

“What Data Is Needed? Why?”

Dr. Khalid Choukri
(Evaluations and Language Resource Association)

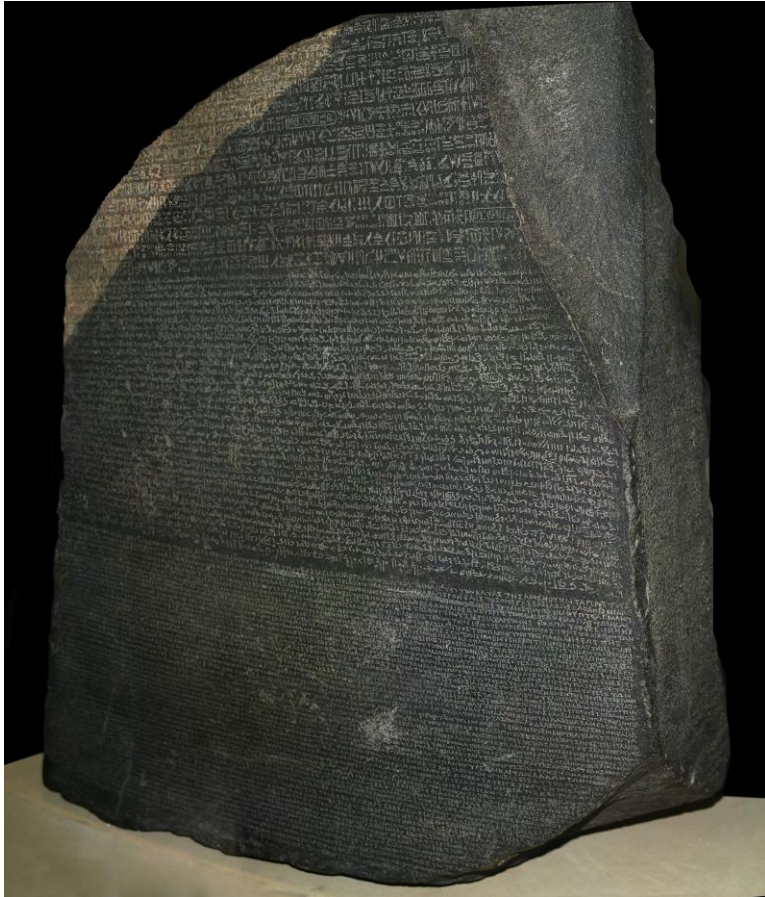


- From previous session, we have seen the predominant approach of data-driven paradigm
 - We “**learn**” from existing data
- How are Language Resources produced:
 - from documents and data to valuable Language Resources for MT
 - Why it is important that you help us with the data you have / you know about
- *The focus is on data in all languages (EU/CEF).*

What Data



wiseGEEK



*Tableau des Signes Phonétiques
des écritures hiéroglyphique et Démotique des anciens Égyptiens*

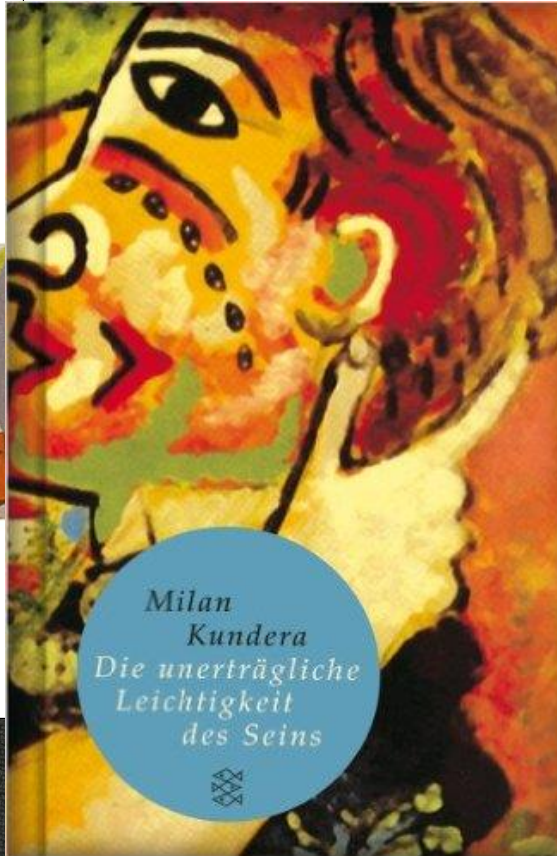
Lettre Grecque	Signes Démotiques	Signes hiéroglyphiques
A	u. w.	
B	4. x.	
Γ	κ. γ.	
Δ	κ. ζ.	
E	ι.	
Z		
H	III. IIII. IIII.	
Θ		
I	∩. III.	
K	κ. κ. κ. κ. κ.	
Λ	λ. λ. λ.	
M	μ. μ.	
N	ν. ν. ν. ν. ν.	
Ξ	ξ.	
O	ο. ο. ο. ο. ο.	
Π	π. π. π. π. π.	
P	ρ. ρ. ρ.	
Σ	σ. σ. σ. σ. σ. σ. σ. σ. σ.	
T	τ. τ. τ. τ.	
Υ		
Φ	φ.	
Χ	χ.	
Ψ		
Ω		
TO.		



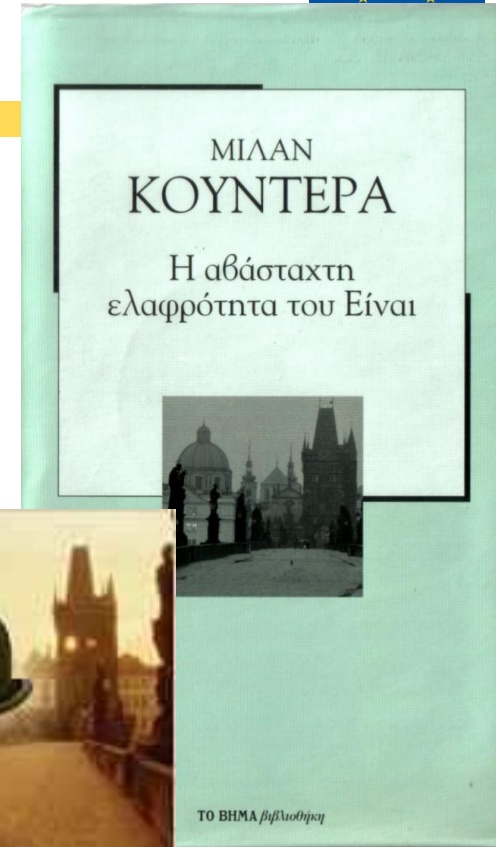
Kundera
L'insoutenable
légèreté de l'être



folio

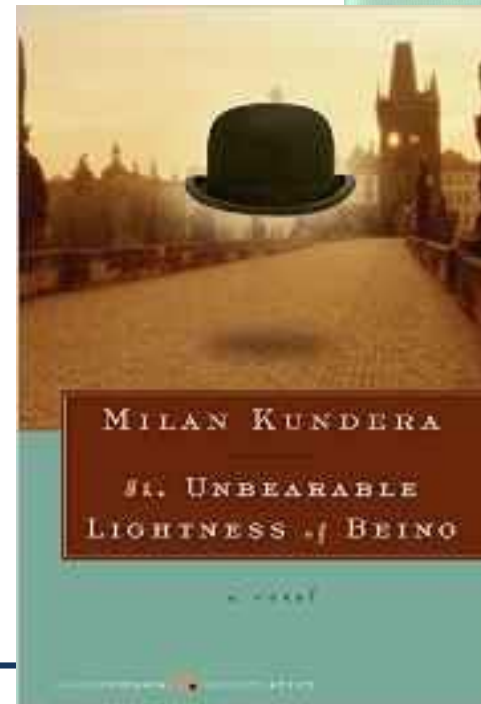


Milan
Kundera
Die unerträgliche
Leichtigkeit
des Seins



ΜΙΛΑΝ
ΚΟΥΝΤΕΡΑ
Η αβάσταχτη
ελαφρότητα του Είναι

ΤΟ ΒΗΜΑ βιβλιοθήκη



MILAN KUNDERA
II. UNBEARABLE
LIGHTNESS OF BEING

What types of data? Translation



wiseGEEK

What types of data? “Aligned” Translation



English



French

What types of data? “Aligned” Translation



GENESIS

The Story of Creation

1 In the beginning, when God created the universe, ²the 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. ³Then God commanded, “Let there be light” – and light appeared. ⁴God was pleased with what he saw. Then he separated the light from the darkness, ⁵and he named the light “Day” and the darkness “Night”. Evening passed and morning came – that was the first day.

⁶⁻⁷Then 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. ⁸He named the dome “Sky”. Evening passed and morning came – that was the second day.

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

¹⁴Then 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; ¹⁵they will shine in the sky to give light to the earth” – and it was done. ¹⁶So 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. ¹⁷He placed the lights in the sky to shine on the earth, ¹⁸to rule over the day and the night, and to separate light from darkness. And God was pleased with what he saw. ¹⁹Evening 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.

²La 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. ³Alors Dieu dit: “Que la lumière paraisse!” et la lumière parut. ⁴Dieu constata que la lumière était une bonne chose, et il sépara la lumière de l'obscurité. ⁵Dieu nomma la lumière jour et l'obscurité nuit. Le soir vint, puis le matin; ce fut la première journée.

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

⁹Dieu dit encore: “Que les eaux qui sont au-dessus du ciel se rassemblent en un lieu unique pour que le continent paraisse!” Et cela se réalisa. ¹⁰Dieu nomma le continent terre et la masse des eaux mer, et il constata que c'était une bonne chose. ¹¹Dieu 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. ¹²La 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. ¹³Le soir vint, puis le matin; ce fut la troisième journée.

¹⁴Dieu 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; ¹⁵et que du haut du ciel elles éclairent la terre!” Et cela se réalisa. ¹⁶Dieu 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. ¹⁷Il les plaça dans le ciel pour éclairer la terre, ¹⁸pour présider au jour et à la nuit, et pour séparer la lumière de l'obscurité. Dieu constata que c'était une bonne chose. ¹⁹Le soir vint, puis le matin; ce fut la quatrième journée.

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!!

highly ...
previous level in time or space.

