

META-NORD Overview

Andrejs Vasiljevs
Tilde, Latvia
andrejs@tilde.com

META-FORUM 2012
Brussels, June 20-21, 2012

META-NORD

Baltic and Nordic Branch of the META-NET.

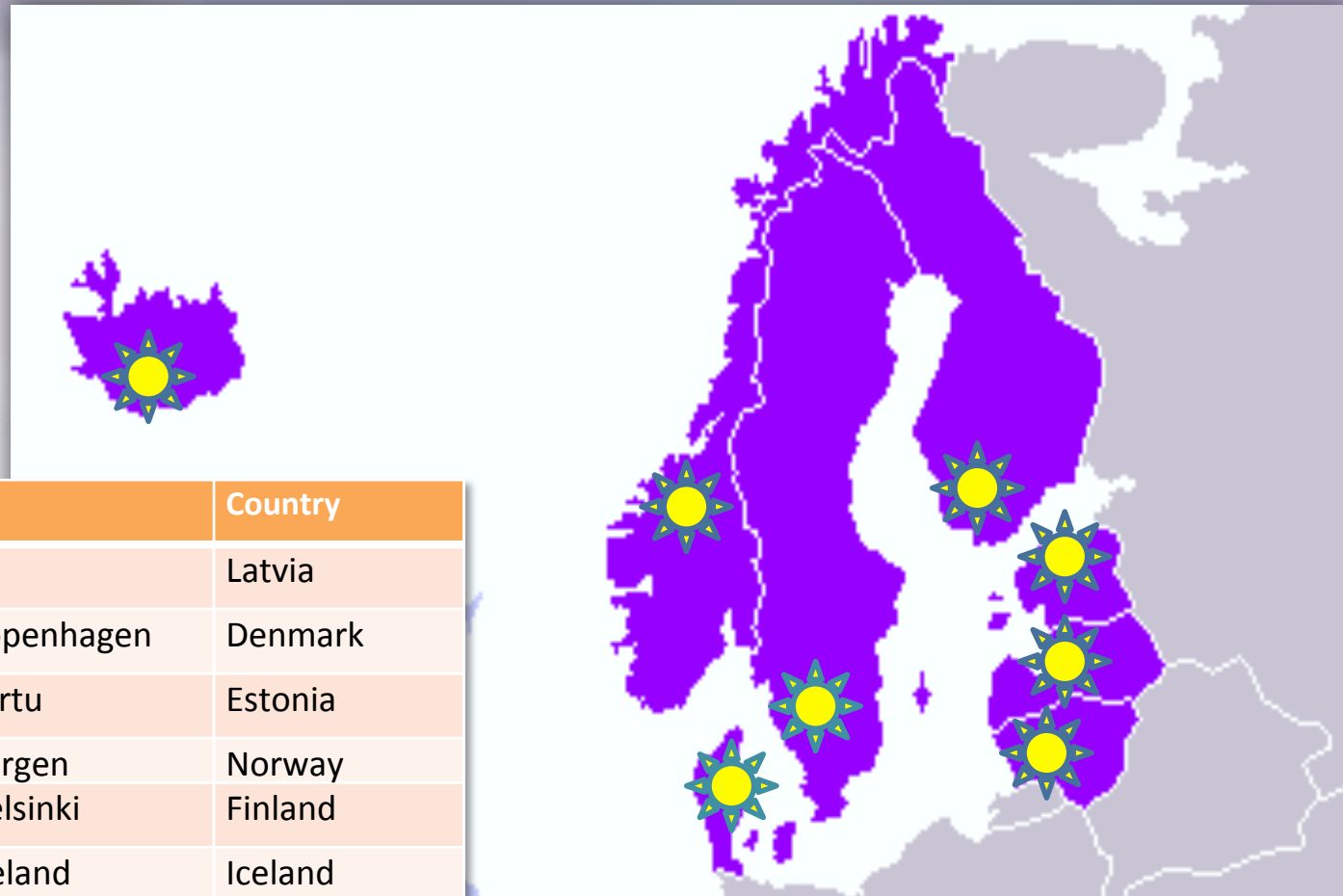
The META-NORD is establishing an open linguistic infrastructure in the:

- **Baltic** - Estonia, Latvia and Lithuania
- and **Nordic countries** - Denmark, Finland, Iceland, Norway and Sweden.

