

Towards a technology for dictionary intermediated dynamic alignment of multilingual corpora. A vision and some concrete steps

**Dan Cristea^{1,2}, Eveline Wendl-Vogt³,
Mihaela Onofrei², 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.wendl-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

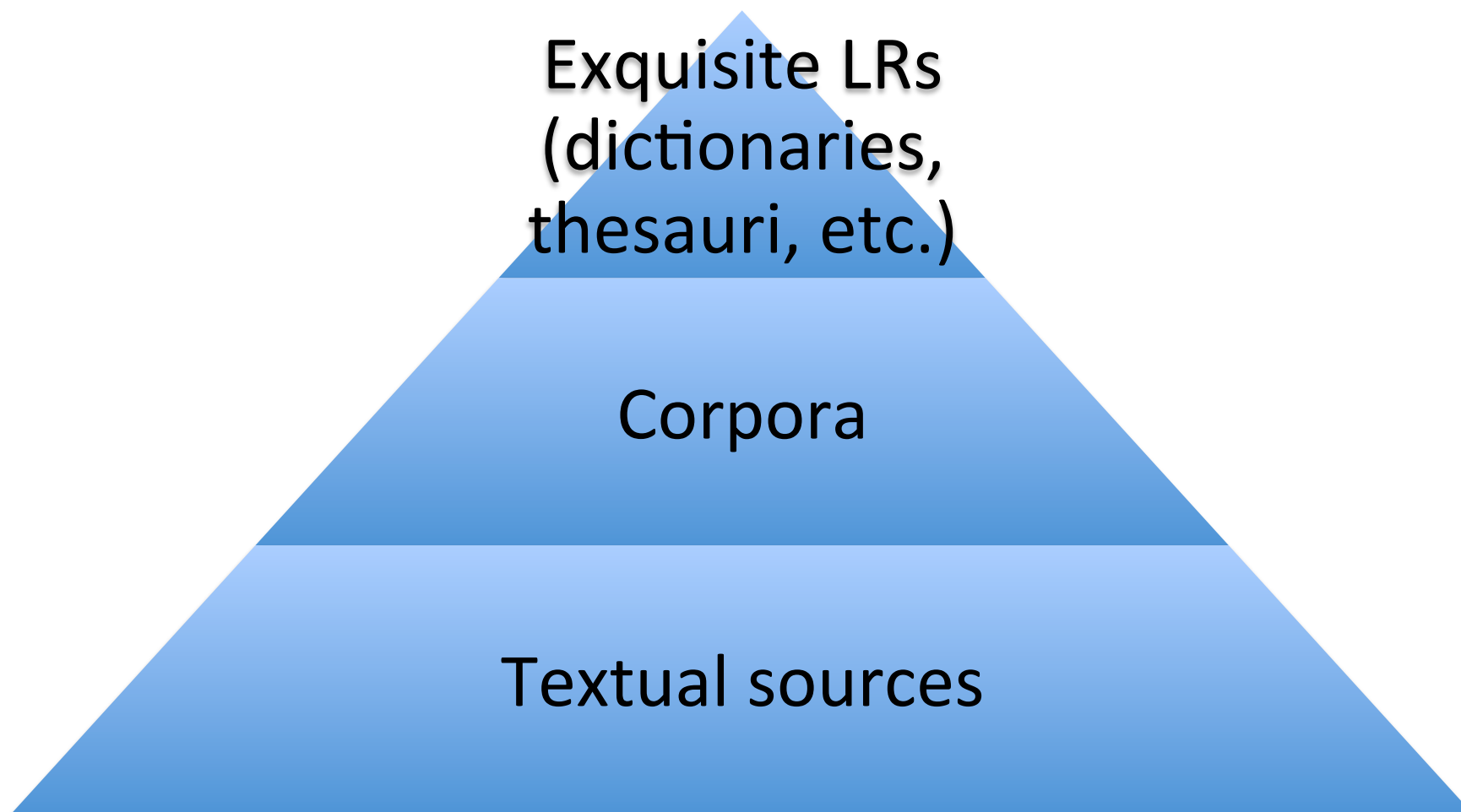
- Long running goals:
 1. Acquire linguistic corpora, on a continuous basis and in more languages
 2. Align them with dictionaries and other LR
 3. Align them among languages
 - technologically
 - content-based

In this talk...

