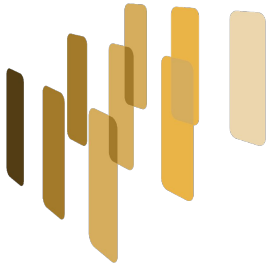
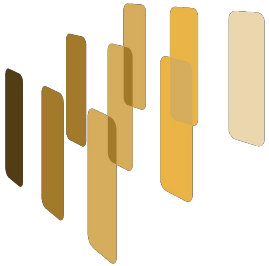


EUROPEAN LANGUAGE GRID

Language Technology Market and Components Taxonomy

Gerhard Backfried (Sail Labs),
Penny Labropoulou (ILSP),
Artem Revenko (SWC),
Thomas Thurner (SWC)

Workshop @ LT Innovate



EUROPEAN LANGUAGE GRID

An Overview of the European Language Grid Project and Initiative

The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 825627 (ELG).

Gerhard Backfried (SAIL LABS)
gerhard.backfried@sail-labs.com

8th Language Technology Industry Summit, Brussels, Belgium, 25 June 2018

EUROPEAN LANGUAGE GRID

Primary Objectives

1. Establish the **ELG** as the **primary platform for LT in Europe**
2. ELG as a platform for **commercial and non-commercial LTs**, both **functional and non-functional**
3. Enable the European LT community to **upload services and data sets** into the ELG, to **deploy** them and to **connect** with, and make use of those resources made available by others
3. Unleash enormous potential for **innovation** (**pilots** etc.)
4. Establish the European Language Grid as the **primary European market place for LT to connect demand and supply**
5. Help establish the **Multilingual Digital Single Market (DSM)**

EUROPEAN LANGUAGE GRID

Consortium

Relevant Initiatives

- META-NET and META
- Cracking the Language Barrier
- ELRC, CEF AT Tools & Services
- CLARIN
- LT Innovate
- Big Data Value Association
- AI4EU, HumanE AI
- EOSC and several others

https://www.european-language-grid.eu/

EUROPEAN LANGUAGE GRID

Objectives

- Establishment of the **Multilingual DSM**
- **Scalable Platform** and **Market Place** for LT companies & Research
- **Vibrant and active community** around the ELG
- Being a product *from* the European LT community *for* the European LT community, **new resources and services** will be added continuously
- ELG as an **infrastructural** antidote to digital language extinction
- **Strengthen European LT sector** (vs. US and China)

EUROPEAN LANGUAGE GRID

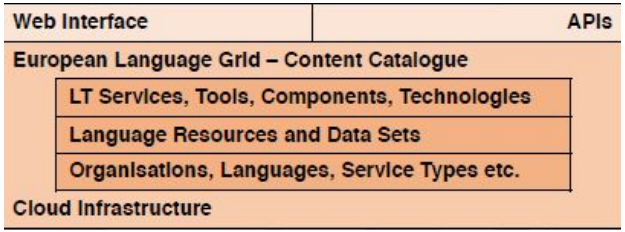
Concept and Methodology

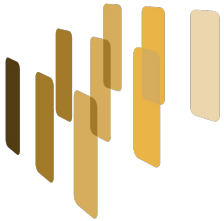
- **Grid Platform**
  - **Catalogue** of functional services, data sets, tools, technologies, models etc.
  - **Catalogue** of LT companies, research organisations, service and application types, languages etc.
  - **Information** about conference and training events, pilot projects, open calls and other pieces of content.
- **Grid Content**
- **Grid Community**

EUROPEAN LANGUAGE GRID

Concept and Methodology

- **Grid Platform**
- **Grid Content**
  - **Content:** tools/services, data resources (corpora, lexica, terminological lists, models, etc.)
  - **Functional Content:**
    - running services
    - download and integrate into other systems
    - upload and make available
  - **Non-functional Content:**
    - data resources (non-running)
    - catalogue of companies, LT business areas, projects, etc.
  - **LT-as-a-Service**:
    - create an image of a LT service locally that can be run on a different system through a VM.
    - Easy and efficient for LT providers to offer LT services & products
- **Grid Community**





EUROPEAN LANGUAGE GRID

Concept and Methodology

- **Grid Platform**
- **Grid Content**
- **Grid Community**
  - **Stakeholders**: LT provider and buyer companies, industry, research centres, universities, administrations, NGOs etc.
  - **15-20 pilot projects (starting in 03/2020)**
    - https://www.european-language-grid.eu/open-calls/
  - Organisation of **conferences and events** (training, presentations, publications, social media, blog posts etc.)
  - **European LT Board** to establish an international, pan-European body, in which LT-related matters can be discussed and coordinated

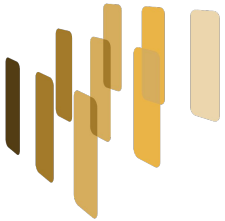
EUROPEAN LANGUAGE GRID Platform Catalogue - 4 Layers

	Type	Description	ELG Consortium brings in
1)	**ELG Language Technology Tools, Services – functional Grid content**	Containerised services (from, e.g., tokenisation or POS tagging to complex workflows); can be uploaded and deployed through the ELG; can be integrated in other systems through the ELG	GATE, GATECloud, DKT, UDPipe, Tilde's and Edinburgh's MT services, Weblicht, SAIL's ASR, KWS, sentiment (polarity) detection, age & gender detection tools, etc.
2)	**ELG Language Technology Tools, Services – functional Grid content, remotely invoked**	Remote APIs (REST) can be registered, described, searched and integrated with the help of the ELG platform	Research and commercial services (e.g. GATE Cloud with 65+ NLP services, UDPipe), TILDE's and UEDIN's MT services, SAIL's and EXPSYS services, etc.
3)	**ELG LRs/LTs – non-functional Grid content**	Upload, describe, search, download of corpora, source code programs, models etc.	Data resources from META-SHARE, ELRC-SHARE, ELRA (e.g. corpora, lexica, models etc.)
4)	**ELG Meta-Information**	ELG catalogue entries, e.g., of an LT provider company (no language processing functionality or code available in ELG, no data)	CRACKER, ELRA, ELRC (NAPs etc.), META-SHARE, CLARIN VLO, LingHub, LT World etc.

