ELRC Reykjavik Workshop 11.11.2016.

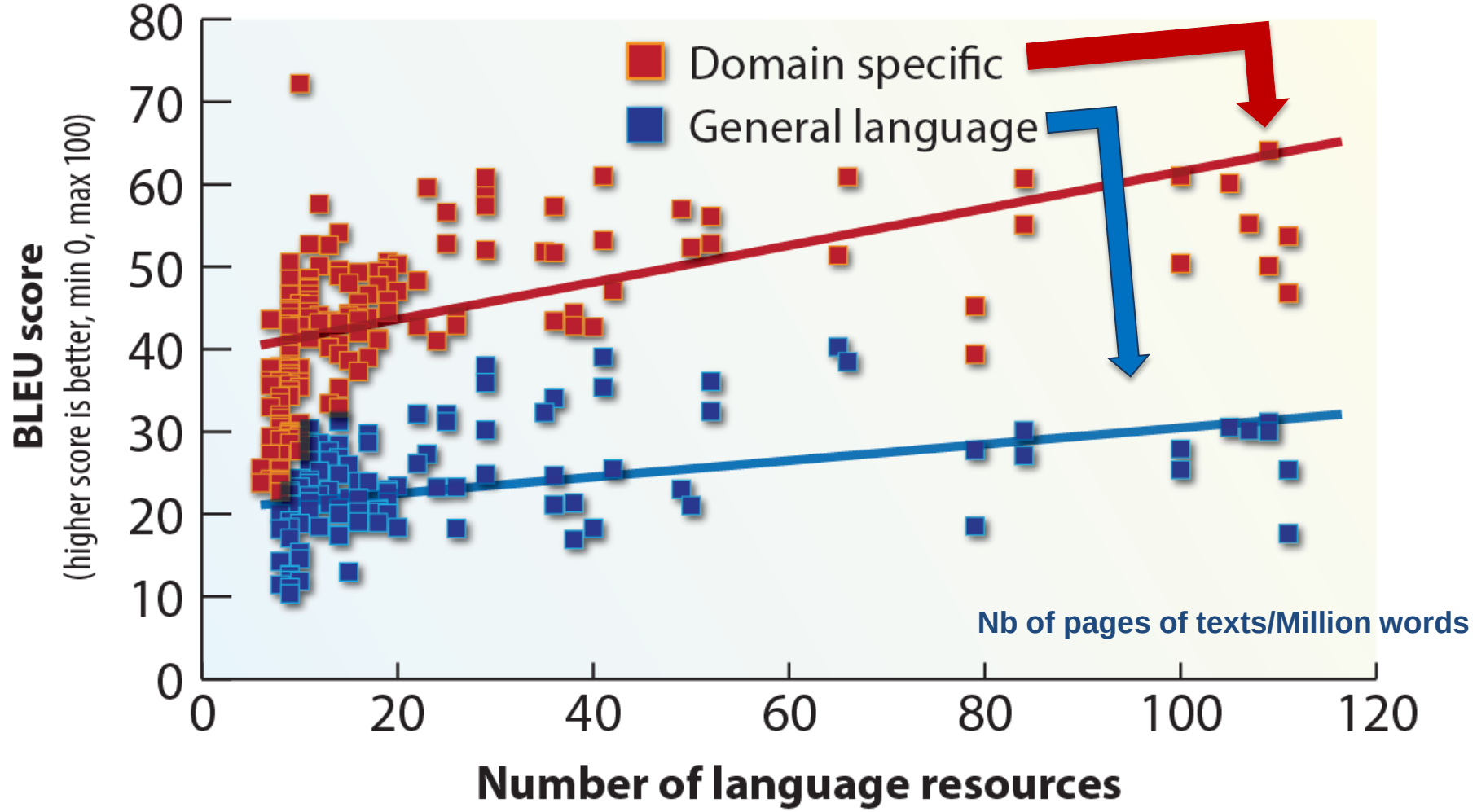
# What data is needed for automated translation?

Andrejs Vasiļjevs, ELRC/Tilde | Credits: Khalid Choukri, ELRA, Josef van Genabith, DFKI

European Language Resource Coordination
Connecting Europe Facility

European Language Resource Coordination
Connecting Europe Facility

# Impact of number of language resources on Statistical MT quality



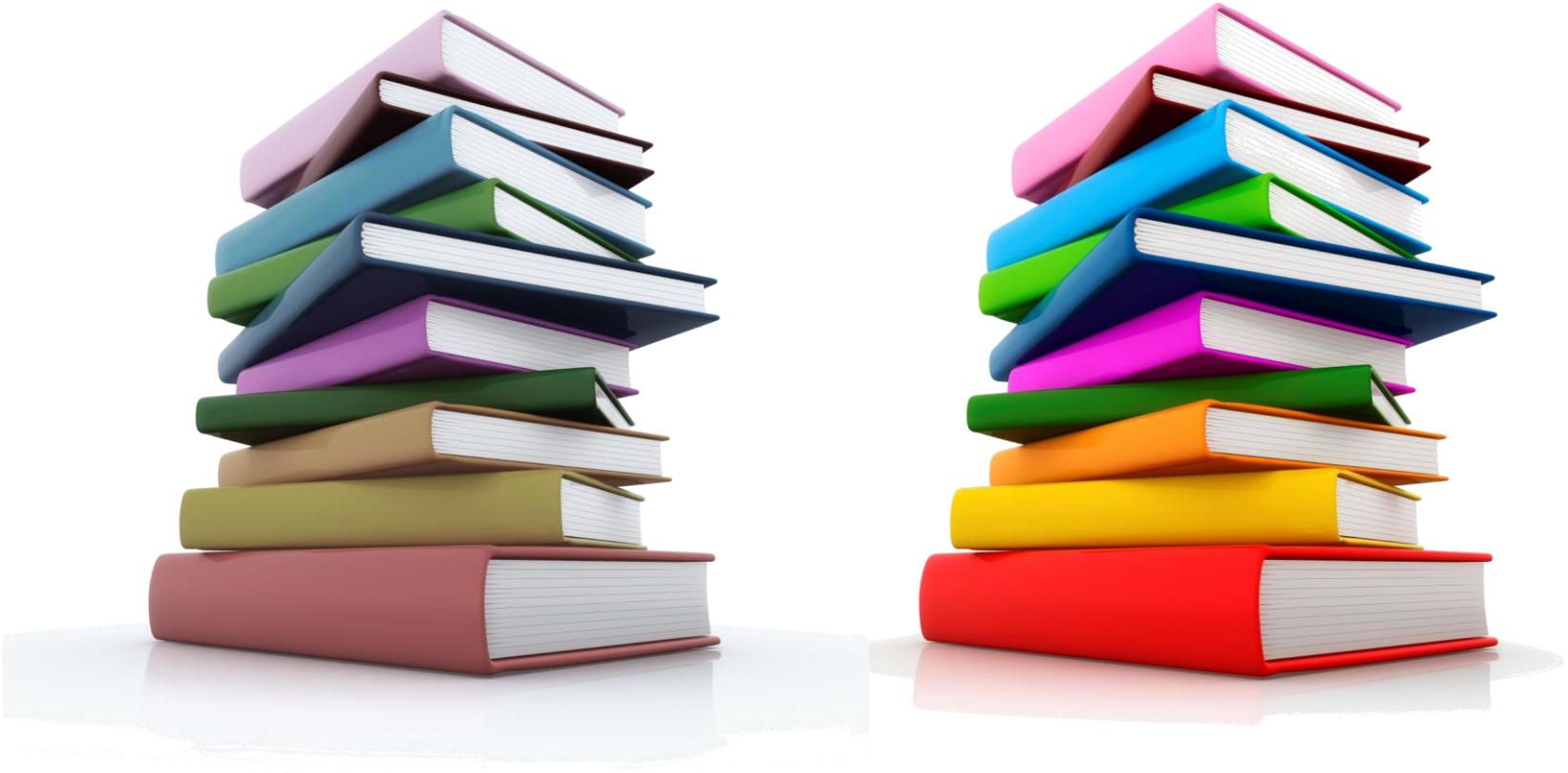
European Language Resource Coordination
Connecting Europe Facility

# What Data?

- Anything that contains “words”
- Preferences for “sentences”, even for sentences expressed in multiple languages
- Examples: reports, speeches, documents, web pages, brochures, etc.
- Bags of “words”, “sentences”, multiple bags