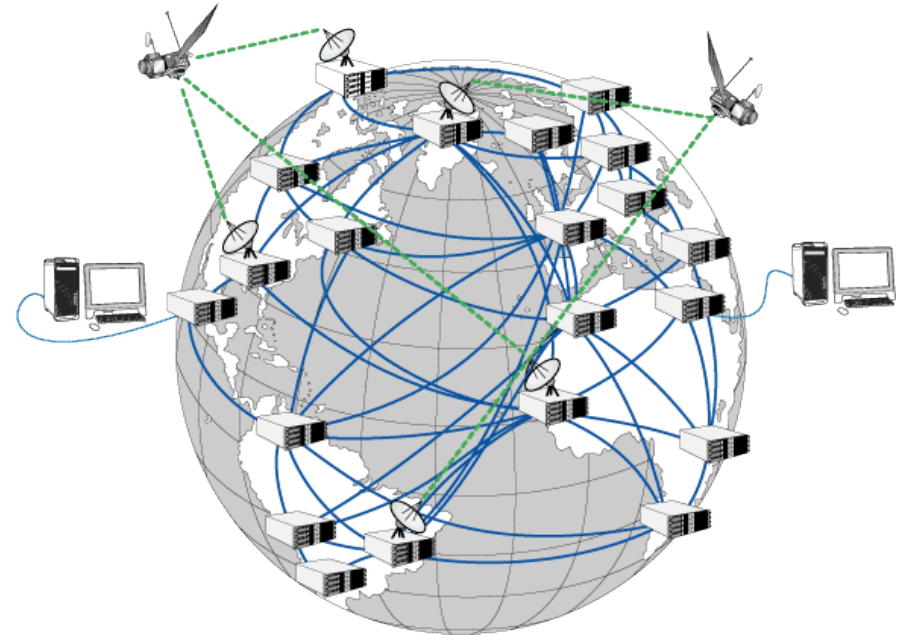
ID	FR	ES	EL
6905	abandon scolaire	abandono escolar	διακοπή της σχολικής φοίτησης
920	abats	despojo	παραπροϊόντα σφαγίων
1857	abattage d'animaux	sacrificio de animales	σφαγή ζώων
6621	abrogation	derogación	κατάργηση
5075	Abruzzes	Abruzos	Αβρουζία
5339	absentéisme	absentismo	συστηματική απουσία από την εργασία
5984	abstentionnisme	abstencionismo	αποχή
2	abus de confiance	abuso de confianza	απιστία
25	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



- Archives

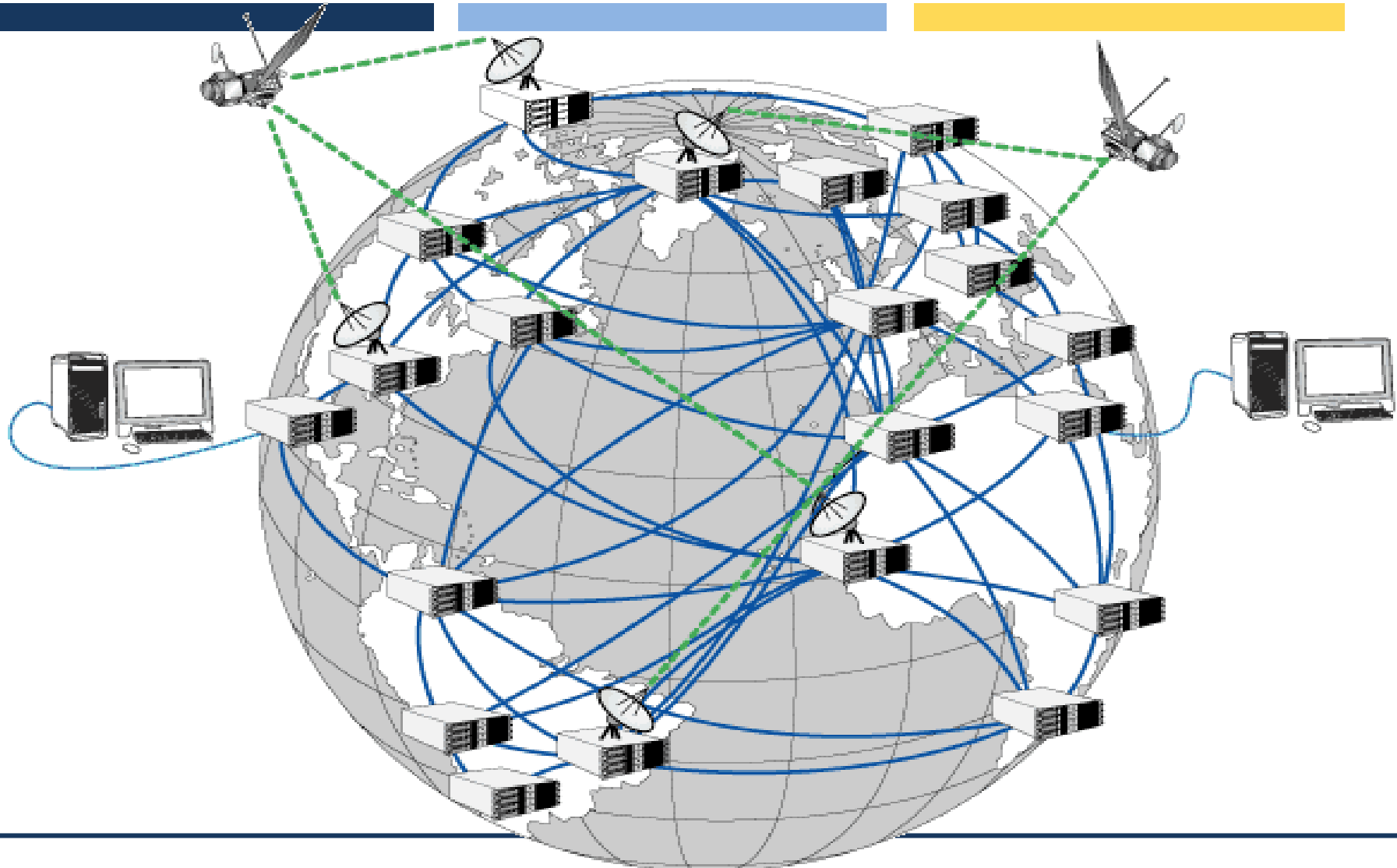


- Internet



Digital word ... Internet





Of course need for digital textual data !!



Various Formats





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



- 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

- 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

Many web sites...



GREECE
ALL TIME CLASSIC

Ελλάδα Προορισμοί Αξιοθέατα-Δραστηριότητες Διάθεση για Newsletter



visitgreece.gr highlights




**Χρήσιμες πληροφορίες
για τις διακοπές σας**

Search   

connect and experience





-  Download banners
-  Download wallpapers
-  Download guides
-  Download brochures
-  Download maps
-  Sign up to our newsletter
-  eBook

Εξερευνήστε την Ελλάδα





- | | |
|---------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
|  Πολιτισμός |  Ελεύθερος χρόνος |
|  Περιηγήσεις |  Γαστρονομία |
|  Θρησκεία |  Meetings |
|  City Break |  Δραστηριότητες |

... are rich in multilingual content

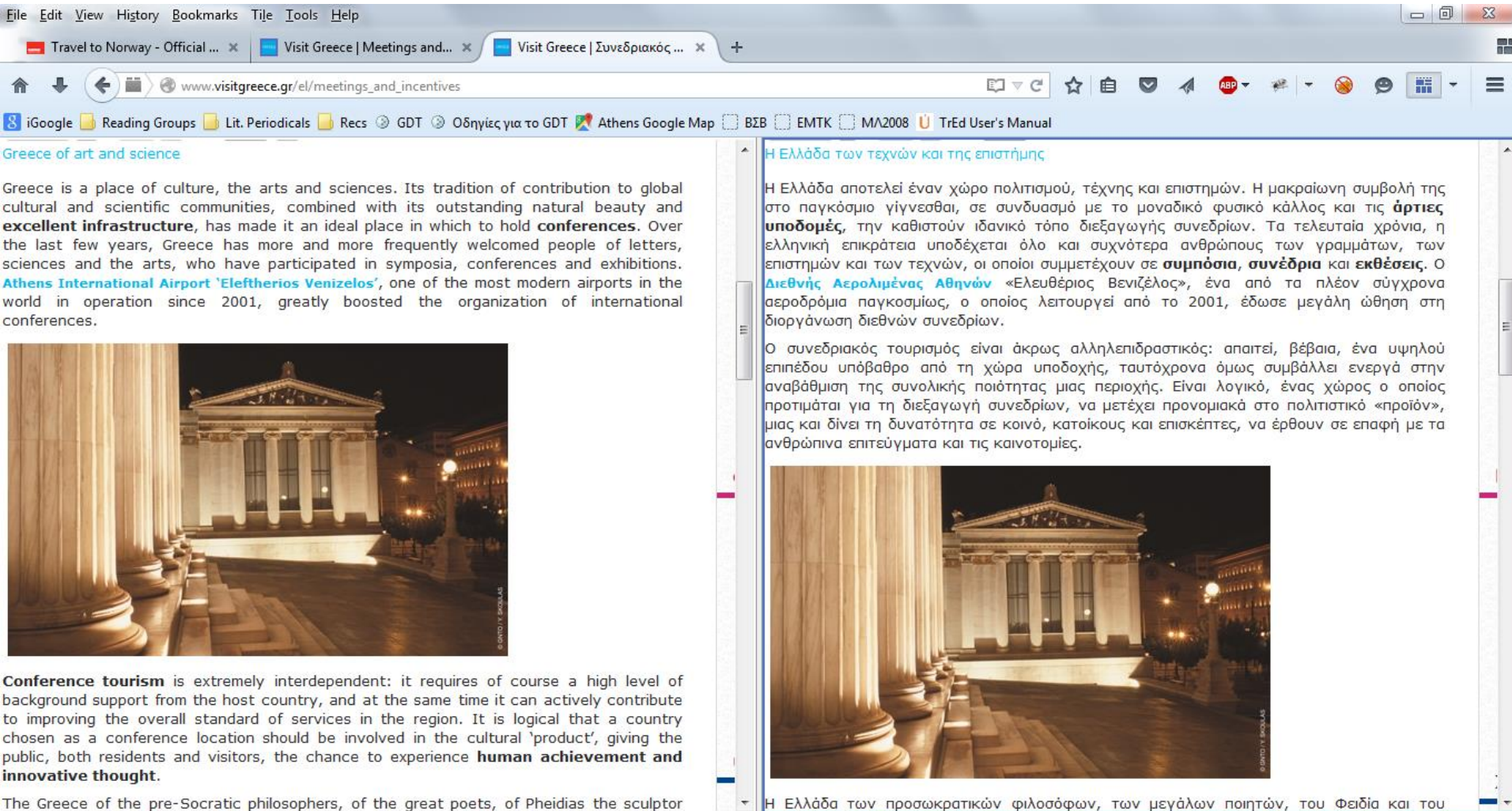


Search   

connect and experience

How can we obtain this content...

File Edit View History Bookmarks Title Tools Help
 Travel to Norway - Official ... x Visit Greece | Meetings and... x Visit Greece | Συνεδριακός ... x +
 www.visitgreece.gr/el/meetings_and_incentives
 iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map ΒΣΒ EMTK ΜΑ2008 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.

Η Ελλάδα των τεχνών και της επιστήμης
 Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. Η μακραίωνη συμβολή της στο παγκόσμιο γίνεσθαι, 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.

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 προσωκρατικών φιλοσόφων, των μεγάλων ποιητών, του Φειδία και του

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, μακροχρόνια συμβολή της στο παγκόσμιο γίγνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις άριστες 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 υψηλό επίπεδο 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.

