European Language Resource Coordination
Connecting Europe Facility

“What Data Is Needed? Why?”

Dr. Khalid Choukri
(Evaluations and Language Resource Association)

What types of data? Translation

- From previous session, we have seen the predominant approach of data-driven paradigm
  - We **“learn”** from existing data

- How are Language Resources produced:
  - from documents and data to valuable Language Resources for MT
  - Why it is important that you help us with the data you have / you know about

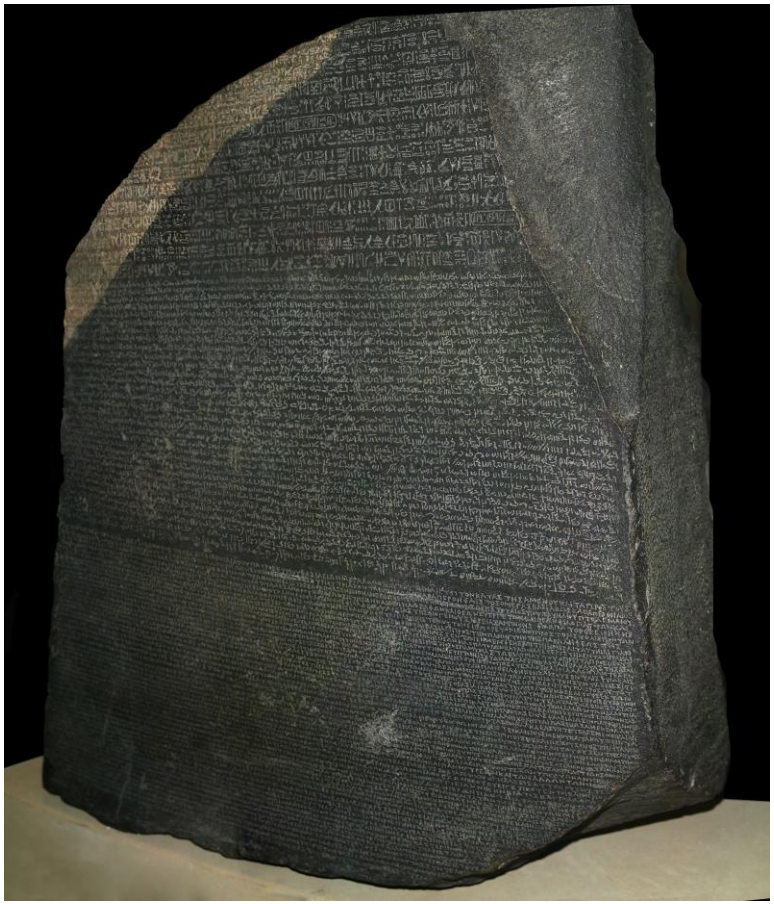
- *The focus is on data in all languages (EU/CEF).*

European Language Resource Coordination
Connecting Europe Facility

What Data

European Language Resource Coordination
Connecting Europe Facility

Translations & Automated Translations



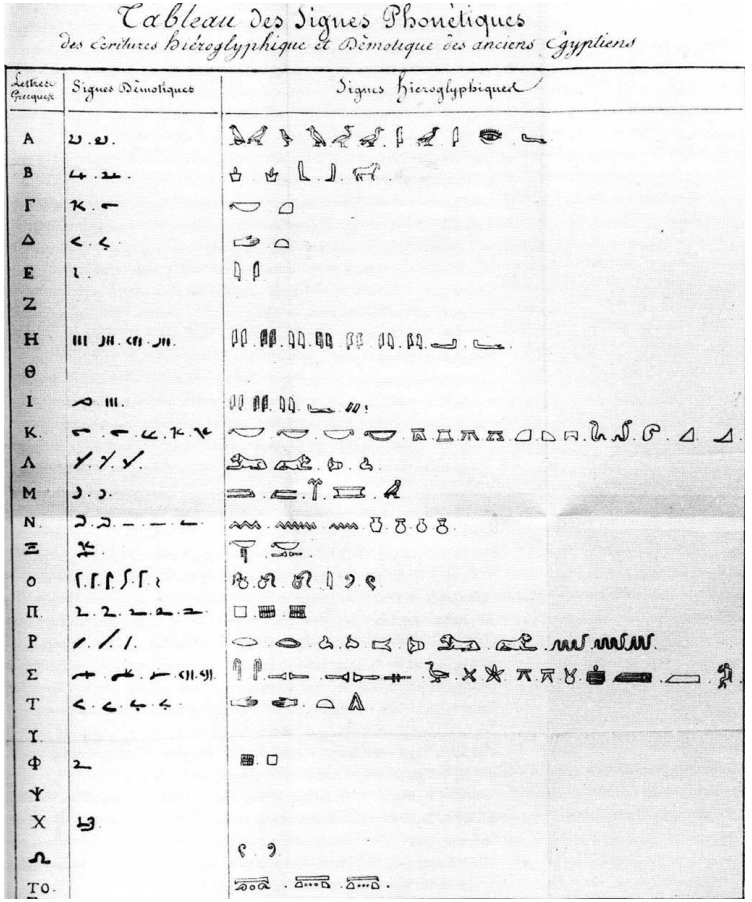
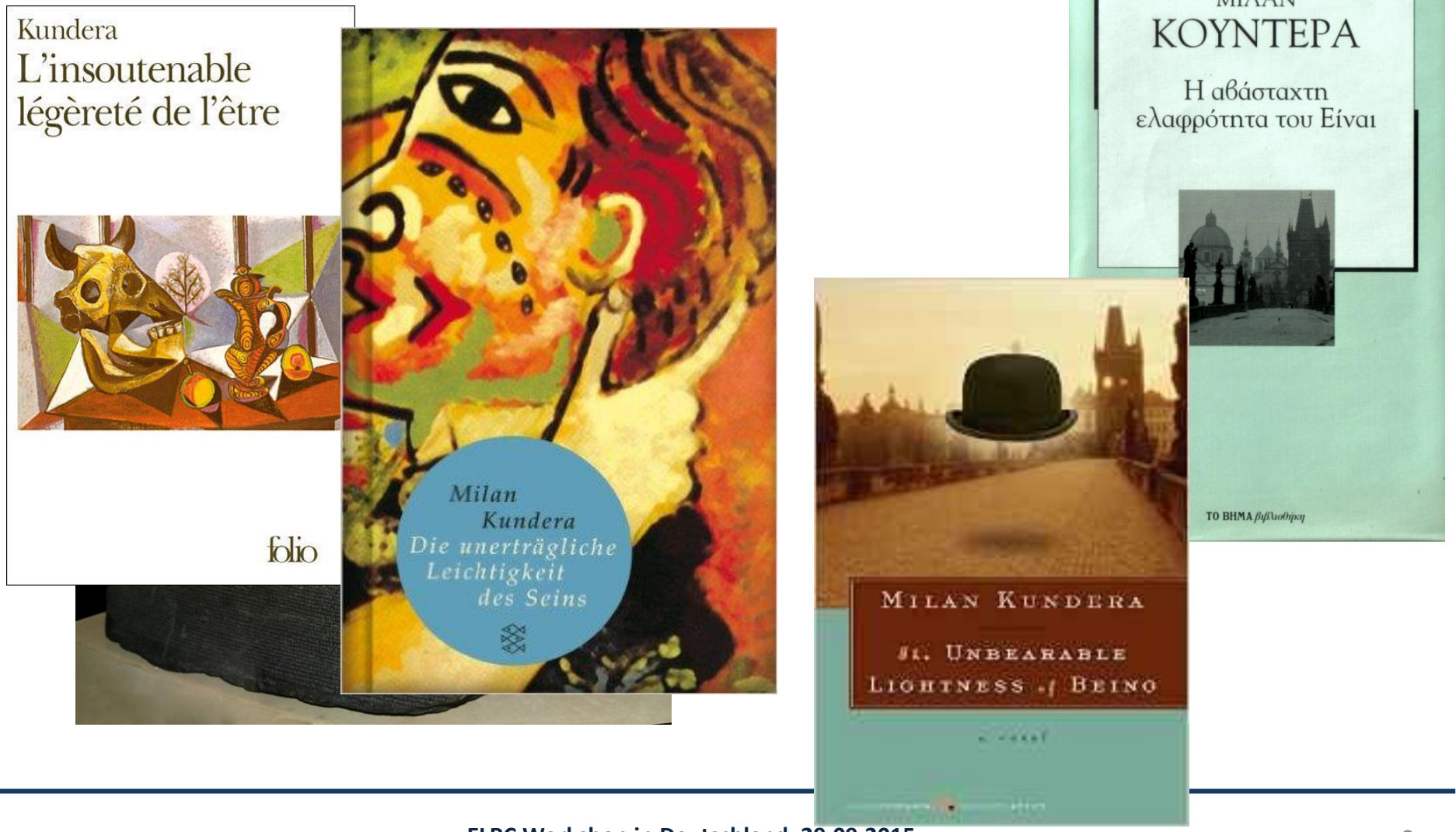


Tableau des Signes Phonétiques
des écritures hiéroglyphique et démotique des anciens Égyptiens

Lettres Grecques	Signes Démotiques	Signes hieroglyphiques
A		
B		
Γ		
Δ		
E		
Z		
H		
Θ		
I		
K		
Λ		
M		
N		
Ξ		
O		
Π		
P		
Σ		
T		
Υ		
Φ		
Ψ		
X		
Ω		
To		

European Language Resource Coordination
Connecting Europe Facility

Translations



European Language Resource Coordination
Connecting Europe Facility

What types of data? Translation

European Language Resource Coordination
Connecting Europe Facility

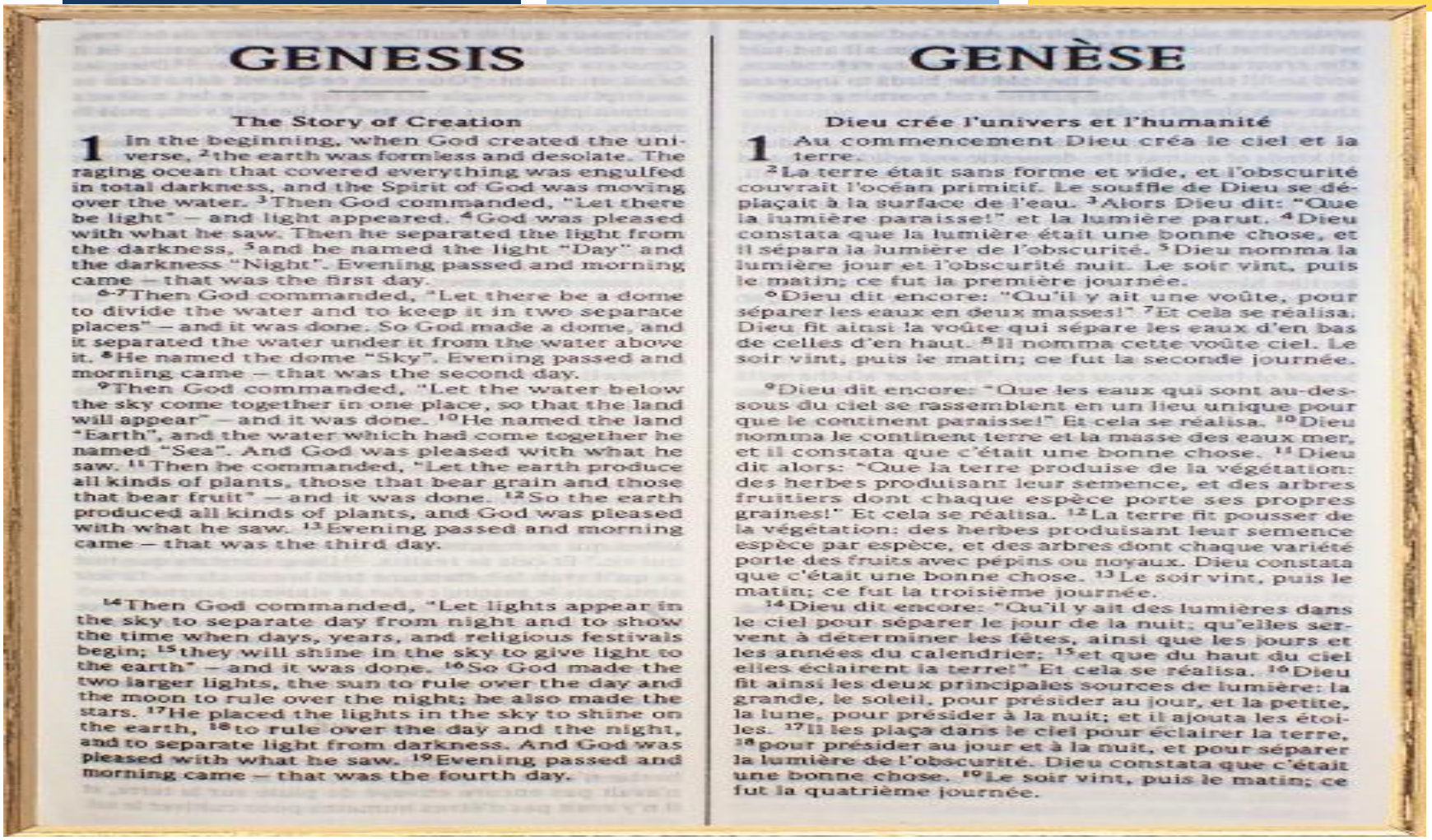
What types of data?
“Aligned” Translation

English

French

European Language Resource Coordination
Connecting Europe Facility

What types of data?
"Aligned" Translation



GENESIS

The Story of Creation

1 In the beginning, when God created the universe, 2 the earth was formless and desolate. The raging ocean that covered everything was engulfed in total darkness, and the Spirit of God was moving over the water. 3 Then God commanded, "Let there be light" — and light appeared. 4 God was pleased with what he saw. Then he separated the light from the darkness, 5 and he named the light "Day" and the darkness "Night". Evening passed and morning came — that was the first day.