Main objectives

- **Describe the national landscape**
Language Whitepapers
- **Collect resources in the Baltic and Nordic countries** and document, link and upgrade them to agreed standards and guidelines
- **Collaborate**
with the META-NET network of excellence and other partner projects
- **Help build and operate**
broad, non-commercial, community-driven, inter-connected repositories, exchanges, and facilities
- **Mobilize national and regional actors,**
public bodies and funding agencies by raising awareness

META-NORD Geography



Partner	Country
Tilde	Latvia
University of Copenhagen	Denmark
University of Tartu	Estonia
University of Bergen	Norway
University of Helsinki	Finland
University of Iceland	Iceland
Institute of Lithuanian Language	Lithuania
University of Gothenburg	Sweden



One of the key members in the META- NET family

META-NET Project family:

- T4ME – Core parts
- CESAR
- METANET4U
- META-NORD

Strong cooperation and interdependency:

- META-SHARE platform
- Working Groups
- Language Whitepapers
- Metadata formats
- IPR management
- Dissemination

Focus

- Focus on European languages with **less than 10 million speakers**
 - EU official languages – Danish, Finnish, Swedish, Estonian, Latvian and Lithuanian
 - Languages of the European Economic Area – Icelandic and Norwegian
- For many META-NORD languages only **limited high-quality language resources** are currently available
- Non-textual resources have been created only for some META-NORD languages

Horizontal actions

WordNets

monolingual WordNets and cross-linked pilots
Danish, Estonian, Finnish and Icelandic

Treebanks

treebanks integrated on a uniform platform and linked across languages using parallel multilingual treebanking
Danish, Estonian, Finnish, Icelandic and Norwegian

Terminology

distributed terminology resources across languages and domains consolidated across *META-NORD and other META-NET languages*

Specific targets

- Facilitate **availability of BLARK resources** for META-NORD languages
- **Provide expertise** to the META-NET in the fields where META-NORD partners have outstanding expertise
- **Develop and document methodologies** for building language resources for under-resourced languages with focus on **semi-automatic/machine assisted resource generation**
- Facilitate **knowledge transfer** between CLARIN and META-NORD, especially on standards and intellectual property rights (IPR)