File Edit View History Bookmarks Tile Tools Help

Travel to Norway - Official ... x Visit Greece | Meetings and... x Visit Greece | Συνεδριακός ... x Sentence alignment for 103.xml... x +

abumatran.eu/~vpapa/data/EN-EL/crawled_data/visitgreece_20150825_154605/eac25a8b-87cd-4b08-b045-571ccb003af6/xml/1234_103_u.tmx.html

iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map BZB EMTK MA2008 TrEd User's Manual

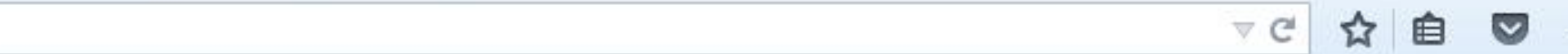
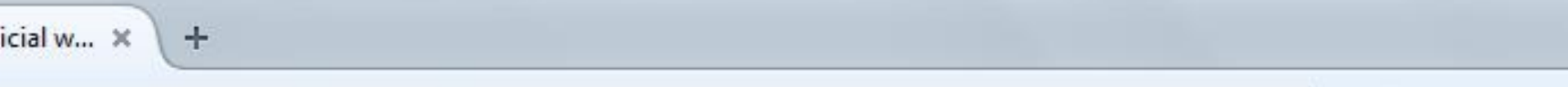
Sentence alignment for 103.xml (en) - 1234.xml (el)

#	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 γίνεσθαι, 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.
	Athens International Airport (Eleftherios Venizelos), one of the most modern airports in	



- 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

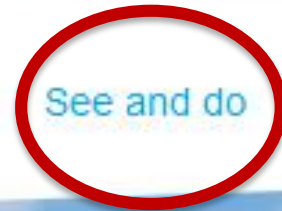
A Journey in the meandering lines of Internet



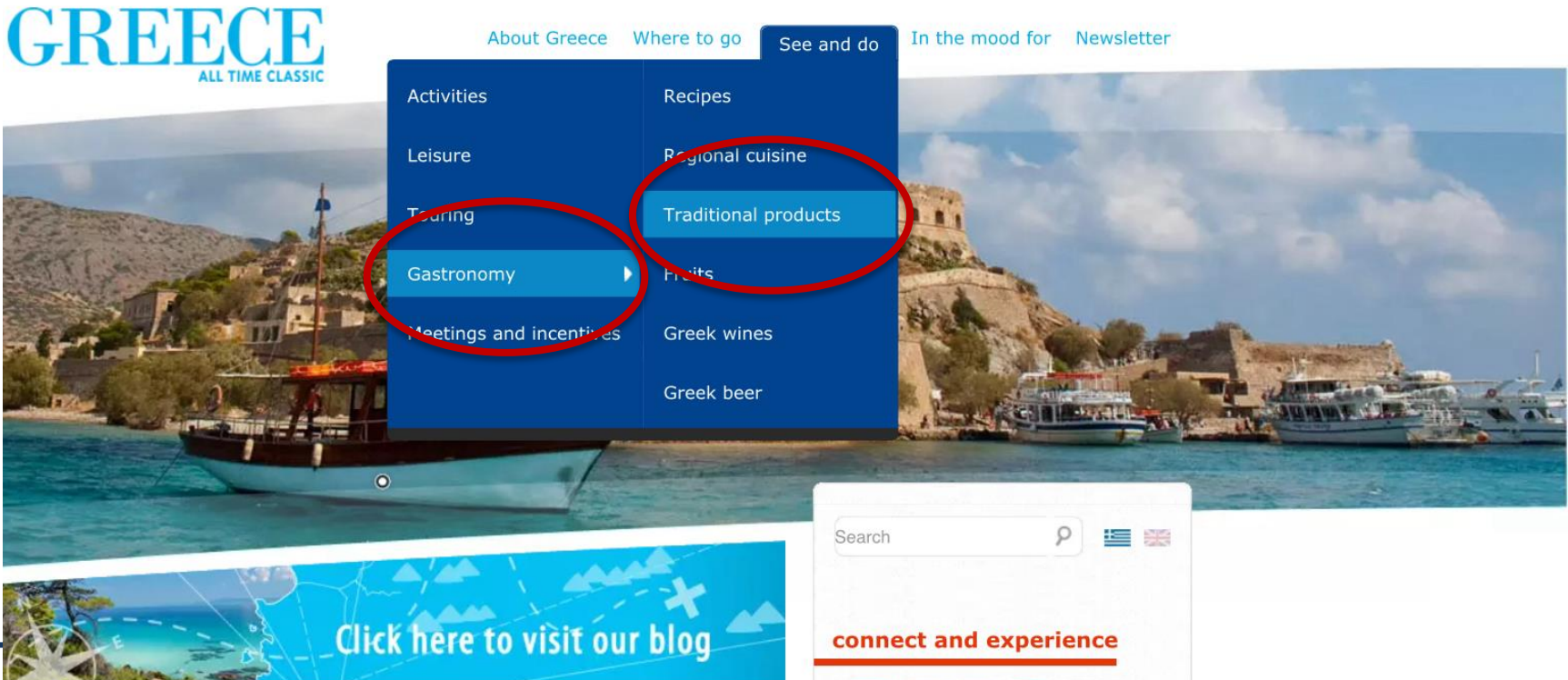
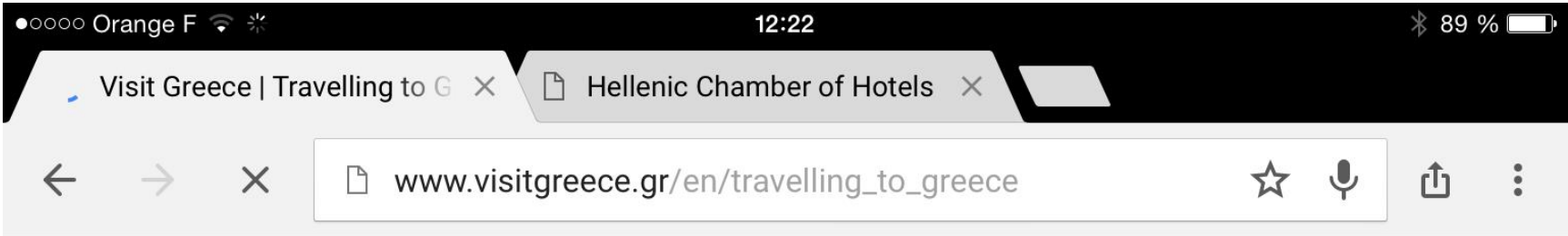
DT ⓘ Οδηγίες για το GDT Athens Google Map ΒΣΒ EMTK ΜΛ2008 TrEd User's Manual

CE
TIME CLASSIC

[About Greece](#) [Where to go](#) [See and do](#) [In the mood for](#) [Ne](#)



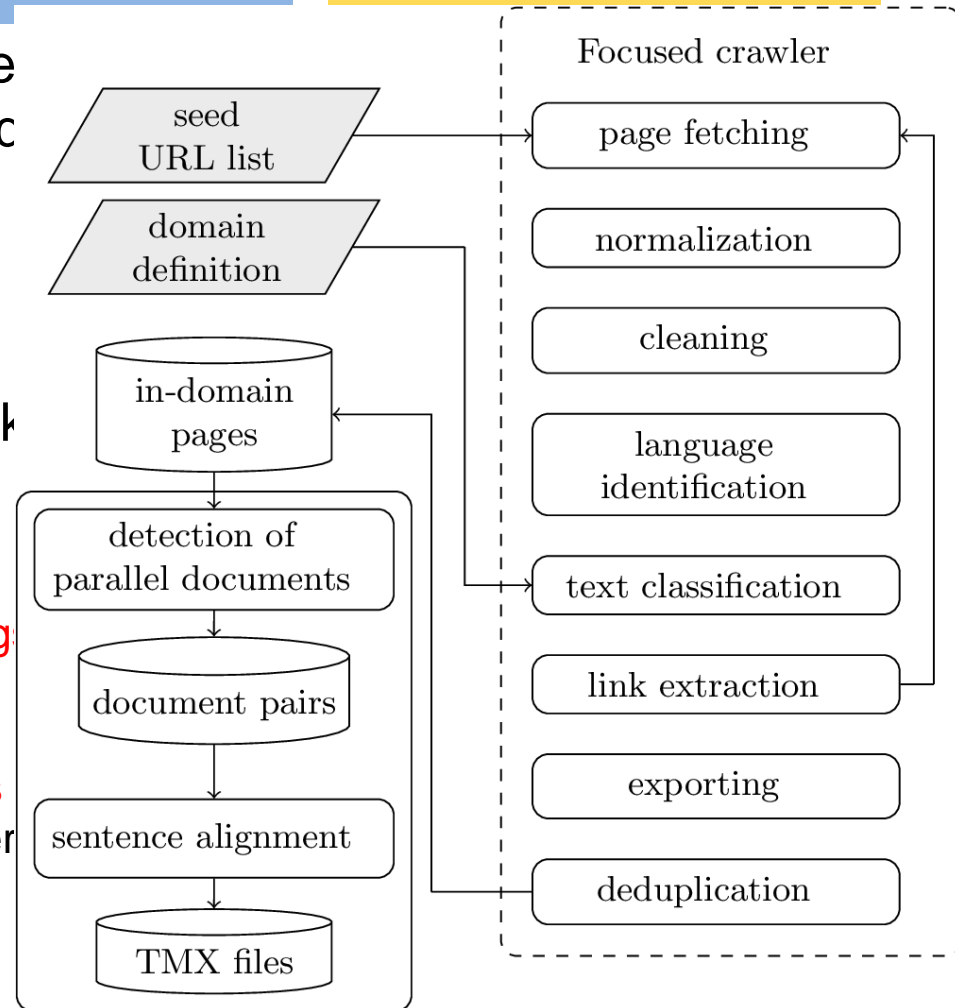
(automatically) Follow all referenced links