6-7 Then God commanded, "Let there be a dome to divide the water and to keep it in two separate places" — and it was done. So God made a dome, and it separated the water under it from the water above it. 8 He named the dome "Sky". Evening passed and morning came — that was the second day.

9 Then God commanded, "Let the water below the sky come together in one place, so that the land will appear" — and it was done. 10 He named the land "Earth", and the water which had come together he named "Sea". And God was pleased with what he saw. 11 Then he commanded, "Let the earth produce all kinds of plants, those that bear grain and those that bear fruit" — and it was done. 12 So the earth produced all kinds of plants, and God was pleased with what he saw. 13 Evening passed and morning came — that was the third day.

14 Then God commanded, "Let lights appear in the sky to separate day from night and to show the time when days, years, and religious festivals begin; 15 they will shine in the sky to give light to the earth" — and it was done. 16 So God made the two larger lights, the sun to rule over the day and the moon to rule over the night; he also made the stars. 17 He placed the lights in the sky to shine on the earth, 18 to rule over the day and the night, and to separate light from darkness. And God was pleased with what he saw. 19 Evening passed and morning came — that was the fourth day.

GENÈSE

Dieu crée l'univers et l'humanité

1 Au commencement Dieu créa le ciel et la terre.
2 La terre était sans forme et vide, et l'obscurité couvrait l'océan primitif. Le souffle de Dieu se déplaçait à la surface de l'eau. 3 Alors Dieu dit: "Que la lumière paraisse!" et la lumière parut. 4 Dieu constata que la lumière était une bonne chose, et il sépara la lumière de l'obscurité. 5 Dieu nomma la lumière jour et l'obscurité nuit. Le soir vint, puis le matin; ce fut la première journée.

6 Dieu dit encore: "Qu'il y ait une voûte, pour séparer les eaux en deux masses!" 7 Et cela se réalisa. Dieu fit ainsi la voûte qui sépare les eaux d'en bas de celles d'en haut. 8 Il nomma cette voûte ciel. Le soir vint, puis le matin; ce fut la seconde journée.

9 Dieu dit encore: "Que les eaux qui sont au-dessous du ciel se rassemblent en un lieu unique pour que le continent paraisse!" Et cela se réalisa. 10 Dieu nomma le continent terre et la masse des eaux mer, et il constata que c'était une bonne chose. 11 Dieu dit alors: "Que la terre produise de la végétation: des herbes produisant leur semence, et des arbres fruitiers dont chaque espèce porte ses propres graines!" Et cela se réalisa. 12 La terre fit pousser de la végétation: des herbes produisant leur semence espèce par espèce, et des arbres dont chaque variété porte des fruits avec pépins ou noyaux. Dieu constata que c'était une bonne chose. 13 Le soir vint, puis le matin; ce fut la troisième journée.

14 Dieu dit encore: "Qu'il y ait des lumières dans le ciel pour séparer le jour de la nuit; qu'elles servent à déterminer les fêtes, ainsi que les jours et les années du calendrier; 15 et que du haut du ciel elles éclairent la terre!" Et cela se réalisa. 16 Dieu fit ainsi les deux principales sources de lumière: la grande, le soleil, pour présider au jour, et la petite, la lune, pour présider à la nuit; et il ajouta les étoiles. 17 Il les plaça dans le ciel pour éclairer la terre, 18 pour présider au jour et à la nuit, et pour séparer la lumière de l'obscurité. Dieu constata que c'était une bonne chose. 19 Le soir vint, puis le matin; ce fut la quatrième journée.

Comparable Collections

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

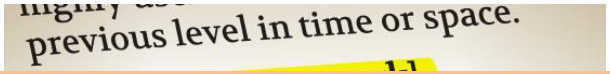
Una telecomunicación es toda transmisión y recepción de señales de cualquier naturaleza, típicamente electromagnéticas, que contengan signos, sonidos, imágenes o, en definitiva, cualquier tipo de información que se desee comunicar a cierta distancia.

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

German page is slightly different but these are (never) translations of one source!!

Dictionaries / Terminologies /Ontologies



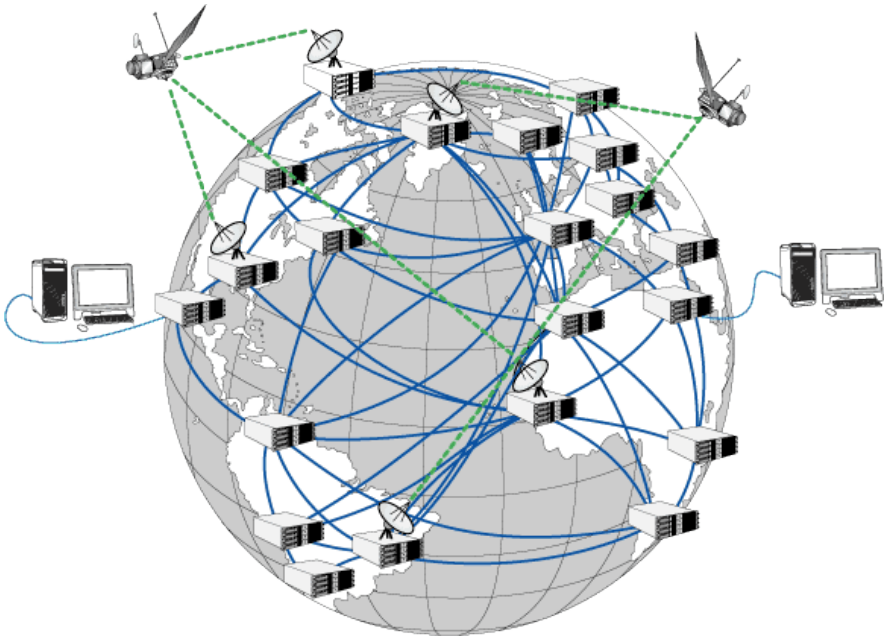
ID	FR	ES	EL
6905	abandon scolaire	abandono escolar	διακοπή της σχολικής φοίτησης
920	abats	despojo	παραπροϊόντα σφαγίων
1857	abattage d'animaux	sacrificio de animales	σφαγή ζώων
6621	abrogation	derogación	κατάργηση
5075	Abruzzes	Abruzos	Αβρουζία
5339	absentéisme	absentismo	συστηματική απουσία από την εργασία
5984	abstentionnisme	abstencionismo	αποχή
2	abus de confiance	abuso de confianza	απιστία
[illegible]	abus de droit	abuso de derecho	κατάχρηση δικαιώματος
[illegible]	abus de pouvoir	abuso de poder	κατάχρηση εξουσίας
[illegible]	accès à l'éducation	acceso a la educación	πρόσβαση στην εκπαίδευση
[illegible]	accès à l'emploi	acceso al empleo	πρόσβαση στην αγορά εργασίας

European Language Resource Coordination
Connecting Europe Facility

Where can we find such data? Digital World

- Archives

- Internet

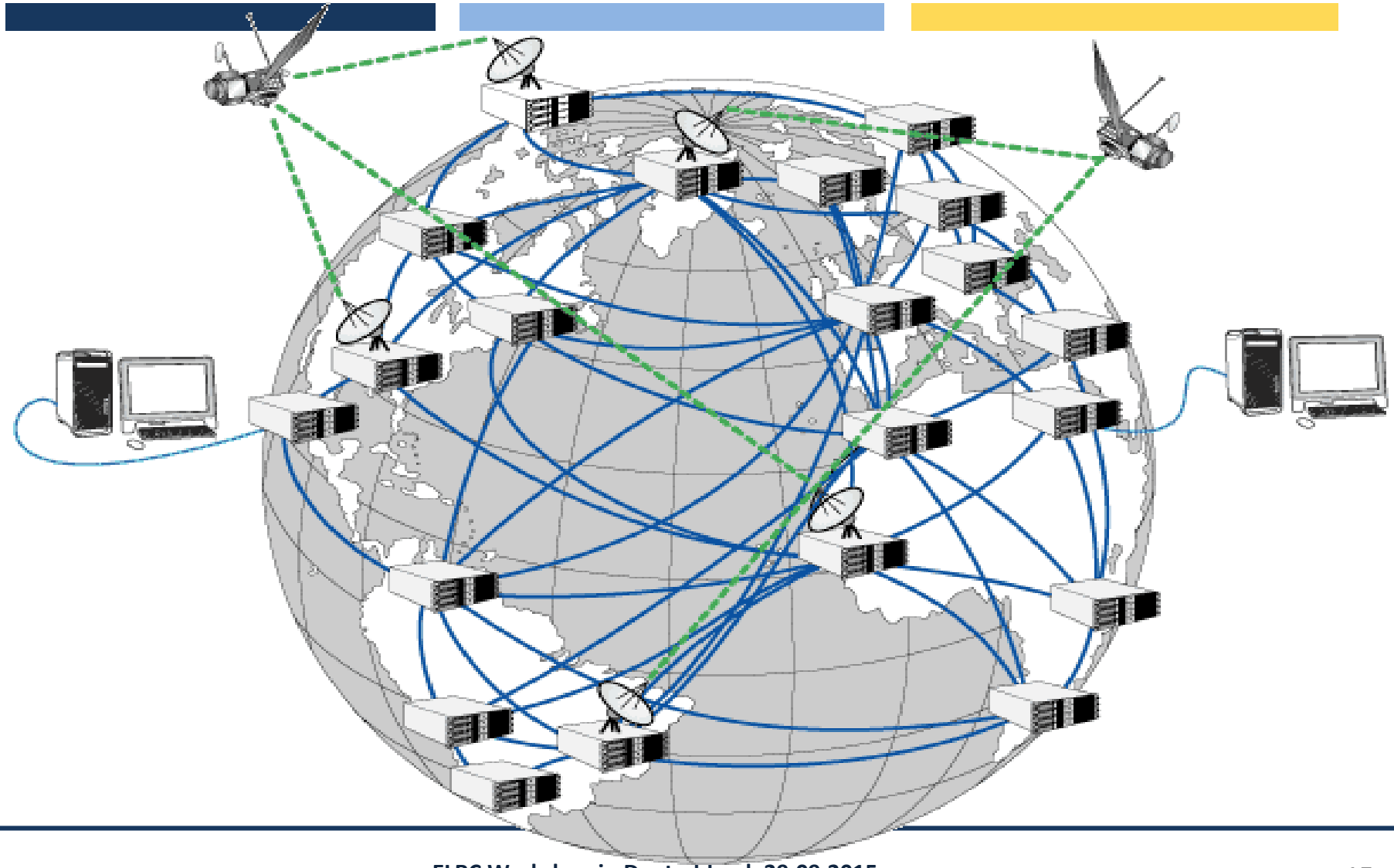


European Language Resource Coordination
Connecting Europe Facility

Digital word … Internet

European Language Resource Coordination
Connecting Europe Facility

Internet Era & Digital Data

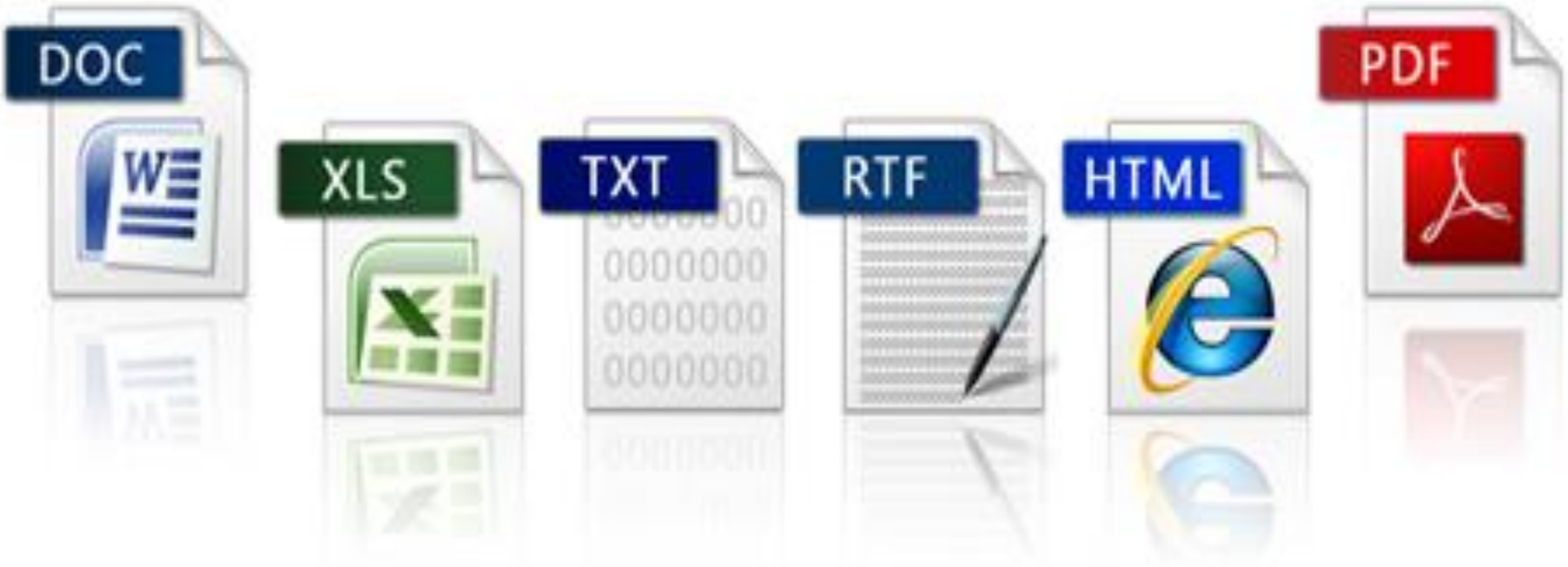


European Language Resource Coordination
Connecting Europe Facility

Of course need for digital textual data !!

