



EUROPEAN LANGUAGE DATA SPACE



**Αναπτύσσοντας μια αγορά γλωσσικών δεδομένων
και υπηρεσιών μέσα από μια κοινή ευρωπαϊκή
προσπάθεια**

Στέλιος Πιπερίδης (ΕΚ Αθηνά & Κοινοπραξία LDS)

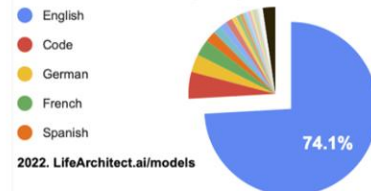
spip@athenarc.gr

08-05-2025 Γλωσσικά Δεδομένα για την Ανάπτυξη Γλωσσικής Τεχνολογίας και Τεχνητής Νοημοσύνης
<https://language-data-space.ec.europa.eu>

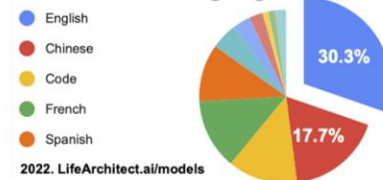
Τεχνητή Νοημοσύνη – Μεγάλα Γλωσσικά Μοντέλα (LLMs)

- Πρωτοφανείς δυνατότητες: Τα μεγάλα γλωσσικά μοντέλα αποτελούν την πιο ανατρεπτική ανακάλυψη στον τομέα της ΤΝ στη σύγχρονη ιστορία (GPT-3, ChatGPT, GPT-4, Claude, Gemini κ.λπ.)
- Τα LLMs εκπαιδεύονται πάνω σε τεράστιες ποσότητες δεδομένων και προαιρετικά σε δεδομένα εικόνας, βίντεο, ήχου, κ.λπ., δηλαδή σε πολυτροπικά δεδομένα
- Η πολυγλωσσία κάνει τα πάντα πολύ πιο δύσκολα (ανισορροπία δεδομένων): Οι γλώσσες της Ευρώπης σε μεγάλο βαθμό διαθέτουν περιορισμένους πόρους, εκτός από τα Αγγλικά
- Πρωτοφανείς ευκαιρίες:
 - Η παγκόσμια αγορά ΓΤ/ΕΦΓ αναμένεται να φτάσει τα 439,85 δισ. δολάρια μέχρι το 2030
 - Η παγκόσμια αγορά Παραγωγικής Τεχνητής Νοημοσύνης (Gen AI) αναμένεται να φτάσει τα 1,3 τρισ. δολάρια μέχρι το 2032
- Είναι απαραίτητη μία συντονισμένη προσπάθεια για τη συλλογή δεδομένων από όλες τις γλώσσες της Ευρώπης, ώστε να μπορέσουν να αναπτυχθούν LLMs σύμφωνα με τις ανάγκες και τον πολιτισμό μας...
- ... και να κάνουμε τη διαφορά με ευρωπαϊκά δεδομένα και ευρωπαϊκούς ενδιαφερόμενους φορείς
- Ήδη οι επενδύσεις ανέρχονται σε πολλά δισεκατομμύρια ...

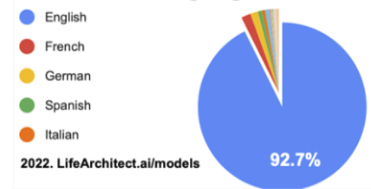
PaLM - 122 languages



BLOOM - 46 languages



GPT-3 - 90 languages



Ευρωπαϊκές Πρωτοβουλίες

- Ευρωπαϊκές πρωτοβουλίες για την ανάπτυξη LLMs
 - Μεγάλα ερευνητικά έργα σχεδόν σε κάθε χώρα, π.χ. Ισπανία, Δανία, Ιταλία, Γερμανία, κ.ά.
 - Εταιρείες σε πολλές χώρες, π.χ., Φινλανδία (Silo.ai), Γαλλία (Mistral), Γερμανία (Aleph Alpha)
 - Έργα χρηματοδοτούμενα από τη ΕΕ, π.χ., HPLT, TrustLLM, OpenEuroLLM, LLMs4EU
 - Νέα πανευρωπαϊκή πρωτοβουλία: ALT-EDIC
 - Νέα πρωτοβουλία της ΕΕ: Εργοστάσια TN (AI Factories), στενά συνδεδεμένα με εθνικά κέντρα Υπερυπολογιστικών Υπηρεσιών/HPC και την KE EuroHPC
- Προκλήσεις:
 - Υποδομές Υπερυπολογιστικών Υπηρεσιών (ποσότητα, πρόσβαση και ευχρηστία)
 - Ταχύτητα μεγάλων τεχνολογικών εταιρειών σε ΗΠΑ και Ασία σε σύγκριση με την ταχύτητα της Ευρώπης
 - Διαθεσιμότητα δεδομένων για όλες τις ευρωπαϊκές γλώσσες – *μακράν το πιο κρίσιμο εμπόδιο αυτή τη στιγμή*

Μακρά Ιστορία του Διαμοιρασμού Γλωσσικών Δεδομένων

The META-SHARE website features a dark header with the logo and navigation links: LEARN, DISCOVER, PARTICIPATE, CONNECT, and LOGIN. The main content area has a dark background with the text 'Search & exchange language resources' and a description of the network. A prominent orange button says 'JOIN OUR NETWORK NOW'. Below this is a search bar and a link to 'OR LEARN MORE'. At the bottom, four statistics are displayed with orange diagonal lines: 4,481 users, 2,887 language resources, 32% text corpora, and 27,630 number of downloads.