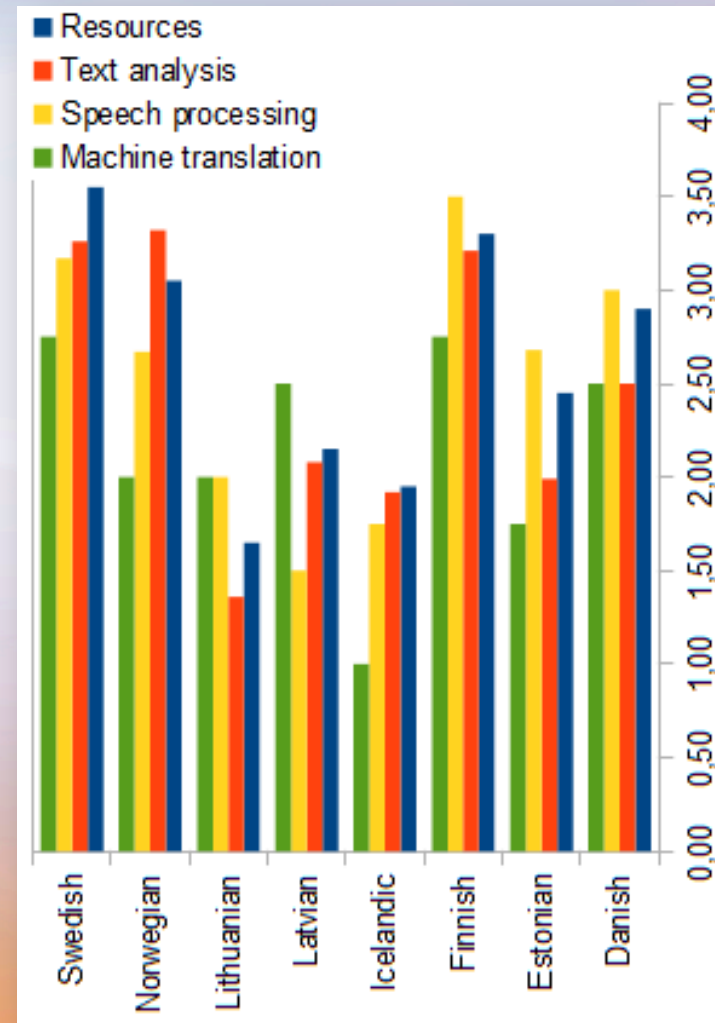
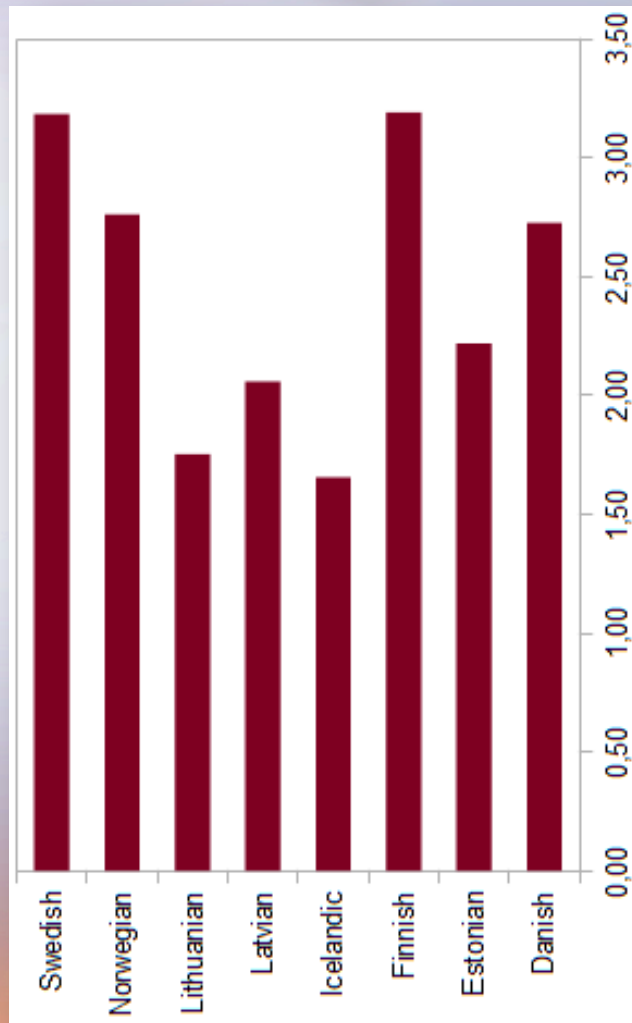
Findings of Language Whitepapers

- Reasonably positive results for the most basic LRs - tokenizers, PoS taggers, morphological analyzers/generators, syntactic parsers, reference corpora, and lexicons/terminologies.
- There are parallel corpora, speech corpora, and grammar resources for some META-NORD languages, though these are limited in size and functionality.
- There is a lack of resources in more advanced fields, such as LR&Ts for sentence and text semantics, information retrieval, language generation, and multimodal data.

LT Comparative availability

	Excellent	Good	Moderate	Fragmentary	Weak/No
Speech Processing			Finnish	Danish, Estonian, Norwegian, Swedish	Icelandic, Latvian, Lithuanian
Machine Translation					Danish, Estonian, Finnish, Icelandic, Latvian, Lithuanian, Norwegian, Swedish
Text Analysis				Danish, Finnish, Norwegian, Swedish	Estonian, Icelandic, Latvian, Lithuanian
Resources			Swedish	Danish, Estonian, Finnish, Norwegian	Icelandic, Latvian, Lithuanian

