European Language Resource Coordination
Connecting Europe Facility

Welche Daten werde benötigt um CEF.AT zu optimieren?

Bartholomäus Wloka (Universität Wien)
Basierend auf der Präsentation von Dr. Khalid Choukri (Evaluations and Language resource Distribution Agency)

European Language Resource Coordination
Connecting Europe Facility

Daten

- Im Vordergrund: ***data-driven paradigm***
  - MÜ Systeme brauchen Daten
  - Fokus für ELRC: Daten in allen Sprachen (EU/CEF)

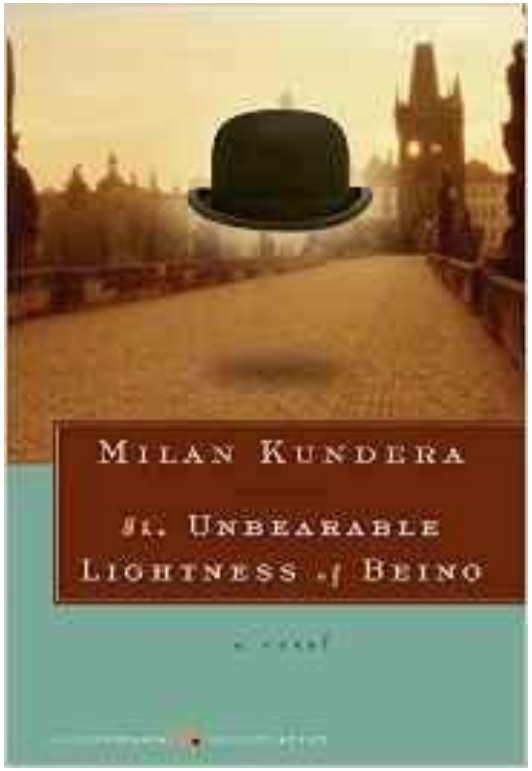
- Daten, d.h. Sprachresourcen werden produziert aus:
  - Dokumenten & jeglichen Textsammlungen
  - Eine Hilfe von Ihrer Seite ist wichtig

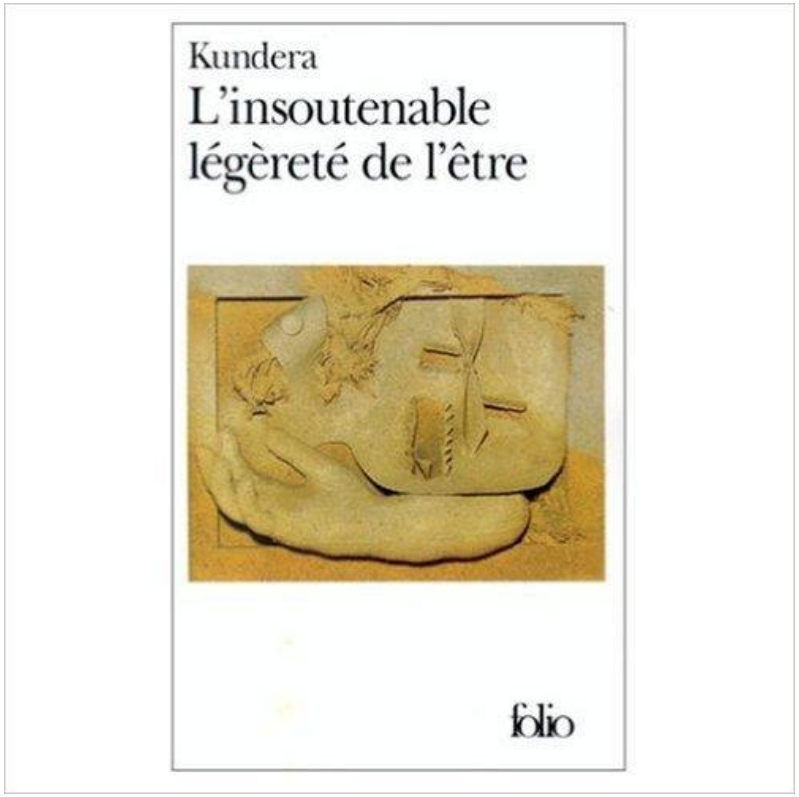
Was sind Daten für MÜ?