META-SHARE

LEARN • DISCOVER • PARTICIPATE • CONNECT • LOGIN

Search & exchange language resources

META-SHARE is an open and secure network of repositories for sharing and exchanging language data, tools and related web services

Share your own resources!

JOIN OUR NETWORK NOW

Already a member? [Log in](#)

Search the META-SHARE inventory

OR LEARN MORE

4,481 users

2,887 language resources

32% text corpora

27,630 number of downloads

The CLARIN VLO homepage has a light grey header with the logo and navigation links: Search, Contributors, and Help. The main content area features the title 'CLARIN Virtual Language Observatory' and a welcome message. It includes a search bar and a link to 'See all records'. Below the search bar, it states 'Search through 1,030,321 records'.

CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or **continue** to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

See all records Take a quick tour

Search through 1,030,321 records

The ELRC-SHARE Repository homepage features a header with the logo and navigation links: Home, Browse Resources, Help, About, Register, and Login. The main content area has a background image of flags and the title 'ELRC-SHARE Repository'. It includes a search bar and a welcome message. Below the search bar, it states 'The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources Coordination and considered useful for feeding the CEF Automated Translation (CEF AT) platform.' and a link to 'register (new user) or login (returning user) and go on to contribute resources'.

ELRC-SHARE Repository

Type in your keywords, please...

Welcome to the ELRC-SHARE repository!

The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources Coordination and considered useful for feeding the CEF Automated Translation (CEF AT) platform.

If you want to contribute resources, all you have to do is [register](#) (new user) or [login](#) (returning user) and go on to contribute resources.

The European Language Grid (ELG) homepage features a header with the logo and navigation links: Browse Resources, Information, Catalogue, Documentation & Media, and About. The main content area has a background image of flags and the title 'Language Technologies'. It includes a search bar and a welcome message. Below the search bar, it states 'Discover, try out, use and download LT services and resources for all European languages.' and a link to 'Browse ELG and find the LT services, resources, developers and providers you are looking for.'.

European Language Grid

Language Technologies

Discover, try out, use and download LT services and resources for all European languages.

Browse ELG and find the LT services, resources, developers and providers you are looking for.

Search the catalogue

Search

Order by: Resource Name A-Z



Αξιοποιεί τα υφιστάμενα επιτεύγματα



Ανταποκρίνεται στις απαιτήσεις των γλωσσικών δεδομένων για την πολύγλωσση Ευρώπη στην εποχή της Τεχνητής Νοημοσύνης



Υποστηρίζει την οικονομία γλωσσικών δεδομένων

Κοινός Ευρωπαϊκός Χώρος Γλωσσικών Δεδομένων



- Τύπος δράσης: σύμβαση προμήθειας (CNECT/LUX/2022/OP/0026)
- Προϋπολογισμός: 6 εκατ. € (+ 2 εκατ. € εφόσον ανανεωθεί) – διάρκεια: 36 μήνες (+ 12 μήνες εφόσον ανανεωθεί)
- Στόχος: Ανάπτυξη και λειτουργία μιας αξιόπιστης και ασφαλούς ευρωπαϊκής υποδομής και αγοράς για τη συλλογή, τη δημιουργία, τον διαμοιρασμό και την επαναχρησιμοποίηση πολύγλωσσων και πολυτροπικών γλωσσικών δεδομένων
- Το LDS λειτουργεί ως μία αγορά (marketplace) για όλους τους οργανισμούς – εμπορικούς και δημόσιους
- Το LDS προσφέρει πρόσβαση σε γραφεία υποστήριξης (helpdesks) για ερωτήσεις και διευκρινίσεις σχετικά με νομικά και τεχνικά ζητήματα
- Βασικά χαρακτηριστικά: πλαίσιο διακυβέρνησης, τεχνική υποδομή, ανοικτότητα, προώθηση
- Ενδιαφερόμενοι φορείς: βιομηχανία, ερευνητική κοινότητα, δημόσια διοίκηση, πολιτιστικοί φορείς, ΜΚΟ και πολίτες
- Οι τέσσερις βασικοί εταίροι του έργου έχουν συμμετάσχει σε πολλά έργα
- Η τεχνική ανάπτυξη βασίζεται στα έργα ELG, ELRC-SHARE, META-SHARE

Lead Partner and Coordinator
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Partners and Operation Leads
R.C. “Athena”, Institute for Language and Speech Processing
Evaluations and Language Resources Distribution Agency
TILDE
Main Subcontractors
3pc GmbH Neue Kommunikation
CLARIN ERIC
Big Data Value Association (Data, AI and Robotics) AISBL