European Language Resource Coordination
Connecting Europe Facility

Various Formats



European Language Resource Coordination
Connecting Europe Facility

Documented Data (Meta-data)

Dublin Core Metadata Element Set

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

European Language Resource Coordination
Connecting Europe Facility

How LRs are produced

- Let us see some examples of raw data (html with tables, pictures, etc.) and how they become LRs
  - Discover & identify sources
  - Clear IPR and Get the data (Download, Harvest, Crawl, …)
  - Clean the data (e.g. detect and remove the “boilerplate”, “templates”, pictures, html tags, etc., convert format)
  - Example of tools (Boilerpipe)
  - Document the data
  - Align the translations when identified and break into “sentences”
  - Compute some alignment confidence
  - Share

A Language Resource Factory

- How can this process be turned into a factory of LR production (Automation of the Procedure)
- Some simple illustrations
- We rather start from the Digital word
  - OCR may be considered for the less-resourced languages

European Language Resource Coordination
Connecting Europe Facility

Many web sites…

GREECE
ALL TIME CLASSIC

Ελλάδα Προορισμοί Αξιοθέατα-Δραστηριότητες Διάθεση για Newsletter

visitgreece.gr highlights

European Language Resource Coordination
Connecting Europe Facility

… are rich in multilingual content

European Language Resource Coordination
Connecting Europe Facility

How can we obtain this content…

File Edit View History Bookmarks Tile Tools Help

Travel to Norway - Official ... | Visit Greece | Meetings and... | Visit Greece | Συνεδριακός ...

www.visitgreece.gr/el/meetings_and_incentives

iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map ΒΣΒ ΕΜΤΚ ΜΛ2008 TrEd User's Manual

Greece of art and science

Greece is a place of culture, the arts and sciences. 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**. 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. **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.

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. 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**.

The Greece of the pre-Socratic philosophers, of the great poets, of Pheidias the sculptor

Η Ελλάδα των τεχνών και της επιστήμης

Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. Η μακραίωνη συμβολή της στο παγκόσμιο γίγνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις **άρτιες υποδομές**, την καθιστούν ιδανικό τόπο διεξαγωγής συνεδρίων. 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**. 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.

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. 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.

Η Ελλάδα των προσωκρατικών φιλοσόφων, των μεγάλων ποιητών, του Φειδία και του

European Language Resource Coordination
Connecting Europe Facility

Greece of art and science

Greece is a place of culture, the arts and sciences. 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**. 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. 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.

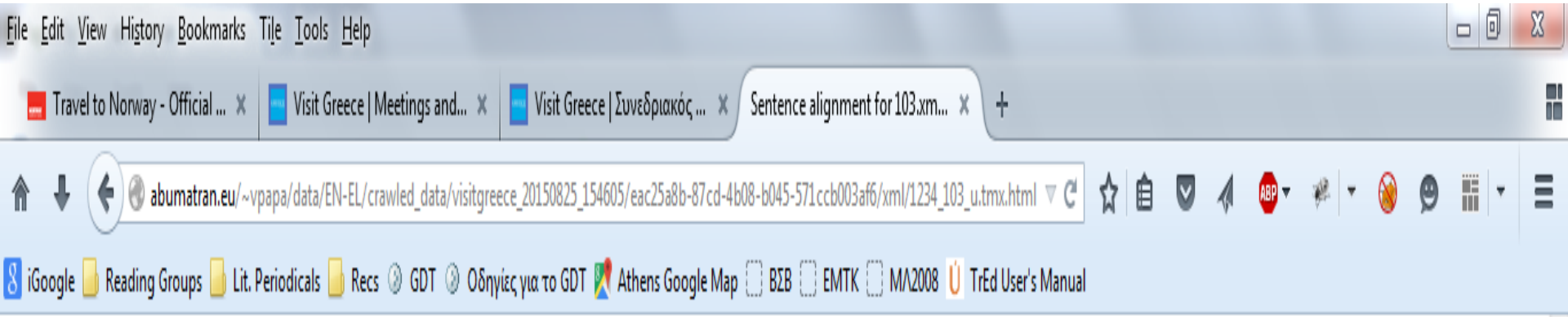
Η Ελλάδα των τεχνών και της επιστήμης

Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. μακραίωνη συμβολή της στο παγκόσμιο γίγνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις **άρτιες υποδομές**, την καθιστούν ιδανικό τόπο διεξαγωγής συνεδρίων. 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**. 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.

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. 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

European Language Resource Coordination
Connecting Europe Facility

… and convert it to valuable Language Resources for Machine Translation?



Sentence alignment for 103.xml (en) - 1234.xml (el)

#	en	el
1	Greece of art and science	Η Ελλάδα των τεχνών και της επιστήμης
2	Greece is a place of culture, the arts and sciences.	Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών.
3	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.
4	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.

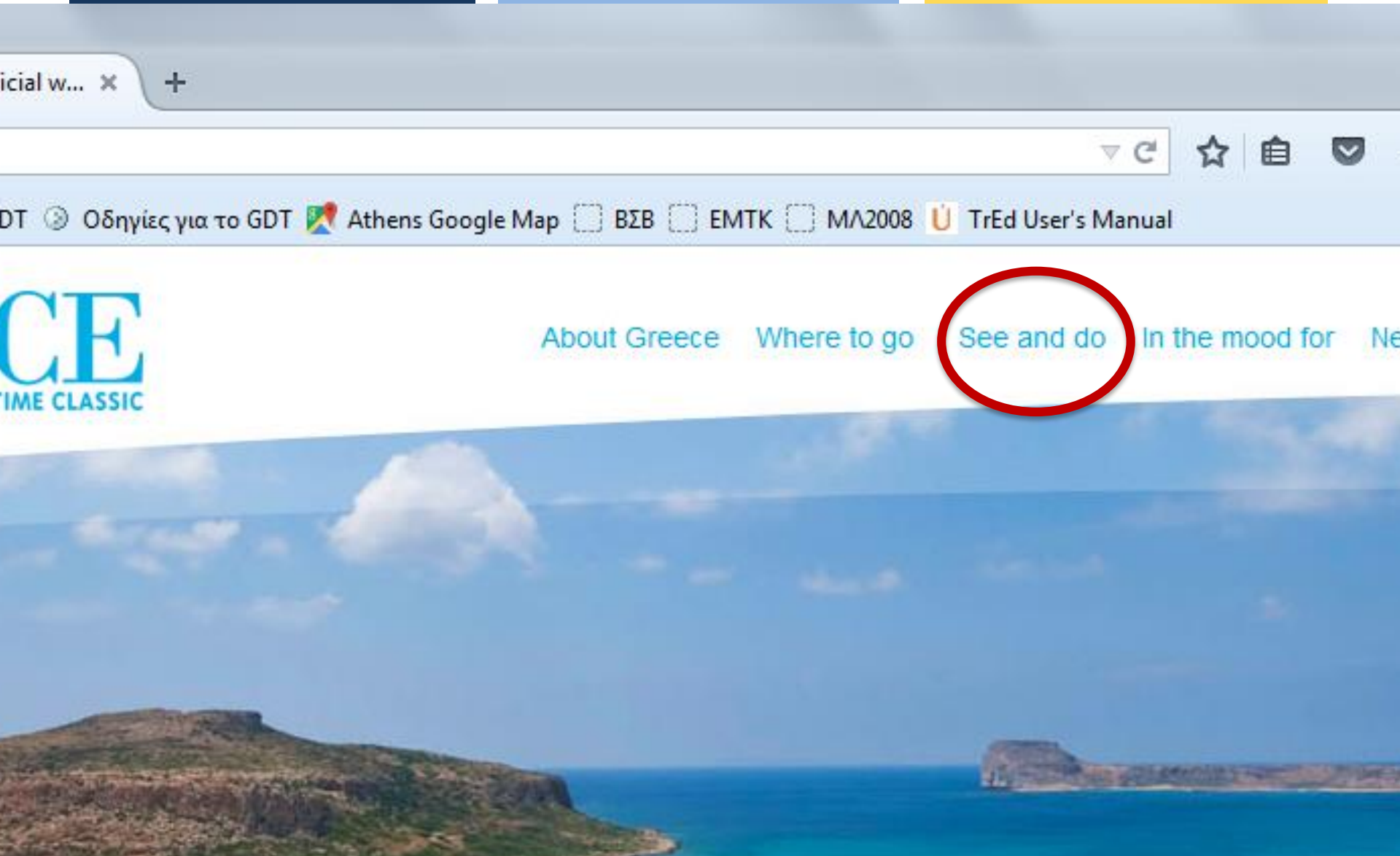
European Language Resource Coordination
Connecting Europe Facility

From a web page to the Factory

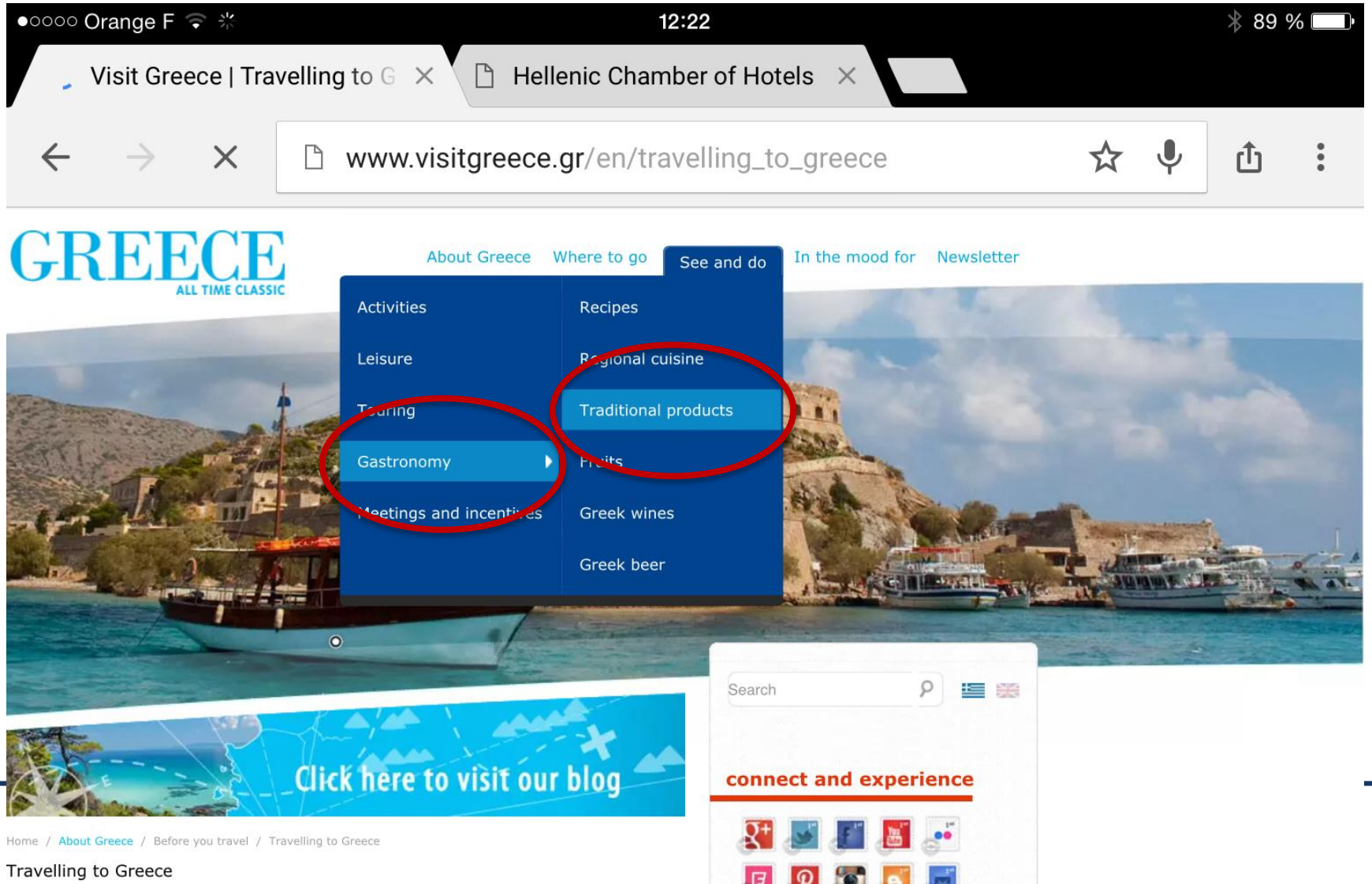
- How does this process scale up:
  - Identify a “useful” source (good candidate for multilingual data)
  - **Review and visit all the links (the URLs referenced in each page)**
    - **“Click on each link” and move forward**
- Get each page and its “potentially” associated one in the other language
- Identify the “**domains”, “genre”,** etc. if possible
- Get rid of the “noise” (ads, format, boilerplate, etc.)
- Align (documents/files, chapters, paragraphs, sentences,)
- Check accuracy of alignment
- Use …. And share