- Alles was Wörter und Sätze beinhaltet, v.A. in mehreren Sprachen
  - Berichte
  - Vorträge
  - Webinhalt
  - Broschüren
  - …
- Wird zu: *Bag of Words*

Was sind Daten für MÜ?

Übersetzungen





European Language Resource Coordination
Connecting Europe Facility

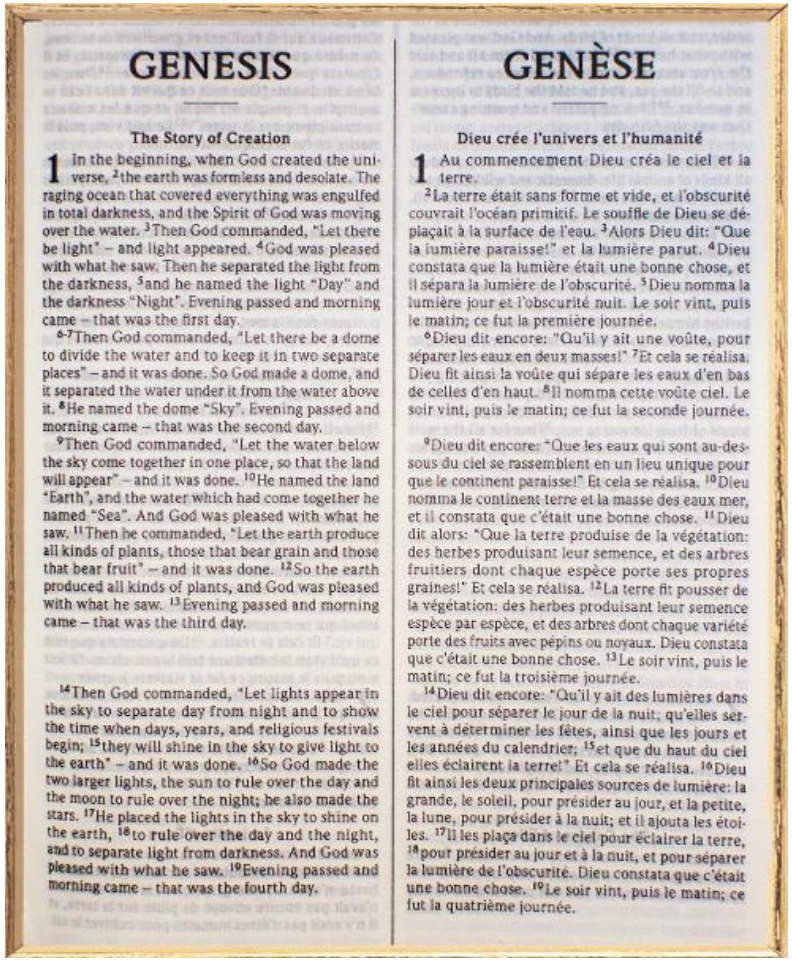
Alignierte Übersetzungen

English

French

European Language Resource Coordination
Connecting Europe Facility

Alignierte Übersetzungen



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 Corpora

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.