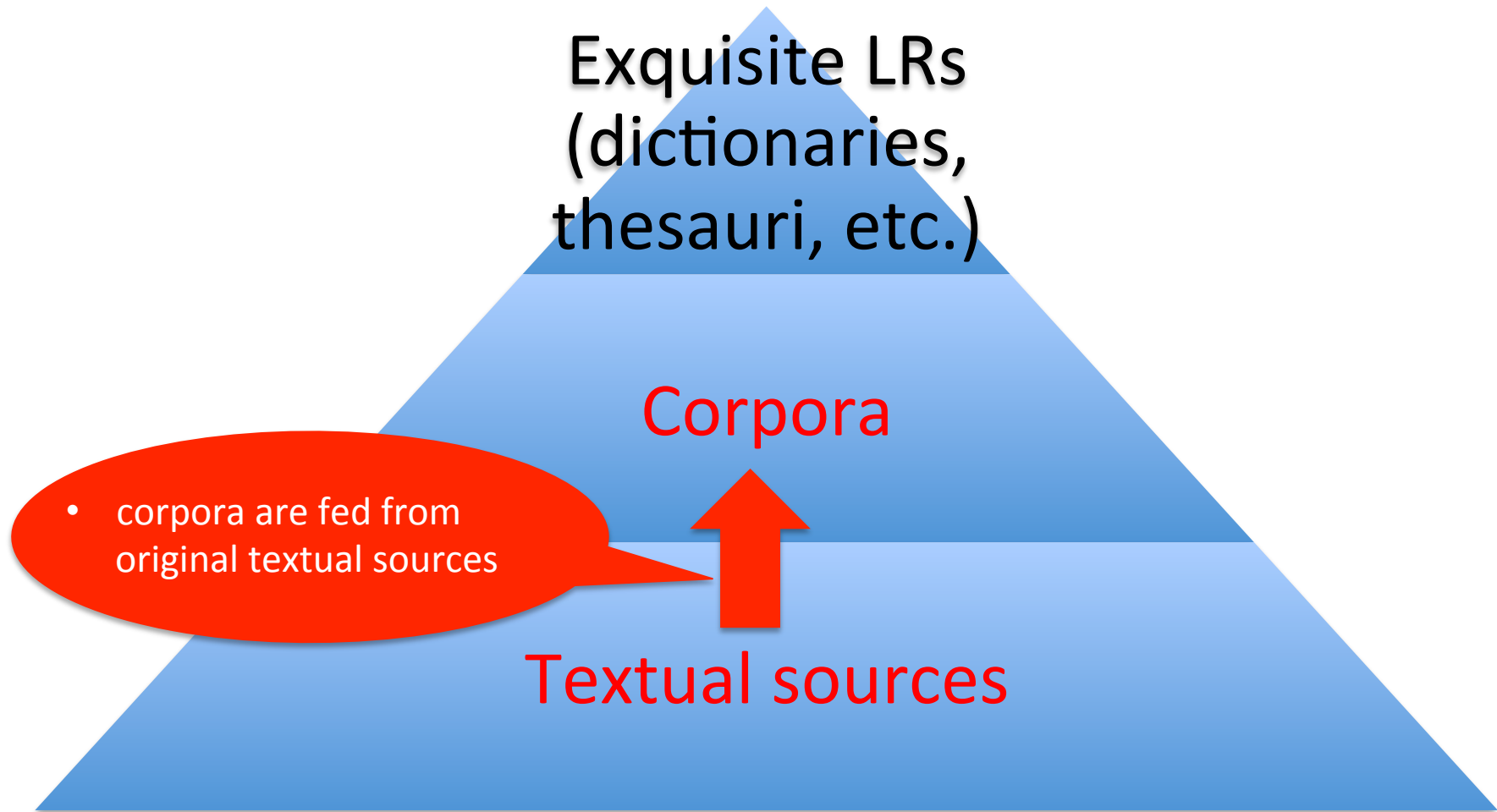
- A methodology of continuously acquiring a language corpus
- Show a pair of languages whose corpora are in the process of being technologically synchronised \leq the first stage of alignment
- Show immediate applications and think for the future