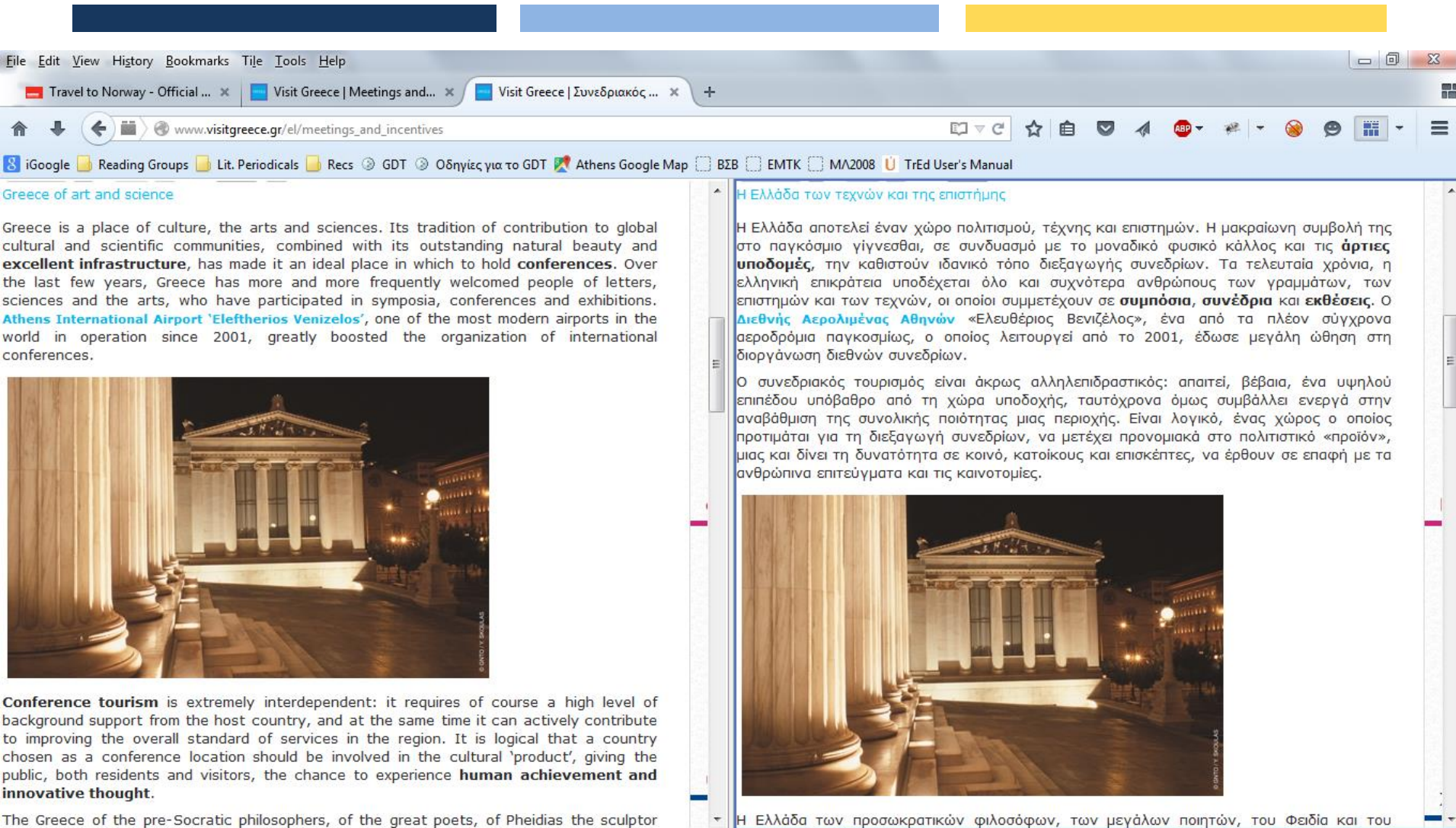




- <http://portal.elda.org/> <http://portal.elda.org/en/>
 - <http://portal.elda.org/news/rss/>
 - <http://portal.elda.org/login/>
 - <http://portal.elda.org/en/login/>
 - <http://portal.elda.org/reset/>
 - <http://portal.elda.org/about/elra/contact/>
 - <http://portal.elda.org/en/about/elra/contact/>
 - <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/>

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





File Edit View History Bookmarks Title Tools Help


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

www.visitgreece.gr/el/meetings_and_incentives

iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map ΒΣΒ EMTK ΜΑ2008 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 γίνεσθαι, 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

... identify the language of each crawled page ...



The screenshot shows a web browser with two tabs open: "Visit Greece | Meetings and..." and "Visit Greece | Συνεδριακός...". The address bar shows the URL: `http://abuma...xml/1234.xml`. The browser's search bar contains "abumatran.eu/~vpapa/data/EN-EL/crawled_data/visitgreece_20150825_154605/eac25a8b-87cd-4b08-b045-571ccb003af6/xml/1234.xml".

The left pane displays XML metadata for the English page. A red box highlights the following tags:

```
</fileDesc>
- <profileDesc>
  - <langUsage>
    <language iso639="en"/>
  </langUsage>
```

The right pane displays XML metadata for the Greek page. A red box highlights the following tags:

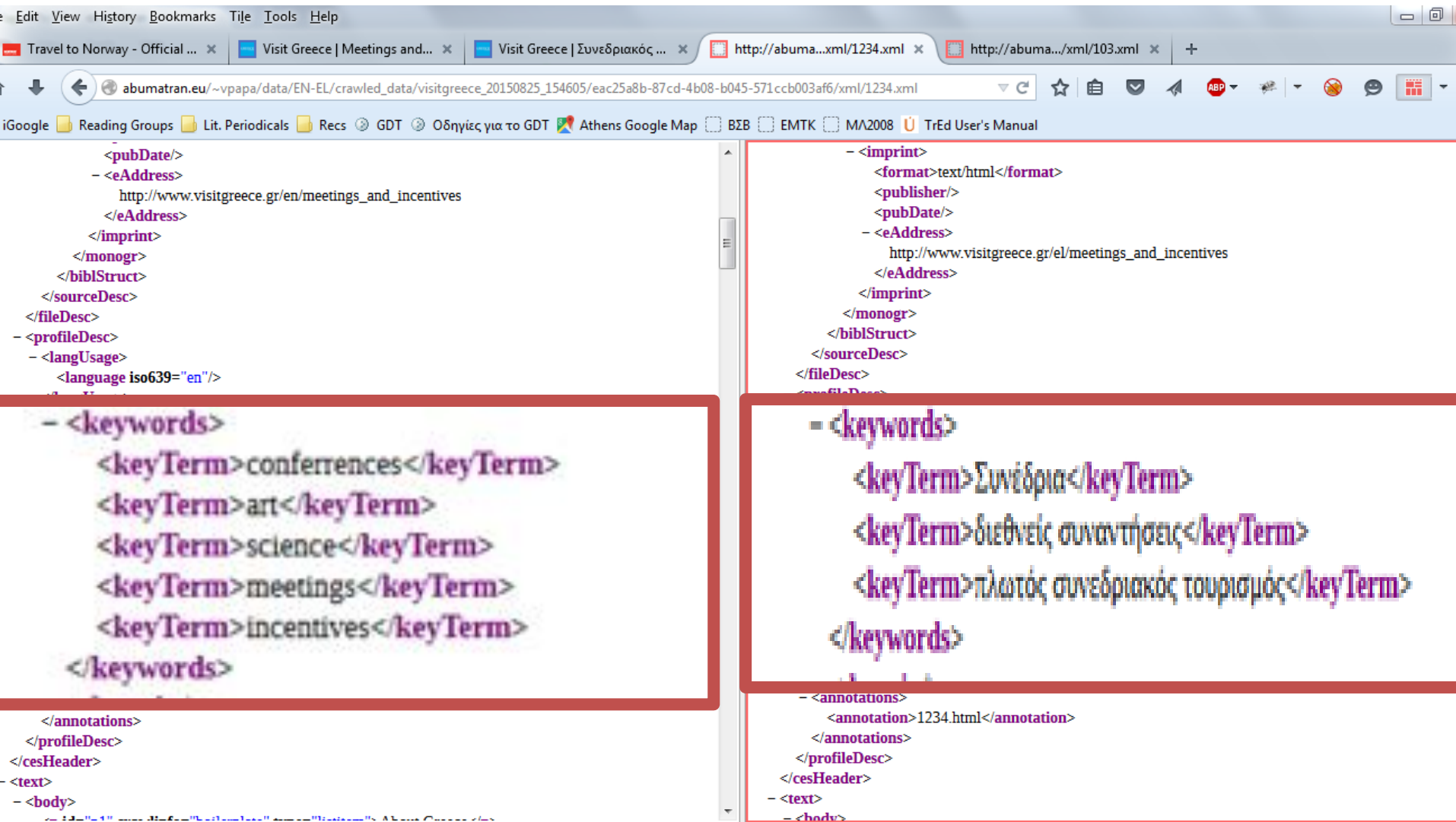
```
- <profileDesc>
  - <langUsage>
    <language iso639="el"/>
  </langUsage>
```

... identify the language of each crawled page



```
Source of: http://www.visitnorway.com/uk/getting-here-and-around/ - Mozilla Firefox
File Edit View Help
49
50 <link rel="alternate" hreflang="da" href="http://www.visitnorway.com/dk/transportmuligheder/" />
51
52 <link rel="alternate" hreflang="es" href="http://www.visitnorway.com/es/como-llegar-y-como-
.....
4 <link rel="alternate" hreflang="fr"
en-norvege/" />
5
59
60 <link rel="alternate" hreflang="sv" href="http://www.visitnorway.com/se/transport/" />
61
62 <link rel="alternate" hreflang="it" href="http://www.visitnorway.com/it/arrivare-e-muoversi/" />
63
64 <link rel="alternate" hreflang="ru" href="http://www.visitnorway.com/ru/getting-here-and-
around/" />
65
66 <link rel="alternate" hreflang="en-US" href="http://www.visitnorway.com/us/getting-here-and-
around/" />
67
68 <link rel="alternate" hreflang="pl" href="http://www.visitnorway.com/pl/transport-w-norwegii/"
/>
69
70 <link rel="alternate" hreflang="zh-CN" href="http://www.visitnorway.com/cn/getting-here-and-
around/" />
71
72 <link rel="alternate" hreflang="pt-BR" href="http://www.visitnorway.com/br/como-chegar-
e-mover-se/" />
73
74
75
76 <link rel="stylesheet" href="//d1fv7ceopli51a.cloudfront.net/bundles/styles/visit-
```

... extract several types of data descriptors (metadata)



```
<pubDate/>
- <eAddress>
  http://www.visitgreece.gr/en/meetings_and_incentives
</eAddress>
</imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
- </profileDesc>
- </langUsage>
  <language iso639='en'/>
- <keywords>
  <keyTerm>conferences</keyTerm>
  <keyTerm>art</keyTerm>
  <keyTerm>science</keyTerm>
  <keyTerm>meetings</keyTerm>
  <keyTerm>incentives</keyTerm>
</keywords>
</annotations>
</profileDesc>
</cesHeader>
- <text>
- <body>
  <!-- ... -->
- </body>


- <imprint>
  <format>text/html</format>
  </publisher/>
  </pubDate/>
  - </eAddress>
    http://www.visitgreece.gr/el/meetings_and_incentives
  </eAddress>
</imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
</profileDesc>
</cesHeader>
</text>
- <body>
  <!-- ... -->
- </body>
```




File Edit View History Bookmarks Title Tools Help
 Sentence alignment for 16.xml... x hreflang - Google Search x Search results | EuroVoc x Visit Greece | Agritourism i... x rural tourism | EuroVoc x Visit Greece | Αγροτουρισ... x

www.visitgreece.gr/en/nature/agrotourism/agritourism_in_greece

iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map BZB EMTK MA2008 TrEd User's Manual

This site is part of  Multilingual Thesaurus of the European Union

Europa > EuroVoc homepage > Domains and MT > rural tourism

Content language:
 (en) English

Simple search

■ **Advanced search**

Browse

■ Browse the subject-oriented version

Download

■ By domain
 ■ Permuted alphabetical
 ■ Multilingual list
 ■ Alphabetical index
 ■ EuroVoc SKOS/RDF

rural tourism

UF *agritourism*
farm holidays

BT2 leisure
 RT agricultural holding [5616]

URI <http://eurovoc.europa.eu/3341>

Has Exact Match
Rural tourism (ECLAS)
agritourism (GEMET)

Has Close Match
country lodge (GEMET)

Harvest grapes to make **wine** or **tsipouro**, or pick **fruits, herbs**, and **mushrooms**. Take part in re about **bee-keeping** by having your own hands-on experience.