Συμβολή στο επερχόμενο οικοσύστημα Κοινών Ευρωπαϊκών Χώρων Δεδομένων

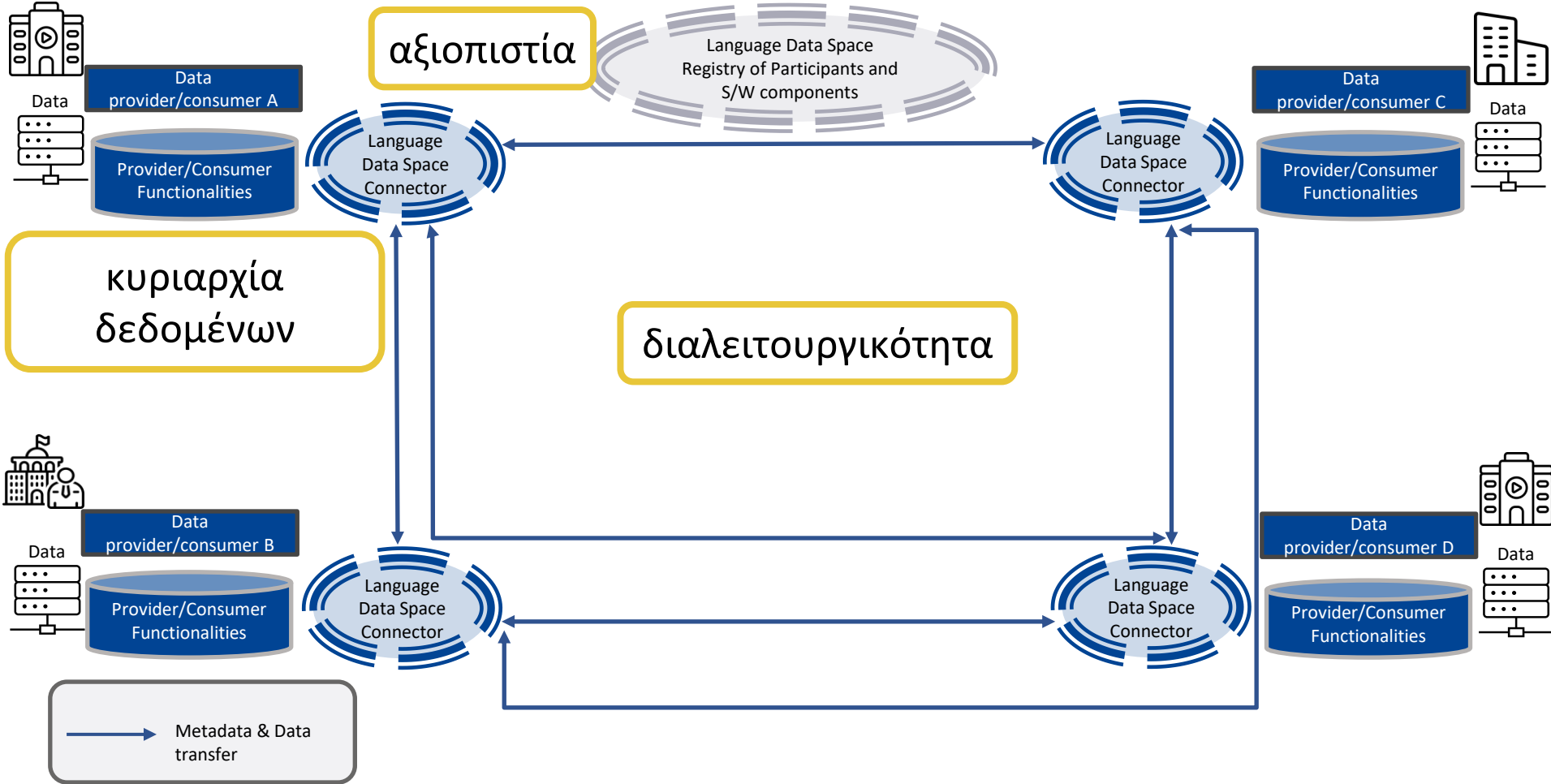
κυριαρχία των
δεδομένων

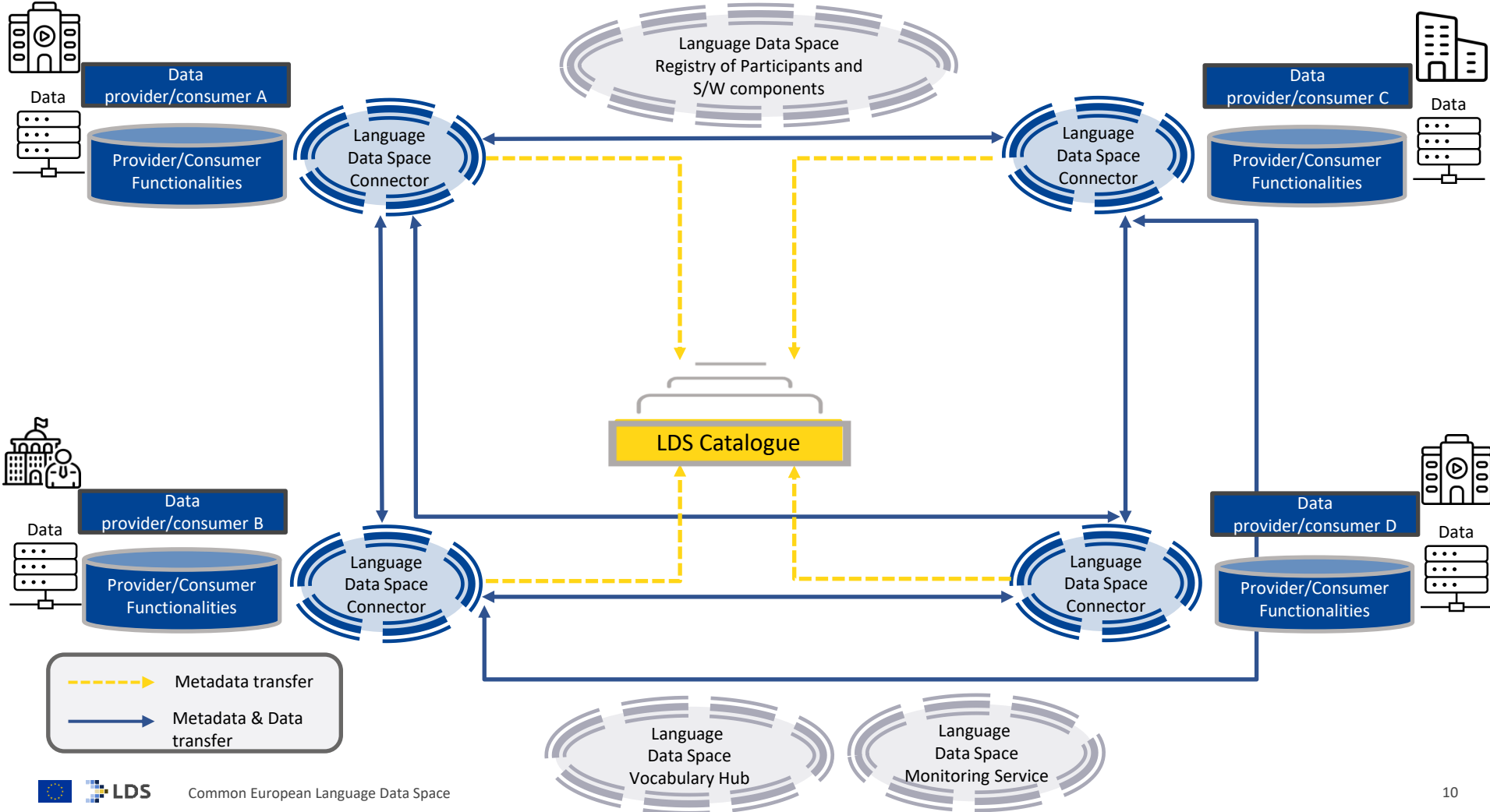
αξιοπιστία

διαλειτουργικότητα

Αρχιτεκτονική LDS

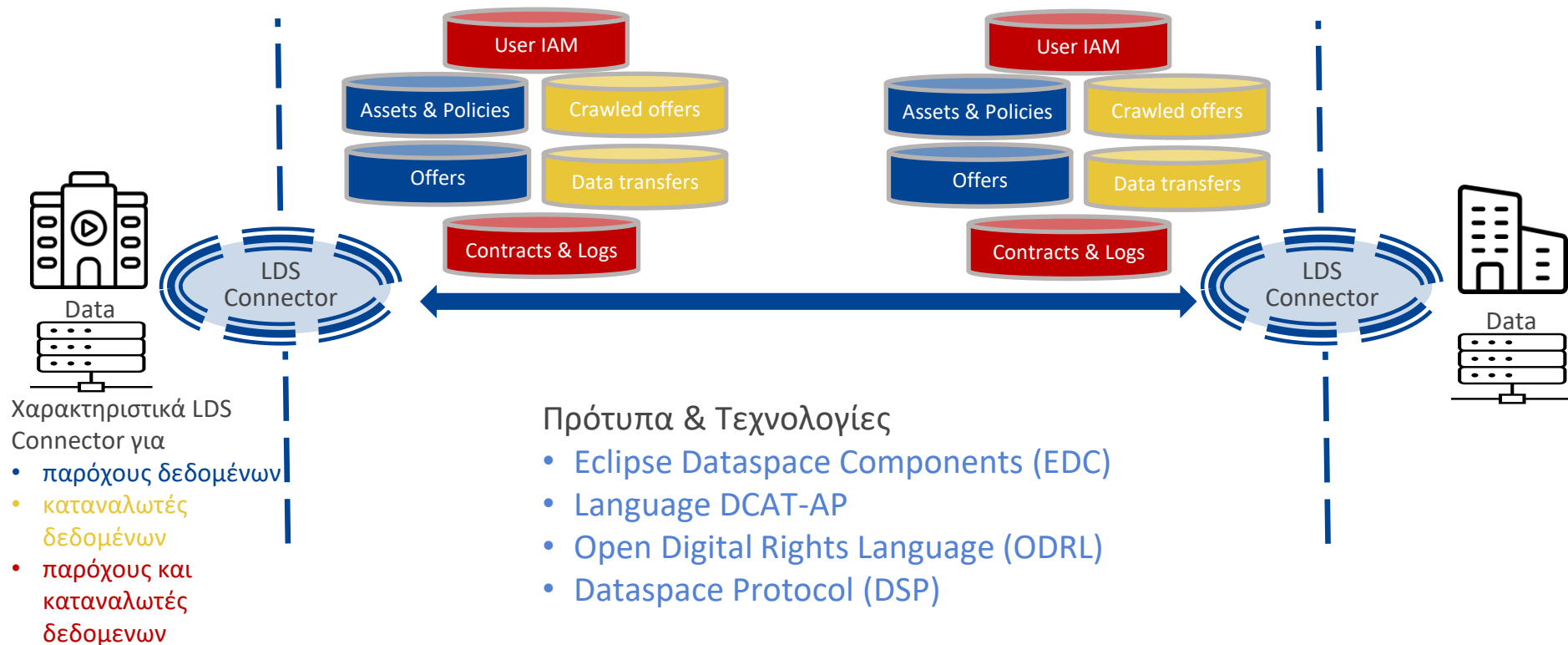
- Στόχος 1: παροχή τεχνολογικής υποστήριξης για συμμόρφωση με τις **αρχές των Χώρων Δεδομένων**
 - κυριαρχία, αξιοπιστία, διαλειτουργικότητα
- Στόχος 2: παροχή τεχνολογικής υποστήριξης για τις **ροές εργασίας του LDS** που έχουν σχεδιαστεί
 - για παρόχους & καταναλωτές δεδομένων
 - για το Συμβούλιο Διακυβέρνησης του LDS