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

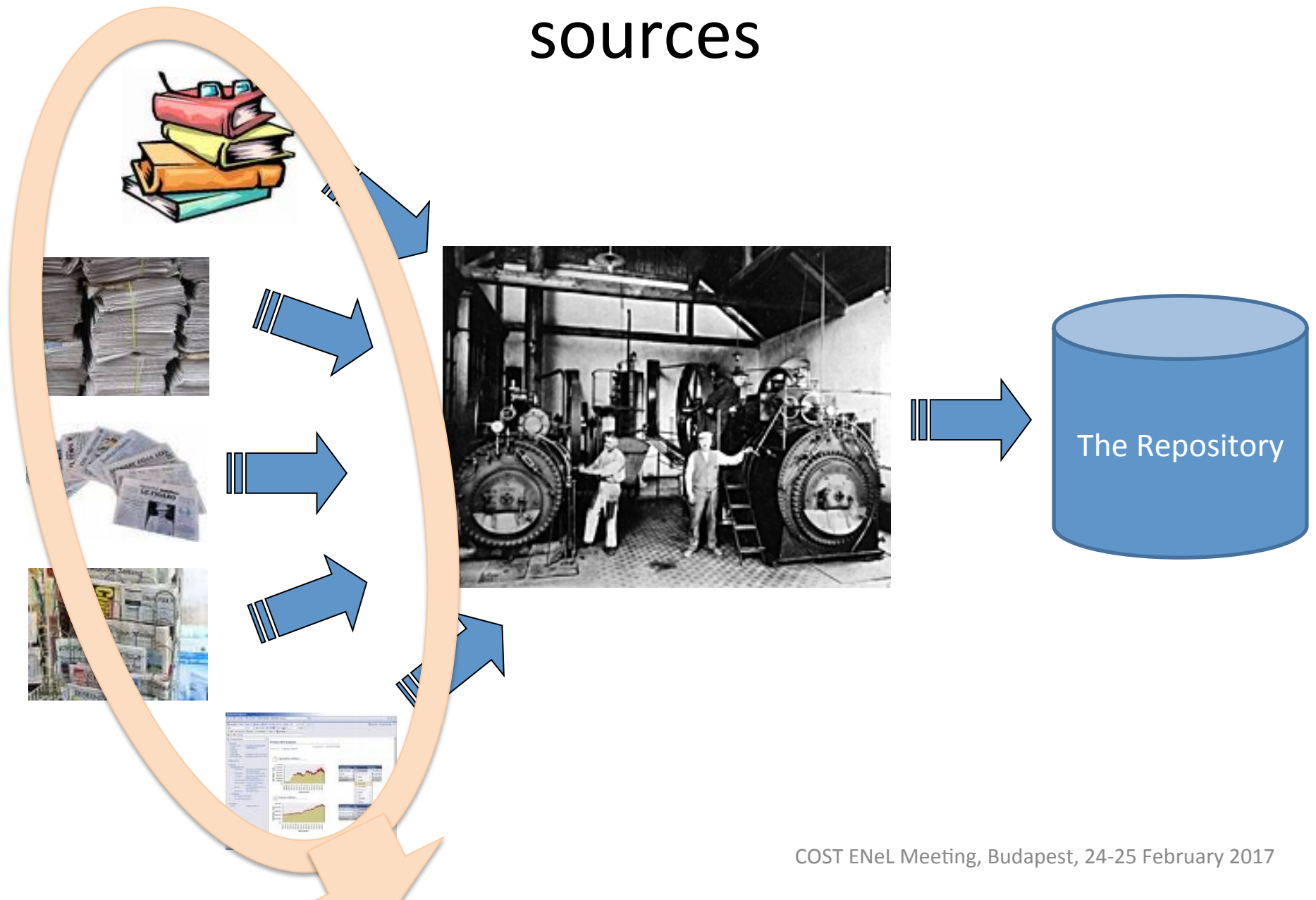
Pyramid of sublimation of LRs for a language