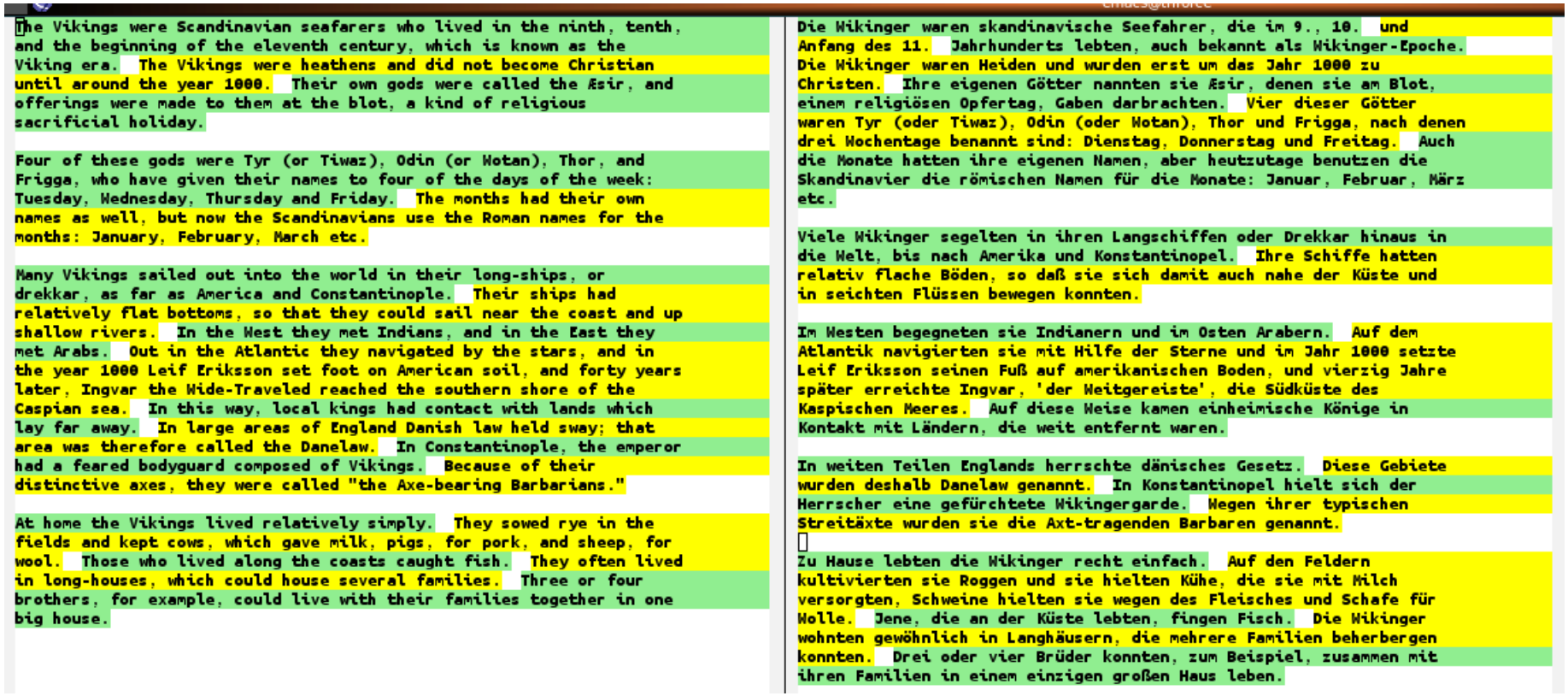
Por metonimia, también se denomina telecomunicación (o telecomunicaciones, indistintamente) a la disciplina que estudia, diseña, desarrolla y explota aquellos sistemas que permiten dichas comunicaciones; de forma análoga, la ingeniería de telecomunicaciones resuelve los problemas técnicos asociados a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

European Language Resource Coordination
Connecting Europe Facility

Alignierte Übersetzungen



The Vikings were Scandinavian seafarers who lived in the ninth, tenth, and the beginning of the eleventh century, which is known as the Viking era. The Vikings were heathens and did not become Christian until around the year 1000. Their own gods were called the Æsir, and offerings were made to them at the blot, a kind of religious sacrificial holiday.

Four of these gods were Tyr (or Tiwaz), Odin (or Wotan), Thor, and Frigga, who have given their names to four of the days of the week: Tuesday, Wednesday, Thursday and Friday. The months had their own names as well, but now the Scandinavians use the Roman names for the months: January, February, March etc.

Many Vikings sailed out into the world in their long-ships, or drekkar, as far as America and Constantinople. Their ships had relatively flat bottoms, so that they could sail near the coast and up shallow rivers. In the West they met Indians, and in the East they met Arabs. Out in the Atlantic they navigated by the stars, and in the year 1000 Leif Eriksson set foot on American soil, and forty years later, Ingvar the Wide-Traveled reached the southern shore of the Caspian sea. In this way, local kings had contact with lands which lay far away. In large areas of England Danish law held sway; that area was therefore called the Danelaw. In Constantinople, the emperor had a feared bodyguard composed of Vikings. Because of their distinctive axes, they were called "the Axe-bearing Barbarians."