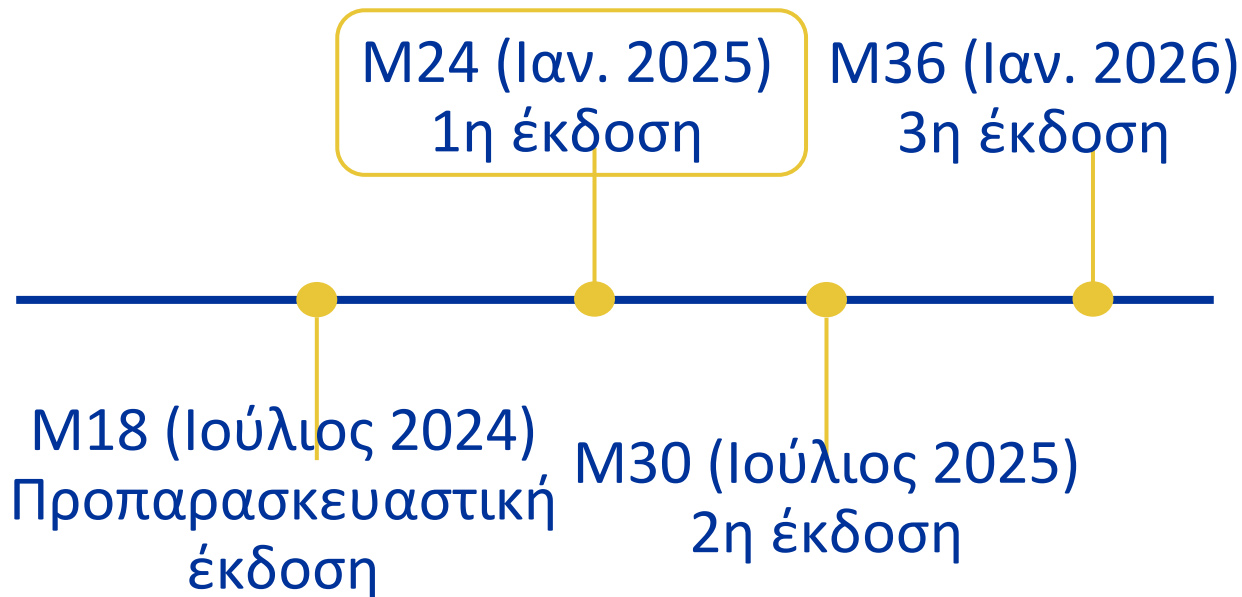


LDS Connector

- Πόρος (Asset): Δεδομένα/Αρχεία+Μεταδεδομένα
- Προσφορά (Offer): Πόρος (Asset)+Πολιτική διάθεσης



Προγραμματισμός Εκδόσεων LDS





**EUROPEAN
LANGUAGE
DATA SPACE**

v1.0.0-beta

LDS v1.0.0-beta

1. Αίτηση συμμετοχής και ένταξη μελών στο LDS

- Ο νόμιμος εκπρόσωπος του Οργανισμού υποβάλλει ηλεκτρονική αίτηση (<https://ldssetup.ilsp.gr>) και αποδέχεται του Όρους Χρήσης του LDS
 - Τα δεδομένα προστίθενται στο **Μητρώο συμμετεχόντων του LDS**
 - Το **Συμβούλιο Διακυβέρνησης του LDS** αξιολογεί την αίτηση (διασφαλίζοντας ότι πληρούνται τα κριτήρια επιλεξιμότητας) → εγκρίνει ή απορρίπτει την αίτηση (αναφέροντας τους λόγους)
 - **Εφόσον η αίτηση εγκριθεί**, ο συμμετέχων εγκαθιστά & παραμετροποιεί τον **LDS Connector** στην υπολογιστική υποδομή του Οργανισμού του
- ➔ Έκδοση κλειδιών πιστοποίησης ταυτότητας
- **Καταχώριση** του **LDS Connector** στο Μητρώο του LDS
 - Παραμετροποίηση των λύσεων αποθήκευσής δεδομένων