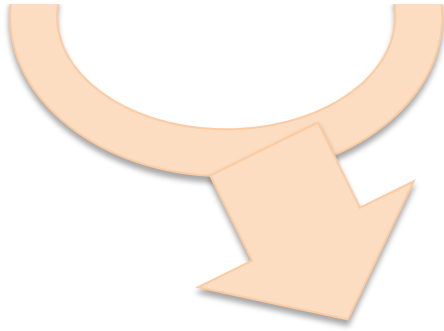


The basic acquisition link



Collecting and processing textual sources



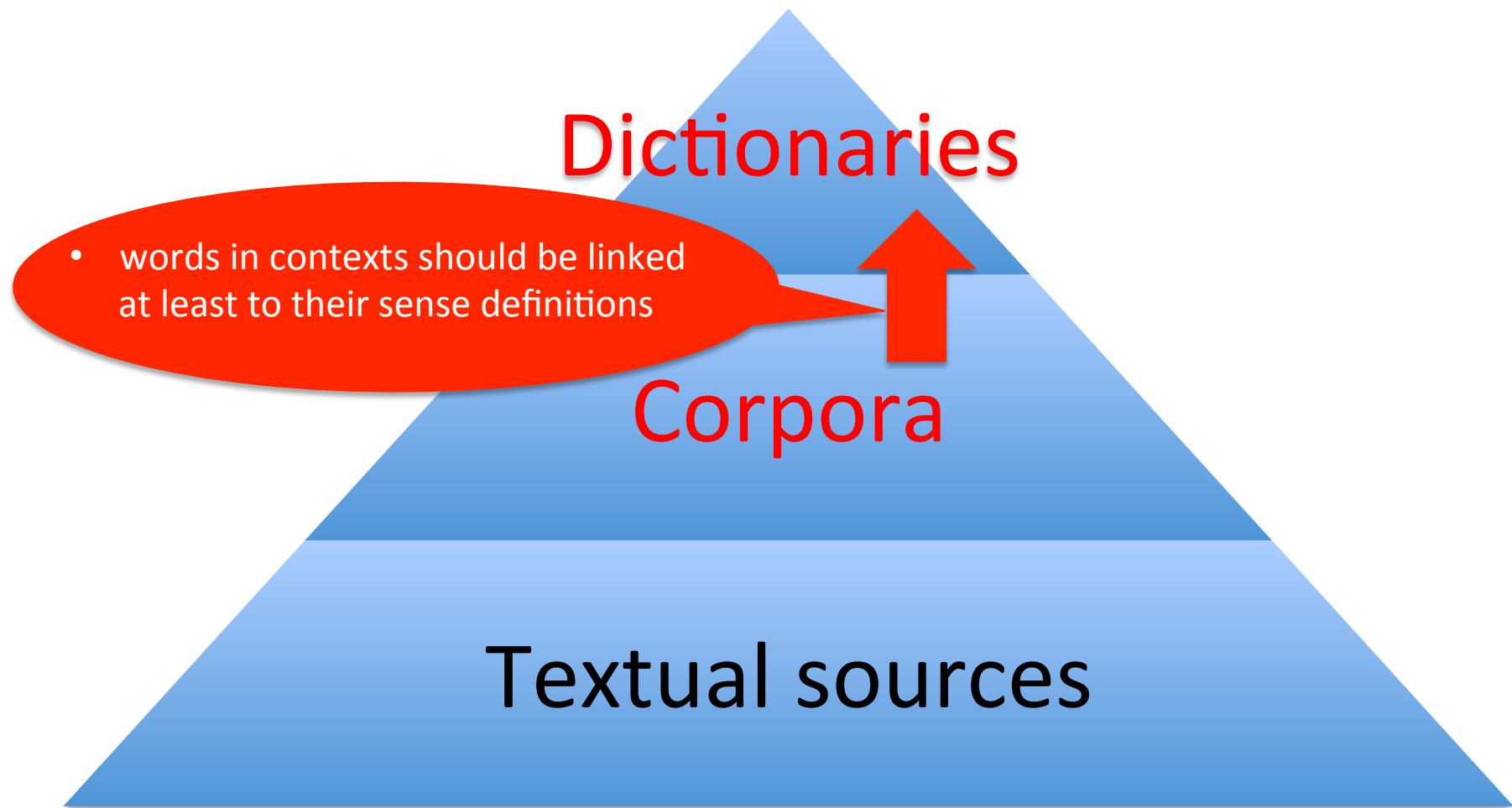


Providers

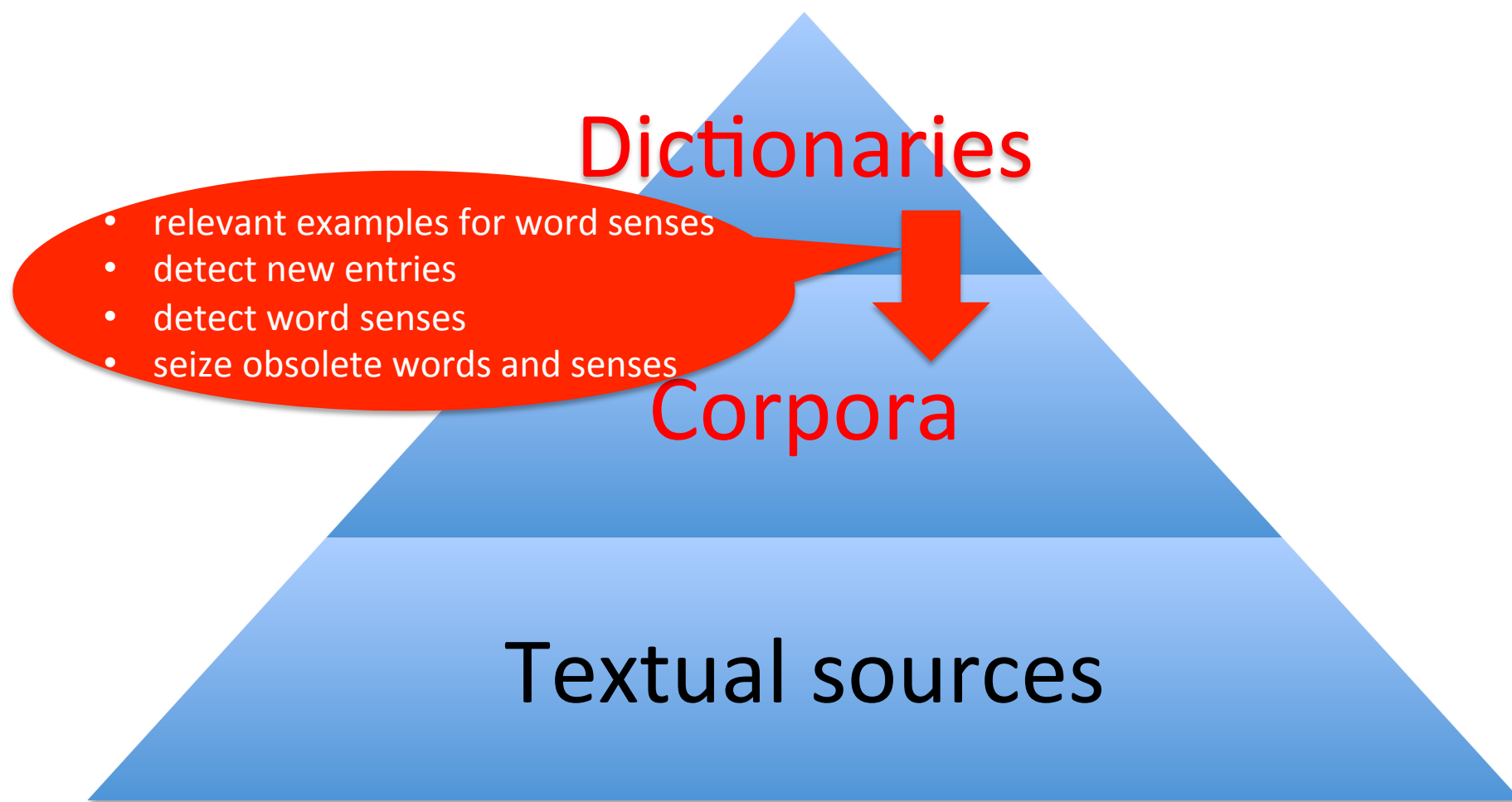
- Editing houses
- Public and particular persons producing printings for public consume
- Recording houses and studios
- Universities, research institutes
- etc.