European Language Resource Coordination
Connecting Europe Facility

A Journey in the meandering lines of Internet



(automatically) Follow all referenced links

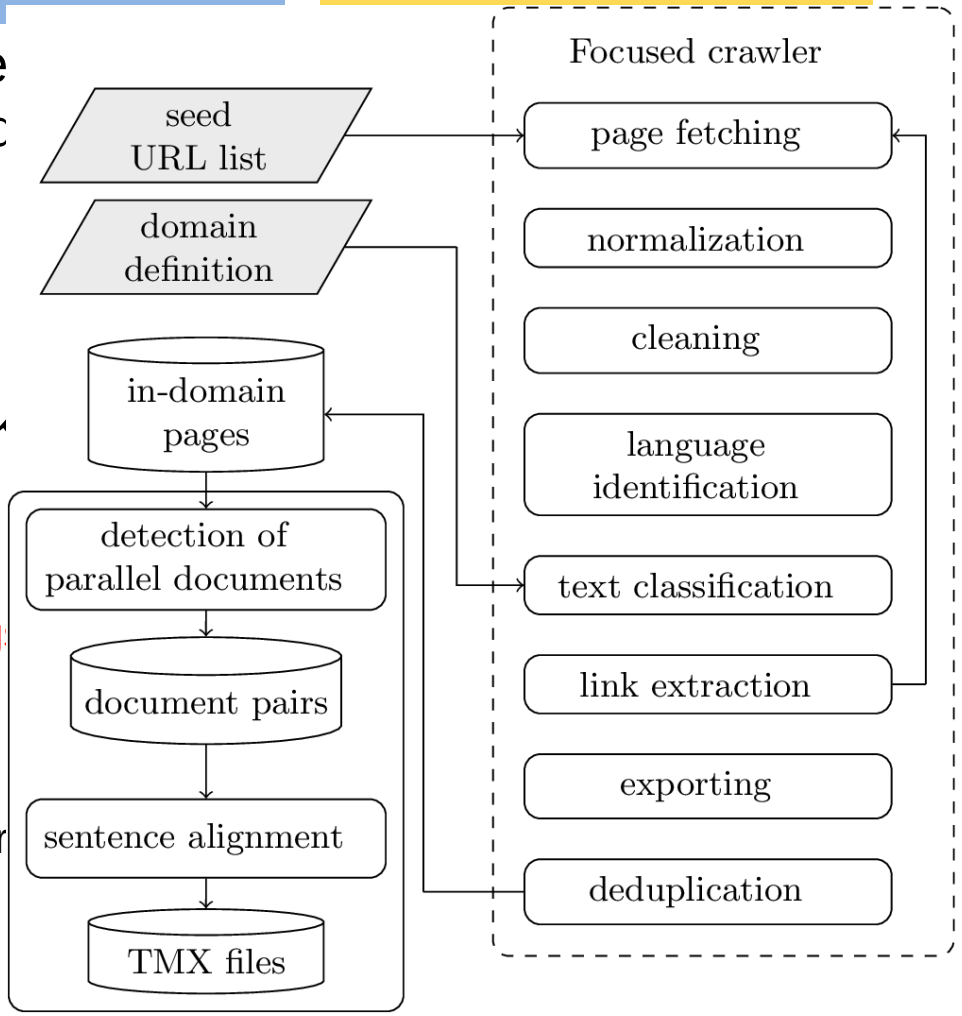


referenced links … Automated process

- http://portal.elda.org/ http://portal.elda.org/en/
  - http://portal.elda.org/news/rss/
  - http://portal.elda.org/login/
    - http://portal.elda.org/en/login/
      - http://portal.elda.org/reset/
  - http://portal.elda.org/about/elra/contact/
    - http://portal.elda.org/en/about/elra/contact/
      - http://portal.elda.org/tag/85/
        - http://portal.elda.org/en/tag/85/
      - http://portal.elda.org/tag/86/
        - http://portal.elda.org/en/tag/86/

ILSP Focused Crawler

- Research prototype for acquiring ge
  or domain-specific, monolingual anc
  bilingual corpora
- Input:
  - Domain definitions (lists of terms)
  - **Seed URLs**
- Modules (open source libraries/toolk
  - Page Fetching/Text Extraction
  - Normalization and Metadata Extraction
  - Boilerplate Detection (Boilerpipe)
  - Language Detection (covering > 50 lang
  - Text Classification
  - Exact and near de-duplication
  - Detection of pairs of parallel documents
  - Sentence alignment (Hunalign and other
- **Generates lists of**
  - document pairs and
  - segment pairs in TMX files



European Language Resource Coordination
Connecting Europe Facility

… it integrates technologies to crawl (part of a /multiple pages) website…

File Edit View History Bookmarks Tile Tools Help

Travel to Norway - Official ... | Visit Greece | Meetings and... | Visit Greece | Συνεδριακός ...

www.visitgreece.gr/el/meetings_and_incentives

iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map ΒΣΒ ΕΜΤΚ ΜΛ2008 TrEd User's Manual

Greece of art and science

Greece is a place of culture, the arts and sciences. 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**. 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. **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.

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. 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**.

The Greece of the pre-Socratic philosophers, of the great poets, of Pheidias the sculptor

Η Ελλάδα των τεχνών και της επιστήμης

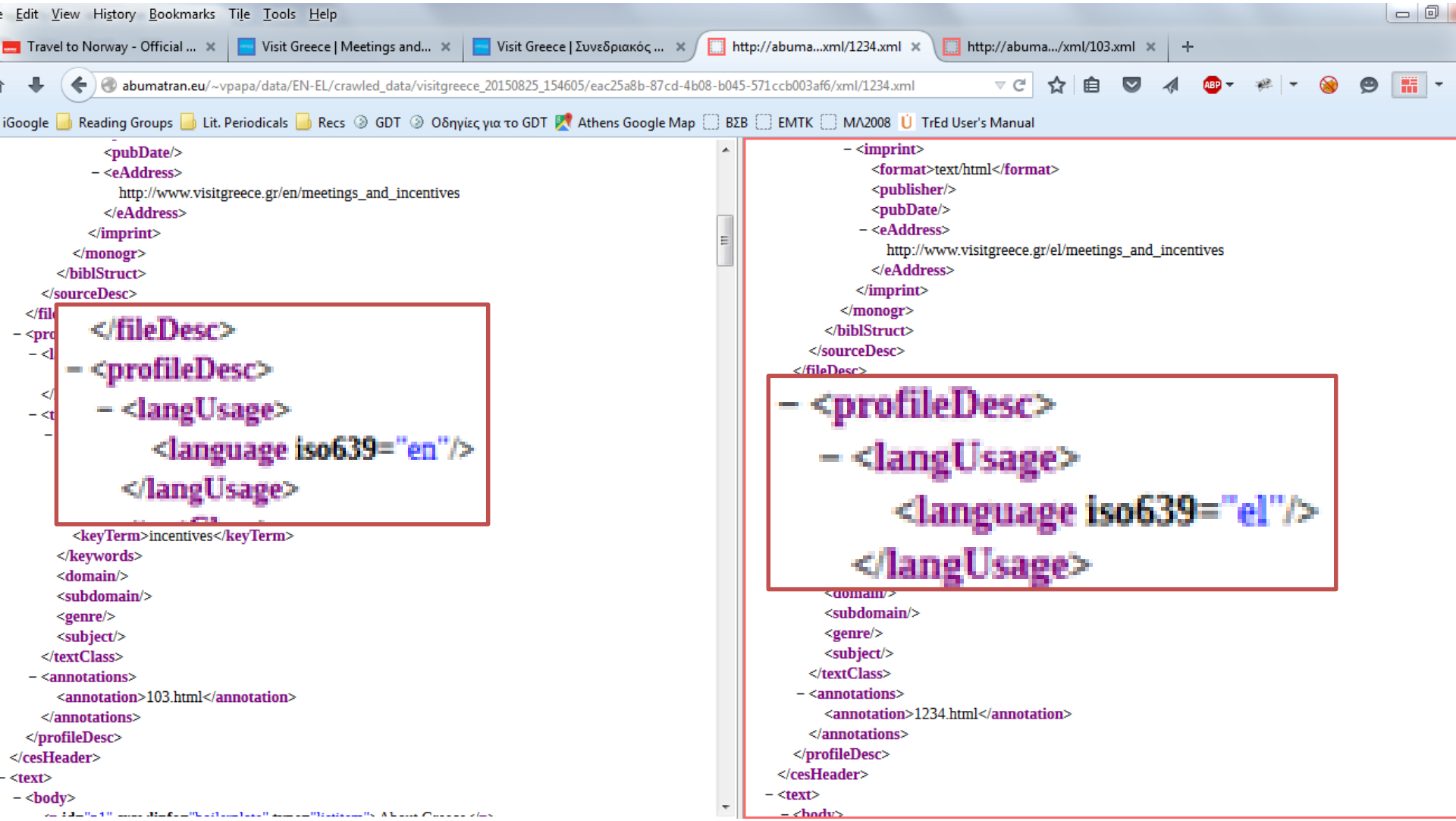
Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. Η μακραίωνη συμβολή της στο παγκόσμιο γίγνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις **άρτιες υποδομές**, την καθιστούν ιδανικό τόπο διεξαγωγής συνεδρίων. Τα τελευταία χρόνια, η ελληνική επικράτεια υποδέχεται όλο και συχνότερα ανθρώπους των γραμμάτων, των επιστημών και των τεχνών, οι οποίοι συμμετέχουν σε **συμπόσια**, **συνέδρια** και **εκθέσεις**. 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.

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. 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.