European Language Resource Coordination
Connecting Europe Facility

# What types of data?
# "Aligned" Translation

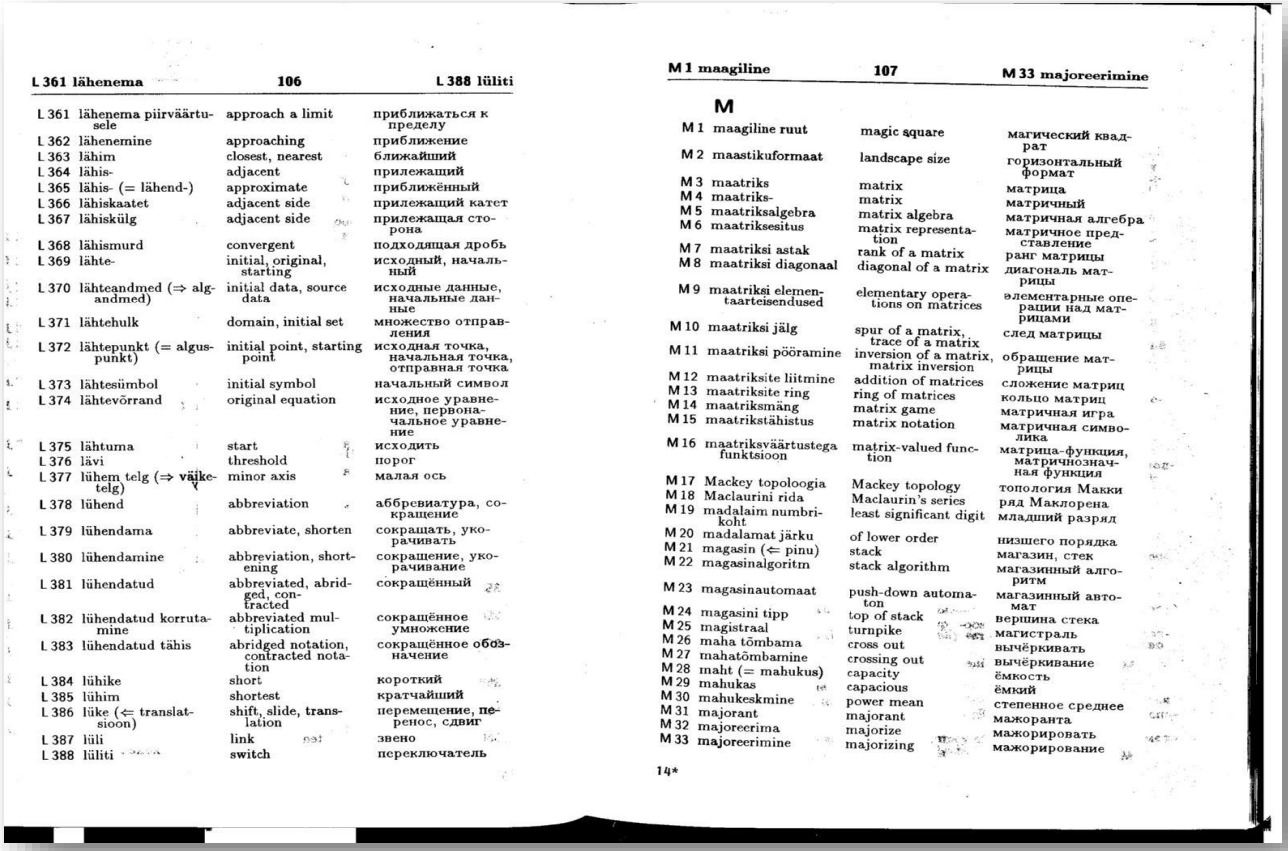


English

Estonian

European Language Resource Coordination
Connecting Europe Facility

# Printed sources are hard to process



L 361 lähenema 106 L 388 lüliti

| | | | |
|---|---|---|---|
| L 361 | lähenema piirväärtusele | approach a limit | приближаться к пределу |
| L 362 | lähenemine | approaching | приближение |
| L 363 | lähim | closest, nearest | ближайший |
| L 364 | lähis- | adjacent | прилежащий |
| L 365 | lähis- (= lähend-) | approximate | приближённый |
| L 366 | lähiskaatet | adjacent side | прилежащий катет |
| L 367 | lähiskülg | adjacent side | прилежащая сторона |
| L 368 | lähismurd | convergent | подходящая дробь |
| L 369 | lähte- | initial, original, starting | исходный, начальный |
| L 370 | lähteandmed (⇒ algandmed) | initial data, source data | исходные данные, начальные данные |
| L 371 | lähtehulk | domain, initial set | множество отправления |
| L 372 | lähtepunkt (= alguspunkt) | initial point, starting point | исходная точка, начальная точка, отправная точка |
| L 373 | lähtesümbol | initial symbol | начальный символ |
| L 374 | lähtevõrrand | original equation | исходное уравнение, первоначальное уравнение |
| L 375 | lähtuma | start | исходить |
| L 376 | lävi | threshold | порог |
| L 377 | lühem telg (⇒ väike-telg) | minor axis | малая ось |
| L 378 | lühend | abbreviation | аббревиатура, сокращение |
| L 379 | lühendama | abbreviate, shorten | сокращать, укорачивать |
| L 380 | lühendamine | abbreviation, shortening | сокращение, укорачивание |
| L 381 | lühendatud | abbreviated, abridged, contracted | сокращённый |
| L 382 | lühendatud korrutamine | abbreviated multiplication | сокращённое умножение |
| L 383 | lühendatud tähis | abridged notation, contracted notation | сокращённое обозначение |
| L 384 | lühike | short | короткий |
| L 385 | lühim | shortest | кратчайший |
| L 386 | lüke (⇐ translatsioon) | shift, slide, translation | перемещение, перенос, сдвиг |
| L 387 | lüli | link | звено |
| L 388 | lüliti | switch | переключатель |

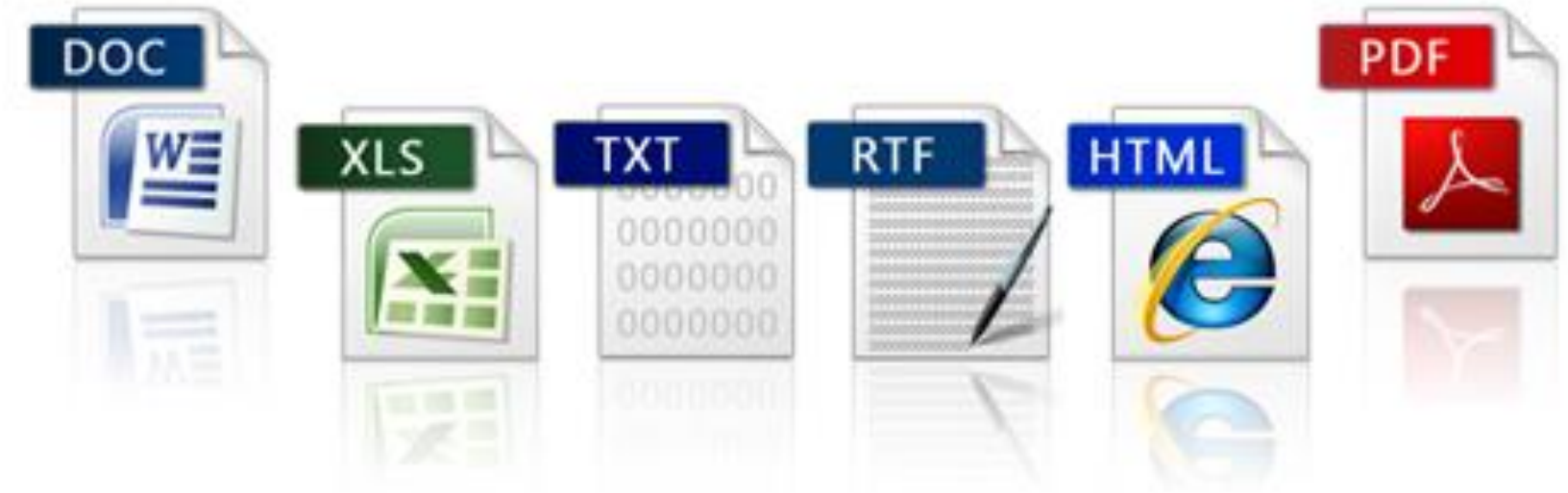
M 1 maagiline 107 M 33 majoreerimine

M

| | | | |
|---|---|---|---|
| M 1 | maagiline ruut | magic square | магический квадрат |
| M 2 | maastikuformaat | landscape size | горизонтальный формат |
| M 3 | maatriks | matrix | матрица |
| M 4 | maatriks- | matrix | матричный |
| M 5 | maatriksalgebra | matrix algebra | матричная алгебра |
| M 6 | maatriksesitus | matrix representation | матричное представление |
| M 7 | maatriksi astak | rank of a matrix | ранг матрицы |
| M 8 | maatriksi diagonaal | diagonal of a matrix | диагональ матрицы |
| M 9 | maatriksi elementaarteisendused | elementary operations on matrices | элементарные операции над матрицами |
| M 10 | maatriksi jälg | spur of a matrix, trace of a matrix | след матрицы |
| M 11 | maatriksi pööramine | inversion of a matrix, matrix inversion | обращение матрицы |
| M 12 | maatriksite liitmine | addition of matrices | сложение матриц |
| M 13 | maatriksite ring | ring of matrices | кольцо матриц |
| M 14 | maatriksmäng | matrix game | матричная игра |
| M 15 | maatrikstähistus | matrix notation | матричная символика |
| M 16 | maatriksväärtustega funktsioon | matrix-valued function | матрица-функция, матричнозначная функция |
| M 17 | Mackey topoloogia | Mackey topology | топология Макки |
| M 18 | Maclaurini rida | Maclaurin's series | ряд Маклорена |
| M 19 | madalaim numbrikoht | least significant digit | младший разряд |
| M 20 | madalamat järku | of lower order | низшего порядка |
| M 21 | magasin (⇐ pinu) | stack | магазин, стек |
| M 22 | magasinalgoritm | stack algorithm | магазинный алгоритм |
| M 23 | magasinautomaat | push-down automaton | магазинный автомат |
| M 24 | magasini tipp | top of stack | вершина стека |
| M 25 | magistraal | turnpike | магистраль |
| M 26 | maha tõmbama | cross out | вычёркивать |
| M 27 | mahatõmbamine | crossing out | вычёркивание |
| M 28 | maht (= mahukus) | capacity | ёмкость |
| M 29 | mahukas | capacious | ёмкий |
| M 30 | mahukeskmine | power mean | степенное среднее |
| M 31 | majorant | majorant | мажоранта |
| M 32 | majoreerima | majorize | мажорировать |
| M 33 | majoreerimine | majorizing | мажорирование |