Willing to donate their textual data!

Linking corpora onto dictionaries



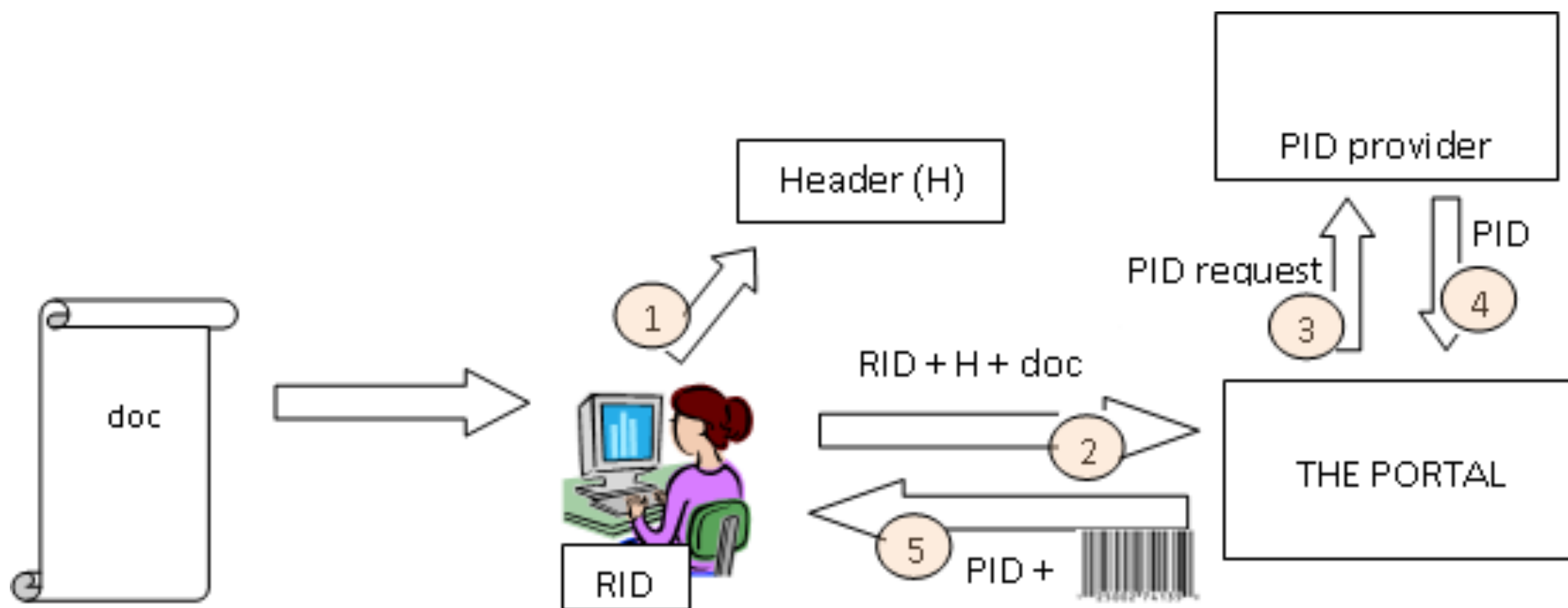
Linking dictionaries onto corpora



In a perfect world...