Η Ελλάδα των προσωκρατικών φιλοσόφων, των μεγάλων ποιητών, του Φειδία και του

European Language Resource Coordination
Connecting Europe Facility

… identify the language of each crawled page …

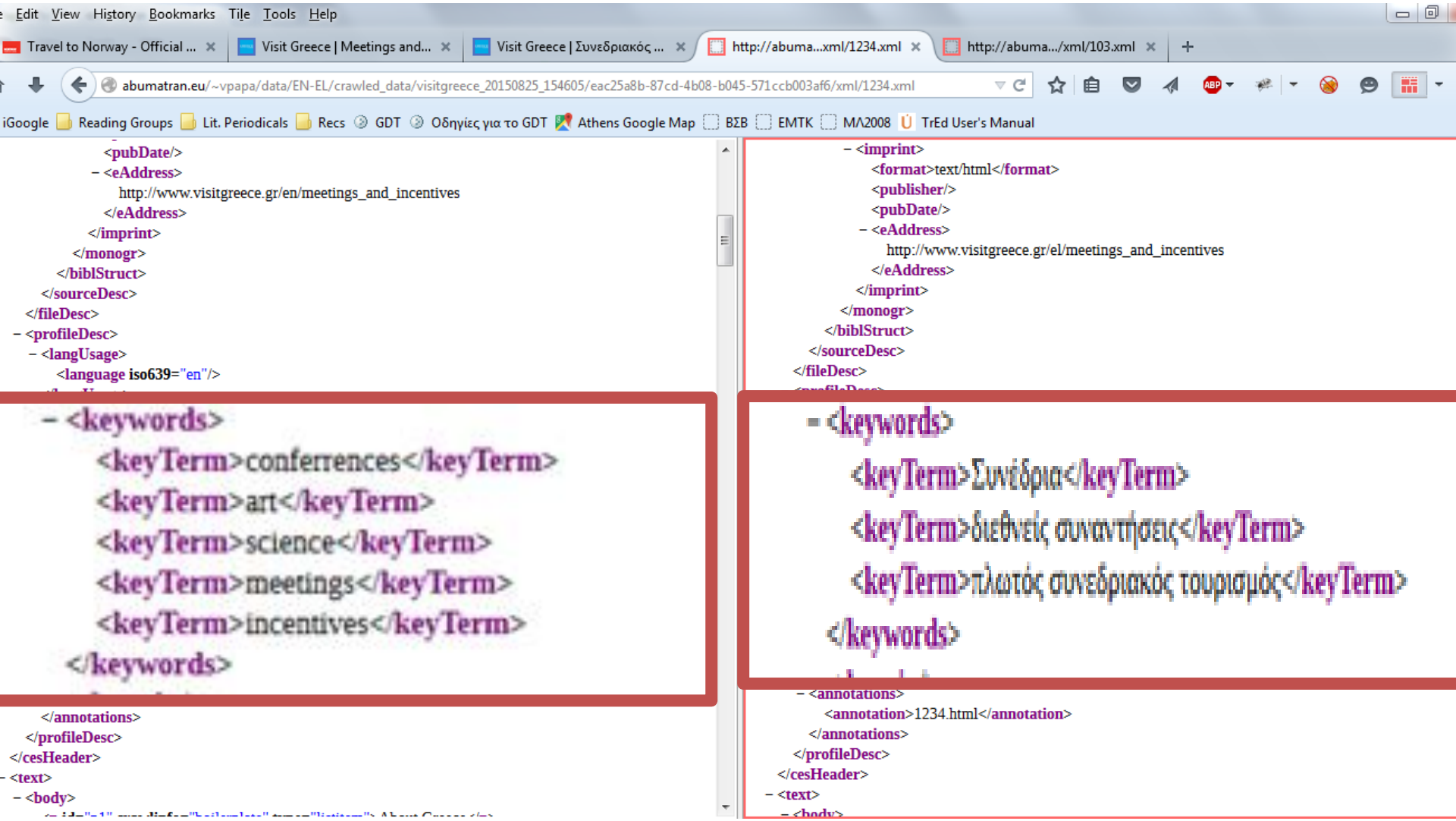


European Language Resource Coordination
Connecting Europe Facility

… identify the language of each crawled page

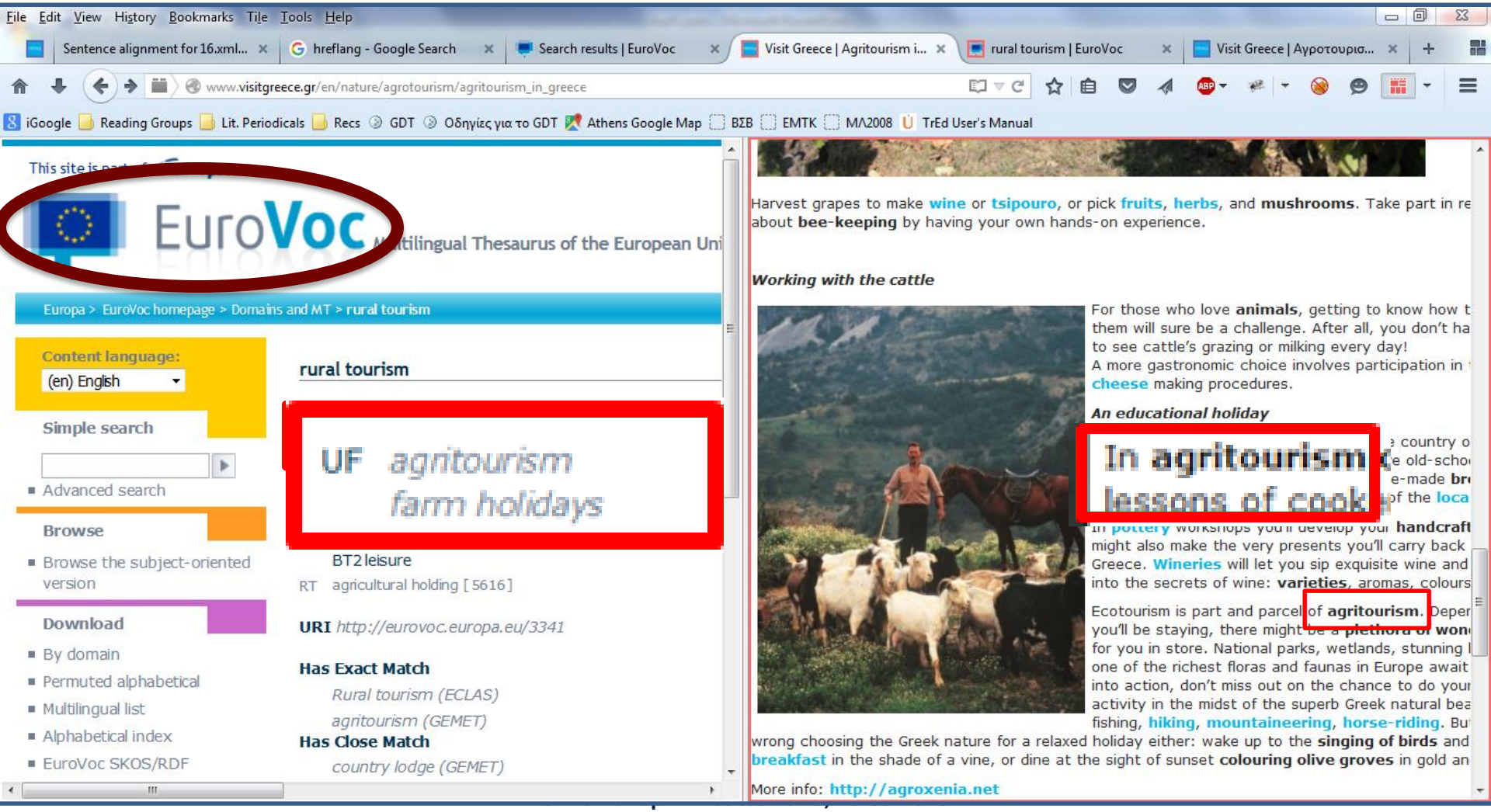
European Language Resource Coordination
Connecting Europe Facility

… extract several types of data descriptors (metadata)



European Language Resource Coordination
Connecting Europe Facility

… and optionally classify each page as relevant or not to a user-defined domain

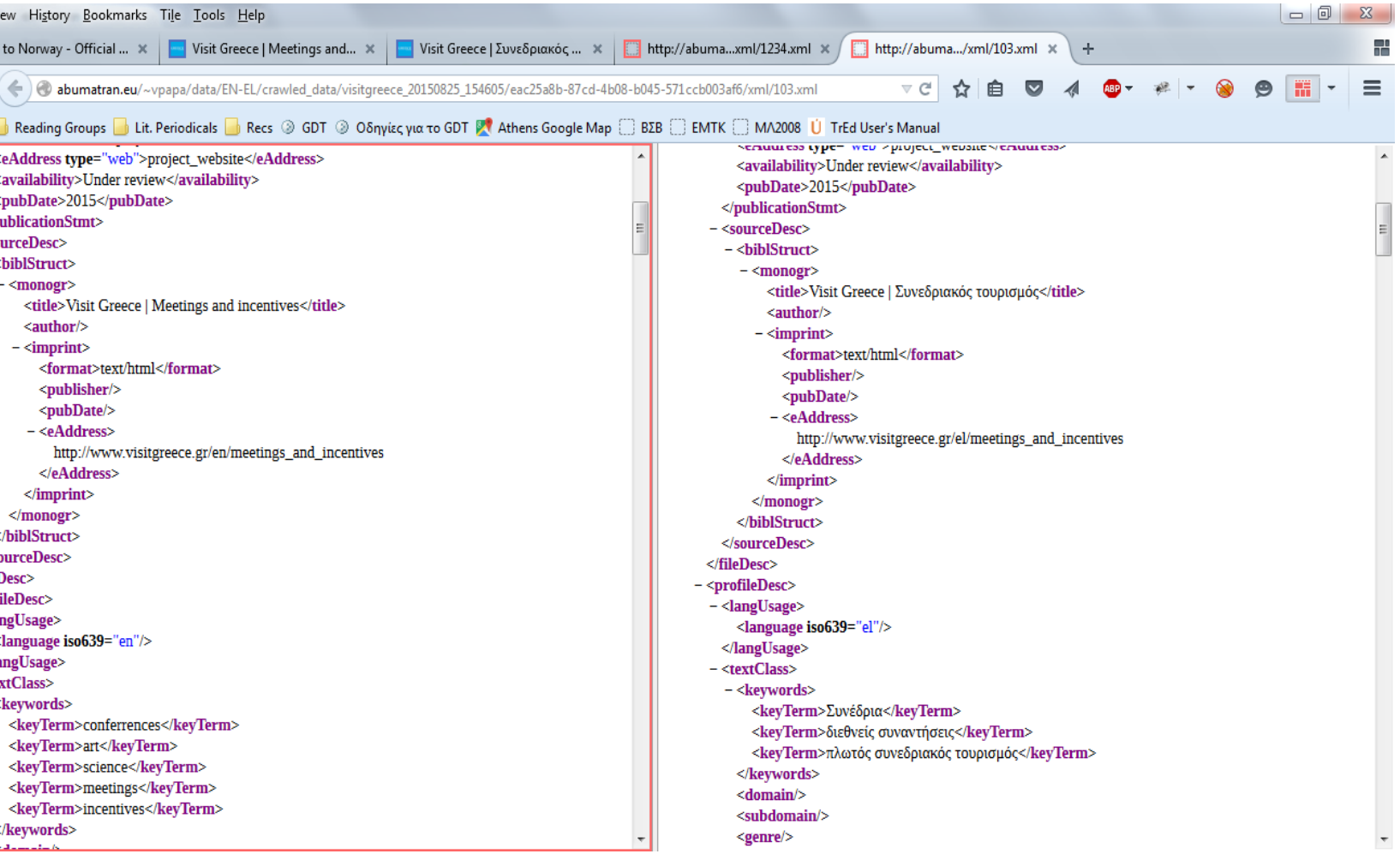


European Language Resource Coordination
Connecting Europe Facility

It can detect boilerplate text …

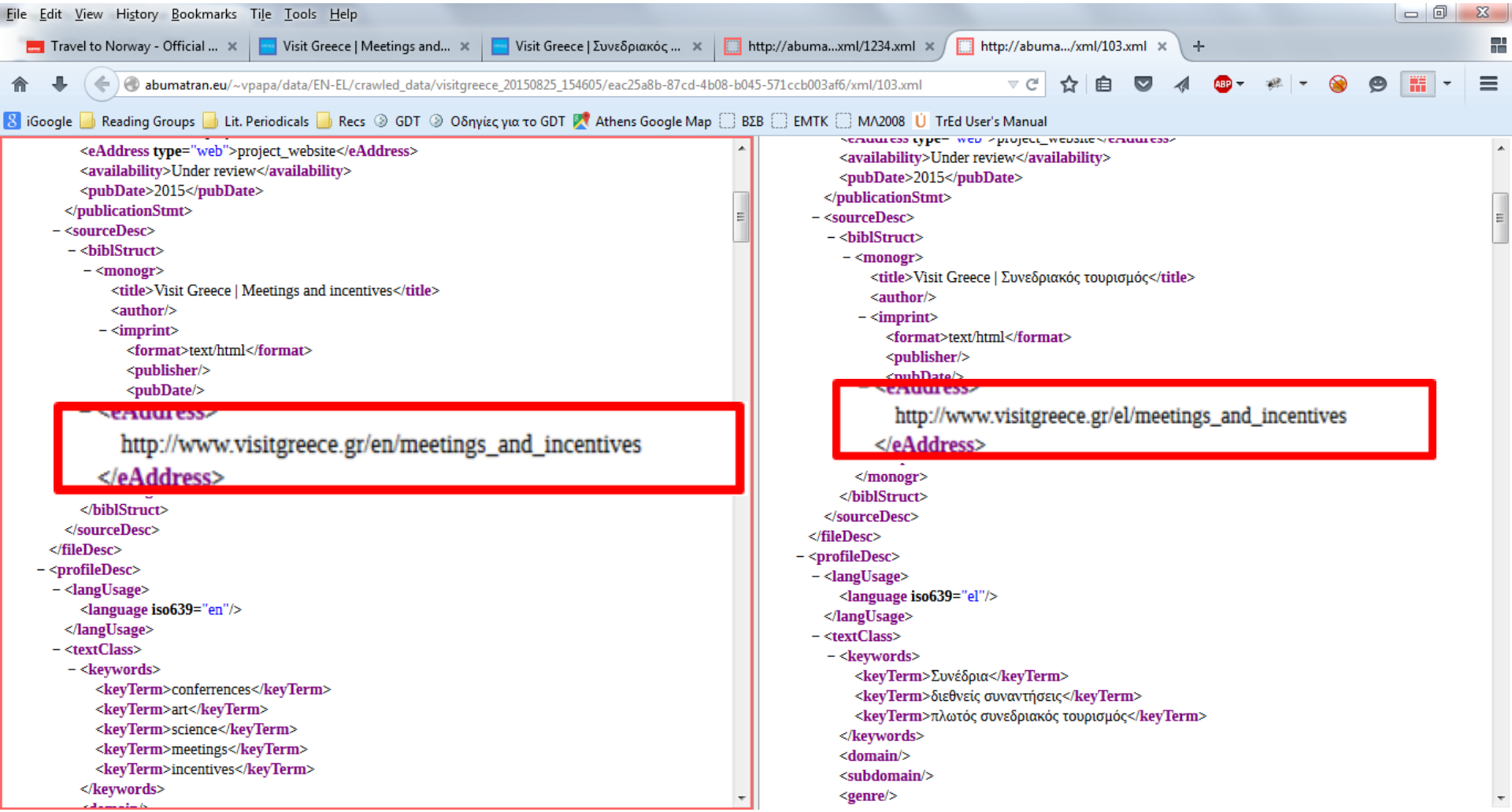
European Language Resource Coordination
Connecting Europe Facility