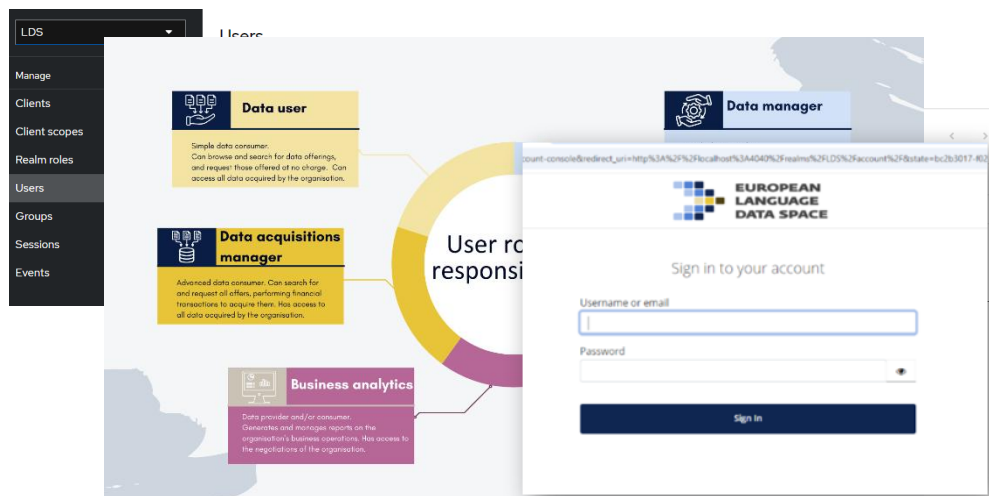
[Demo 1: Apply for a membership in the LDS - YouTube](#)



2. Διαχείριση Ταυτότητας & Πρόσβασης Χρήστη (στον LDS Connector του συμμετέχοντα Οργανισμού)



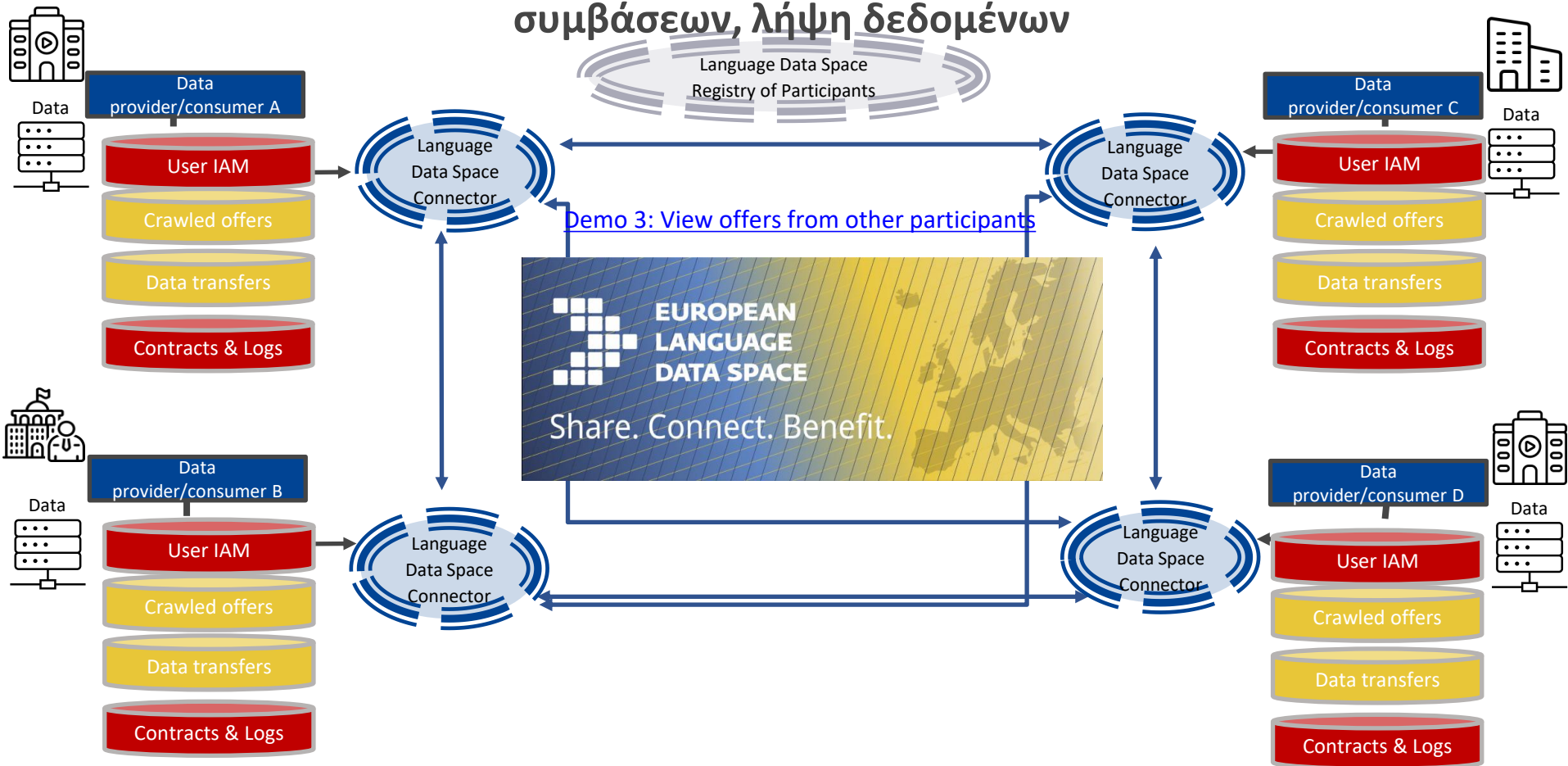
- **Λύση Keycloak**
- **Εγγραφή** χρηστών από διαχειριστή (με δυνατότητα χρήσης active directory)
- Ανάθεση προκαθορισμένων **ρόλων χρηστών** LDS για παρόχους & καταναλωτές με απλά και προηγμένα δικαιώματα (με δυνατότητα ανάθεσης πολλαπλών ρόλων)