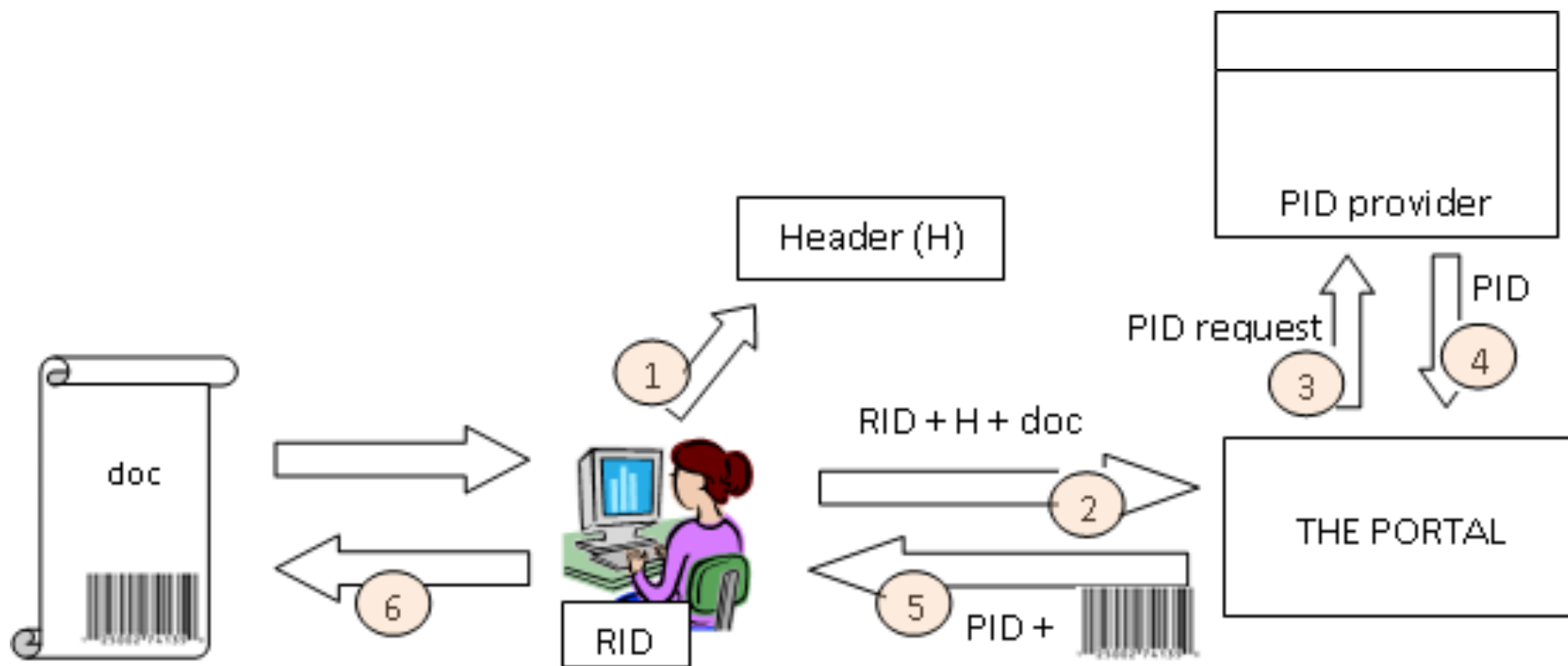
- Language corpora are being accumulated daily, to keep with language evolution
 - Links between words and dictionaries should be permanently updated =>
 - new entries and/or new senses in dictionaries, with corresponding definitions, are being generated when necessary
 - new words and their senses are 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 korpusstechnologisch
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 Iași

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)

(Tufiş *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)

AntConc 3.4.3w (Windows) 2014

File Global Settings Tool Preferences Help

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Concordance Hits 10796

Hit KWIC File

1 _JJ life_NN1 Peers_NN2 is_VBZ to_TO be_VBI made_VVN at_II a_AT1 meeting_NN1 LOB_A.txt

2 _NN1 and_CC he_PPHS1 is_VBZ to_TO be_VBI backed_VVN by_II Mr._NNB Will_NP1 LOB_A.txt

3 _AT House_NN1 of_IO Lords_NP should_VM be_VBI abolished_VVN and_CC that_DD1 Labour_NN1 LOB_A.txt

4 _NN1 Peers_NN2 and_CC Peeresses_NN2 have_VH0 been_VBN created_VVN _YSTP Most_DAT Labour_JJ LOB_A.txt

5 _NP1 conference_NN1 in_II London_NP1 has_VHZ been_VBN boycotted_VVN by_II the_AT two_MC LOB_A.txt

6 te_NN1 Banking_NN1 Committee_NN1 _YCOM which_DDQ is_VBZ headed_VVN by_II another_DD1 Southern_JJ LOB_A.txt