At home the Vikings lived relatively simply. They sowed rye in the fields and kept cows, which gave milk, pigs, for pork, and sheep, for wool. Those who lived along the coasts caught fish. They often lived in long-houses, which could house several families. Three or four brothers, for example, could live with their families together in one big house.

Die Wikinger waren skandinavische Seefahrer, die im 9., 10. und Anfang des 11. Jahrhunderts lebten, auch bekannt als Wikinger-Epoche. Die Wikinger waren Heiden und wurden erst um das Jahr 1000 zu Christen. Ihre eigenen Götter nannten sie Æsir, denen sie am Blot, einem religiösen Opfertag, Gaben darbrachten. Vier dieser Götter waren Tyr (oder Tiwaz), Odin (oder Wotan), Thor und Frigga, nach denen drei Wochentage benannt sind: Dienstag, Donnerstag und Freitag. Auch die Monate hatten ihre eigenen Namen, aber heutzutage benutzen die Skandinavier die römischen Namen für die Monate: Januar, Februar, März etc.

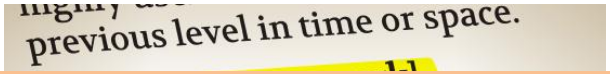
Viele Wikinger segelten in ihren Langschiffen oder Drekkar hinaus in die Welt, bis nach Amerika und Konstantinopel. Ihre Schiffe hatten relativ flache Böden, so daß sie sich damit auch nahe der Küste und in seichten Flüssen bewegen konnten.

Im Westen begegneten sie Indianern und im Osten Arabern. Auf dem Atlantik navigierten sie mit Hilfe der Sterne und im Jahr 1000 setzte Leif Eriksson seinen Fuß auf amerikanischen Boden, und vierzig Jahre später erreichte Ingvar, 'der Weitgereiste', die Südküste des Kaspischen Meeres. Auf diese Weise kamen einheimische Könige in Kontakt mit Ländern, die weit entfernt waren.

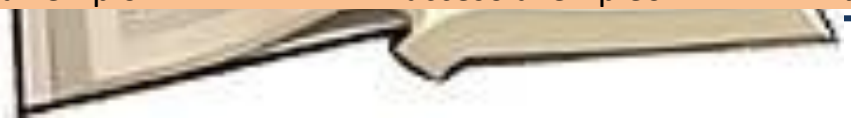
In weiten Teilen Englands herrschte dänisches Gesetz. Diese Gebiete wurden deshalb Danelaw genannt. In Konstantinopel hielt sich der Herrscher eine gefürchtete Wikingergarde. Wegen ihrer typischen Streitäxte wurden sie die Axt-tragenden Barbaren genannt.

Zu Hause lebten die Wikinger recht einfach. Auf den Feldern kultivierten sie Roggen und sie hielten Kühe, die sie mit Milch versorgten, Schweine hielten sie wegen des Fleisches und Schafe für Wolle. Jene, die an der Küste lebten, fingen Fisch. Die Wikinger wohnten gewöhnlich in Langhäusern, die mehrere Familien beherbergen konnten. Drei oder vier Brüder konnten, zum Beispiel, zusammen mit ihren Familien in einem einzigen großen Haus leben.