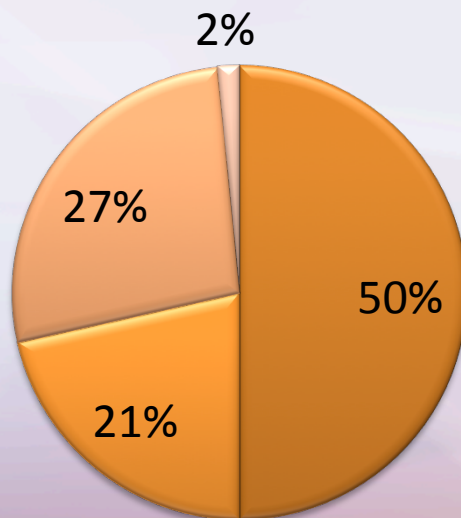
LT Comparative availability



Language Resources Identified

155 LR identified and registered

- Partners' resources
- 3rd parties - available
- 3rd parties- pot. available
- 3rd parties - unclear



Resources in the First Batch

Resources in First Batch	
36	Lexical resources
16	Corpora
6	Treebanks
5	Resources for speech
3	WordNet
1	Tool
67	Total

- 109 LRTs fit well for language technology development
- 47 are multilingual
- 103 are well maintained

Some of the LRs in the Pipeline

Latvian-English legislation corpus of Republic of Latvia
 Corpus of Latvian literature
 Danish [wordnet](#)
 Copenhagen Dependency [Treebanks](#)
 The Copenhagen Danish-English Dependency Treebank
 The Comprehensive Corpus of Estonian Treebank
 Estonian [WordNet](#)
 Corpora of morphologically disambiguated texts
 Corpora with shallow syntactic annotation
 English-Estonian and Estonian-English parallel corpus
 Semantically disambiguated corpus
 The database of Estonian verbal multi-word expressions
 Morph syntactic disambiguator and shallow parser
[Leksikografisk bokmålskorpus](#) Searchable
[Det nynorske tekstkorpuset](#) Searchable
[Akustisk database for norsk \(NST\)](#)
 NHH [Termbase](#) Searchable

Norwegian-Vietnamese digital dictionary Searchable
 NST lexicon
[Stadsnamnsamlinga](#) Searchable
 Oslo-Bergen tagger
 Terminology database [Snorre](#)
 International Computer Archive of Modern and Medieval English Downloadable
 International Computer Archive of Modern and Medieval English Searchable
 Norwegian Newspaper corpus Searchable
 Translation Corpus Aligner 2 [Sofie](#) Treebank
[Acquis communautaire](#) Written corpora (old literary Finnish)
 Finnish [TreeBank](#)
 Cross-lingually linked resource
 Cross-lingually linked resource
 Cross-lingually linked resource
 Helsinki Finite-State Transducer Technology
 Finnish [WordNet](#)
 Samples of Spoken Finnish ([Suomen kielen näytteitä](#))
 Speech and EGG (electroglottography) simultaneous recordings

([Puheen ja EGG:n samanaikaiset tallenteet](#))
 Open Source (Finnish) Morphology
[Morfessor](#)
 National Semantic Web Ontology Project in Finland
 TKK Voice Source Analysis and [Parametrisation](#) Toolkit
 Corpus of early modern Finnish ([Varhaisnykysuomen korpus](#))
 Finnish literature classics ([Suomalaisen kirjallisuuden klassikoita](#))
 Up-to-date word list of modern Finnish ([Ajantasainen nykysuomen sanalista](#))
 Frequency list of words in written Finnish ([Kirjoitetun suomen kielen sanojen taajuuslista](#))
[CombiTagger](#)
 IceNLP - Tagger, Parser, [Lemmatizer](#)
[Apertium-is-en](#) Translation System
 Icelandic Frequency Dictionary Corpus (web version)
 Balanced Tagged Icelandic Corpus (web version)
 Icelandic Parsed Historical Corpus
 The [Jensson](#) Corpus
 The Thor Corpus