7 _GE nomination_NN1 before_CS it_PPH1 can_VM be_VBI considered_VVN by_II the_AT full_JJ LOB_A.txt

8 doubt_NN1 _YSTP For_CS the_AT Tories_NN2 were_VBDR massed_VVN in_II31 answer_II32 to_II33 LOB_A.txt

9 welfare_NN1 food_NN1 scheme_NN1 _YSTP It_PPH1 was_VBDZ maintained_VVN during_II the_AT war_NN1 . LOB_A.txt

10 _PPH1 seemed_VVD it_PPH1 could_VM not_XX be_VBI carried_VVN on_RP _YSTP When_CS Mr._ LOB_A.txt

11 continue_VVI without_IW either_RR development_NN1 being_VBG limited_VVN or_CC an_AT1 adjustment_NN1 LOB_A.txt

12 _VBG limited_VVN or_CC an_AT1 adjustment_NN1 being_VBG made_VVN in_II financing_NN1 _YSTP LOB_A.txt

13 _GE moves_NN2 towards_II the_AT Six_MC was_VBDZ taken_VVN as_II a_AT1 friendly_JJ LOB_A.txt

14 _AT Prime_JJ Minister_NN1 to_II Paris_NP1 was_VBDZ dropped_VVN _YSTP Instead_RR Mr._NNB LOB_A.txt

15 Lord_NNB Privy_NP1 Seal_NN1 _YCOM who_PNQS is_VBZ charged_VVN with_IW the_AT conduct_NN1 LOB_A.txt

16 1 _YCOM that_CST they_PPHS2 wish_VV0 to_TO be_VBI kept_VVN in_II touch_NN1 in_II LOB_A.txt

17 gotiating_NN1 team_NN1 _YSTP The_AT team_NN1 is_VBZ composed_VVN of_IO experienced_JJ negotiato LOB_A.txt

18 eries_NN2 _YSTP The_AT Foreign_JJ Office_NN1 is_VBZ represented_VVN by_II Sir_NNB Roderick_NP1 LOB_A.txt

19 _AT Six_MC decide_VV0 negotiations_NN2 should_VM be_VBI held_VVN _YSTP Some_DD of_IO the_ LOB_A.txt

20 _YSTP Some_DD of_IO the_AT problems_NN2 were_VBDR reviewed_VVN yesterday_RT at_II a_AT1 LOB_A.txt

Search Term ☒ Words ☐ Case ☐ Regex Search Window Size 50

_VB *_VFN Advanced

Start Stop Sort

Kwic Sort

☒ Level 1 1R ☒ Level 2 2R ☒ Level 3 3R

Clone Results

Applications of DRUKOLA

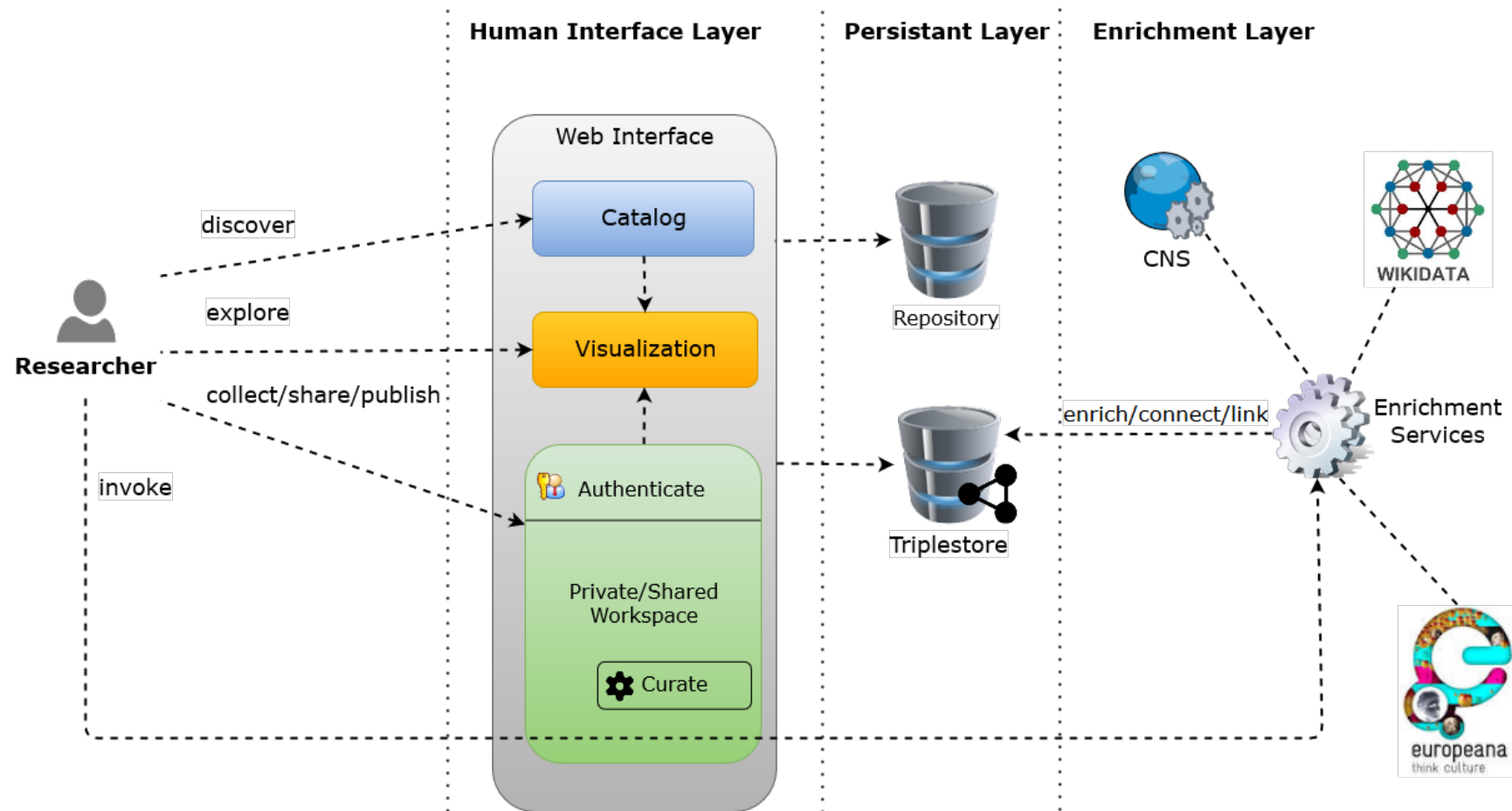
- In a wide range of areas:
 - corpus linguistics
 - computational linguistics
 - applied linguistics, cross-linguistic studies
 - applied computer science
 - research infrastructure development
 - research politics
 - (L1-specific L2-teaching and learning)
 - (machine translation)