3. Πάροχοι δεδομένων – Δημιουργία πόρων, πολιτικών διάθεσης και προσφορών



4. Καταναλωτές δεδομένων – αναζήτηση και εντοπισμός, διαπραγμάτευση, σύναψη συμβάσεων, λήψη δεδομένων



Τεχνική υποστήριξη

- Επικοινωνία με το LDS https://language-data-space.ec.europa.eu/help_en
- Συχνές ερωτήσεις (FAQs) https://language-data-space.ec.europa.eu/help/faq_en
- [Τεχνικό εγχειρίδιο LDS](#)
 - για τεχνικούς διαχειριστές
 - για τους χρήστες του LDS (παρόχους & καταναλωτές)
 - για τα μέλη του Συμβουλίου Διακυβέρνησης του LDS



Share

You will be able to share and monetise your language data, language models and other language resources through a single platform, taking European values and compliance with EU regulations fully into account.



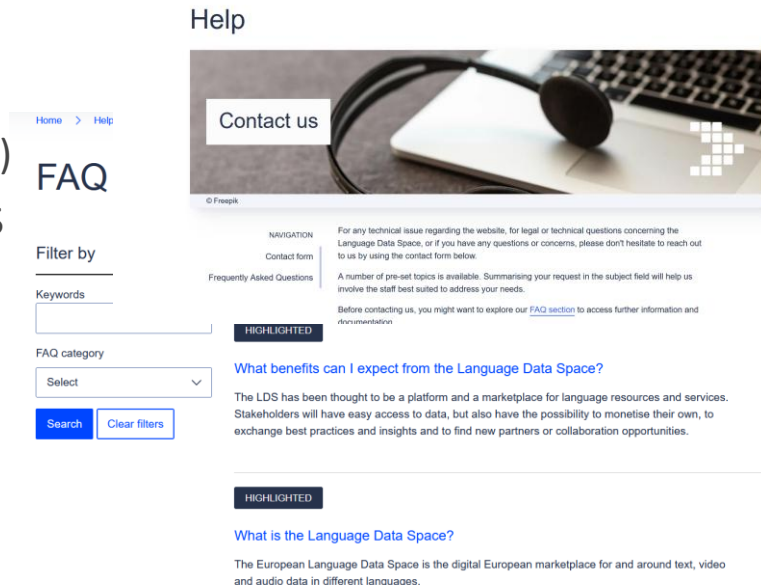
Connect

You will be able to connect and exchange with other stakeholders through the European Language Data Space.



Benefit

You will be supported by the European Language Data Space in the development of multilingual and multimodal language technologies and language-centric AI.



Υλοποίηση των Αρχών του Κοινού Ευρωπαϊκού Χώρου Γλωσσικών Δεδομένων

Πλαίσιο αξιοπιστίας

- οι αιτήσεις αξιολογούνται από το Συμβούλιο Διακυβέρνησης του LDS
- απαιτείται αποδοχή των Όρων Χρήσης του LDS
- οι συναλλαγές γίνονται μόνο μεταξύ πιστοποιημένων και εξουσιοδοτημένων connectors
- διαχείριση ταυτότητας και πρόσβασης χρήστη (IAM) σε επίπεδο οργανισμού
- μεταφορά δεδομένων μόνο μετά από σύναψη σύμβασης (μεταξύ γνωστών συμβαλλομένων μερών)

Κυριαρχία των δεδομένων

- τα δεδομένα παραμένουν στο χώρο αποθήκευσης του συμμετέχοντα
- κάθε συμμετέχων ορίζει τη δική του πολιτική δεδομένων
- οι πολιτικές εφαρμόζονται πριν από την έγκριση αιτήματος για χρήση/μεταφορά δεδομένων

Πρόελευση & ιχνηλασιμότητα

- οι συμβάσεις αποθηκεύονται στους connectors και των δύο συμβαλλομένων μερών
- καταγραφή & παρακολούθηση των συναλλαγών τοπικά & κεντρικά

