (Reykjavik, Iceland), 1235-1236.
- Chiarcos, C., Cimiano, P., Declerck, T., McCrae, J.P. (2013). Linguistic Linked Open Data (LLOD) - Introduction and Overview. In: Christian Chiarcos, Philipp Cimiano, Thierry Declerck, John P. McCrae (eds.): 2nd Workshop on Linked Data in Linguistics, Pages i-xi, Pisa, Italy.
- Cosma, R., Cristea, D., Kupietz, M., Tufiș, D., Witt, A. (2016). *DRuKoLa – Towards Contrastive German-Romanian Research based on Comparable Corpora*, in Proceedings of the Workshop on the Challenges in the Management of Large Corpora (CMLC-4), Language Resources and Evaluation Conference (LREC), 28 May 2016, Portoroz.
- Cristea, D. 2011. Romanian Linguistic Resources on Very Large Scale. *Computer Science Journal of Moldova*. 19, 2 (56), 130-145.
- Cristea, D., Bolea, C., Bibiri, A.-D., Scutelnicu, L.-A., Moruz, A.M. and Pistol, L. 2015. *Metadata of a Huge Corpus of Contemporary Romanian Data and Organization of the Work*. In Proceedings of the 7th Balkan Conference on Informatics, Craiova, 2-4 September, 2015.
- Declerck, T., Wandl-Vogt, E., Mörth, K., Resch, C. (2014). Towards a Unified Approach for Publishing Regional and Historical Language Resources on the Linked Data Framework, in *Proceedings of Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (OCCURL-2014)*, co-located with LREC 2014. 26-31 May, Reykjavik.