EUROPEAN LANGUAGE GRID

Impact

- **Impact on Business and Industry**
  - Tackle fragmentation
  - Enable EU SMEs to expand their business online across many languages
  - Open up markets and foster growth
  - Connect demand and supply
  - Reduce costs
  - Reduce time-to-market
- **Impact on Innovation**
  - Fast and efficient experimenting with new methods and technologies
  - Rapid markets development and penetration
  - Business growth
- **Impact on the Digital Single Market**
  - Elimination of language barriers (unlock DSM)
  - Products & services in more languages, taking advantage of the DSM



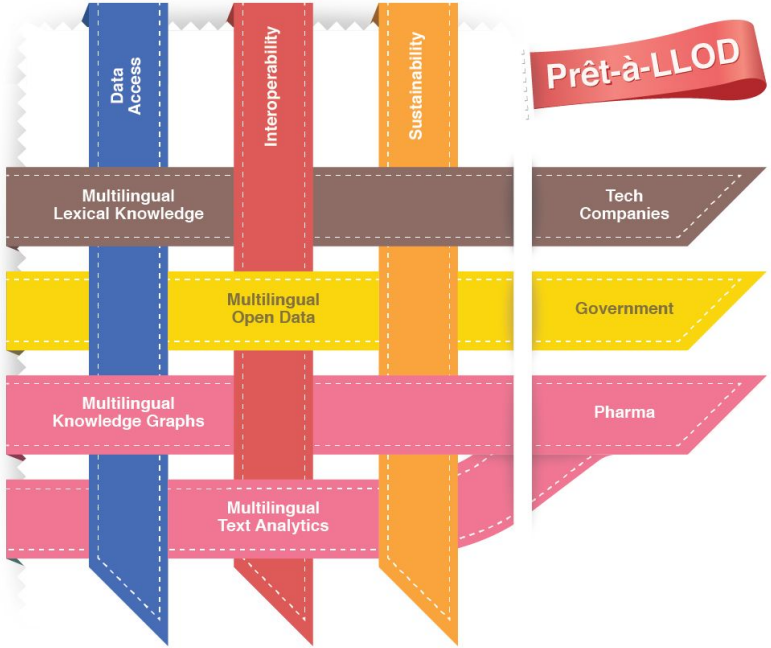
EUROPEAN LANGUAGE GRID

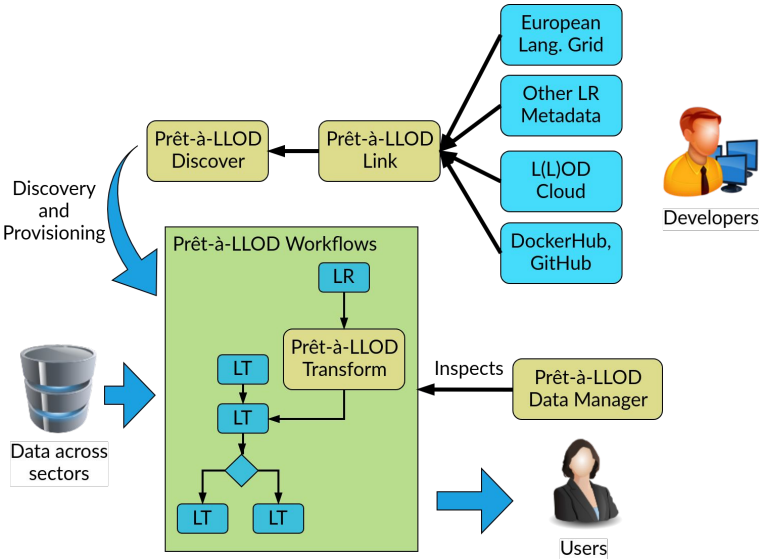
Unique Selling Proposition

- The **ELG** will be the **primary platform for LT in Europe**, uniting a network of European experts in the field
- **One-stop shop** for LT in Europe
- As a **marketplace** for the **European LT business space**, it will strengthen Europe's position in this field and create new jobs and incentives for high potential talent to stay in Europe

Prêt-à-LLOD

Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors





NUI Galway OÉ Gaillimh

Universidad Zaragoza

German Research Center for Artificial Intelligence

Universität Bielefeld

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

OXFORD UNIVERSITY PRESS

SEMANTIC WEB COMPANY linking data to knowledge

Derilinx

Prêt-à-LLOD

Objectives:

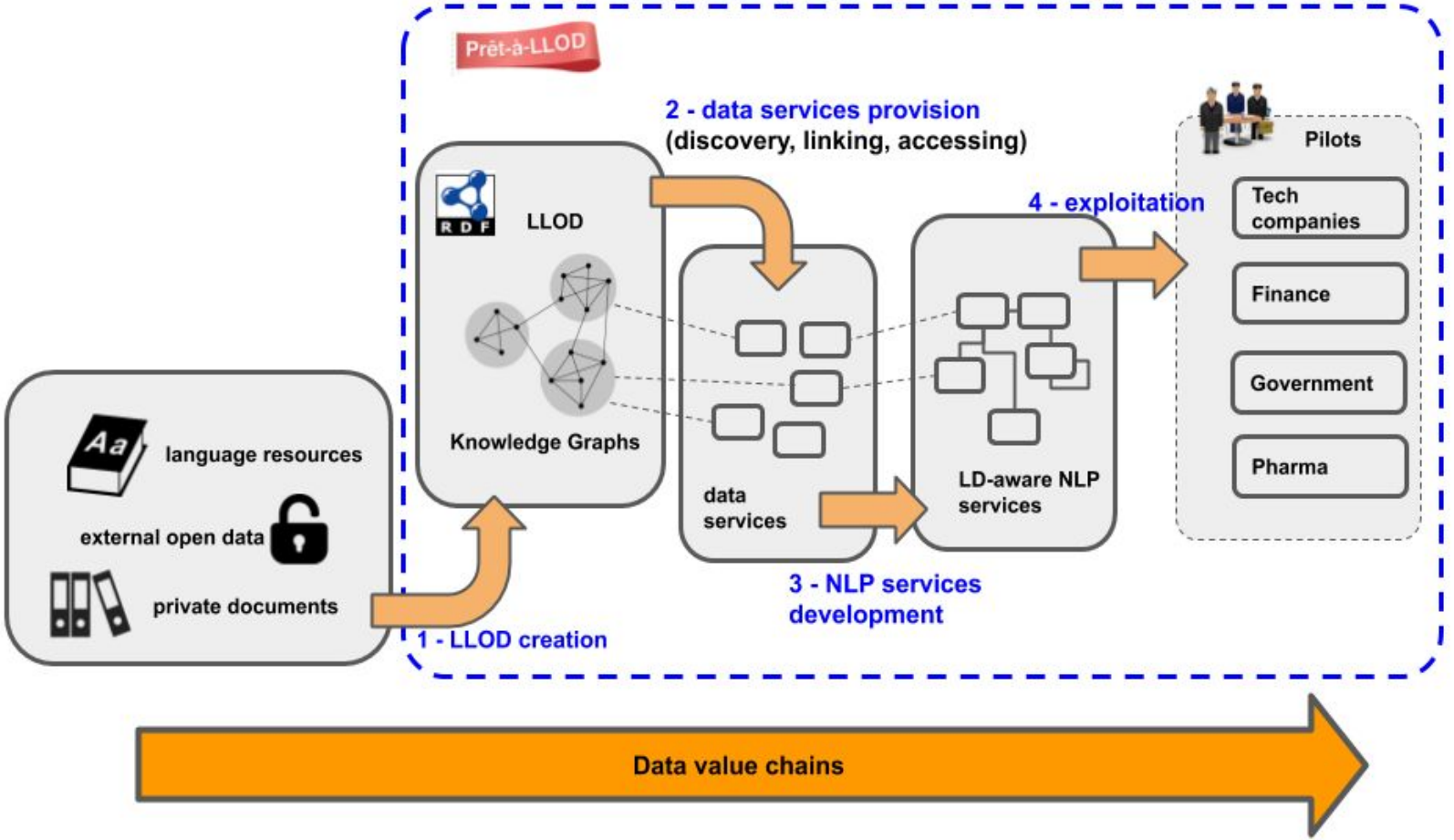
- Multilingual cross-sectoral **data access**
- **Interoperable** language technology services and language data
- **Sustainability** of language technologies and language resources