Working with the cattle

For those who love **animals**, getting to know how t them will sure be a challenge. After all, you don't ha to see cattle's grazing or milking every day! A more gastronomic choice involves participation in **cheese** making procedures.

An educational holiday

In agritourism lessons of cook

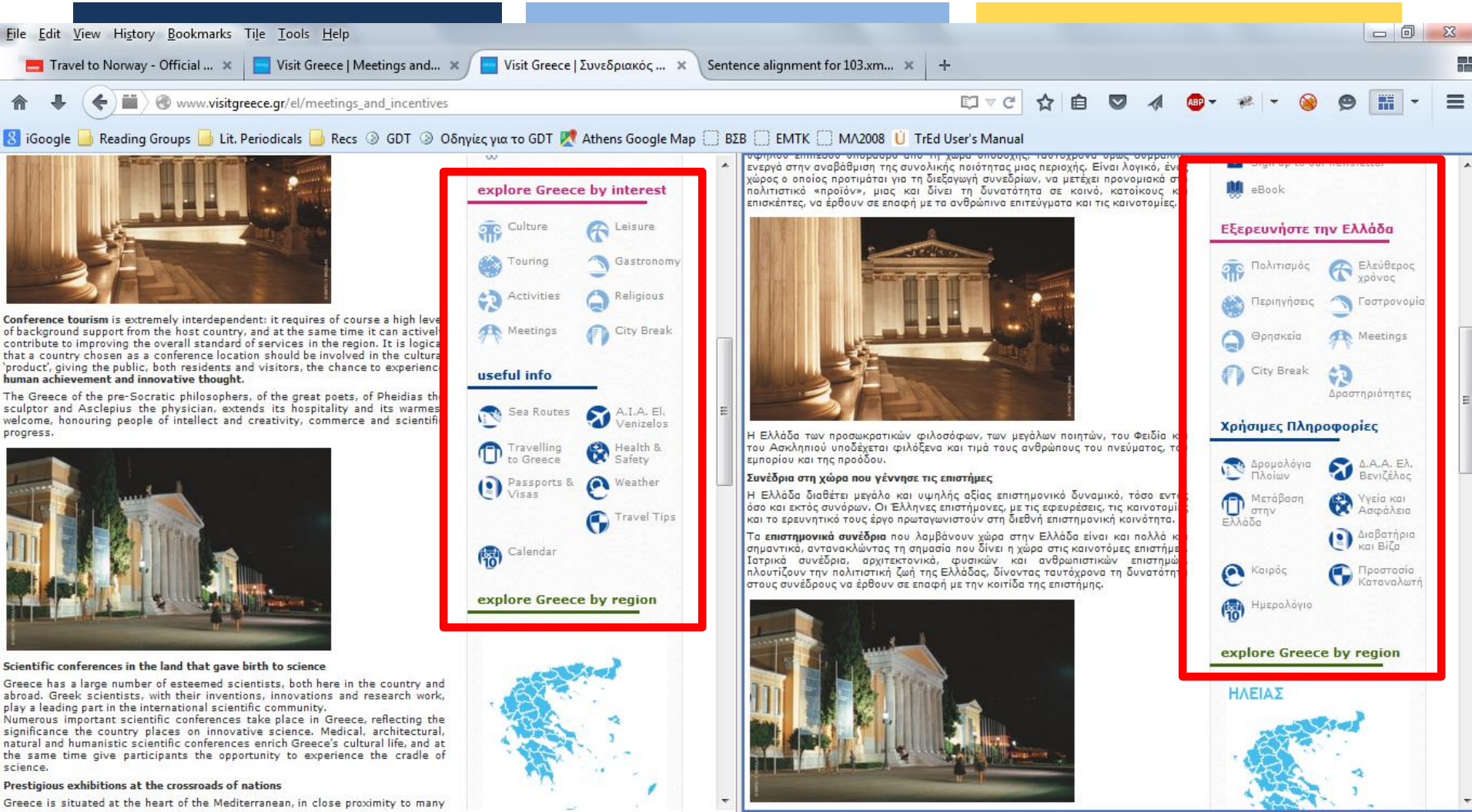
in **pottery** workshops you'll develop your **handcraft** might also make the very presents you'll carry back Greece. **Wineries** will let you sip exquisite wine and into the secrets of wine: **varieties, aromas, colours**

Ecotourism is part and parcel of **agritourism**. Deper you'll be staying, there might be a **plethora of won** for you in store. National parks, wetlands, stunning l one of the richest floras and faunas in Europe await into action, don't miss out on the chance to do your activity in the midst of the superb Greek natural bea fishing, **hiking, mountaineering, horse-riding**. Bu

wrong choosing the Greek nature for a relaxed holiday either: wake up to the **singing of birds** and **breakfast** in the shade of a vine, or dine at the sight of sunset **colouring olive groves** in gold an

More info: <http://agrozenia.net>

It can detect boilerplate text ...

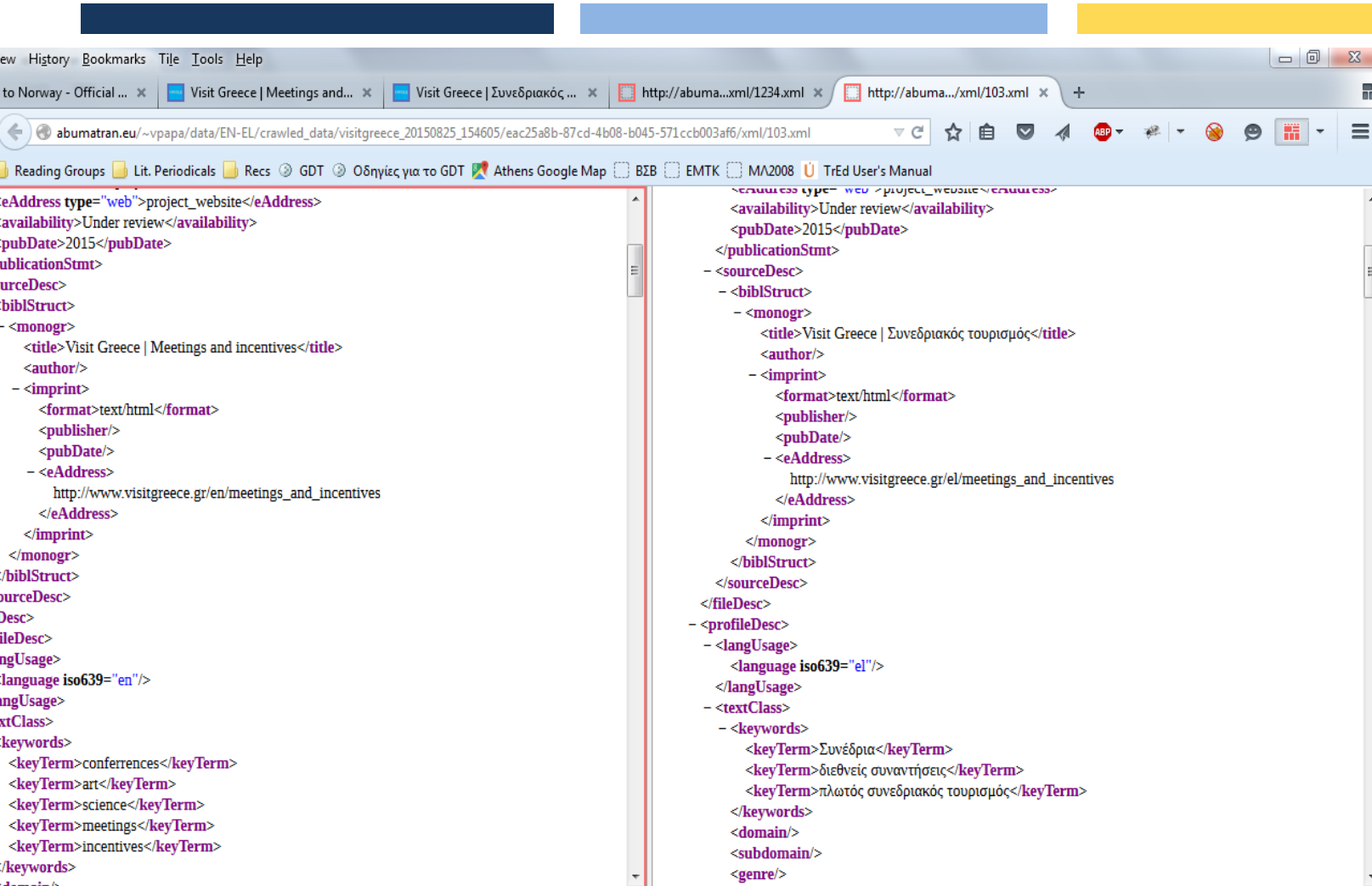
File Edit View History Bookmarks Tile Tools Help
 Travel to Norway - Official ... Visit Greece | Meetings and... Visit Greece | Συνεδριακός ... Sentence alignment for 103.xm...
 www.visitgreece.gr/el/meetings_and_incentives
 iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map ΒΣΒ EMTK ΜΑ2008 TrEd User's Manual

explore Greece by interest
 Culture Leisure
 Touring Gastronomy
 Activities Religious
 Meetings City Break
useful info
 Sea Routes A.I.A. El. Venizelos
 Travelling to Greece Health & Safety
 Passports & Visas Weather
 Calendar Travel Tips
explore Greece by region

Εξερευνήστε την Ελλάδα
 Πολιτισμός Ελεύθερος χρόνος
 Περιηγήσεις Γαστρονομία
 Θρησκεία Meetings
 City Break Δραστηριότητες
Χρήσιμες Πληροφορίες
 Δρομολόγια Πλοίων Δ.Α.Α. Ελ. Βενιζέλος
 Μετάβαση στην Ελλάδα Υγεία και Ασφάλεια
 Διαβατήρια και Βίζα
 Καίρος Προστασία Καταναλωτή
 Ημερολόγιο
explore Greece by region
 ΗΑΕΙΑΣ

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 and Asclepius the physician, extends its hospitality and its warm welcome, honouring people of intellect and creativity, commerce and scientific progress.
 Scientific conferences in the land that gave birth to science
 Greece has a large number of esteemed scientists, both here in the country and abroad. Greek scientists, with their inventions, innovations and research work, play a leading part in the international scientific community. Numerous important scientific conferences take place in Greece, reflecting the significance the country places on innovative science. Medical, architectural, natural and humanistic scientific conferences enrich Greece's cultural life, and at the same time give participants the opportunity to experience the cradle of science.
 Prestigious exhibitions at the crossroads of nations
 Greece is situated at the heart of the Mediterranean, in close proximity to many