The Broadcast News RUV-1 Corpus
 Parliament Speech Corpus
[Hjal](#) Speech Corpus
 Pronunciation Dictionary for Icelandic
 Database of Modern Icelandic Inflections (web version)
 Database of Semantic Relations
 Icelandic [WordNet](#) - Pilot Project
[Íslenskur orðsafiður](#) - Large Corpus 8web version)
 Icelandic Term Bank - Terminology (web version)
 ISLEX - Icelandic Dictionary Base (web version)
 Ministry for Foreign Affairs - Translation Centre - Dictionary (web version)
 Modern Lithuanian Dictionary
 Dalin's morphological dictionary
 Old Swedish morphology
 Loan Word Typology list
 Preparatory Action for Linguistic Resources
 Organization for Language Engineering
 Swedish Associative Thesaurus
 Examples from the Swedish Associative Thesaurus

Some Highlights: Treebanks

- Work on Parallel Treebanks for “Sophie’s World”

iness
Signed in as *gunn*. [Sign out](#) | [Edit is](#)

Parallel Sentences

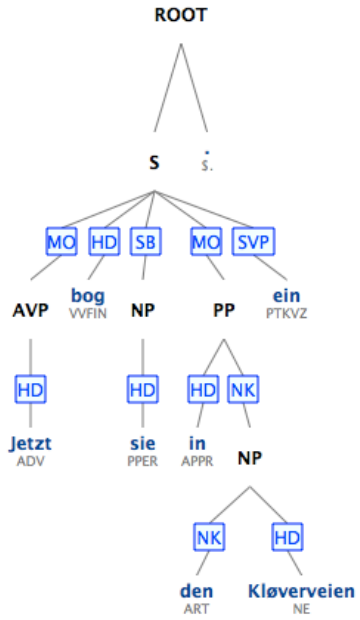
Source treebank: , target treebank:

#12: **Jetzt bog sie in den Kløerveien ein .**

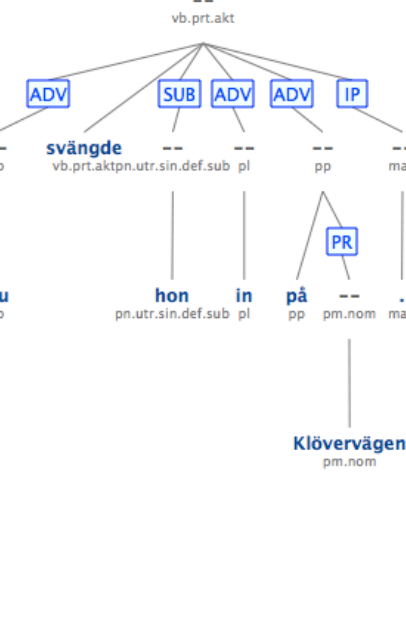
#13: **Nu svängde hon in på Klövervägen .**

hide settings | go to #: | ordered

ROOT



--
vb.prt.akt



Some Highlights: Consolidation of Terminology Resources

Search results

Translations View
Entries View

computer

computer language

computer name

computer science

computer security

computer system

EN	computer language		
DA	computersprog	•	information technology and data processing
FI	konekieli	•	information technology and data processing
LT	kompiuterinė kalbà	•	information and information processing information technology and data processing ...
LV	datorvaloda	•	information and information processing communications ...
	mašīnvaloda	•	information and information processing communications ...
	mašīnkods	•	information and information processing communications ...
SV	datorspråk	•	information technology and data processing

Display options

show source

show domains

show definitions (2)

Filter by domain

communications

information and information processing

information technology and data processing

natural and applied sciences

Select all / Select none

Filter by language

DA (1)