Timeframe: 01.01.2019 - 31.12.2021
Research and Innovation Action:
Domain-specific/challenge-oriented Human Language Technology
10 partners: 6 academic + 4 commercial

Targets:

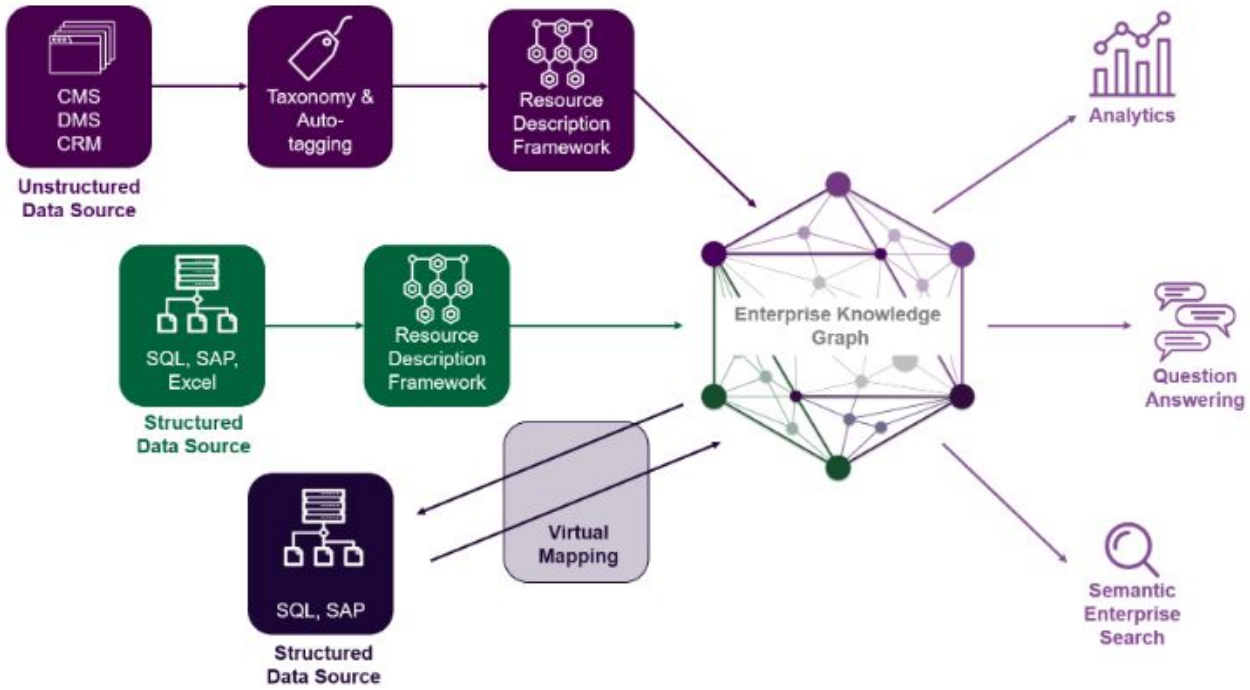
1. Provide European research and language technology industry with a **better access** to and usage of **quality** language resources and tools
2. **Increase** in **the quality and coverage** of multilingual solutions used by industrial players in sectors relevant to the emergence of the DSM
3. **Increase** in **the uptake** of language technologies in Europe in various sectors
4. **Cost savings** for private and public sector' users of language technology solutions

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

What a knowledge graph is good for.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Discover LT-Resources