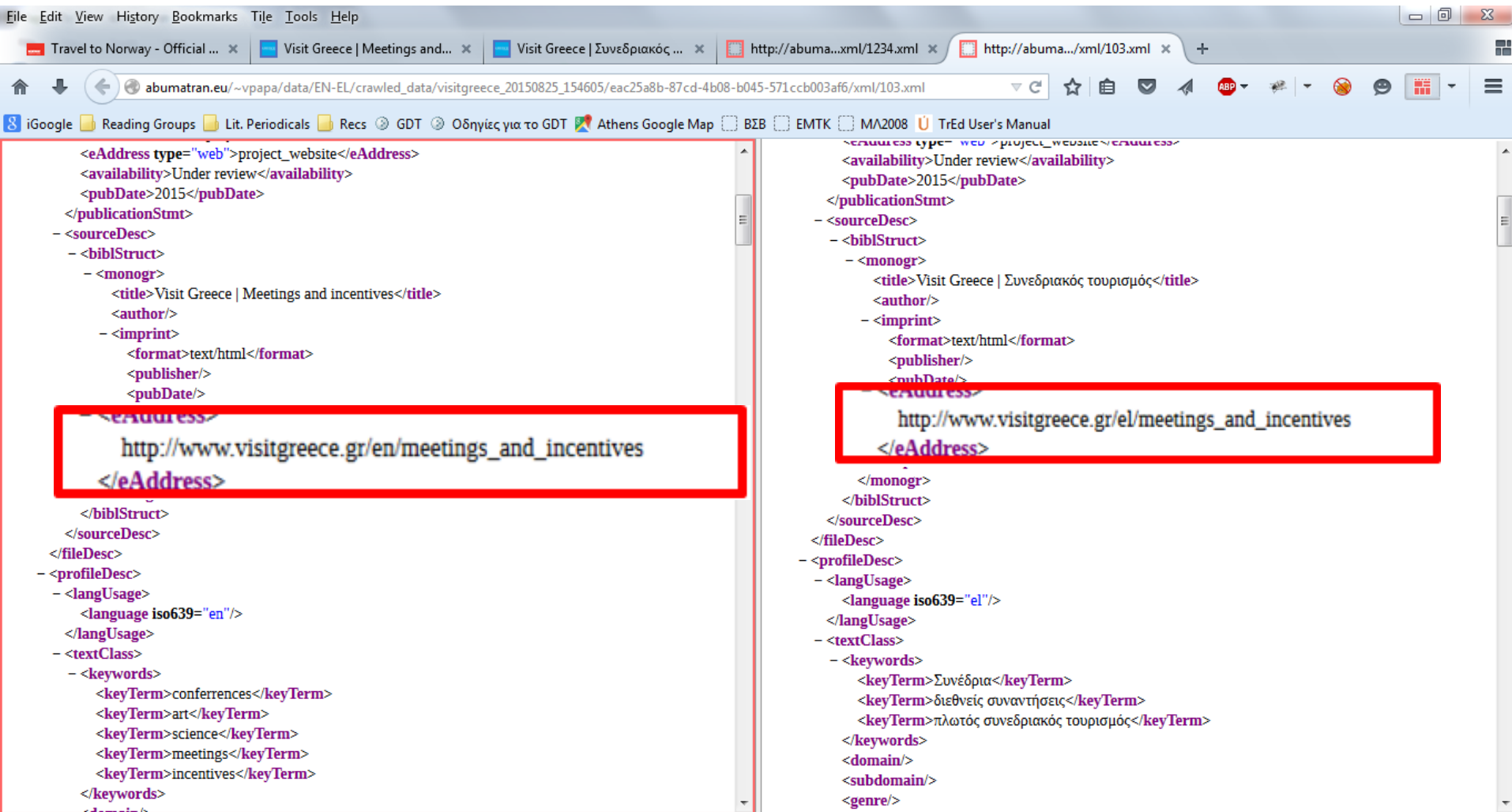
... HTML structure and/or URL similarity to detect document pairs



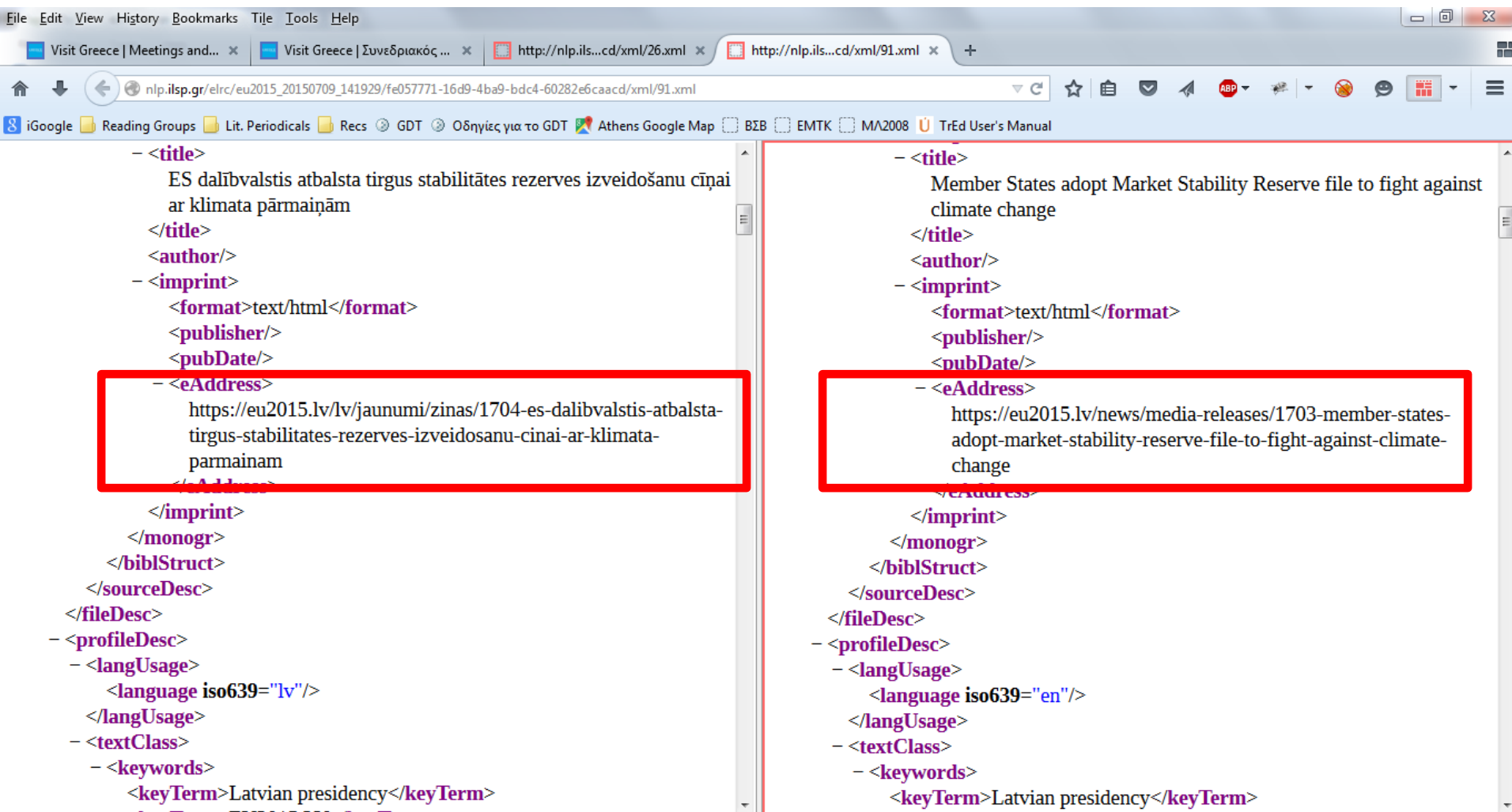
The screenshot shows a web browser with two tabs open: "Visit Greece | Meetings and..." and "Visit Greece | Συνεδριακός...". The address bar shows the URL "http://abuma...xml/103.xml". The browser displays XML code for two documents side-by-side. The left document is in English and the right is in Greek. The XML code is highlighted in red and blue. The left document is in English and the right is in Greek. The XML code is highlighted in red and blue.

```
<eAddress type="web">project_website</eAddress>
<availability>Under review</availability>
<pubDate>2015</pubDate>
<publicationStm
<sourceDesc>
  <biblStruct>
    <monogr>
      <title>Visit Greece | Meetings and incentives</title>
      <author/>
    </monogr>
  </biblStruct>
</sourceDesc>
<fileDesc>
  <profileDesc>
    <langUsage>
      <language iso639="en"/>
    </langUsage>
  </profileDesc>
  <textClass>
    <keywords>
      <keyTerm>conferences</keyTerm>
      <keyTerm>art</keyTerm>
      <keyTerm>science</keyTerm>
      <keyTerm>meetings</keyTerm>
      <keyTerm>incentives</keyTerm>
    </keywords>
  </textClass>
</fileDesc>
</document>
```

```
<eAddress type="web">project_website</eAddress>
<availability>Under review</availability>
<pubDate>2015</pubDate>
<publicationStm
<sourceDesc>
  <biblStruct>
    <monogr>
      <title>Visit Greece | Συνεδριακός τουρισμός</title>
      <author/>
    </monogr>
  </biblStruct>
</sourceDesc>
<fileDesc>
  <profileDesc>
    <langUsage>
      <language iso639="el"/>
    </langUsage>
  </profileDesc>
  <textClass>
    <keywords>
      <keyTerm>Συνέδρια</keyTerm>
      <keyTerm>διεθνείς συναντήσεις</keyTerm>
      <keyTerm>πλωτός συνεδριακός τουρισμός</keyTerm>
    </keywords>
  </textClass>
</fileDesc>
</document>
```