… HTML structure and/or URL similarity to detect document pairs



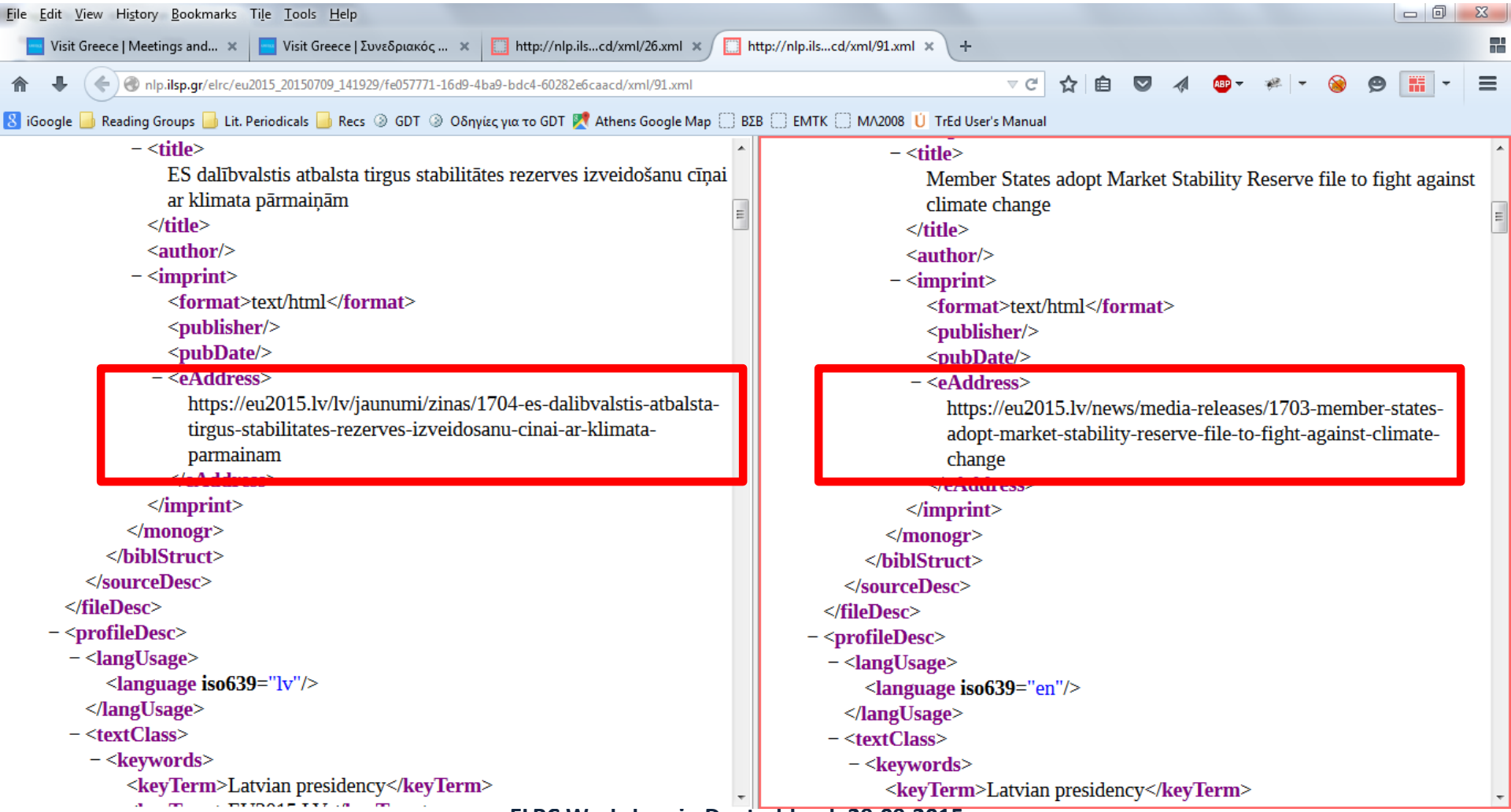
European Language Resource Coordination
Connecting Europe Facility

… HTML structure and/or URL similarity to detect document pairs



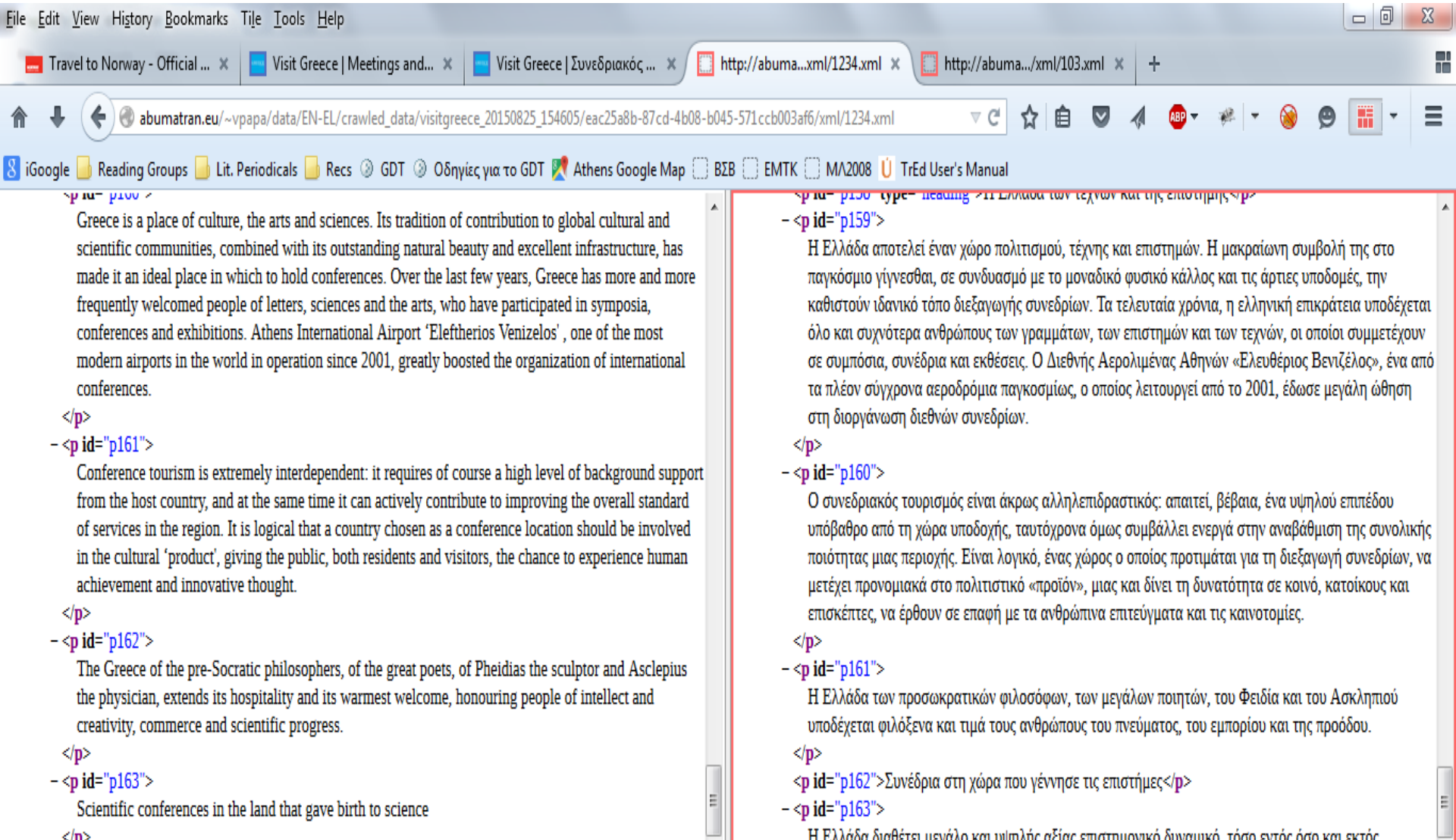
European Language Resource Coordination
Connecting Europe Facility

Sometimes URLs are not enough for finding document pairs…



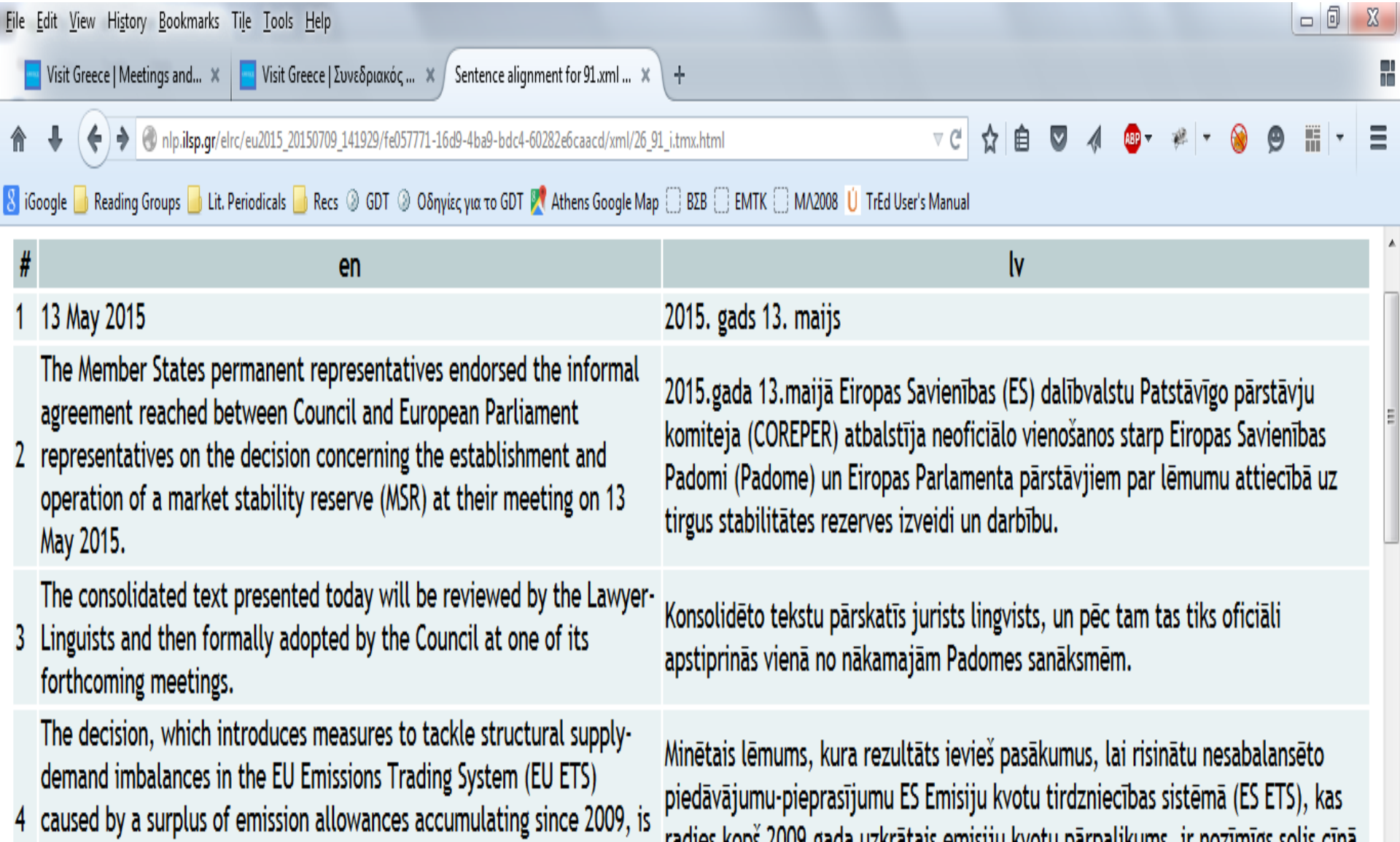
European Language Resource Coordination
Connecting Europe Facility

…. segment content in paragraphs …



European Language Resource Coordination
Connecting Europe Facility

… sentences and align it



#	en	lv
1	13 May 2015	2015. gads 13. maijs
2	The Member States permanent representatives endorsed the informal agreement reached between Council and European Parliament representatives on the decision concerning the establishment and operation of a market stability reserve (MSR) at their meeting on 13 May 2015.	2015.gada 13.maijā Eiropas Savienības (ES) dalībvalstu Patstāvīgo pārstāvju komiteja (COREPER) atbalstīja neoficiālo vienošanos starp Eiropas Savienības Padomi (Padome) un Eiropas Parlamenta pārstāvjiem par lēmumu attiecībā uz tirgus stabilitātes rezerves izveidi un darbību.
3	The consolidated text presented today will be reviewed by the Lawyer-Linguists and then formally adopted by the Council at one of its forthcoming meetings.	Konsolidēto tekstu pārskatīs jurists lingvists, un pēc tam tas tiks oficiāli apstiprinās vienā no nākamajām Padomes sanāksmēm.
4	The decision, which introduces measures to tackle structural supply-demand imbalances in the EU Emissions Trading System (EU ETS) caused by a surplus of emission allowances accumulating since 2009, is	Minētais lēmums, kura rezultāts ieviešpasākumus, lai risinātu nesabalansēto piedāvājumu-pieprasījumu ES Emisiju kvotu tirdzniecības sistēmā (ES ETS), kas radies kopš 2009.gada uzkrātais emisiju kvotu pārpalikums, ir nozīmīgs solis cīņā

European Language Resource Coordination
Connecting Europe Facility

Our tools supports all EU languages!