Prêt-à-LLOD Discovery

A new portal for the discovery of language resources

Prêt-à-LLOD Data Manager

Automatic inference of licensing restrictions across multiple datasets

Prêt-à-LLOD Workflows

Scalable, highly portable workflows for text analytics

Prêt-à-LLOD Transform

Smart AI-driven tool for the conversion of existing data to linked data

Prêt-à-LLOD Link

Semi-automatic procedure for linking datasets

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Categorize LT-Resources

Prêt-à-LLOD Discovery

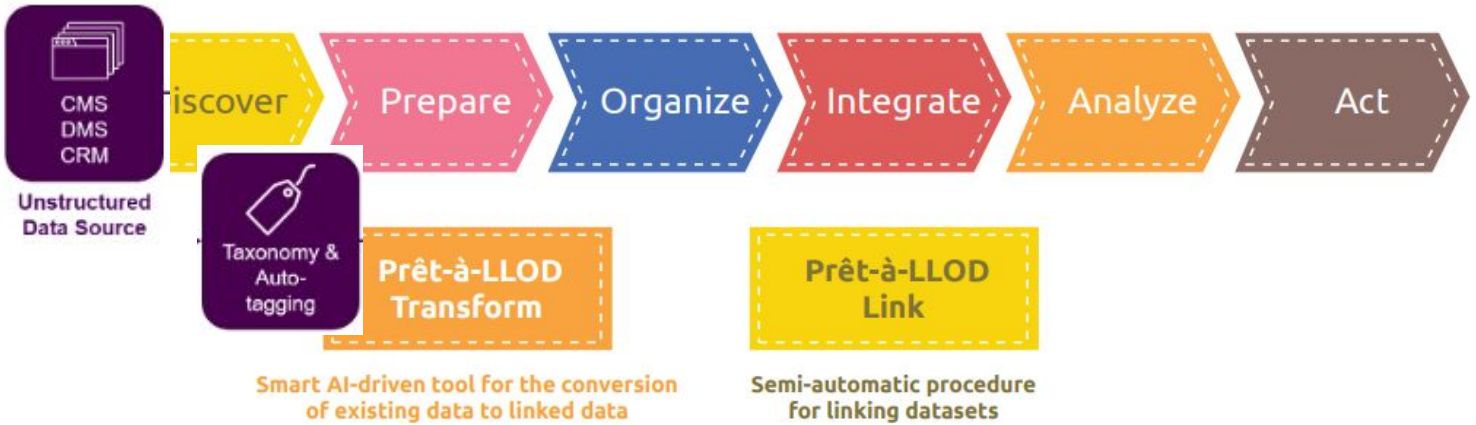
A new portal for the discovery of language resources

Prêt-à-LLOD Data Manager

Automatic inference of licensing restrictions across multiple datasets

Prêt-à-LLOD Workflows

Scalable, highly portable workflows for text analytics



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Catalogue LT-Resources

Prêt-à-LLOD Discovery

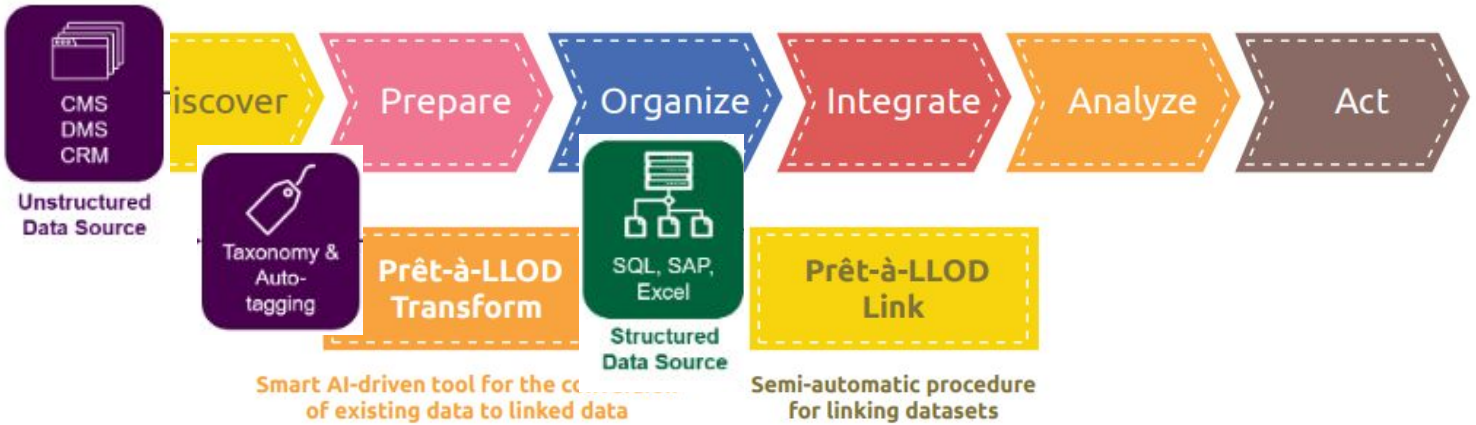
A new portal for the discovery of language resources