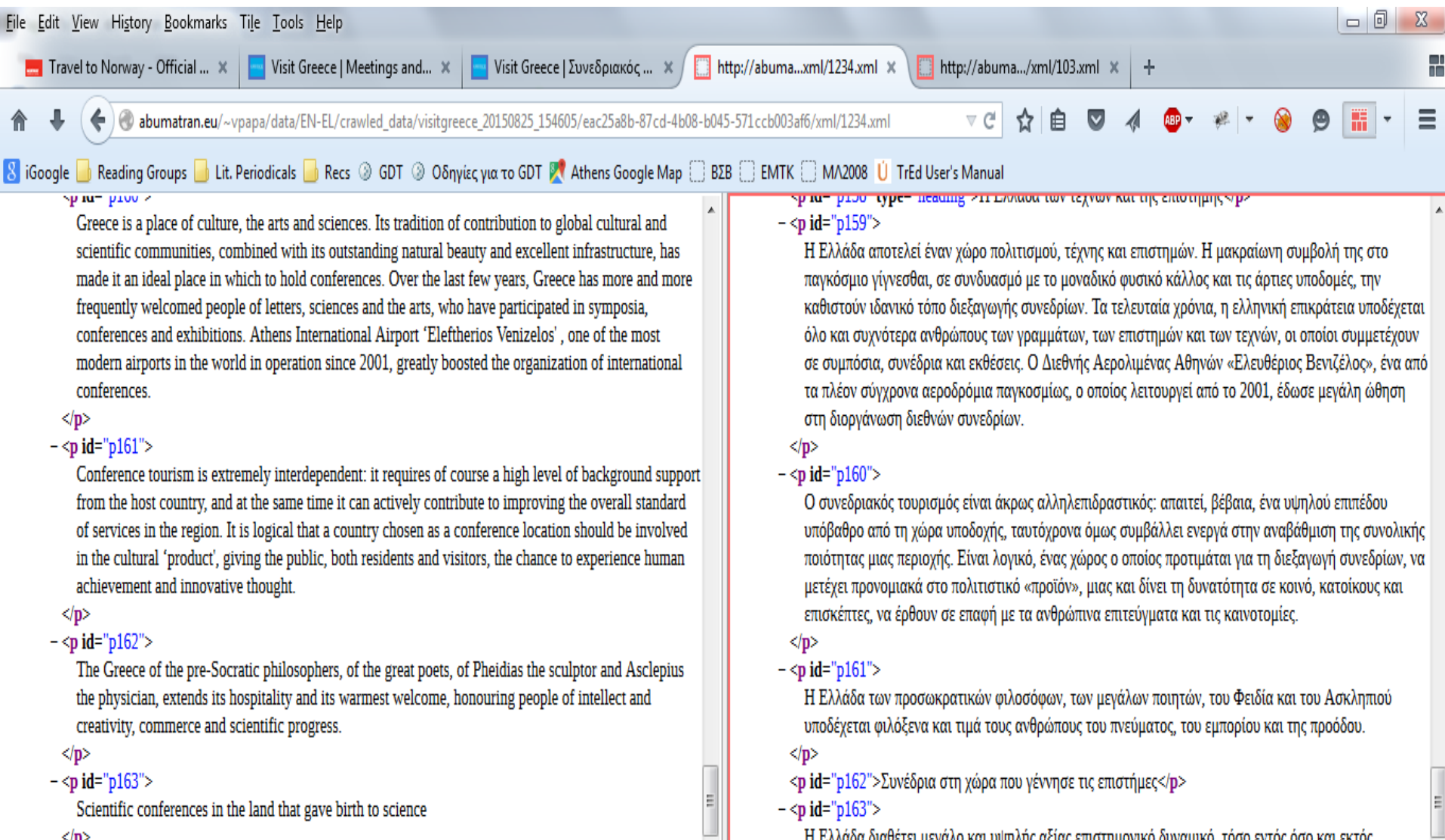
The screenshot shows a web browser with two XML documents open side-by-side. The browser's address bar shows the URL `http://abumatran.eu/~vpapa/data/EN-EL/crawled_data/visitgreece_20150825_154605/eac25a8b-87cd-4b08-b045-571ccb003af6/xml/103.xml`. The left document is from `abumatran.eu` and the right document is from `abuma...xml/103.xml`. Both documents contain an `<eAddress type="web">project_website</eAddress>` element with the value `http://www.visitgreece.gr/en/meetings_and_incentives`, which is highlighted with a red box in both views. The left document also contains a `<title>Visit Greece | Meetings and incentives</title>` element, while the right document contains a `<title>Visit Greece | Συνεδριακός τουρισμός</title>` element. The browser's search bar shows the text "iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map ΒΣΒ ΕΜΤΚ ΜΑ2008 TrEd User's Manual".



The screenshot shows a web browser with two XML documents open. The left document is in Latvian and the right is in English. Both documents have their `<eAddress>` fields highlighted with red boxes. The left document's `<eAddress>` field contains a URL with a Latvian path, while the right document's `<eAddress>` field contains a URL with an English path. The rest of the XML structure, including `<title>`, `<author>`, `<imprint>`, `<publisher>`, `<pubDate>`, `</monogr>`, `</biblStruct>`, `</sourceDesc>`, `</fileDesc>`, `<profileDesc>`, `<langUsage>`, `<textClass>`, and `<keywords>`, is identical in both documents.

```
- <title>
  ES dalībvalstis atbalsta tirgus stabilitātes rezerves izveidošanu cīņai
  ar klimata pārmaiņām
</title>
<author/>
- <imprint>
  <format>text/html</format>
  <publisher/>
  <pubDate/>
- <eAddress>
  https://eu2015.lv/lv/jaunumi/zinas/1704-es-dalibvalstis-atbalsta-
  tirgus-stabilitates-rezerves-izveidosanu-cinai-ar-klimata-
  parmainam
</eAddress>
</imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
- <profileDesc>
- <langUsage>
  <language iso639="lv"/>
</langUsage>
- <textClass>
- <keywords>
  <keyTerm>Latvian presidency</keyTerm>
```

```
- <title>
  Member States adopt Market Stability Reserve file to fight against
  climate change
</title>
<author/>
- <imprint>
  <format>text/html</format>
  <publisher/>
  <pubDate/>
- <eAddress>
  https://eu2015.lv/news/media-releases/1703-member-states-
  adopt-market-stability-reserve-file-to-fight-against-climate-
  change
</eAddress>
</imprint>
</monogr>
</biblStruct>
</sourceDesc>
</fileDesc>
- <profileDesc>
- <langUsage>
  <language iso639="en"/>
</langUsage>
- <textClass>
- <keywords>
  <keyTerm>Latvian presidency</keyTerm>
```

The screenshot shows a web browser window with multiple tabs. The active tab displays an XML document from abumatran.eu. The XML content is shown in a split view, with the left pane displaying the raw XML and the right pane displaying the rendered HTML content. The rendered content consists of several paragraphs of Greek text, each enclosed in a red border. The paragraphs are separated by XML tags like `<p id="p159">` and `</p>`.

Left pane (raw XML):

```
<p id="p158">
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.
</p>
<p id="p161">
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.
</p>
<p id="p162">
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.
</p>
<p id="p163">
Scientific conferences in the land that gave birth to science
</p>
```

Right pane (rendered HTML):

```
<p id="p159">
Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. Η μακραίωνη συμβολή της στο παγκόσμιο γίνεσθαι, 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.
</p>
<p id="p160">
Ο συνεδριακός τουρισμός είναι άκρως αλληλεπιδραστικός: απαιτεί, βέβαια, ένα υψηλό επίπεδο 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.
</p>
<p id="p161">
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 πρόοδου.
</p>
<p id="p162">
Συνέδρια στη χώρα που γέννησε τις επιστήμες</p>
<p id="p163">
Η Ελλάδα διαθέτει μεγάλο και υψηλής αξίας επιστημονικό δυναμικό τόσο εντός όσο και εκτός
```



File Edit View History Bookmarks Title Tools Help

Visit Greece | Meetings and... x Visit Greece | Συνεδριακός... x Sentence alignment for 91.xml... x +

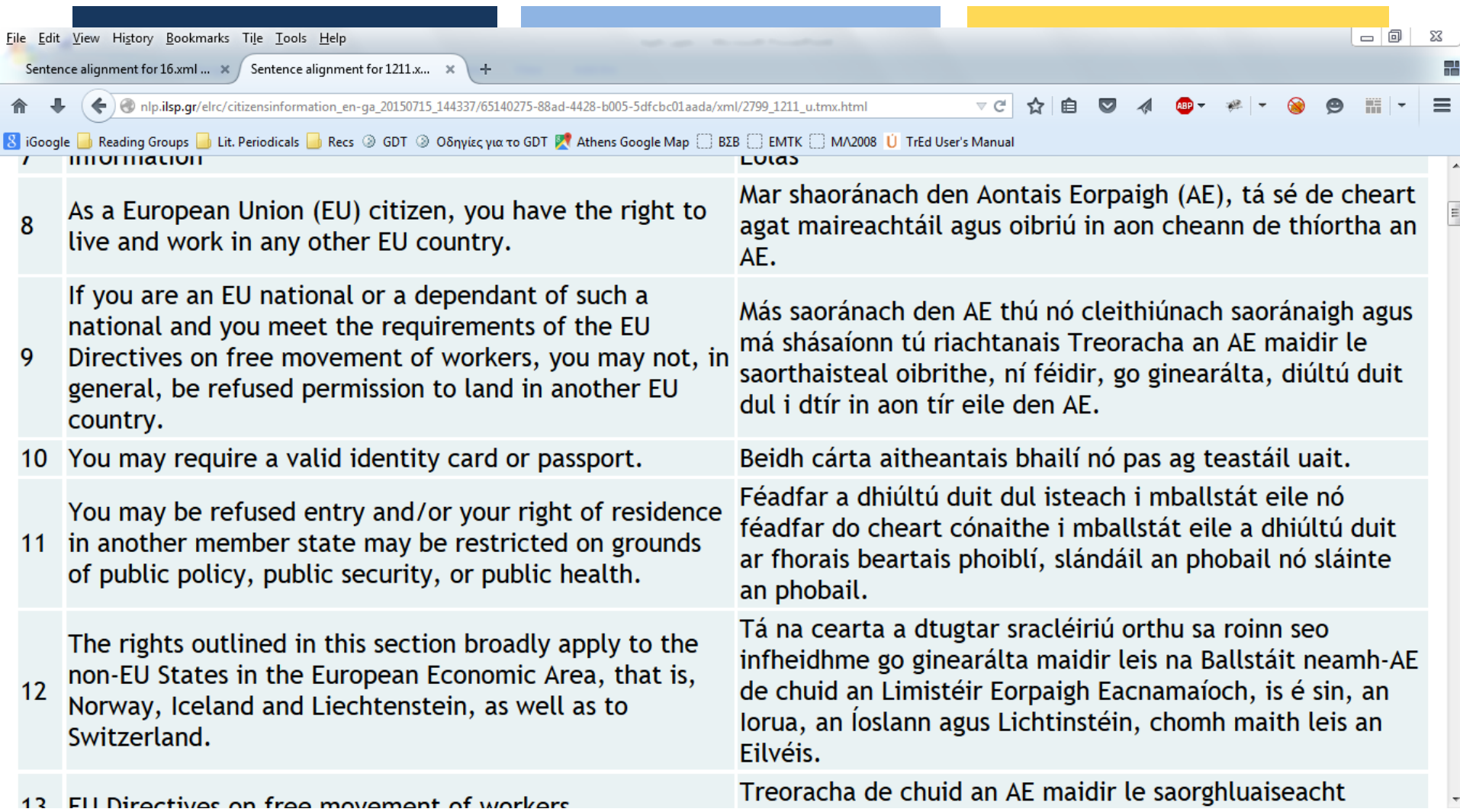
nlp.ilsp.gr/elrc/eu2015_20150709_141929/fe057771-16d9-4ba9-bdc4-60282e6caacd/xml/26_91_i.tmx.html

iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map BZB EMTK MA2008 TrEd User's Manual

#	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 ievieš pasākumus, 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



EN-GA



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

it supports all EU languages!

EN-FR



File	Edit	View	History	Bookmarks	Title	Tools	Help
Sentence alignment for 16.xml ... x Sentence alignment for 1211.x... x +							
nlp.ilsp.gr/elrc/cleiss_fr_20150711_100255/a9f72a1e-b072-44c8-8785-0885ea439e32/xml/16_24_h.tmx.html							
iGoogle Reading Groups Lit. Periodicals Recs GDT Οδηγίες για το GDT Athens Google Map BZB EMTK MA2008 TrEd User's Manual							
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 deux ans, en cas de maladie ou de maternité, l'assuré doit 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

It supports all EU languages!