Case Study:

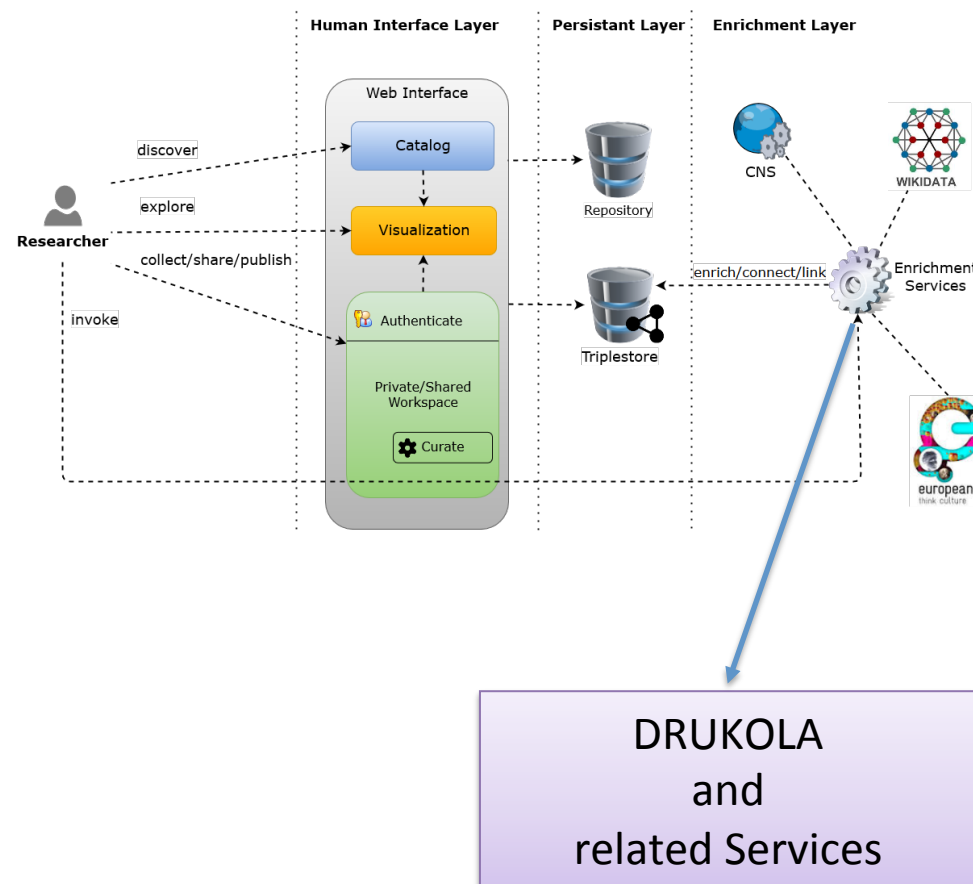
Biodiversity and Linguistic Diversity

- 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., Tufiş, 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.