14*

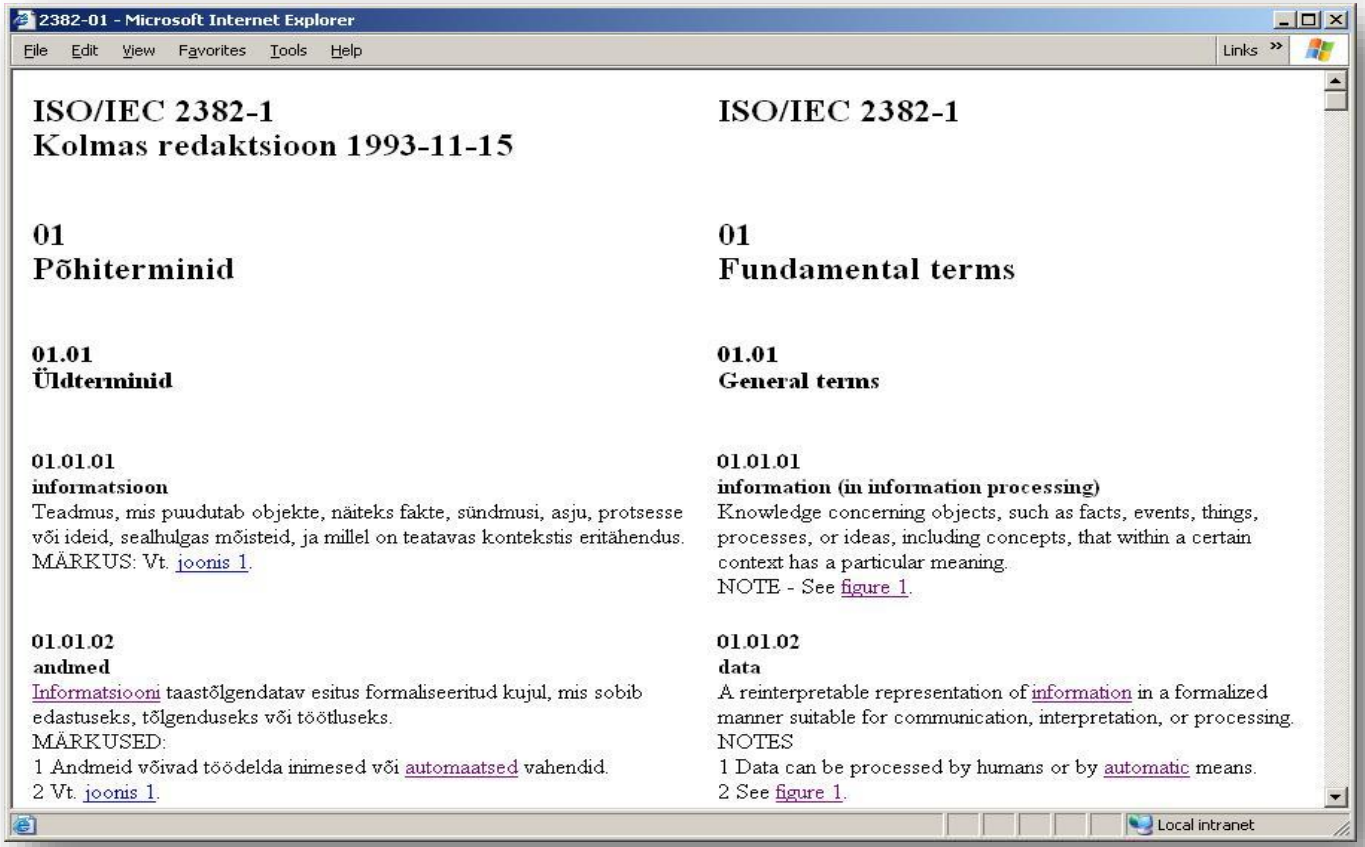
European Language Resource Coordination
Connecting Europe Facility




# Digital textual data in various formats – valuable language resources

European Language Resource Coordination
Connecting Europe Facility

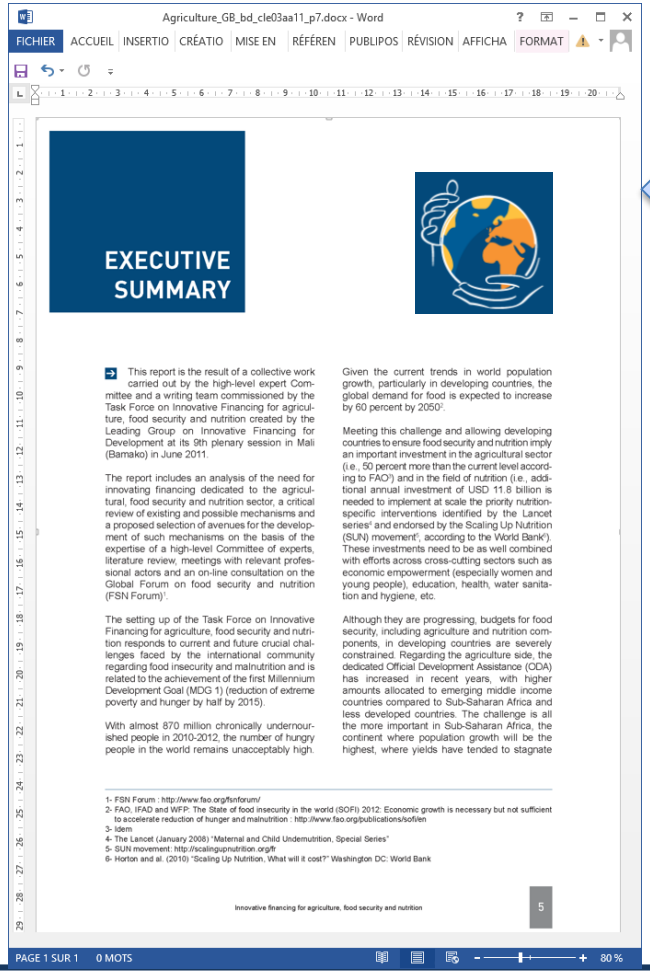
# Example of bilingual parallel document



European Language Resource Coordination
Connecting Europe Facility

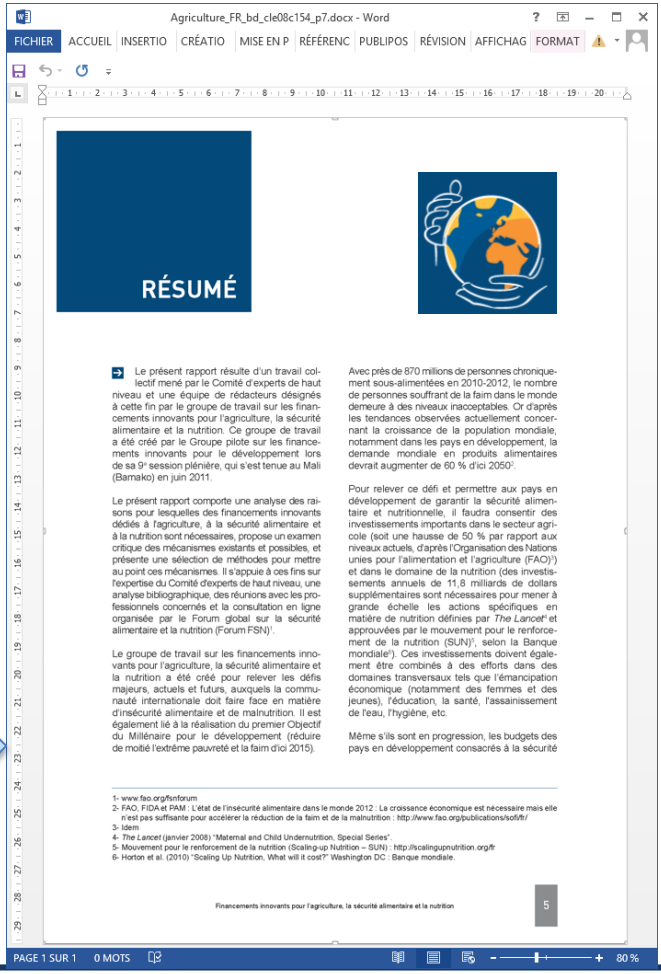
# Processing of Bilingual Documents for Machine Translation

Word docs from http://www.diplomatie.gouv.fr/fr/photos-videos-publications/publications/enjeux-planetaires-cooperation/rapports/article/rapports-du-groupe-pilote, Financements innovants pour l'agriculture, la sécurité alimentaire et la nutrition, Ministère des Affaires étrangères et du Développement international




English version

French version



European Language Resource Coordination
Connecting Europe Facility

# Processing of Bilingual Documents for Machine Translation

## English version – Raw text

Executive Summary
This report is the result of a collective work carried out by the high-level expert Committee and a writing team commissioned by the Task Force on Innovative Financing for agriculture, food security and nutrition created by the Leading Group on Innovative Financing for Development at its 9th plenary session in Mali (Bamako) in June 2011.
The report includes an analysis of the need for innovating financing dedicated to the agricultural, food security and nutrition sector, a critical review of existing and possible mechanisms and a proposed selection of avenues for the development of such mechanisms on the basis of the expertise of a high-level Committee of experts, literature review, meetings with relevant professional actors and an on-line consultation on the Global Forum on food security and nutrition (FSN Forum)1.
The setting up of the Task Force on Innovative Financing for agriculture, food security and nutrition responds to current and future crucial challenges faced by the international community
[...]