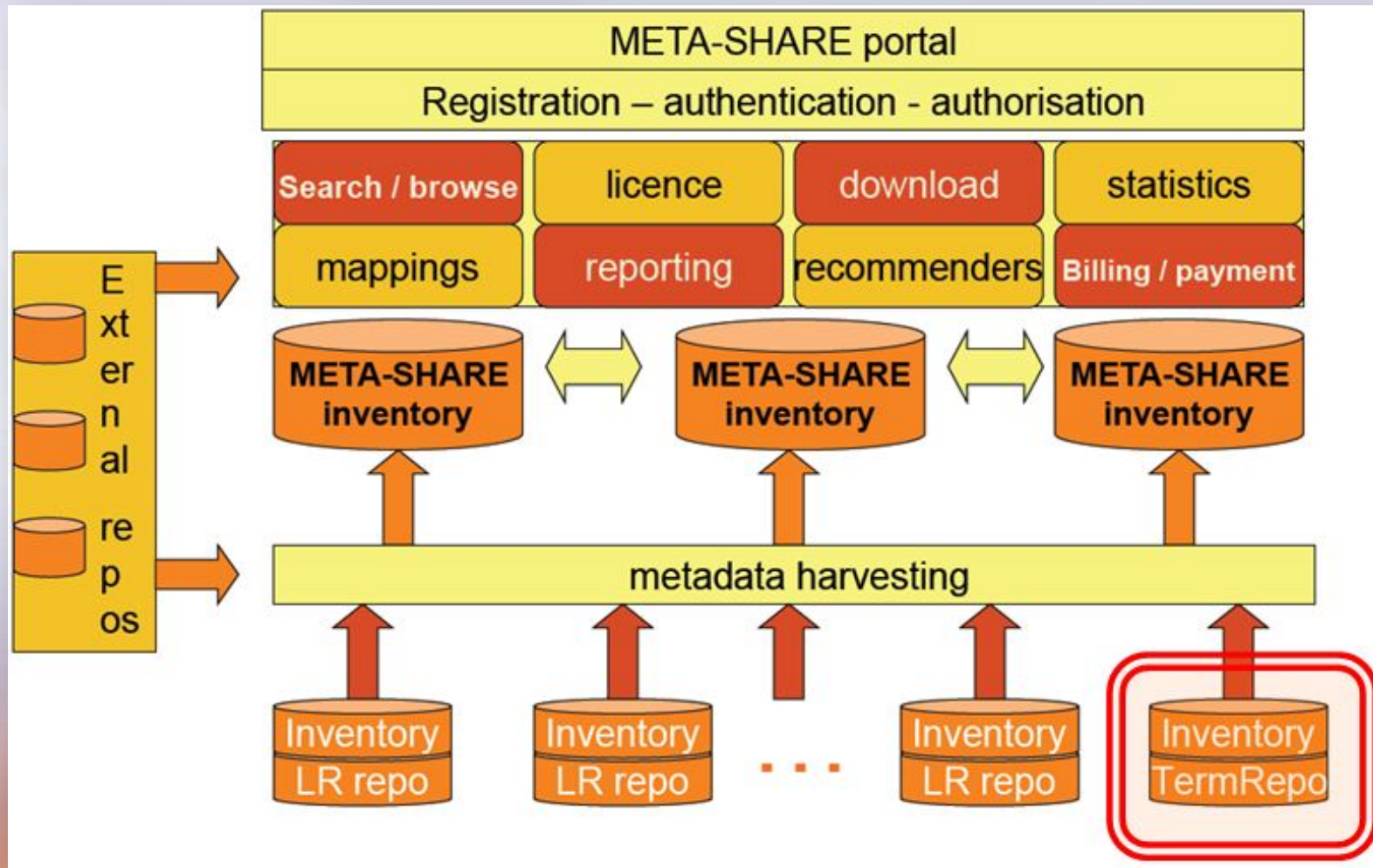
FI (1)

SV (1)

LV (1)