Wörterbücher/ Terminologien /Ontologien

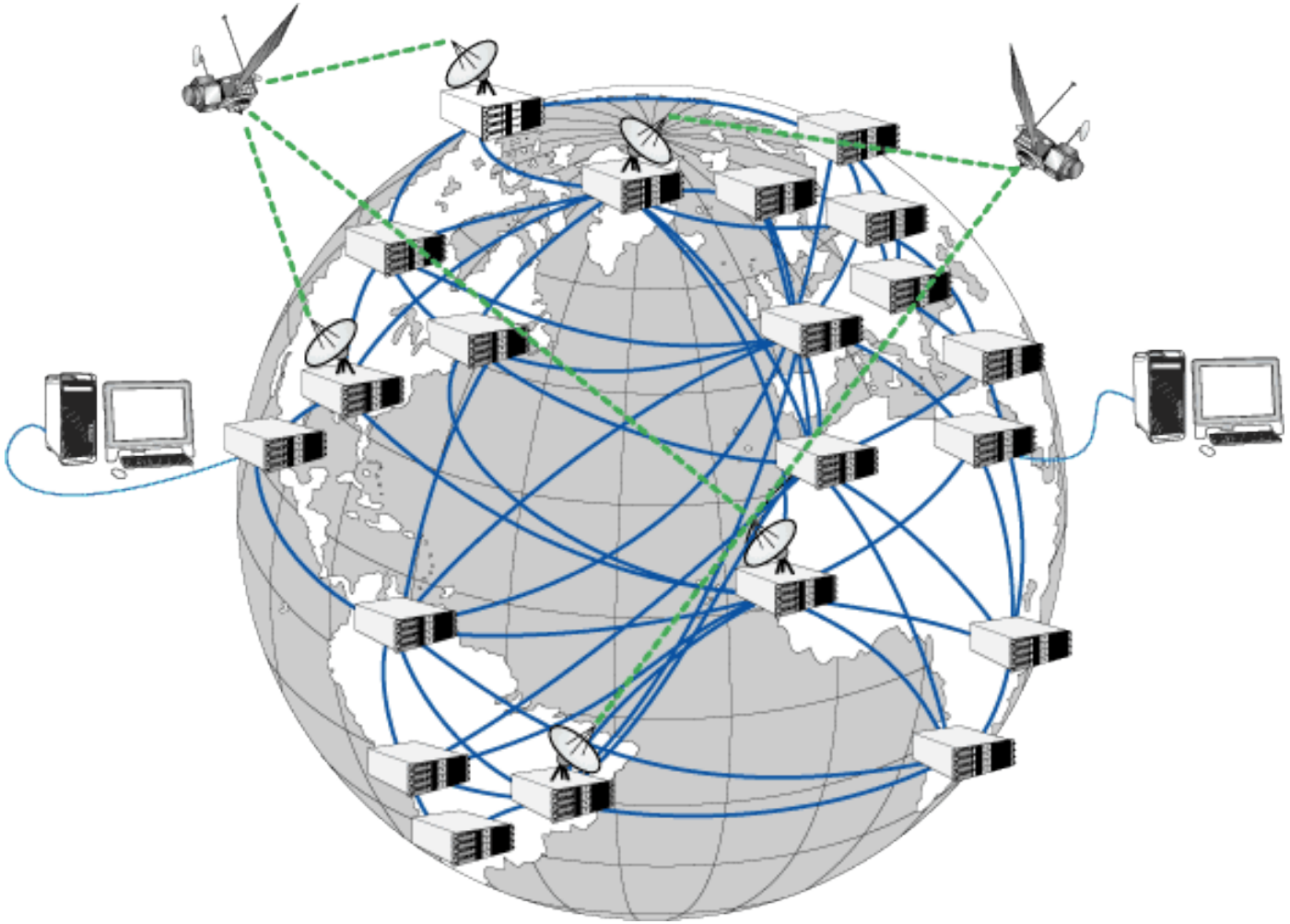


ID	FR	ES	EL
6905	abandon scolaire	abandono escolar	διακοπή της σχολικής φοίτησης
920	abats	despojo	παραπροϊόντα σφαγίων
1857	abattage d'animaux	sacrificio de animales	σφαγή ζώων
6621	abrogation	derogación	κατάργηση
5075	Abruzzes	Abruzos	Αβρουζία
5339	absentéisme	absentismo	συστηματική απουσία από την 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

Wo sind diese Daten zu finden?



European Language Resource Coordination
Connecting Europe Facility

Welche Formate werden benötigt?

Dokumentierte Daten (Metadaten)

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

Wie werden Sprachresourcen erstellt?