## French version – Raw text

Résumé
Le présent rapport résulte d'un travail collectif mené par le Comité d'experts de haut niveau et une équipe de rédacteurs désignés à cette fin par le groupe de travail sur les financements innovants pour l'agriculture, la sécurité alimentaire et la nutrition. Ce groupe de travail a été créé par le Groupe pilote sur les financements innovants pour le développement lors de sa 9e session plénière, qui s'est tenue au Mali (Bamako) en juin 2011.
Le présent rapport comporte une analyse des raisons pour lesquelles des financements innovants dédiés à l'agriculture, à la sécurité alimentaire et à la nutrition sont nécessaires, propose un examen critique des mécanismes existants et possibles, et présente une sélection de méthodes pour mettre au point ces mécanismes. Il s'appuie à ces fins sur l'expertise du Comité d'experts de haut niveau, une analyse bibliographique, des réunions avec les professionnels concernés et la consultation en ligne organisée par le Forum global sur la sécurité alimentaire et la nutrition (Forum FSN)1.
Le groupe de travail sur les financements innovants pour l'agriculture, la sécurité alimentaire et la nutrition a été créé pour relever les défis majeurs, actuels et futurs, auxquels la communauté
[...]

European Language Resource Coordination
Connecting Europe Facility

# Processing of Bilingual Documents for Machine Translation

## Alignement of English and French versions

| | |
|---|---|
| **S1.** Executive Summary | **S1.** Résumé |
| **S2.** This report is the result of a collective work carried out by the high-level expert Committee and a writing team commissioned by the Task Force on Innovative Financing for agriculture, food security and nutrition created by the Leading Group on Innovative Financing for Development at its 9th plenary session in Mali (Bamako) in June 2011. | **S2.** Le présent rapport résulte d'un travail collectif mené par le Comité d'experts de haut niveau et une équipe de rédacteurs désignés à cette fin par le groupe de travail sur les financements innovants pour l'agriculture, la sécurité alimentaire et la nutrition.<br>**S3.** Ce groupe de travail a été créé par le Groupe pilote sur les financements innovants pour le développement lors de sa 9e session plénière, qui s'est tenue au Mali (Bamako) en juin 2011. |
| **S3.** The report includes an analysis of the need for innovating financing dedicated to the agricultural, food security and nutrition sector, a critical review of existing and possible mechanisms and a proposed selection of avenues for the development of such mechanisms on the basis of the expertise of a high-level Committee of experts, literature review, meetings with relevant professional actors and an on-line consultation on the Global Forum on food security and nutrition (FSN Forum)1. | **S4.** Le présent rapport comporte une analyse des raisons pour lesquelles des financements innovants dédiés à l'agriculture, à la sécurité alimentaire et à la nutrition sont nécessaires, propose un examen critique des mécanismes existants et possibles, et présente une sélection de méthodes pour mettre au point ces mécanismes.<br>**S5.** Il s'appuie à ces fins sur l'expertise du Comité d'experts de haut niveau, une analyse bibliographique, des réunions avec les professionnels concernés et la consultation en ligne organisée par le Forum global sur la sécurité alimentaire et la nutrition (Forum FSN)1. |
| **S4.** The setting up of the Task Force on Innovative Financing for agriculture, food security and nutrition responds to current and future crucial challenges faced by the international community [...] | **S6.** Le groupe de travail sur les financements innovants pour l'agriculture, la sécurité alimentaire et la nutrition a été créé pour relever les défis majeurs, actuels et futurs, auxquels la communauté [...] |

European Language Resource Coordination
Connecting Europe Facility

# Sentence aligned data from parallel documents

| en | el |
|---|---|
| Greece of art and science | Η Ελλάδα των τεχνών και της επιστήμης |
| Greece is a place of culture, the arts and sciences. | Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. |
| Its tradition of contribution to global cultural and scientific communities, combined with its outstanding natural beauty and excellent infrastructure, has made it an ideal place in which to hold conferences. | Η μακραίωνη συμβολή της στο παγκόσμιο γίγνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις άρτιες υποδομές, την καθιστούν ιδανικό τόπο διεξαγωγής συνεδρίων. |
| Over the last few years, Greece has more and more frequently welcomed people of letters, sciences and the arts, who have participated in symposia, conferences and exhibitions. | Τα τελευταία χρόνια, η ελληνική επικράτεια υποδέχεται όλο και συχνότερα ανθρώπους των γραμμάτων, των επιστημών και των τεχνών, οι οποίοι συμμετέχουν σε συμπόσια, συνέδρια και εκθέσεις. |
| Athens International Airport 'Eleftherios Venizelos' , one of the most modern airports in the world in operation since 2001, greatly boosted the organization of international conferences. | Ο Διεθνής Αερολιμένας Αθηνών «Ελευθέριος Βενιζέλος», ένα από τα πλέον σύγχρονα αεροδρόμια παγκοσμίως, ο οποίος λειτουργεί από το 2001, έδωσε μεγάλη ώθηση στη διοργάνωση διεθνών συνεδρίων. |
| Conference tourism is extremely interdependent: it requires of course a high level of background support from the host country, and at the same time it can actively contribute to improving the overall standard of services in the region. | Ο συνεδριακός τουρισμός είναι άκρως αλληλεπιδραστικός: απαιτεί, βέβαια, ένα υψηλού επιπέδου υπόβαθρο από τη χώρα υποδοχής, ταυτόχρονα όμως συμβάλλει ενεργά στην αναβάθμιση της συνολικής ποιότητας μιας περιοχής. |
| It is logical that a country chosen as a conference location should be involved in the cultural 'product', giving the public, both residents and visitors, the chance to experience human achievement and innovative thought. | Είναι λογικό, ένας χώρος ο οποίος προτιμάται για τη διεξαγωγή συνεδρίων, να μετέχει προνομιακά στο πολιτιστικό «προϊόν», μιας και δίνει τη δυνατότητα σε κοινό, κατοίκους και επισκέπτες, να έρθουν σε επαφή με τα ανθρώπινα επιτεύγματα και τις καινοτομίες. |
| The Greece of the pre-Socratic philosophers, of the great poets, of Pheidias the sculptor and Asclepius the physician, extends its hospitality and its warmest welcome, honouring people of intellect and creativity, commerce and scientific progress. | Η Ελλάδα των προσωκρατικών φιλοσόφων, των μεγάλων ποιητών, του Φειδία και του Ασκληπιού υποδέχεται φιλόξενα και τιμά τους ανθρώπους του πνεύματος, του εμπορίου και της προόδου. |
| Scientific conferences in the land that gave birth to science | Συνέδρια στη χώρα που γέννησε τις επιστήμες |
| Greece has a large number of esteemed scientists, both here in the country and abroad. | Η Ελλάδα διαθέτει μεγάλο και υψηλής αξίας επιστημονικό δυναμικό, τόσο εντός όσο και εκτός συνόρων. |
| Greek scientists, with their inventions, innovations and research work, play a leading part in the international scientific community. | Οι Έλληνες επιστήμονες, με τις εφευρέσεις, τις καινοτομίες και το ερευνητικό τους έργο πρωταγωνιστούν στη διεθνή επιστημονική κοινότητα. |
| Numerous important scientific conferences take place in Greece, reflecting the significance the country places on innovative science. | Τα επιστημονικά συνέδρια που λαμβάνουν χώρα στην Ελλάδα είναι και πολλά και σημαντικά, αντανακλώντας τη σημασία που δίνει η χώρα στις καινοτόμες επιστήμες. |
| Medical, architectural, natural and humanistic scientific conferences enrich Greece's cultural life, and at the same time give participants the opportunity to experience the | Ιατρικά συνέδρια, αρχιτεκτονικά, φυσικών και ανθρωπιστικών επιστημών, πλουτίζουν την πολιτιστική ζωή της |

European Language Resource Coordination
Connecting Europe Facility

# Example of Comparable Data

**English**

Telecommunication occurs when the exchange of information between two or more entities (communication) includes the use of technology.

Communication technology uses channels to transmit information (as electrical signals), either over a physical medium (such as signal cables), or in the form of electromagnetic waves.

The word is often used in its plural form, telecommunications, because it involves many different technologies.

**Greek**

Με τον γενικό όρο τηλεπικοινωνίες, (telecommunications), χαρακτηρίζεται η κάθε μορφής ενσύρματη ή ασύρματη, ηλεκτρομαγνητική, ηλεκτρική, κ.λπ., ακουστική και οπτική επικοινωνία που πραγματοποιείται ανεξαρτήτως απόστασης.

Στους σύγχρονους καιρούς, αυτή η διαδικασία σχεδόν πάντα περιλαμβάνει την αποστολή ηλεκτρομαγνητικών κυμάτων ή ηλεκτρικών σημάτων από κατάλληλες ηλεκτρονικές συσκευές, όπως το τηλέφωνο ή ο ασύρματος, αλλά παλαιότερα περιελάμβανε τη χρήση ακουστικών σημάτων, όπως τυμπάνων, ή οπτικών, όπως ο σηματοφόρος καπνός ή η λάμψη της φωτιάς.