Διαλειτουργικότητα

- συμβατότητα με το Dataspace Protocol
- υιοθέτηση πρότυπων πρωτοκόλλων για μεταφορά δεδομένων
- ανάπτυξη πρότυπου μοντέλου για περιγραφές (βάσει του μοντέλου DCAT)
- ανάπτυξη μοντέλου ODRL για πολιτικές
- συστάσεις για πρότυπα μοντέλα και μορφότυπα δεδομένων

Παρουσίαση (demo) & δοκιμές (testing) του LDS

Κάντε αίτηση στο:

<https://ldssetup.ilsp.gr/>

για να συμμετάσχετε
στο LDS και να το
δοκιμάσετε!

vanilla

Κατηγορίες Δεδομένων

Class of Data	Typical Size	Providers	Integration into LDS	Relevance, especially for LLM Development
Regular Corpora and Language Resources	Small (MB, GB)	Primarily NLP/LT research: ELG, META-SHARE, CLARIN, ELRA, ELDA etc.	Can be easily integrated by connecting the repositories to LDS	Usually very high quality data and thus relevant for LLMs but not as base data
Web Crawls	Very big (TB, PB)	Common Crawl (and OSCAR-processed CC dumps), Internet Archive dumps etc.	Challenge due to their size (hard to transfer, hard to preprocess, hard to store; must be close to the HPC)	Indispensable due to their size and coverage – but: high level of noise, massive need for pre-processing
New, fresh data from industry and other organisations	Arbitrary size, ideally as large as possible	Publishing houses, media companies, libraries, call centres, broadcasters etc.; also: Media Data Space	Can be easily integrated by connecting these organisations to LDS	High quality data; domain-specific data; data covering specific languages; raw data; processed data; evaluation data; data relevant for LLM development etc.

Ποιότητα δεδομένων

- Συχνή ερώτηση: "Τι γίνεται με την ποιότητα των δεδομένων; Μετράτε ή επιβάλλετε ένα συγκεκριμένο επίπεδο ποιότητας;
- «Όχι».
- Η ποιότητα των γλωσσικών δεδομένων δεν μπορεί να μετρηθεί με γενικό τρόπο.
- Σημείωση: σε άλλους τομείς τα πράγματα είναι διαφορετικά, π.χ. δεδομένα αισθητήρων, όπου υπάρχουν διάφορες πτυχές που σχετίζονται με την ποιότητα, π.χ. ανάλυση, κ.λπ.
- Για κάποια περίπτωση χρήσης, ένα συγκεκριμένο σύνολο δεδομένων μπορεί να είναι πολύ υψηλής ποιότητας και εξαιρετικά χρήσιμο και για ένα άλλο άτομο, το ίδιο σύνολο δεδομένων μπορεί να είναι άχρηστο.
- Το LDS βασίζεται σε υψηλής ποιότητας μεταδεδομένα που παρέχουν επαρκείς λεπτομέρειες σχετικά με το σύνολο δεδομένων που περιγράφουν.
- Στόχος: η ποιότητα ή η σημασία ενός συνόλου δεδομένων θα πρέπει να περιγράφεται μέσω του συνόλου μεταδεδομένων του.

Ομάδα Χρηστών LDS

<https://language-data-space.ec.europa.eu>

• Συναντήσεις και Συνέδρια

- Εναρκτήρια συνάντηση τον Μάρτιο του 2024
- Δεύτερη συνάντηση τον Ιούνιο του 2024
- Τρίτη συνάντηση τον Νοέμβριο του 2024
- Η Ομάδα Χρηστών του LDS συνεχώς μεγαλώνει
- Δοκιμές (testing) του LDS Connector (μέχρι στιγμής έχουν πραγματοποιηθεί 10+ εγκαταστάσεις λογισμού LDS)

Join the LDS user group



© Freepik

The European Language Data Space (LDS) user group members shall actively contribute to and take advantage of the LDS, bringing in their own requirements and validating the emerging LDS infrastructure.

If you are a stakeholder who is in need of language data or if you want to give the language data of your organisation a second life, potentially monetising it, you are welcome to join.

[Click to join](#)



LDS User Group (as of 05 March 2025)

Individual
Members

Unique
Organisations

198

158

• Γίνετε μέλος της Ομάδας Χρηστών του LDS ή ενθαρρύνετε τα μέλη του δικτύου σας να συμμετάσχουν!

- Επικύρωση εννοιών, ιδεών, λογισμικού, πρώτες δοκιμαστικές εγκαταστάσεις του LDS Connector (σε εξέλιξη αυτή τη στιγμή!), πρώτες πιλοτικές ανταλλαγές δεδομένων, έρευνες κ.λπ.



Common European Language Data Space

Σας ευχαριστώ!



A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

Στέλιος Πυπερίδης (ΕΚ Αθηνά & Κοινοπραξία LDS)
spip@athenarc.gr

08-05-2025 Γλωσσικά Δεδομένα για την Ανάπτυξη Γλωσσικής Τεχνολογίας και Τεχνητής Νοημοσύνης
<https://language-data-space.ec.europa.eu>