Prêt-à-LLOD Data Manager

Automatic inference of licensing restrictions across multiple datasets

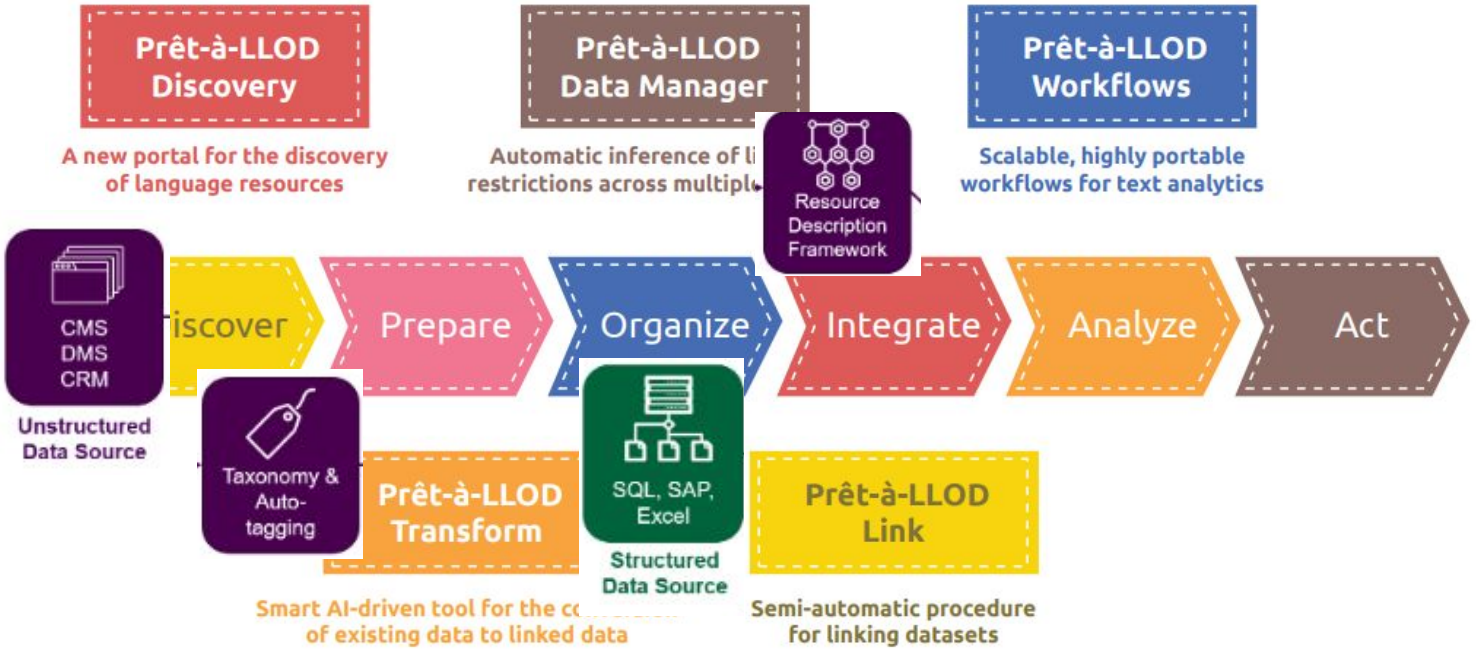
Prêt-à-LLOD Workflows

Scalable, highly portable workflows for text analytics



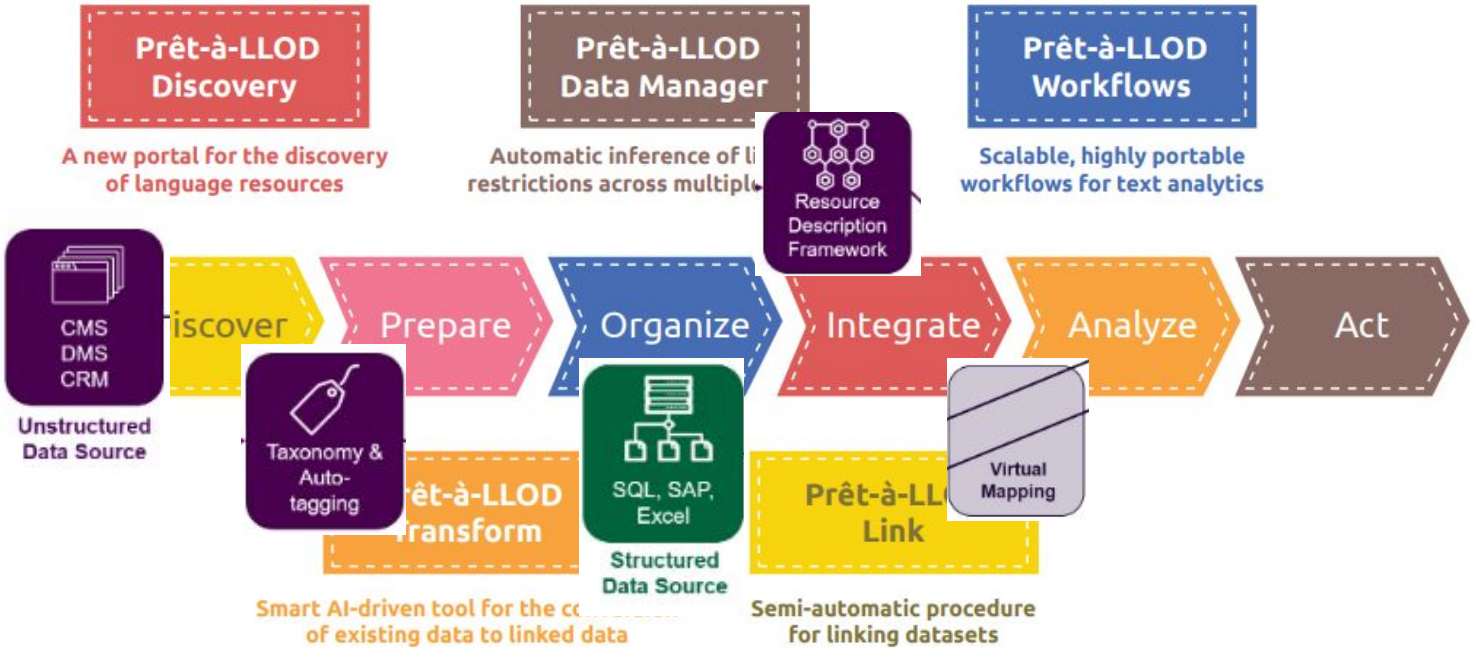
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Store LT Resource Descriptions



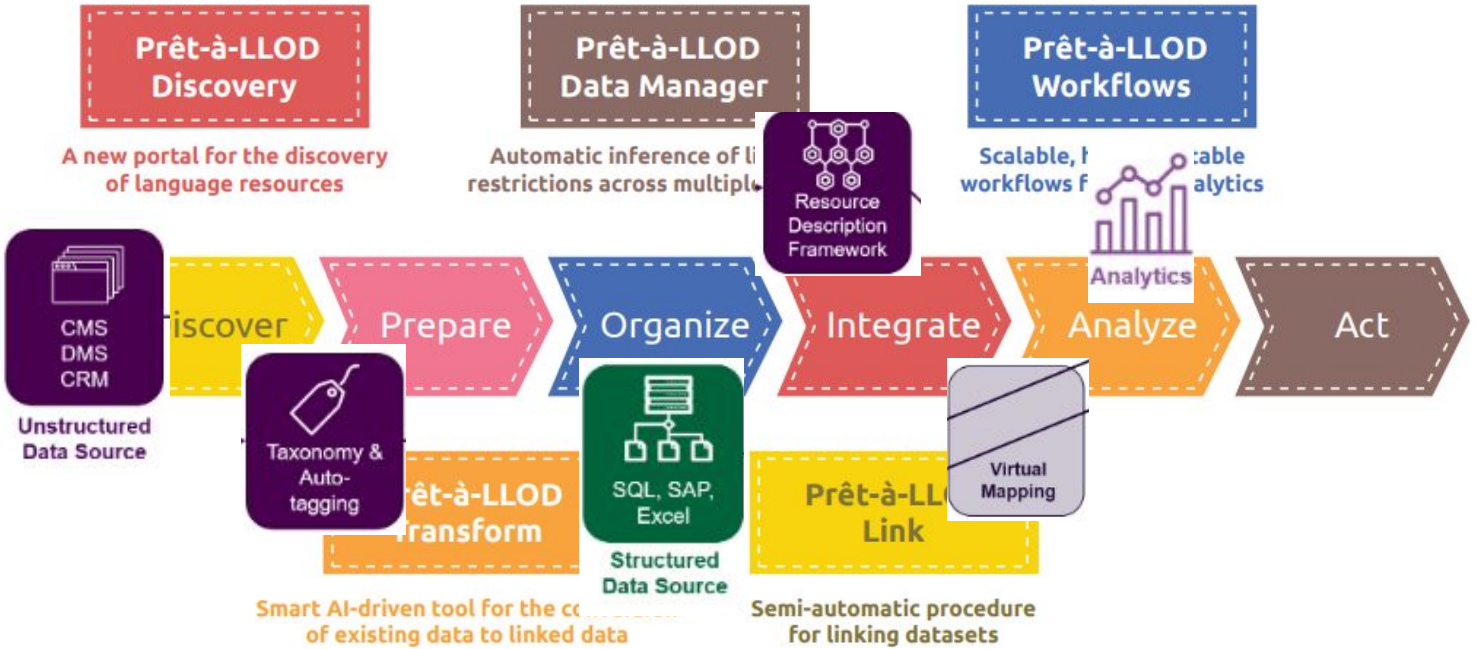
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Map/link various LT Resources