EN-GA

File Edit View History Bookmarks Tile Tools Help

Sentence alignment for 16.xml ... | Sentence alignment for 1211.x...

nlp.ilsp.gr/elrc/citizensinformation_en-ga_20150715_144337/65140275-88ad-4428-b005-5dfcbc01aada/xml/2799_1211_u.tmx.html

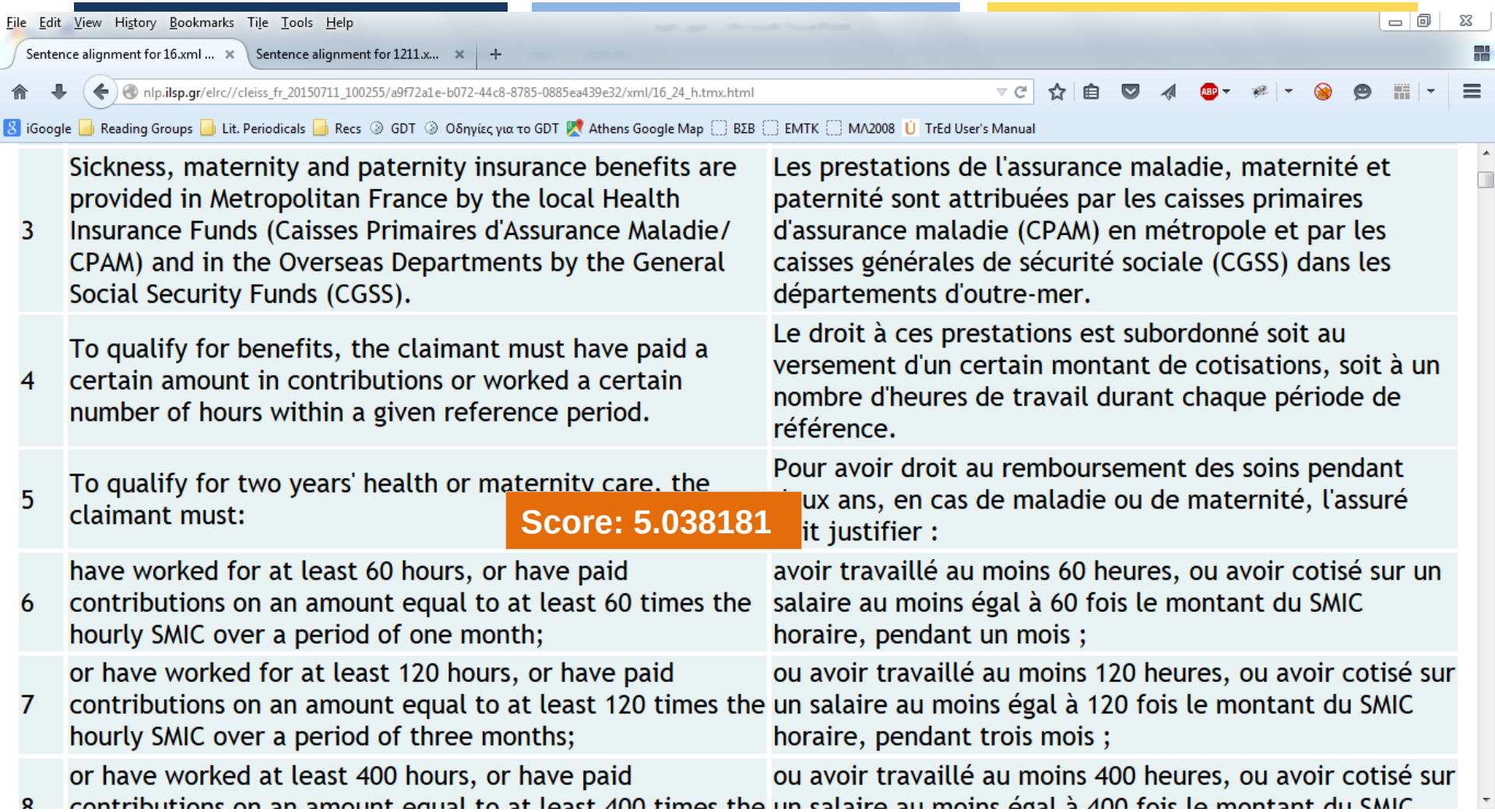
iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map ΒΣΒ ΕΜΤΚ ΜΛ2008 TrEd User's Manual

7	information	Eolas
8	As a European Union (EU) citizen, you have the right to live and work in any other EU country.	Mar shaoránach den Aontais Eorpaigh (AE), tá sé de cheart agat maireachtáil agus oibriú in aon cheann de thíortha an AE.
9	If you are an EU national or a dependant of such a national and you meet the requirements of the EU Directives on free movement of workers, you may not, in general, be refused permission to land in another EU country.	Más saoránach den AE thú nó cleithiúnach saoránaigh agus má shásaíonn tú riachtanais Treoracha an AE maidir le saorthaisteal oibrithe, ní féidir, go ginearálta, diúltú duit dul i dtír in aon tír eile den AE.
10	You may require a valid identity card or passport.	Beidh cárta aitheantais bhailí nó pas ag teastáil uait.
11	You may be refused entry and/or your right of residence in another member state may be restricted on grounds of public policy, public security, or public health.	Féadfar a dhiúltú duit dul isteach i mballstát eile nó féadfar do cheart cónaithe i mballstát eile a dhiúltú duit ar fhorais beartais phoiblí, slándáil an phobail nó sláinte an phobail.
12	The rights outlined in this section broadly apply to the non-EU States in the European Economic Area, that is, Norway, Iceland and Liechtenstein, as well as to Switzerland.	Tá na cearta a dtugtar sracléiriú orthu sa roinn seo infheidhme go ginearálta maidir leis na Ballstáit neamh-AE de chuid an Limistéir Eorpaigh Eacnamaíoch, is é sin, an Iorua, an Íoslann agus Lichtinstéin, chomh maith leis an Eilvéis.
13	EU Directives on free movement of workers	Treoracha de chuid an AE maidir le saorghluaiseacht

European Language Resource Coordination
Connecting Europe Facility

it supports all EU languages!

EN-FR



3	Sickness, maternity and paternity insurance benefits are provided in Metropolitan France by the local Health Insurance Funds (Caisses Primaires d'Assurance Maladie/ CPAM) and in the Overseas Departments by the General Social Security Funds (CGSS).	Les prestations de l'assurance maladie, maternité et paternité sont attribuées par les caisses primaires d'assurance maladie (CPAM) en métropole et par les caisses générales de sécurité sociale (CGSS) dans les départements d'outre-mer.
4	To qualify for benefits, the claimant must have paid a certain amount in contributions or worked a certain number of hours within a given reference period.	Le droit à ces prestations est subordonné soit au versement d'un certain montant de cotisations, soit à un nombre d'heures de travail durant chaque période de référence.
5	To qualify for two years' health or maternity care, the claimant must:	Pour avoir droit au remboursement des soins pendant ux ans, en cas de maladie ou de maternité, l'assuré it justifier :
6	have worked for at least 60 hours, or have paid contributions on an amount equal to at least 60 times the hourly SMIC over a period of one month;	avoir travaillé au moins 60 heures, ou avoir cotisé sur un salaire au moins égal à 60 fois le montant du SMIC horaire, pendant un mois ;
7	or have worked for at least 120 hours, or have paid contributions on an amount equal to at least 120 times the hourly SMIC over a period of three months;	ou avoir travaillé au moins 120 heures, ou avoir cotisé sur un salaire au moins égal à 120 fois le montant du SMIC horaire, pendant trois mois ;
8	or have worked at least 400 hours, or have paid contributions on an amount equal to at least 400 times the	ou avoir travaillé au moins 400 heures, ou avoir cotisé sur un salaire au moins égal à 400 fois le montant du SMIC

Score: 5.038181

European Language Resource Coordination
Connecting Europe Facility

It supports all EU languages!

EN-LV

File Edit View History Bookmarks Tile Tools Help

Sentence alignment for 16.xml ... | Sentence alignment for 1211.x... | Sentence alignment for 873.xm...

nlp.ilsp.gr/elrc/fold_lv_20150709_144037/61e8d498-c38c-49ba-885a-dcf47367d993/xml/873_373_i.tmx.html

iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map ΒΣΒ ΕΜΤΚ ΜΛ2008 TrEd User's Manual

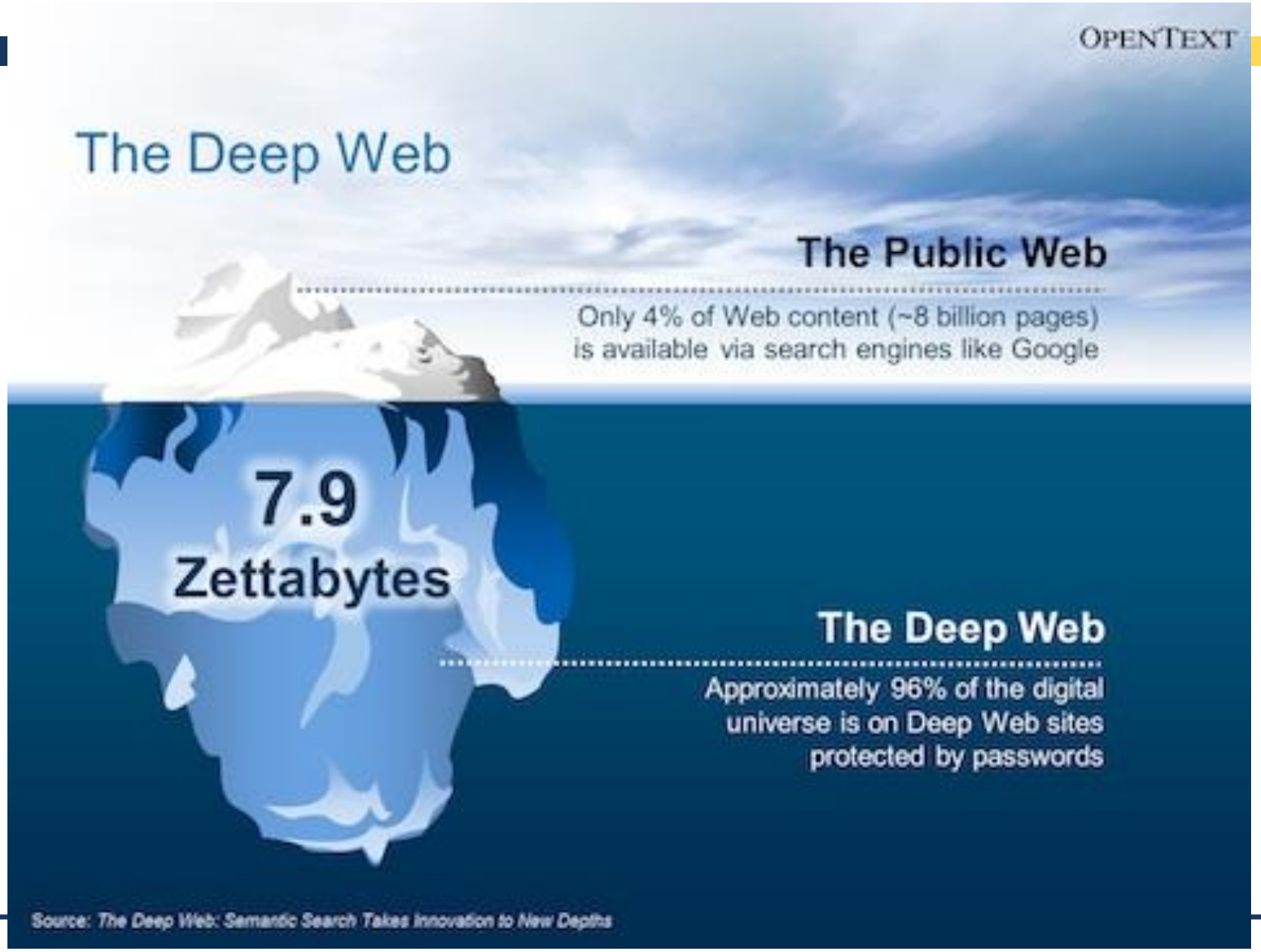
7	Edgars:	Aizeks:
8	- How did you come up with the name for the company and what does it mean to you? Isaac:	- Mēs vienmēr esam ticējuši sadarbībai darba procesā, kurā daudzi cilvēki kopīgi veido vienu ideju.
9	- We've always believed in the collaborative approach to working, that many people are contributing to one idea. Also, our ideas, the things we build are never really finished until they are out in the real world and being used by many people.	Turklāt mūsu idejas, lietas, ko būvējam, nekad nav pilnīgi pabeigtas, kamēr tās nav izgājušas reālajā pasaulē un pirms tās sāk lietot daudzi cilvēki.
10	So that is the double meaning of «many».	Šī tad arī ir vārda «many» dubultā nozīme.
11	It did take us a long time to come up with a good name and lots of bad names got rejected.	Pagāja laiks, kamēr mēs izdomājām labu nosaukumu, mēs noraidījām daudzus neveiksmīgus vārdus.
12	We actually did an exercise «what we shouldn't call ourselves» [laughs], and the top rejected name was «fluffy».	Izmēģinājām arī vingrinājumu «kā mums nevajadzētu sevi saukt», un saraksta augšgalā bija vārds «pufīgi» [smejas].
13	We wanted to be human.	Mēs gribējām būt cilvēcīgi.
14	Oskars:	Oskars:
15	Isaac:	Aizeks:
16	- There were four of us, we founded it back in 2007	-Mēs bijām četri, un uzņēmumu mēs nodibinājām

A better alternative

- This process can be turned into a factory of LR production (Automation of the Procedure)
  - Identify sources of data
  - Browse through the page links
- BUT *what we can get is the "visible" part, there are many more in your organizations*