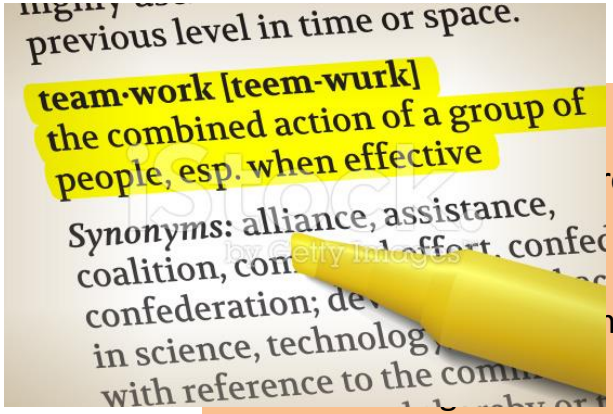
**Spanish**

Una telecomunicación es toda transmisión y recepción de señales de cualquier naturaleza, típicamente electromagnéticas, que contengan signos, sonidos, imágenes o, en definitiva, cualquier tipo de información que se desee comunicar a cierta distancia.

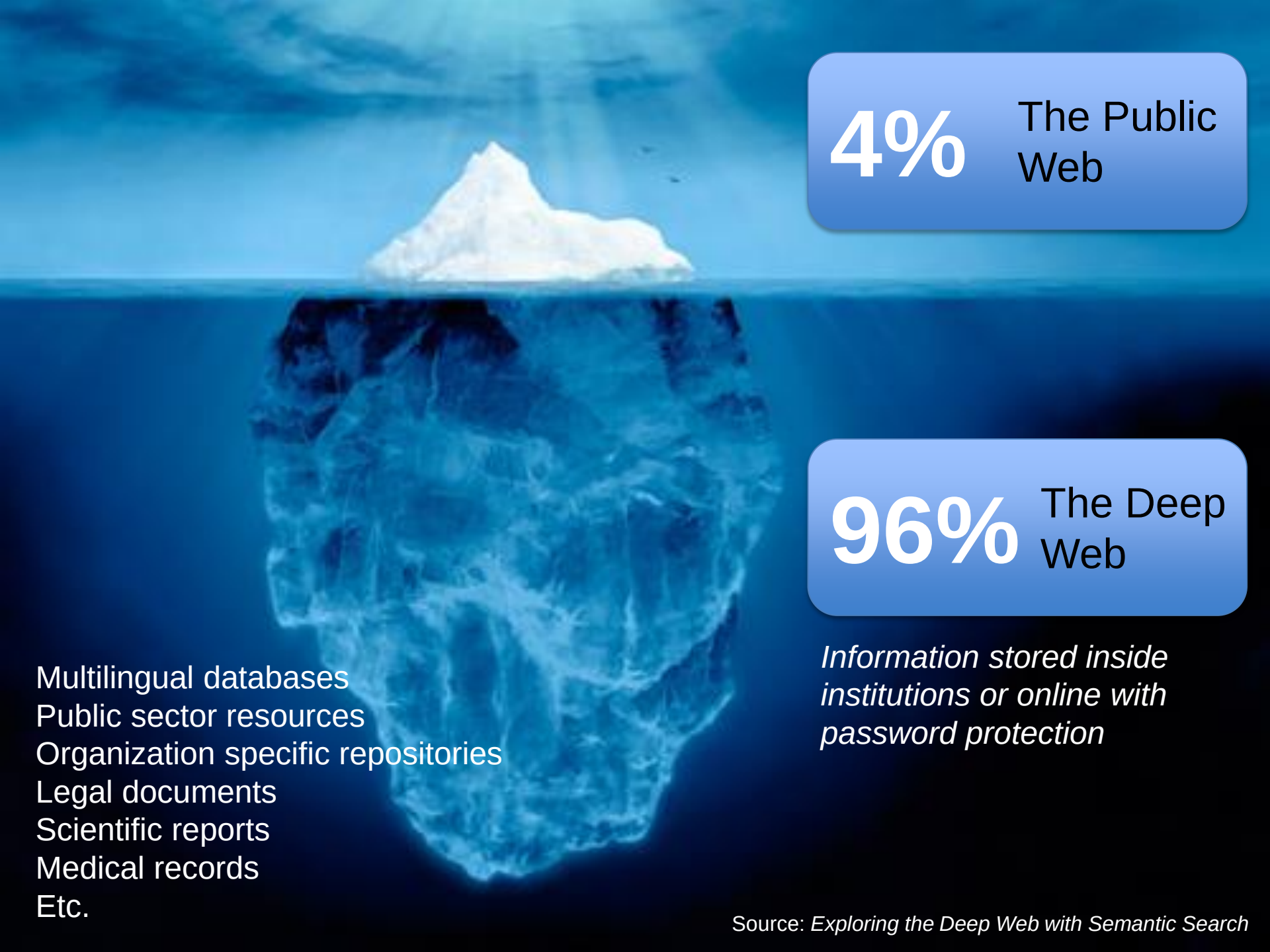
Por metonimia, también se denomina telecomunicación (o telecomunicaciones, indistintamente) a la disciplina que estudia, diseña, desarrolla y explota aquellos sistemas que permiten dichas comunicaciones; de forma análoga, la ingeniería de telecomunicaciones resuelve los problemas técnicos asociados a esta disciplina.

Source: First sentences of articles for Telecommunications in the English, Greek and Spanish Wikipedias

# Dictionaries / Terminologies /Ontologies




| | | ES | EL |
|---|---|---|---|
| | | abandono escolar | διακοπή της σχολικής φοίτησης |
| | | despojo | παραπροϊόντα σφαγίων |
| | | sacrificio de animales | σφαγή ζώων |
| | | derogación | κατάργηση |
| 5075 | Abruzzes | Abruzos | Αβρουζία |
| 5339 | absentéisme | absentismo | συστηματική απουσία από την εργασία |
| 5984 | abstentionnisme | abstencionismo | αποχή |
| 2 | abus de confiance | abuso de confianza | απιστία |
| | abus de droit | abuso de derecho | κατάχρηση δικαιώματος |
| | abus de pouvoir | abuso de poder | κατάχρηση εξουσίας |
| | accès à l'éducation | acceso a la educación | πρόσβαση στην εκπαίδευση |
| | accès à l'emploi | acceso al empleo | πρόσβαση στην αγορά εργασίας |


4%
The Public Web
96%
The Deep Web
Information stored inside institutions or online with password protection
Multilingual databases
Public sector resources
Organization specific repositories
Legal documents
Scientific reports
Medical records
Etc.
Source: Exploring the Deep Web with Semantic Search

European Language Resource Coordination
Connecting Europe Facility

# Public Sector Data Sources

- Visible data e.g. Web
  - Public websites
  - leaflets, brochures, news etc.
- Invisible Data: archives , hidden web, internal repositories
  - translated documents, reports, speeches, meeting minutes etc.
- Data from outsourced translations:
  - From Language Service Providers and freelance translators
  - Translation Memory as part of contract delivery
  - Respective provisions in the contracts for outsourced translations

European Language Resource Coordination
Connecting Europe Facility

# ELRC Support on Legal Aspects

- Legal Issues
  - Legal status determination (accept/reject decision)
  - Accuracy and acceptance of privacy processing (e.g., anonymization)
  - Application of PSI versus need for a License
- Role of the ELRC Consortium
  - Support on practical issues
  - Technical/legal helpdesk, consultancy
  - Model licensing agreements
    - Government Open Licenses
    - Standard Re-use Licenses
    - License interoperability