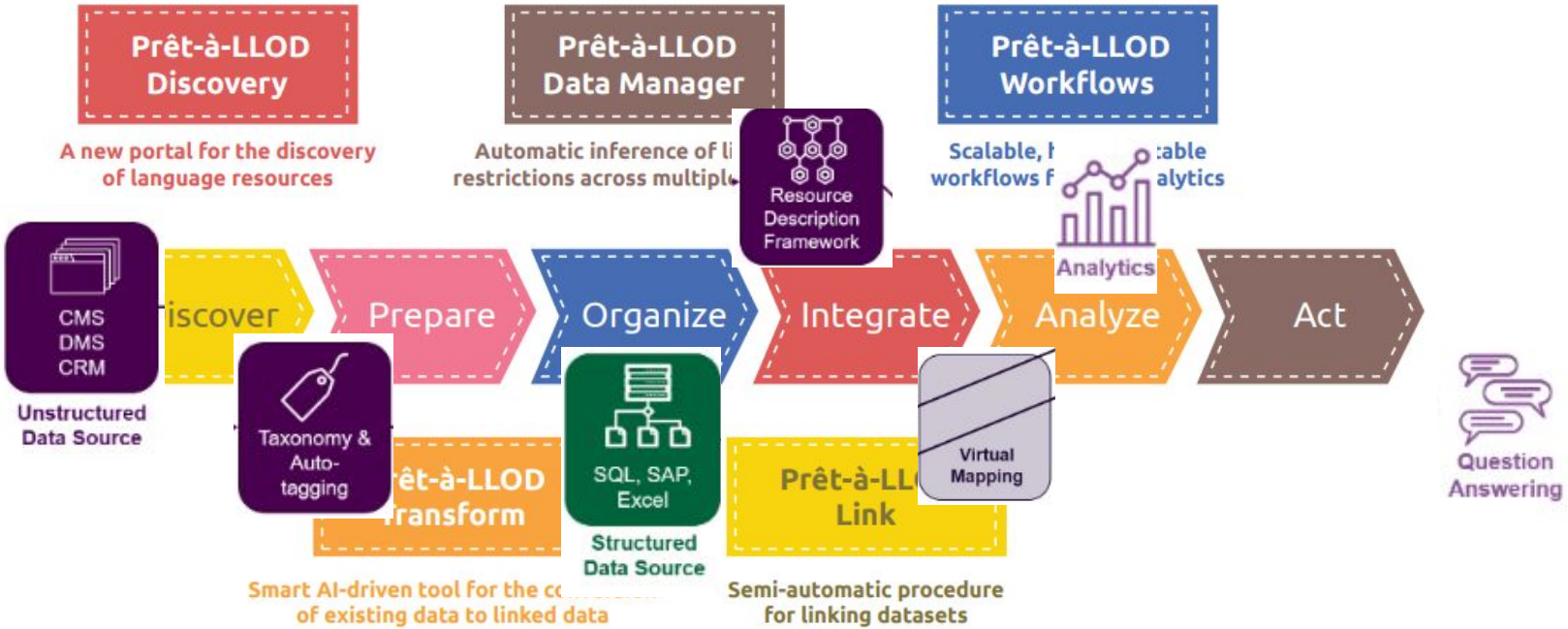
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Analyse contents



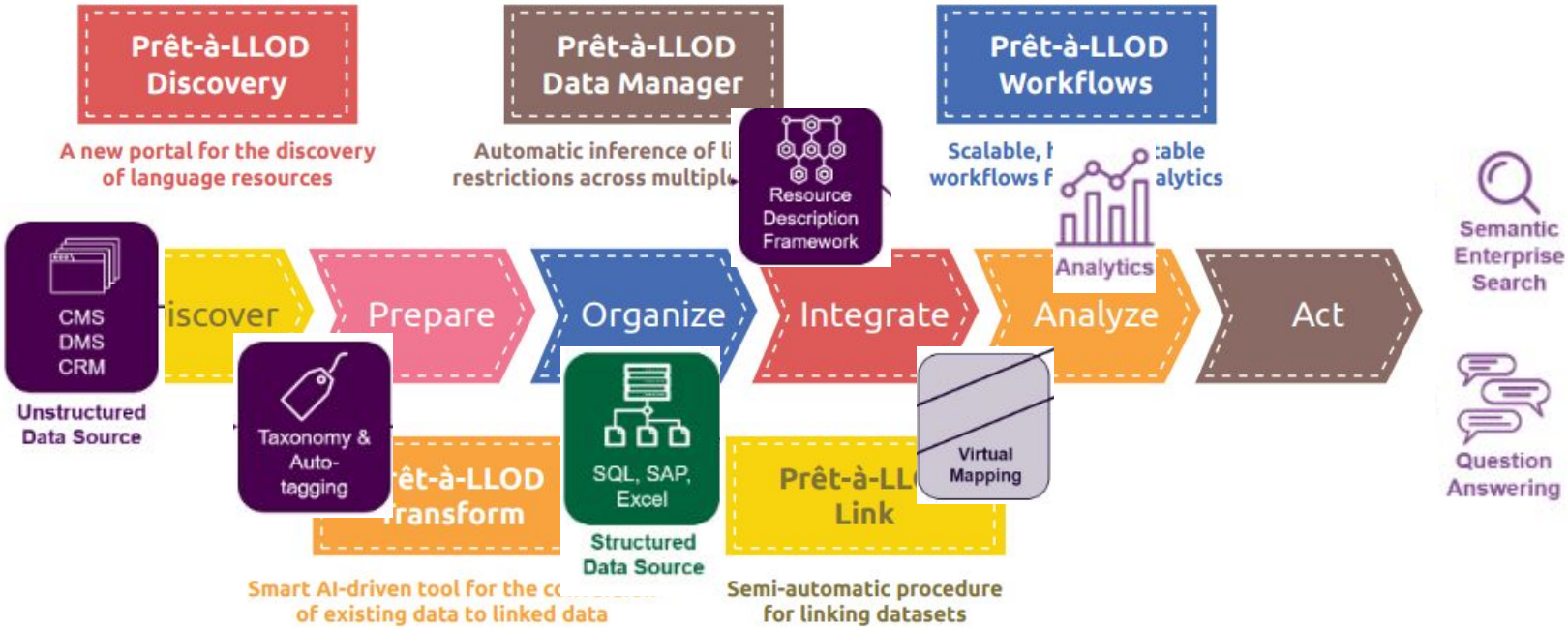
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Drive new applications