EN-LV



	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

- 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*

OPENTEXT

The Deep Web

The Public Web

Only 4% of Web content (~8 billion pages) is available via search engines like Google

7.9
Zettabytes

The Deep Web

Approximately 96% of the digital universe is on Deep Web sites protected by passwords

Source: *The Deep Web: Semantic Search Takes Innovation to New Depths*

Our contributions ... Deep web





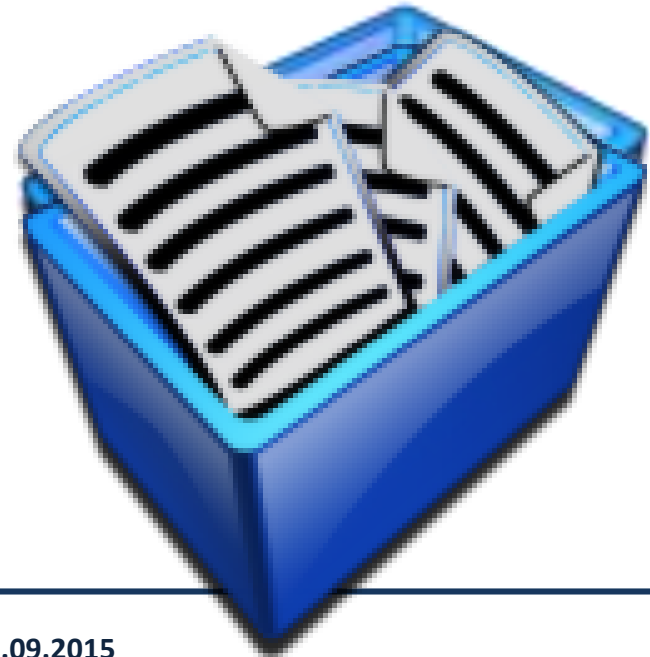
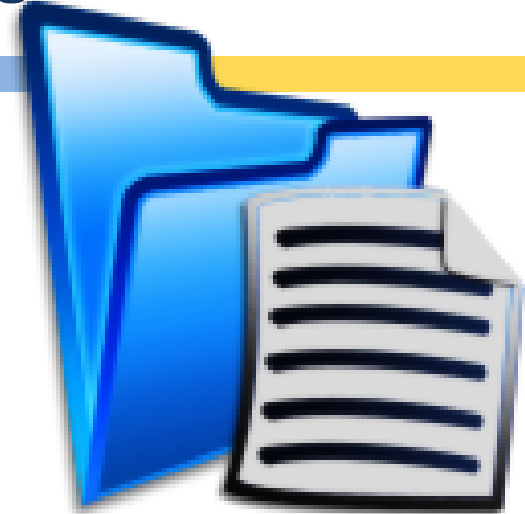
- Such documents exist already:
 - At the various documentation centers (translated reports, leaflets, brochures, speeches, web pages, etc.)
 - At the Language Service Providers (LSP), to whom translation works are subcontracted
- Help us identify and liaise with both sources
 - (see next Panel interactions)

Your involvement is essential so please let us work together



**BRING YOUR OWN
LANGUAGE RESOURCES**

Your involvement is essential so please let us work together



Fichier Accueil Partage Affichage

Volet de visualisation Volet de détails

Très grandes icônes Grandes icônes Icônes moyennes
Petites icônes Liste Détails

Mosaïques Contenu

Trier par Grouper par

Ajouter des colonnes
Ajuster la taille de toutes les colonnes

Cases à cocher des éléments
Extensions de noms de fichiers
Éléments masqués

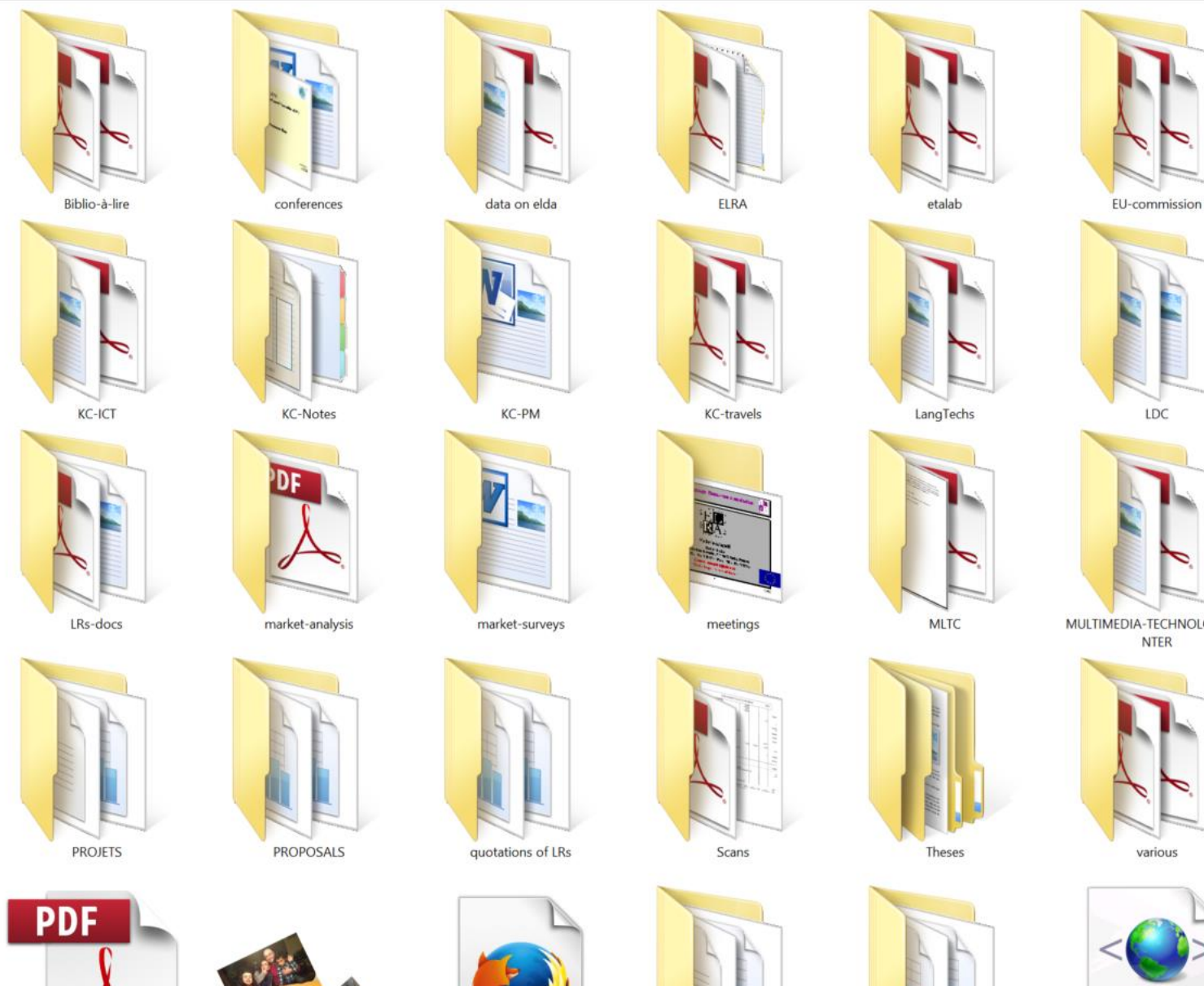
Masquer les éléments sélectionnés

Options

Afficher/Masquer

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
- assurances

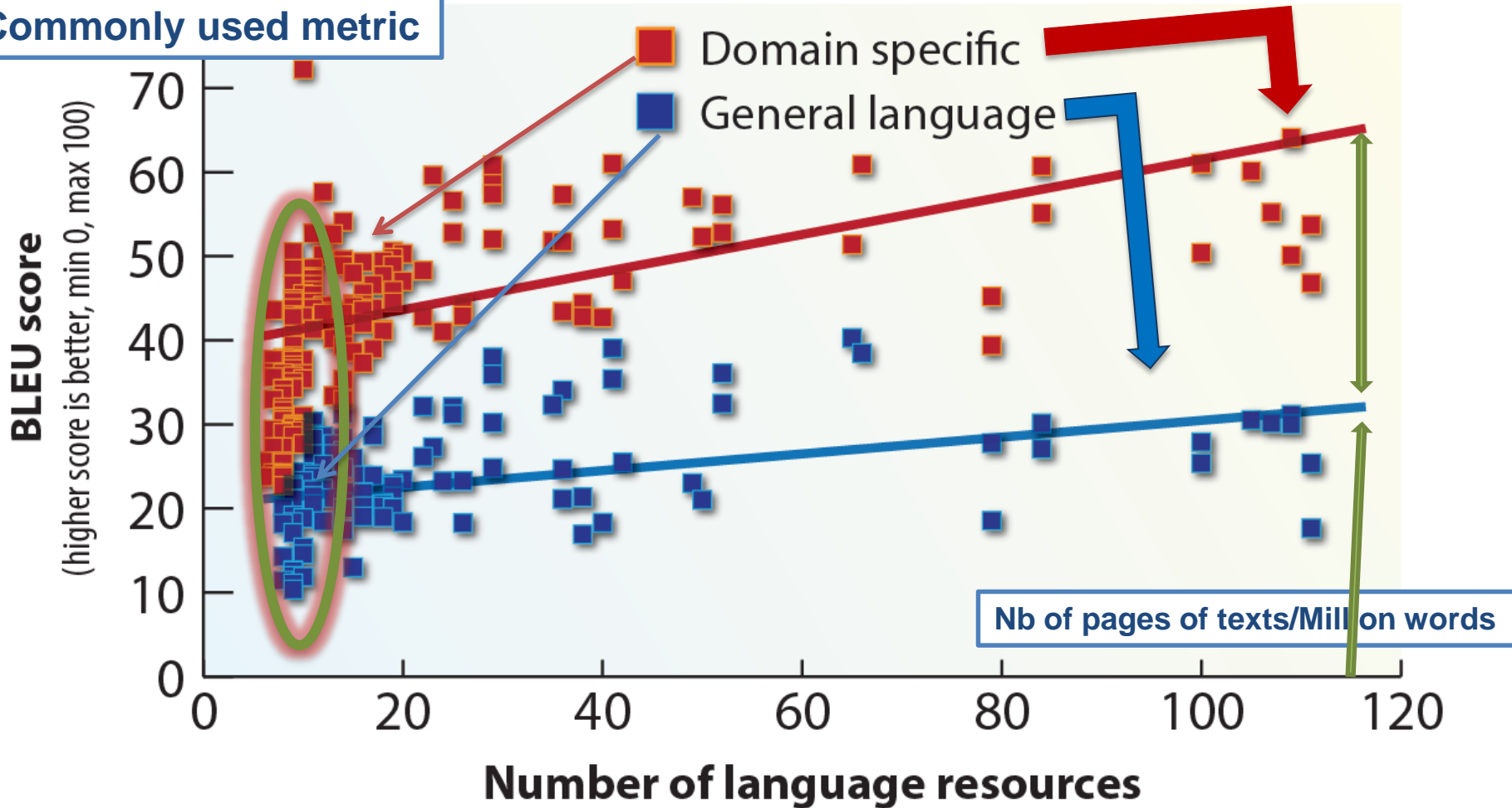


- How to help upload the data
- See the information on the REPOSITORY set-up for this
- **How much data is needed ?**

Impact of number of language resources on Statistical MT quality



A Commonly used metric



Nb of pages of texts/Million words

- 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)*