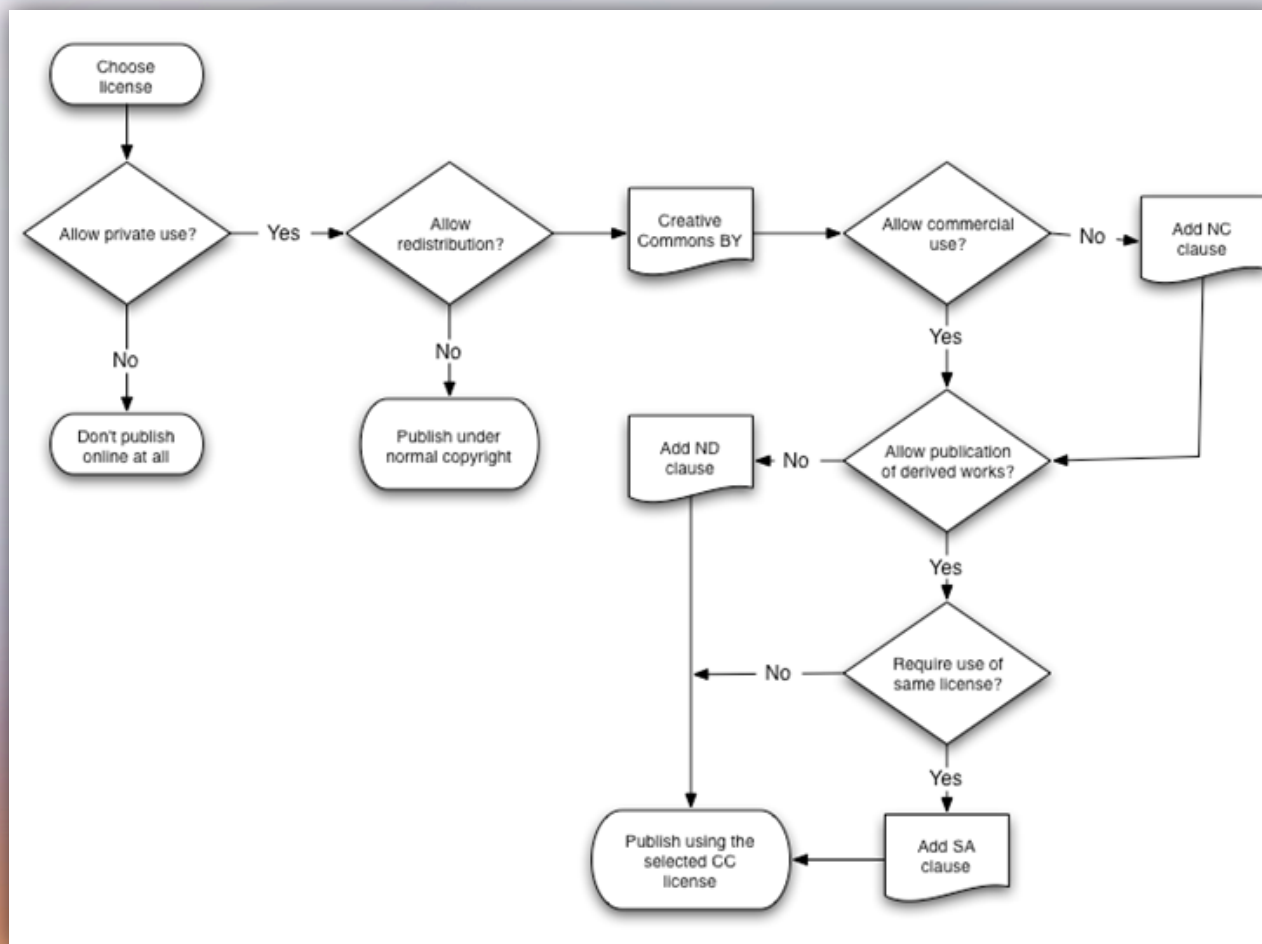
LT (2)

Some Highlights: Work on Dedicated Terminology Node as Part of the META-SHARE



Some Highlights: IPR

- Approach for choosing the right open content license for particular resource



Some Highlights

- META-SHARE nodes operating at the University of Gothenburg, Tilde and University of Helsinki
- Cooperation with FP7 R&D and ICT PSP projects on resource sharing
 - Comparable corpora – ACCURAT, TTC
 - Resources for Machine Translation – LetsMT!, ACCURAT, EASTIN-CL
 - Terminology Resources – TTC, EASTIN-CL
- Cooperation with CLARIN
- Considerable mobilization and awareness in the Nordic and Baltic countries, in particular at the European day of Languages

Sustainability

- Commitments from the partner institutions:
 - Hosting META-SHARE nodes
 - Providing their LRs through META-SHARE
 - Providing technical and user support services
 - Participating in the software development of the platform
 - For at least 24 months
- Participation in the related national programmes
- Integrating and running resource-specific nodes, e.g., collaboration in terminology resources and services with the TaaS project

Thank you!

meta-nord.eu
meta-net.eu

Contact information:
Andrejs Vasiljevs
andrejs@tilde.com

The work within the project **META-NORD** has received funding from the ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme
Grant agreement no 270899