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

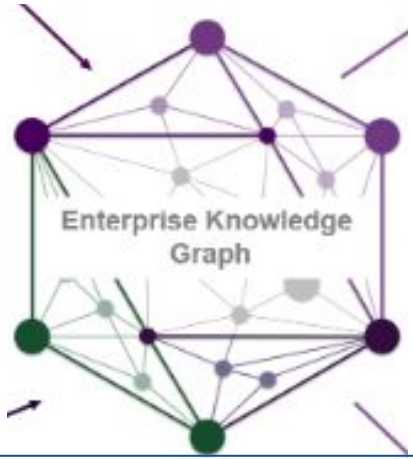
Increase Domain Findability



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

This is why we need a KG

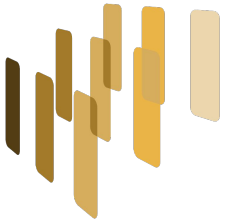
Discover → Prepare → Organize → Integrate → Analyze → Act



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Current Taxonomy

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182



EUROPEAN LANGUAGE GRID

LT taxonomy Role (1): Facilitating Grid operations

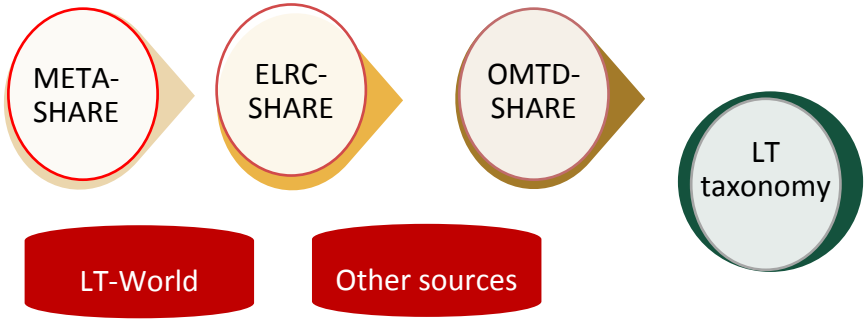
- Making list **contents easier to find**
  - facets: *operation* (LT Tools/Services); *intended application* (Data resources); *business applications* (LT actors, projects, etc.)
  - free text search: label, but backend can also exploit relations (synonyms, broader/narrower concepts)
- Making ELG catalogue **contents interoperable internally**
  - between LT tools/services and processable data resources
  - between LT tools/services and compatible resources (e.g. Machine Learning models, terminological dictionaries, gazetteers, etc.)
  - by matching them through the *operation* metadata element (... and other controlled vocabularies, e.g. data format, language, etc.)
- Making ELG catalogue **contents interoperable externally**
  - with resources in other catalogues/repositories/etc.
  - by matching the taxonomy with concepts of other communities
- **Supporting LR providers**
  - suggesting values from existing concepts (e.g. exploiting synonyms)
  - allowing users to enter their values which can then go through a curation phase

EUROPEAN LANGUAGE GRID

Role (2): Community building

- Through **raising awareness among LT experts**
  - navigating to other resources relevant for the LT activity
  - discovering LT companies, actors, projects, etc. related to a specific LT activity
  - going to the portal and checking out open calls, projects tagged for the specific LT activity
  - getting an overview of the LT activities in relation to various criteria (e.g. activity with most tools/services or companies, emerging LTs with new resources, demand of LT tools/services, etc.)
- Through **training LT-less aware citizens and experts** from other communities
  - dedicated short introductions in the catalogue
  - dedicated pages in the portal content
  - with enlightening definitions in layman terminology
  - linking to training material (videos, publications, webinars, etc.)

EUROPEAN LANGUAGE GRID



Building the LT taxonomy

- User input ⇒ curation by experts
- Values extracted from relevant metadata elements in catalogues
  - META-SHARE: LT focus; LT operations added by providers in a free text element
  - ELRC-SHARE: focus on multilingual Public Domain resources and Machine Translation; operations added in a controlled vocabulary after consulting users
  - OMTD-SHARE: focus on Text and Data Mining; values from users; creation of the OMTD-SHARE ontology (http://w3id.org/meta-share/omtd-share/) which is curated by TDM experts
- Values extracted from relevant information in portals
  - LT-World ontology: used as training material / "thesaurus" concepts
- Work under progress: enrichment, validation, adding definitions, …
- ⇒ Join us! https://www.w3.org/community/ld4lt/

Extensions

Recognizing new methodologies: extracting new terms for free-text descriptions, manuals, papers, etc.

Predicting new links between resources: extracting relations from text to extend the KG

Analysis of trends: mentions in news articles, reviews, blogs (with aspect mining)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

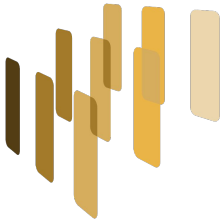
Usage Scenarios

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Workshop Questions

- Which usage of a general LT-Taxonomy do you see in your context?
- Are you aware of similar/past initiatives?
- What are the „must includes“ of such a taxonomy?
- Are you interested to use the results? Where?
- Do you want to be informed of the progress?
- Are you interest to join the development?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

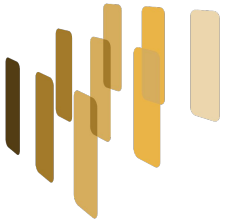


EUROPEAN LANGUAGE GRID

Sample Scenario 1

Corruption?