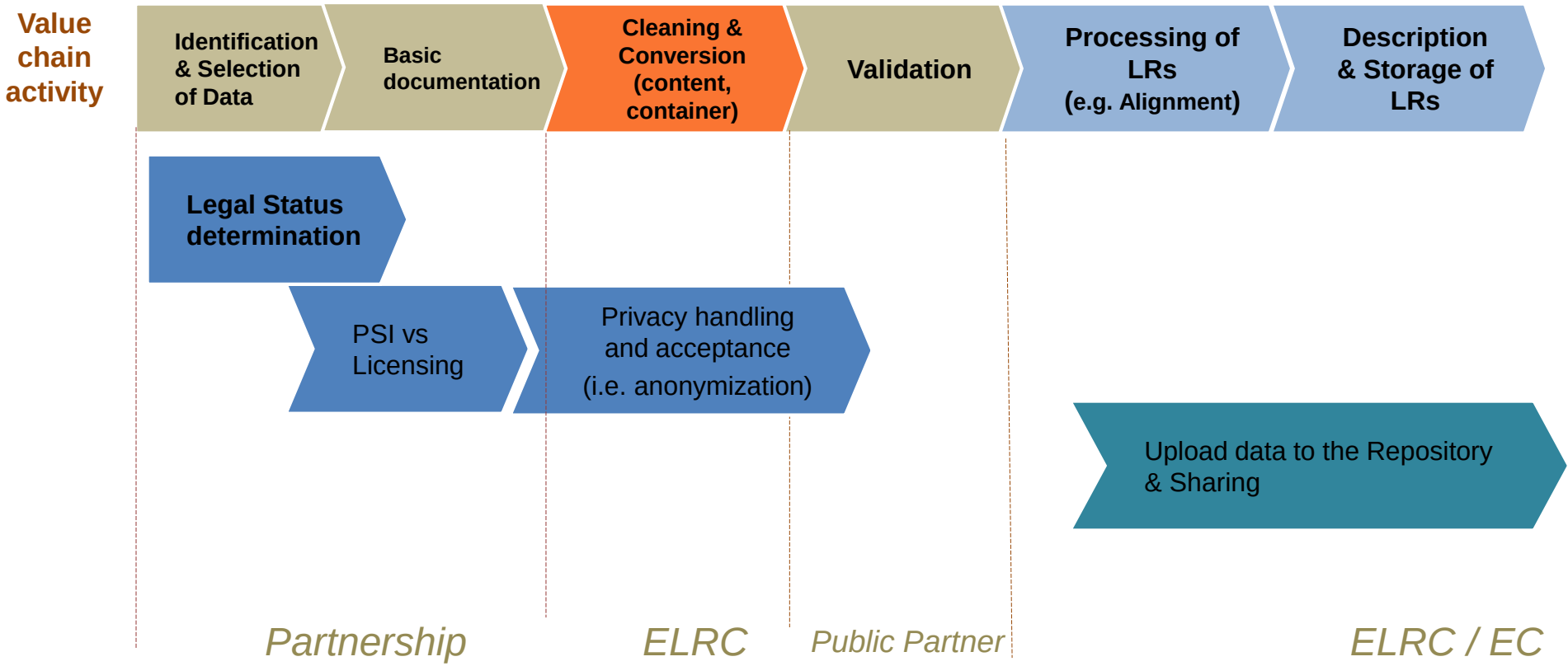
European Language Resource Coordination
Connecting Europe Facility

# Data Anonymization

- Identify a large source of data on individuals, organizations etc.
- Use a Named Entity Recognizer (NER) to find and remove private data (names, locations, dates, birth information, etc.) and replace with generic placeholders.
- Confirm results meet acceptable requirements
  - Reject data if anonymization not accurate as required

European Language Resource Coordination
Connecting Europe Facility

# Illustration of Data Packaging Workflow
# Data ➔LRs (Language Resources)



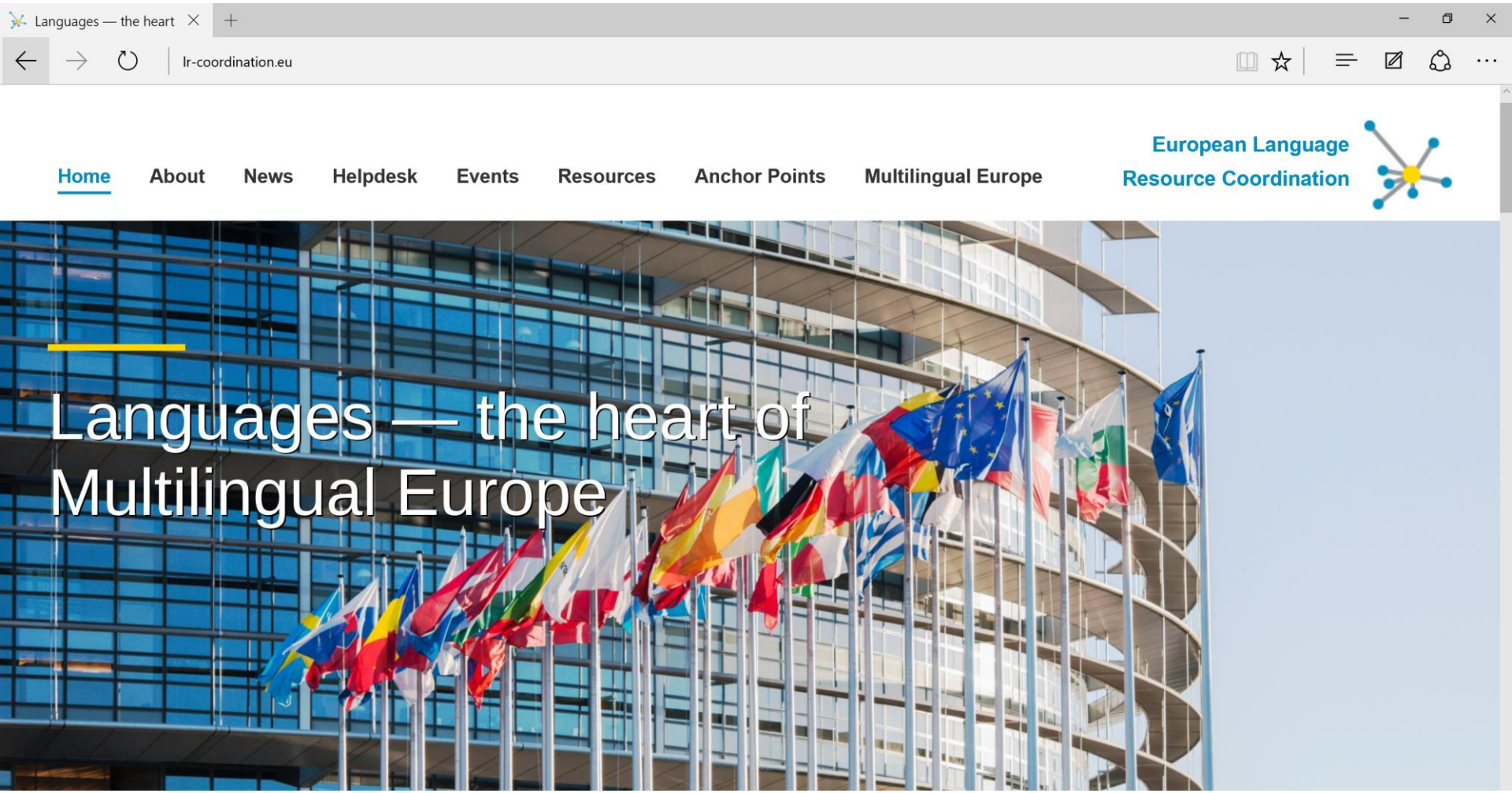
European Language Resource Coordination
Connecting Europe Facility

# Technical aspects dealt by ELRC

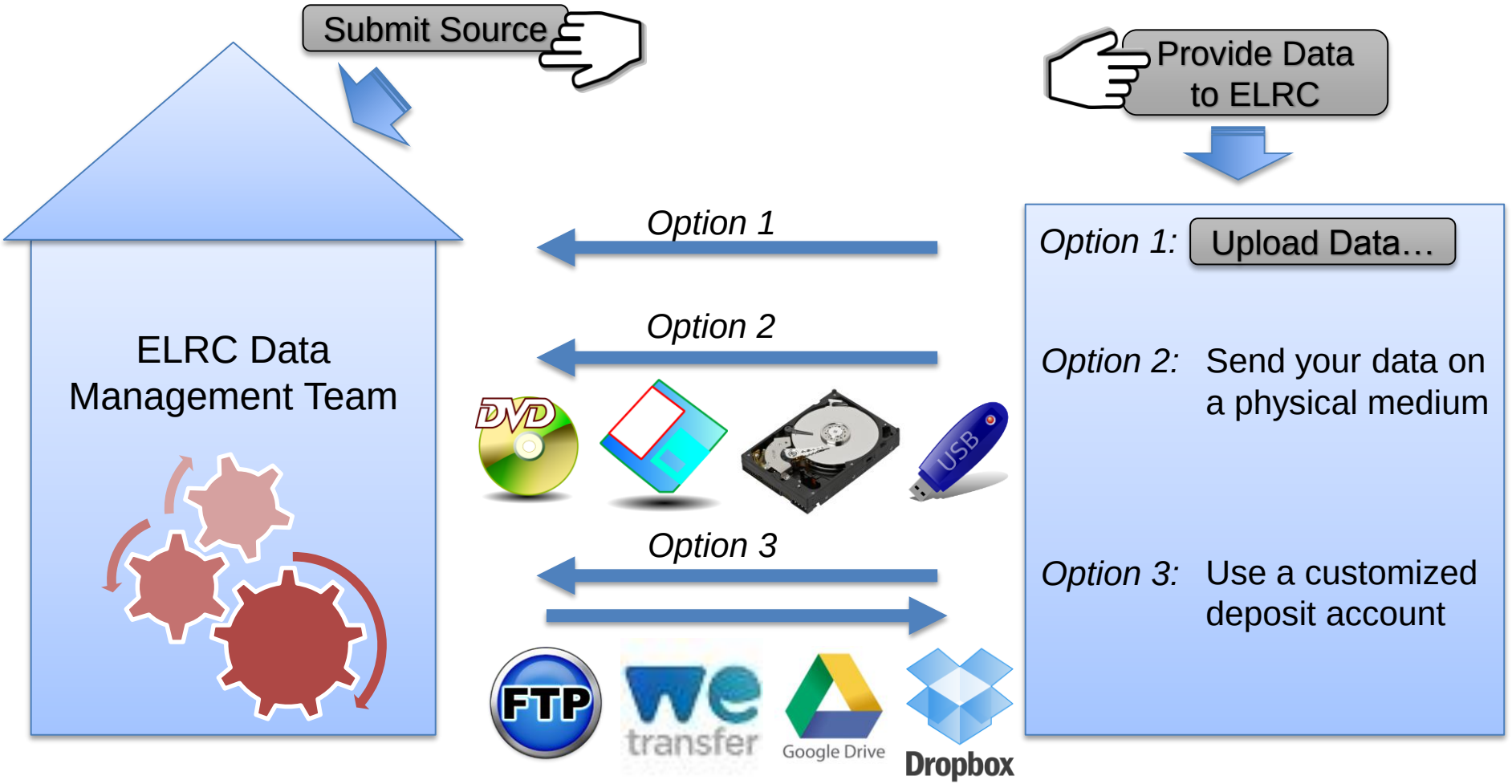
- Cleaning of data format
  *Discarding formatting, encoding character sets e.g. UTF8, formatting features e.g. bold, italic; graphics, ads, tables, html tags, etc.*
- File conversions
  *e.g. conversion to XML, XLIFF, etc.*
- Data preparation for Automated Translation tools
  *e.g. extraction and alignment*
- Validation and Quality Control of the output
  *Language Resource format, content, storage*
- Description of the Language Resource (meta-data)
- Packaging and delivery (data repository with e-sharing) to EC and Owner