European Language Resource Coordination
Connecting Europe Facility

Visible data versus existing ones



European Language Resource Coordination
Connecting Europe Facility

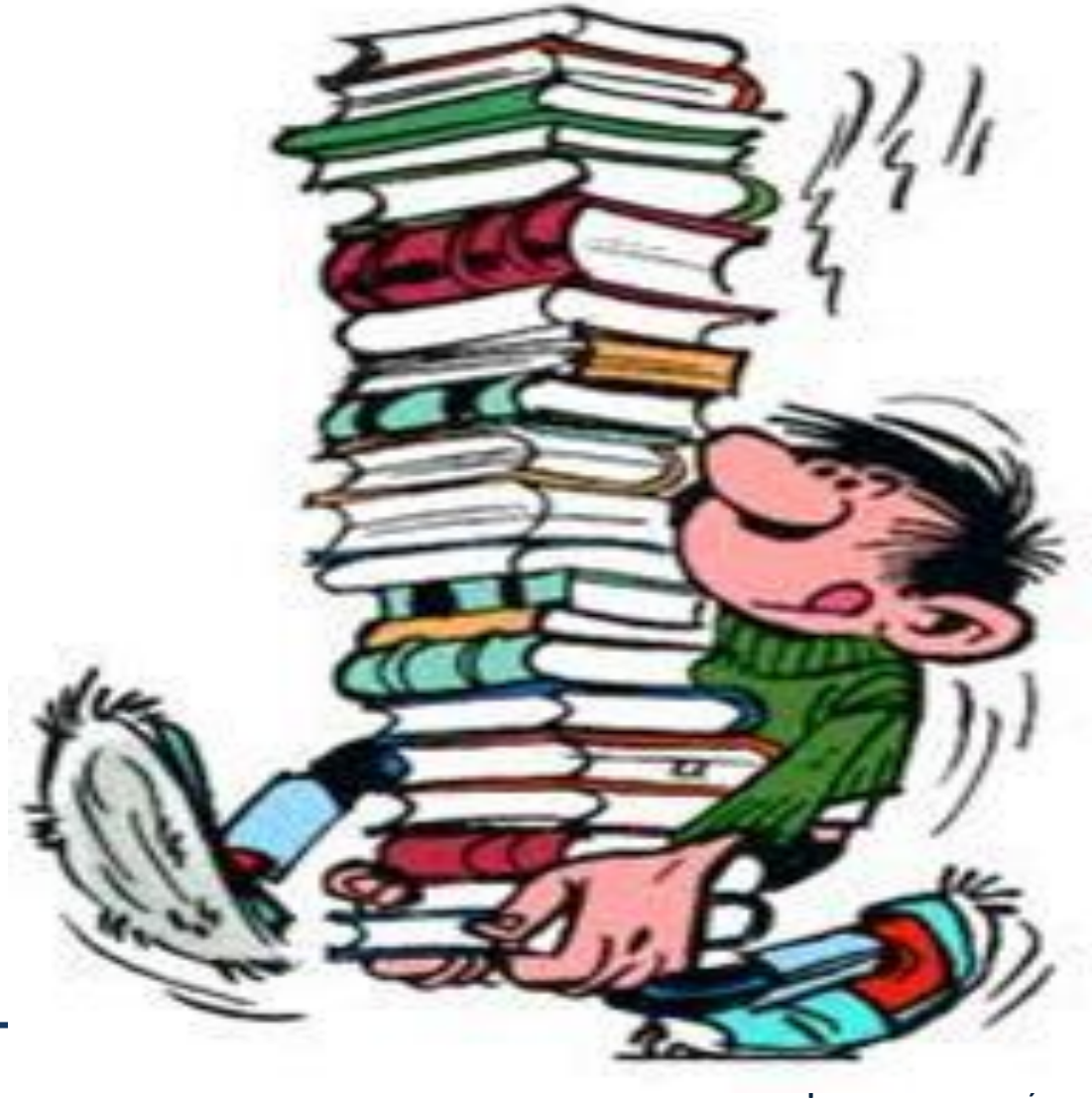
Our contributions … Deep web

European Language Resource Coordination
Connecting Europe Facility

A better alternative

- Such documents exist already:
  - At the various documentation centers (translated reports, leaflets, brochures, speeches, web pages, etc.)
  - At the Language Service Providers (LSP), to whom translation works are subcontracted
- Help us identify and liaise with both sources
  - (see next Panel interactions)

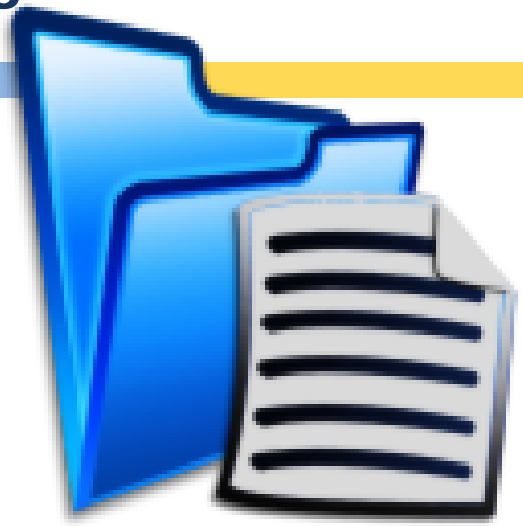
Your involvement is essential so please let us work together

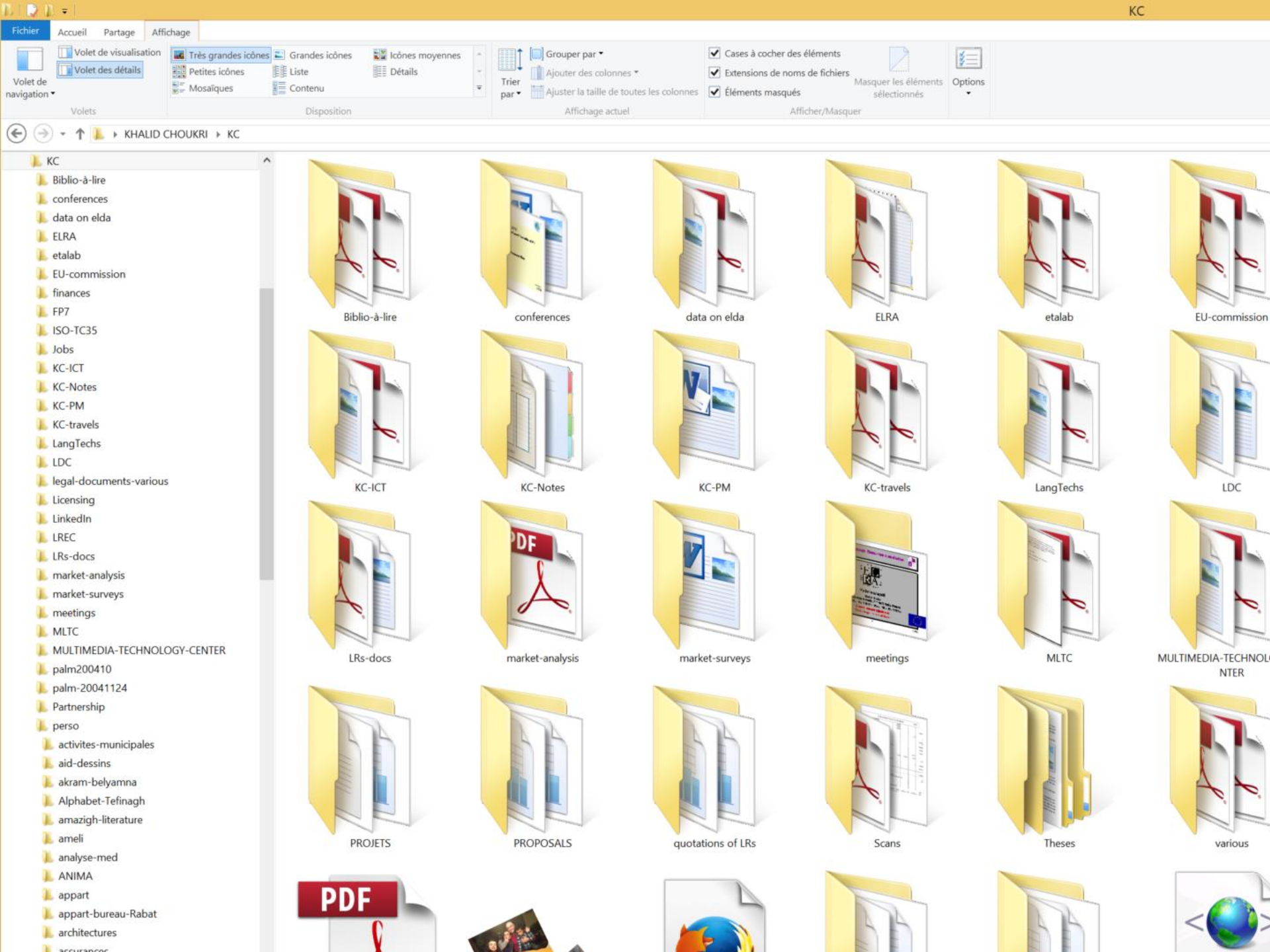


BRING YOUR OWN LANGUAGE RESOURCES

European Language Resource Coordination
Connecting Europe Facility

Your involvement is essential so please let us work together




KC
Fichier
Accueil
Partage
Affichage
Volet de navigation
Volet de visualisation
Volet des détails
Très grandes icônes
Grandes icônes
Icônes moyennes
Petites icônes
Liste
Détails
Mosaïques
Contenu
Volets
Disposition
Trier par
Grouper par
Ajouter des colonnes
Ajuster la taille de toutes les colonnes
Affichage actuel
Cases à cocher des éléments
Extensions de noms de fichiers
Éléments masqués
Masquer les éléments sélectionnés
Afficher/Masquer
Options
KHALID CHOUKRI
KC
KC
Biblio-à-lire
conferences
data on elda
ELRA
etalab
EU-commission
finances
FP7
ISO-TC35
Jobs
KC-ICT
KC-Notes
KC-PM
KC-travels
LangTechs
LDC
legal-documents-various
Licensing
LinkedIn
LREC
LRs-docs
market-analysis
market-surveys
meetings
MLTC
MULTIMEDIA-TECHNOLOGY-CENTER
palm200410
palm-20041124
Partnership
perso
activites-municipales
aid-dessins
akram-belyamna
Alphabet-Tefinagh
amazigh-literature
ameli
analyse-med
ANIMA
appart
appart-bureau-Rabat
architectures
PROJETS
PROPOSALS
quotations of LRs
Scans
Theses
various
PDF

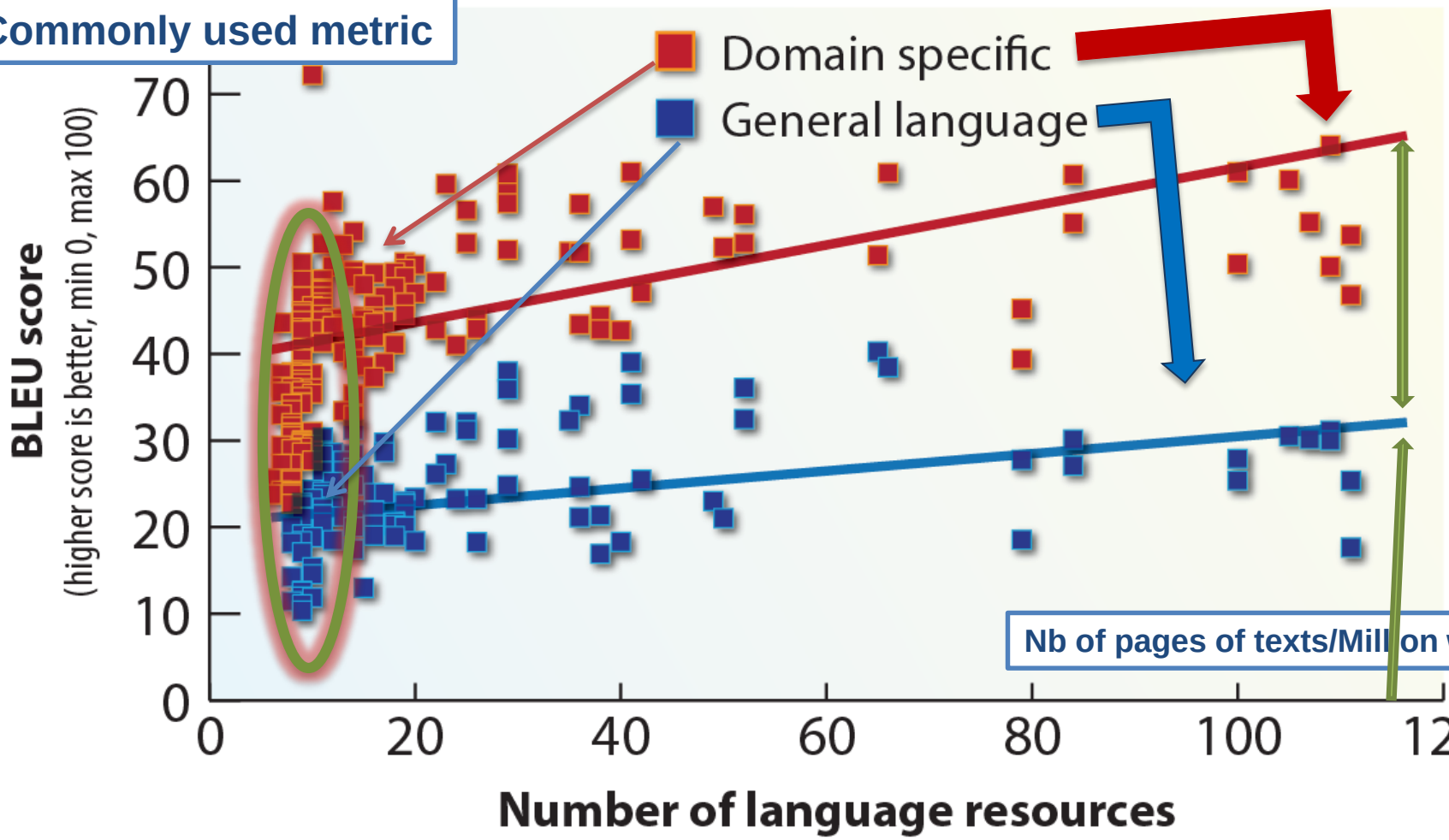
- How to help upload the data
- See the information on the REPOSITORY set-up for this

- **How much data is needed ?**

European Language Resource Coordination
Connecting Europe Facility

Impact of number of language resources on Statistical MT quality

A Commonly used metric



CONCLUSIONS

- How data is produced: **repurposing and repackaging existing data**
- Why is important: the data driven paradigm is very efficient
  - results improve as more data become available
- *Let us not under estimate the value of our resources*
- *How can you contribute and benefit from CEF.AT*
  - *(next sessions)*