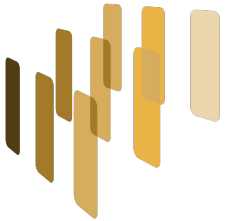
- A CEO of a company in the pharmaceutical business active in Russia sees a video in Russian mentioning their company and some word which sounds like “коррупция” - What is going on?
- This could be linked to a new drug currently being introduced to the Russian market
- The pharmaceutical business has recently been struck by several corruption cases
- Might this be linked to a corruption allegation?
- It’s a video -> requires ASR to "know' what’s being said
- It’s in Russian -> so it needs MT as the CEO doesn’t speak Russian
- It mentions particular terminology (pharmaceutical, medical) so entities need to be identified and also disambiguated -> IE
- *Happy ending*: it’s something mentioning how much the company does against corruption
- *Sad ending*: the company is indeed allegedly involved in corruption case in Russia



EUROPEAN LANGUAGE GRID

Sample Scenario 1

- The CEO is an LT expert and already knows of ELG
- Goes to the catalogue
- Searches in free text box for "ASR services for Russian"
- Gets back results of all tools/services that are classified under "Speech recognition" (since the two terms are synonyms) with input language "Russian"
- Goes on a second search for "Machine translation tools from Russian to English"
- He notices in the facet that there are both "Human aided Machine Translation" tools and "Computer-aided translation"
- Looks at the brief definitions of the two by simply hovering over the two terms and decided to select first the tools/services classified as CAT
- A third search for "NER tools" brings too many results back
- But again he opens up the LT taxonomy values and finds under it "NER of pharmaceutical entities" which is exactly what he 's looking for!
- Luckily all the tools he has selected can be combined together in a pipeline, so he goes on to use them

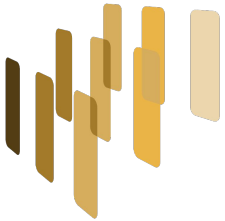


EUROPEAN LANGUAGE GRID

Sample Scenario 2

Language Learning

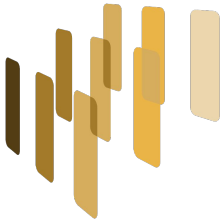
- A Spanish university student of Erasmus who is about to set off for Sweden
- She would like to prepare herself for the exchange by learning some basic Swedish
- Student interacts with e-learning platform using audio
- LT component within the platform computes text-complexity fitting the student's level and progress
- TTS provides pronunciations of texts
- ASR is used to transcribe speech and evaluate the pronunciations
- The student is presented tailored exercises depending on her abilities and according to her pace improving her experience and leading to more efficient language learning



EUROPEAN LANGUAGE GRID

Sample Scenario 2

- The student is totally ignorant of LT
- but the eLearning platform developers have already integrated all the LT tools/services required
- They didn't know much about LT at first, but someone told them about ELG and they had a look around the catalogue
- They had looked for tools/services that can be used for validating text complexity and found them under "Readability annotation"
- They looked for "tools that transcribe speech" and found TTS
- They already knew about "voice recognizers" and their search brought them results with ASR tools/services
- The student is quite impressed with the eLearning platform and decides to find more about LT
- Goes to ELG and checks the training material on the above terms

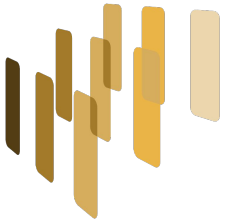


EUROPEAN LANGUAGE GRID

Sample Scenario 3

Aging Population

- An SME in Malta is interested in developing a dialog system for elderly persons for use within home-care
- The Maltese market by itself is not large enough, so they aim at a European solution, starting with Italy and Greece
- The company uses their own dialog-system but lacks input and output
- They are knowledgeable in NLP/ASR technology and want to train and adjust their own ASR models
- They require acoustic resources composed of elderly speakers and in multiple languages
- TTS is required for output
- Audio event detection might be helpful to detect emergency situations
- The SME searches for resources (and partners) on the ELG and finds corpora for Italian and Greek ASR, components for multi-lingual TTS as well as a potential partner offering solutions for Audio Event Detection



EUROPEAN LANGUAGE GRID

Sample Scenario 3

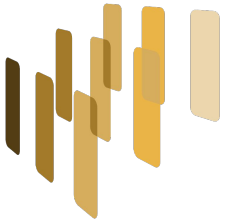
- They are experts in LT and already know all the appropriate terms
- They can use the faceted search and free-text alike
- They look for "multilingual audio corpora" with the desired age group of "participants"
- They go on to search for tools that can be used for "training ML models"
- Look for "TTS" for the desired languages
- And for tools/services that perform "event identification" and get back results for "event extraction"
- For all of them they check that they can get them in a form they can integrate in their final product (e.g. downloadable corpora, ready-to-run trainers), negotiate with providers and finally acquire them

EUROPEAN LANGUAGE GRID

Sample Scenario 4

Cross-language summarization

- An SME in Central Europe would like to create a browser plug-in for summarization of foreign news articles.
- In the age of fake-news and propaganda, interest in news in the languages of neighboring countries and especially of international powers such as Russia or China, has grown substantially.
- As the SME does not have an expert in the field of summarization, they use the grid to find adequate partners for this project.
- Together with the partner, they find that, as the source language might not be known in advance, language identification is required.
- Furthermore, two possible options for cross-lingual summarization are identified:
- use MT to translate to one particular target language and then apply (mono-lingual) summarization in this language or
- employ (multi-lingual) summarization to texts in the original language with subsequent translation using MT into the target language
- After scanning and evaluating the available resources etc. on the ELG, they decide on option 1



EUROPEAN LANGUAGE GRID

Sample Scenario 4

- Looks in the catalogue for companies that are active in the area of "cross-lingual summarization" and gets back results for "cross-language summarization"
- Decides to go up a node and check what other companies are active in "summarization" in general
- Seems like a more promising prospect for finding partners
- Goes to LT tools/services that perform "summarization" and finds quite interesting ones
- Thinks of using these and then translate the summary into other languages
- So, has another look at companies and resources that are active in "Machine Translation" for the languages of interest
- Finds some interesting fully automatic tools that can be integrated in the envisaged solution

Workshop Questions

- Which usage of a general LT-Taxonomy do you see in your context?
- Are you aware of similar/past initiatives?
- What are the „must includes“ of such a taxonomy?
- Are you interested to use the results? Where?
- Do you want to be informed of the progress?
- Are you interest to join the development?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182

Prêt-à-LLOD

ELSE IF 2019

European Language Services - Industry Forum

September 09, 2019
Karlsruhe, Germany

DEADLINE JULY 09, 2019
SUBMIT A TALK

www.pret-a-llod.eu/else-if19

EUROPEAN
LANGUAGE
GRID

Contact

- artem.revenko@semantic.web.com
- Gerhard.Backfried@sail-labs.com
- Penny Labropoulou <penny@ilsp.gr>
- thomas.thurner@semantic-web.com