European Language Resource Coordination
Connecting Europe Facility

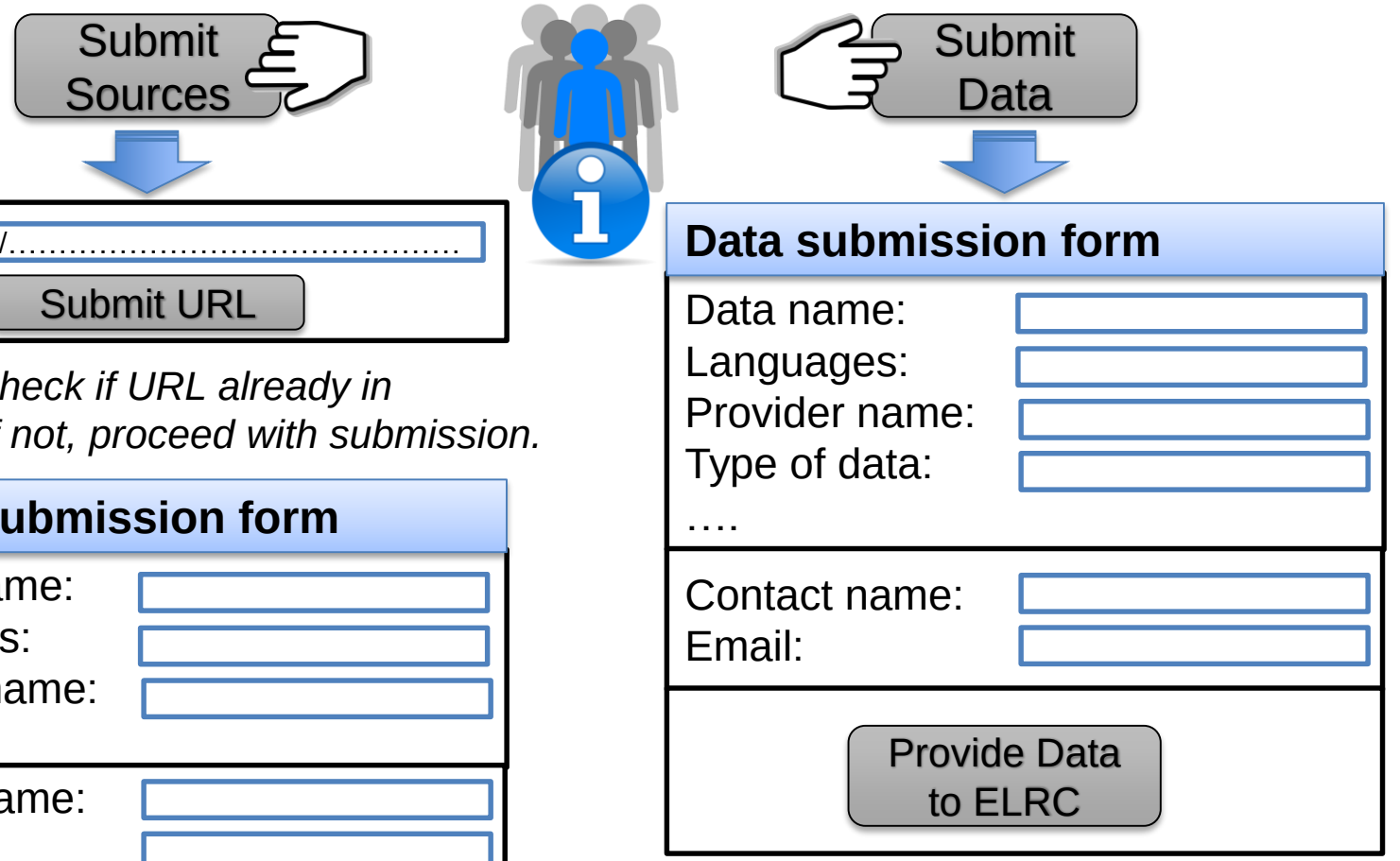
# ELRC Portal: www.lr-coordination.eu

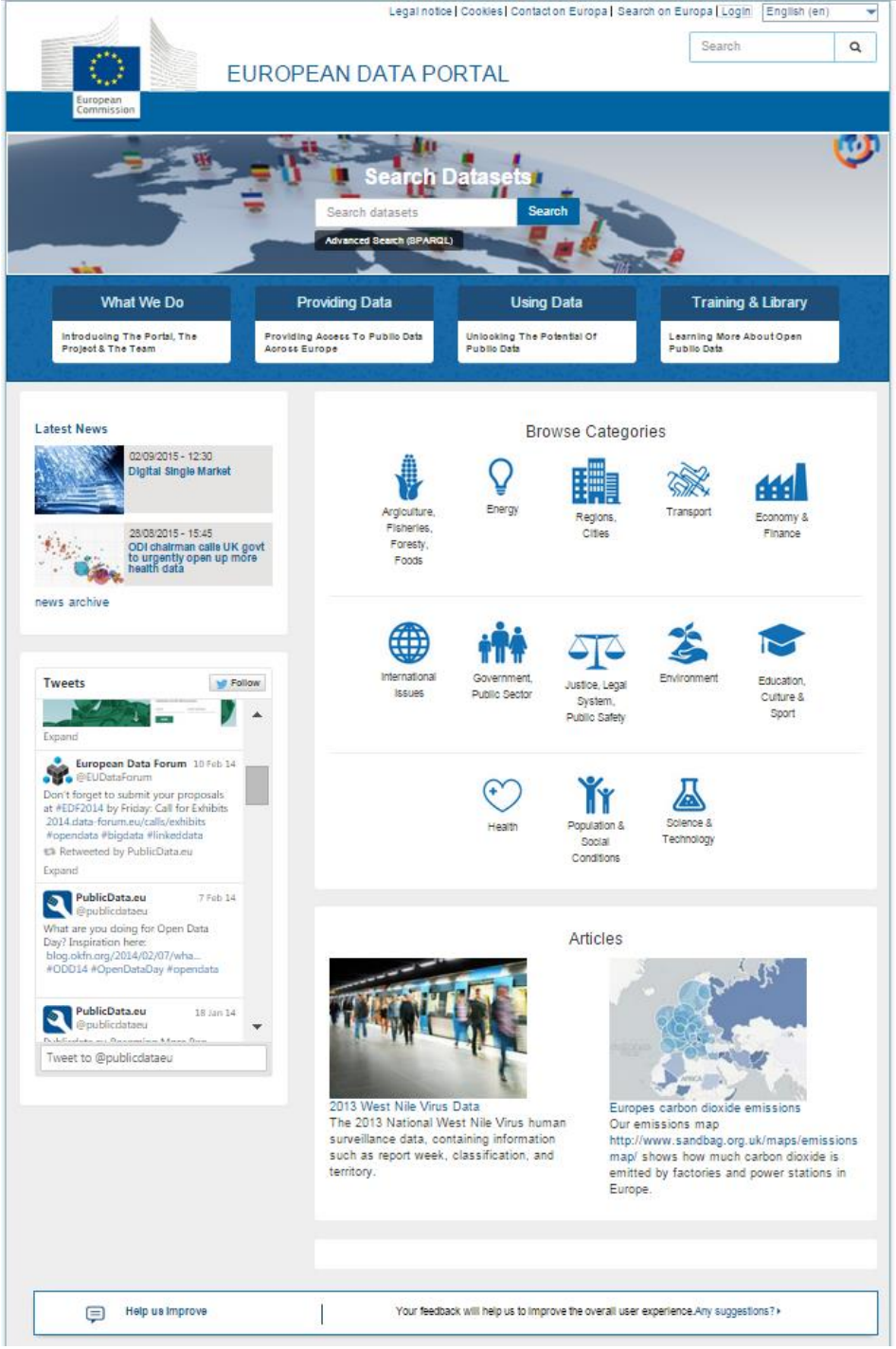


# Options for submitting your data



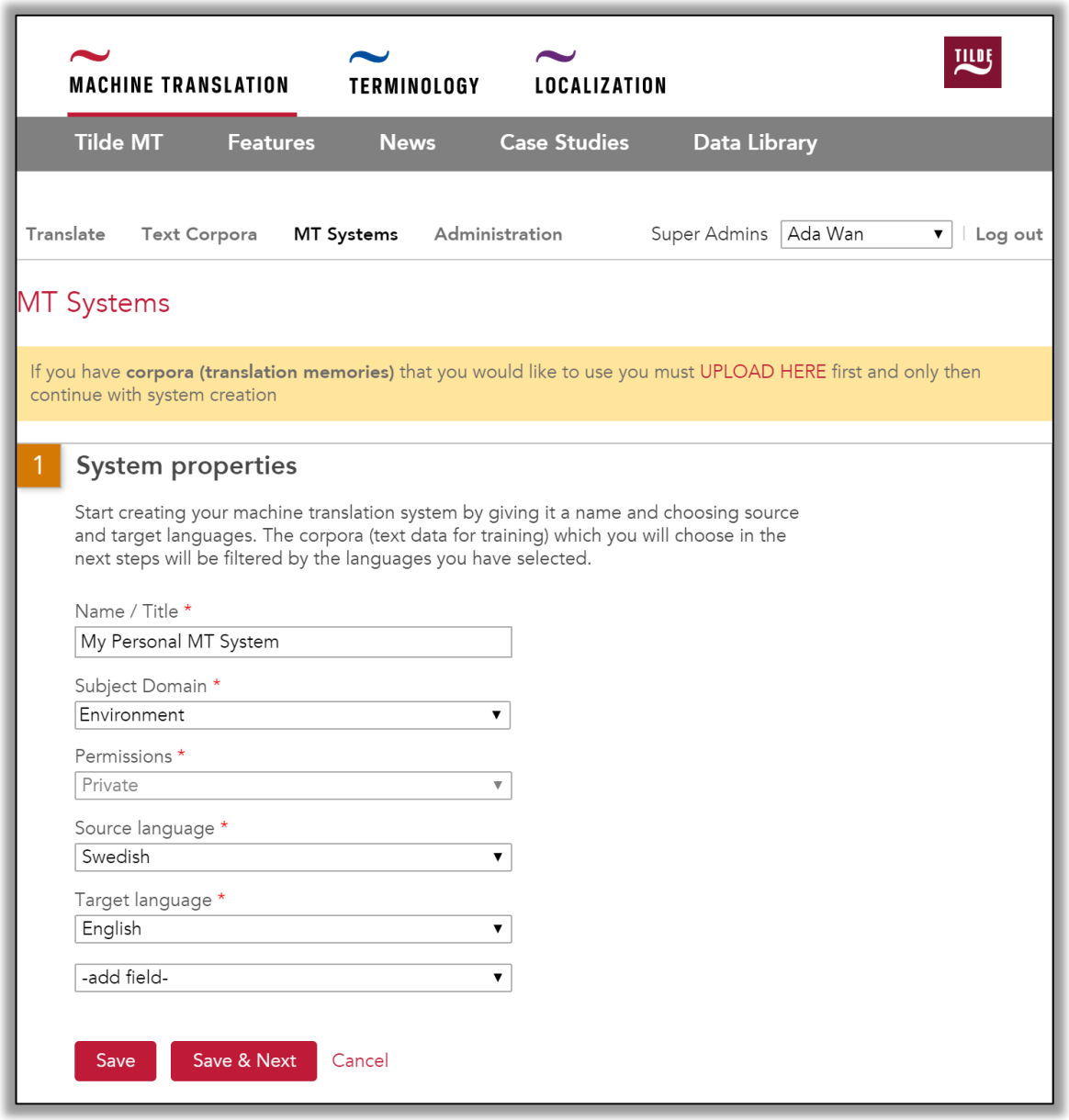
# Data submission to ELRC repository






EUROPEAN DATA PORTAL

# Create your own custom MT with [tilde.com/mt](tilde.com/mt)



# Create you own custom MT with [tilde.com/term](tilde.com/term)

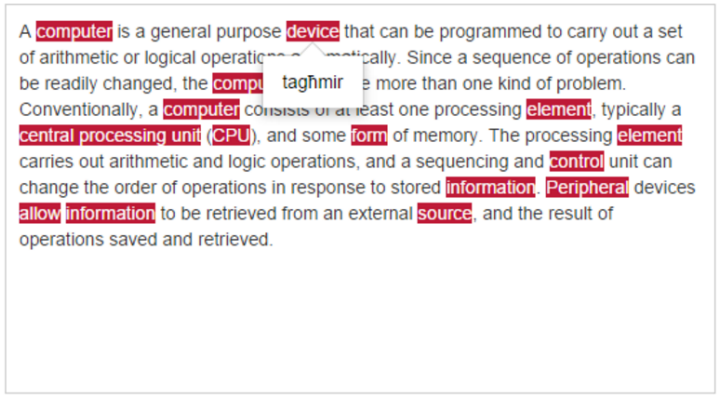
MACHINE TRANSLATION | TERMINOLOGY | LOCALIZATION

## Quick Look up

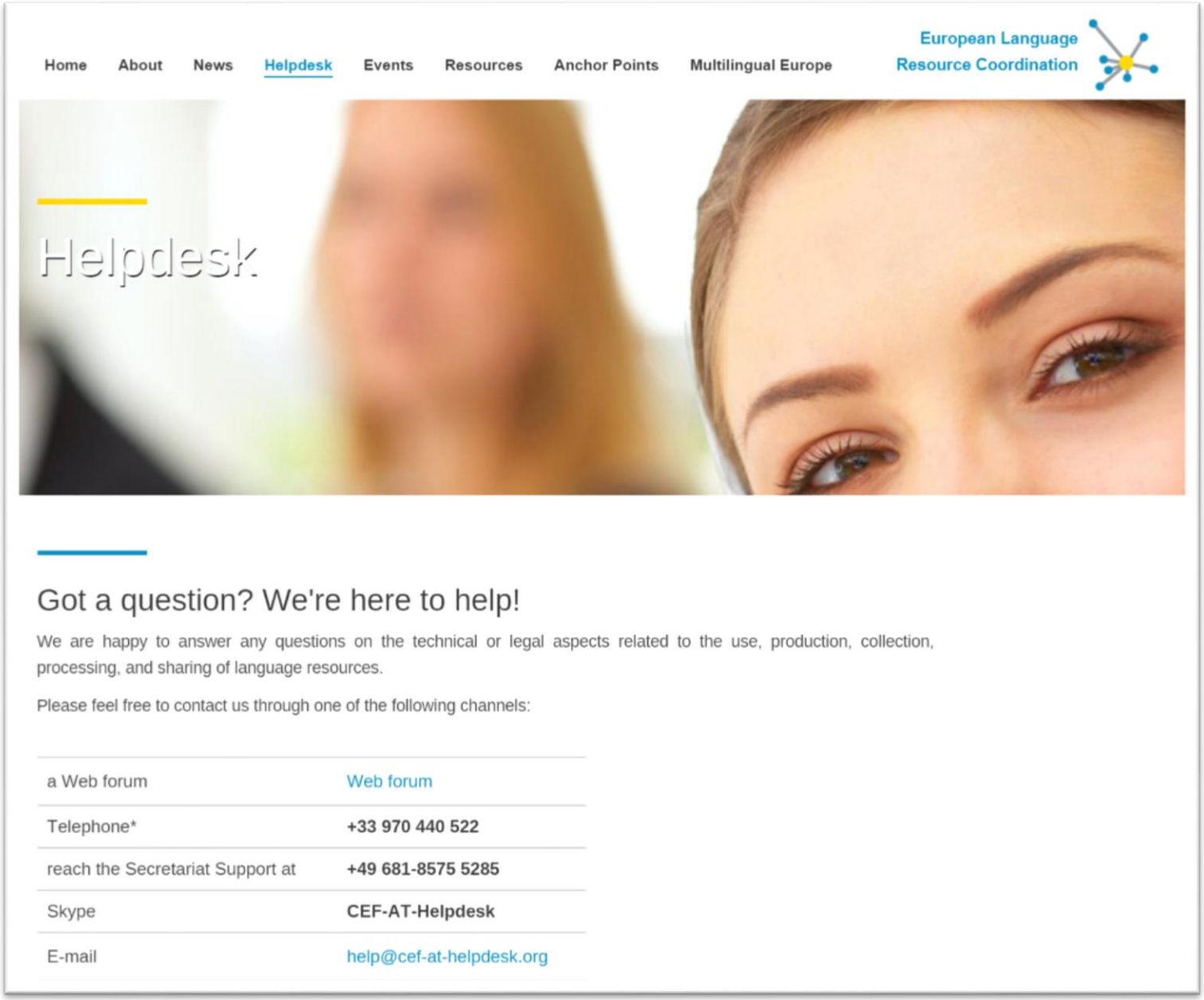
Paste a short text below. Hover over highlighted terms to see translations. Max 2000 characters for unregistered users. Register here.

A computer is a general purpose device that can be programmed to carry out a set of arithmetic or logical operations automatically. Since a sequence of operations can be readily changed, the computer can solve more than one kind of problem. Conventionally, a computer consists of at least one processing element, typically a central processing unit (CPU), and some form of memory. The processing element carries out arithmetic and logic operations, and a sequencing and control unit can change the order of operations in response to stored information. Peripheral devices allow information to be retrieved from an external source, and the result of operations saved and retrieved.

Look up



# ELRC Portal: Helpdesk



European Language Resource Coordination

Connecting Europe Facility

# Takk fyrir!

Looking forward to a fruitful cooperation to make Icelandic machine translation better!

European Language Resource Coordination (ELRC) is a tender operating under the EU's Connecting Europe Facility SMART 2014/1074 programme.