- Beispiele von Rohdaten (HTML mit Bildern, Tabellen, etc.) die in Sprachrecources umgewandelt werden
  - Auffinden und identifizieren der Daten
  - Erwerb der Daten (z.B. mit einem Crawler)
  - Daten säubern (d.h. entfernen von Standardtexten, Menus, Bildern, HTML Tags, etc. und Formatkonvertierung)
  - Beispiel: *Boilerpipe*
  - Dokumentation der Daten
  - Alignierung der Übersetzungen auf Satzebene.
  - Berechnung der Alignierungsqualität
  - Share

Wie werden Sprachressourcen erstellt

La France en Allemagne

Ein Ausgangspunkt sind Webseiten

6 juillet 2015

Elysée : rencontre sur la Grèce avec Angela Merkel

La chancelière allemande Angela Merkel et le président de la République François Hollande auront, lundi 6 juillet au soir à l'Elysée, un entretien suivi d'un diner de travail pour évaluer les conséquences du référendum en Grèce.
Cette rencontre s'inscrit dans le cadre de la coopération permanente entre la France et l'Allemagne pour contribuer à une solution durable en (...)

Réformes en France

Paris Climat 2015 / COP21 - Pour un accord universel

Wie werden Sprachressourcen erstellt

Frankreich in Deutschland

6. Juli 2015

Griechenland: Staatspräsident Hollande empfängt Bundeskanzlerin (...)

Staatspräsident François Hollande empfängt am Abend des 6. Juli Bundeskanzlerin Angela Merkel im Elysée-Palast zu einem Gespräch und einem Arbeitsessen, um die Konsequenzen aus dem Referendum in Griechenland zu erörtern.
Das Treffen findet im Rahmen der ständigen Zusammenarbeit zwischen Frankreich und Deutschland mit dem Ziel statt, zu einer dauerhaften Lösung für Griechenland zu (...)

Reformagenda

Klimakonferenz Paris 2015

Wie identifizieren wir Vergleichsmerkmale?

Griechenland: Staatspräsident Hollande empfängt Bundeskanzlerin Merkel [fr]

Drucken

Google

Facebook

Twitter

Staatspräsident François Hollande empfängt am Abend des 6. Juli Bundeskanzlerin Angela Merkel im Elysée-Palast zu einem Gespräch und einem Arbeitsessen, um die Konsequenzen aus dem Referendum in Griechenland zu erörtern.

Das Treffen findet im Rahmen der ständigen Zusammenarbeit zwischen Frankreich und Deutschland mit dem Ziel statt, zu einer dauerhaften Lösung für Griechenland zu kommen..

Letzte Änderung 06/07/2015

Seitenanfang

- **Elysée : rencontre sur la Grèce avec Angela Merkel [de]**
- Imprimer
- Google
- Facebook
- Twitter
- La chancelière allemande Angela Merkel et le président de la République François Hollande auront, lundi 6 juillet au soir à l'Elysée, un entretien suivi d'un diner de travail pour évaluer les conséquences du référendum en Grèce.
- Cette rencontre s'inscrit dans le cadre de la coopération permanente entre la France et l'Allemagne pour contribuer à une solution durable en Grèce.
- Dernière modification : 06/07/2015
- Haut de page

➢ **RAW DATA TO PROCESS**

Wie identifizieren wir Vergleichsmerkmale?

Griechenland: Staatspräsident Hollande empfängt Bundeskanzlerin Merkel

François Hollande empfängt am Abend des 6. Juli Bundeskanzlerin Angela Merkel im Elysée-Palast zu einem Gespräch und einem Arbeitsessen, um die Konsequenzen aus dem Referendum in Griechenland zu erörtern.

Das Treffen findet im Rahmen der ständigen Zusammenarbeit zwischen Frankreich und Deutschland mit dem Ziel statt, zu einer dauerhaften Lösung für Griechenland zu kommen..

Letzte Änderung 06/07/2015

Seitenanfang

- **Elysée : rencontre sur la Grèce avec Angela Merkel**
-
- La chancelière allemande Angela Merkel et le président de la République François Hollande auront, lundi 6 juillet au soir à l'Elysée, un entretien suivi d'un diner de travail pour évaluer les conséquences du référendum en Grèce.
- Cette rencontre s'inscrit dans le cadre de la coopération permanente entre la France et l'Allemagne pour contribuer à une solution durable en Grèce.
- Dernière modification : 06/07/2015
- Haut de page

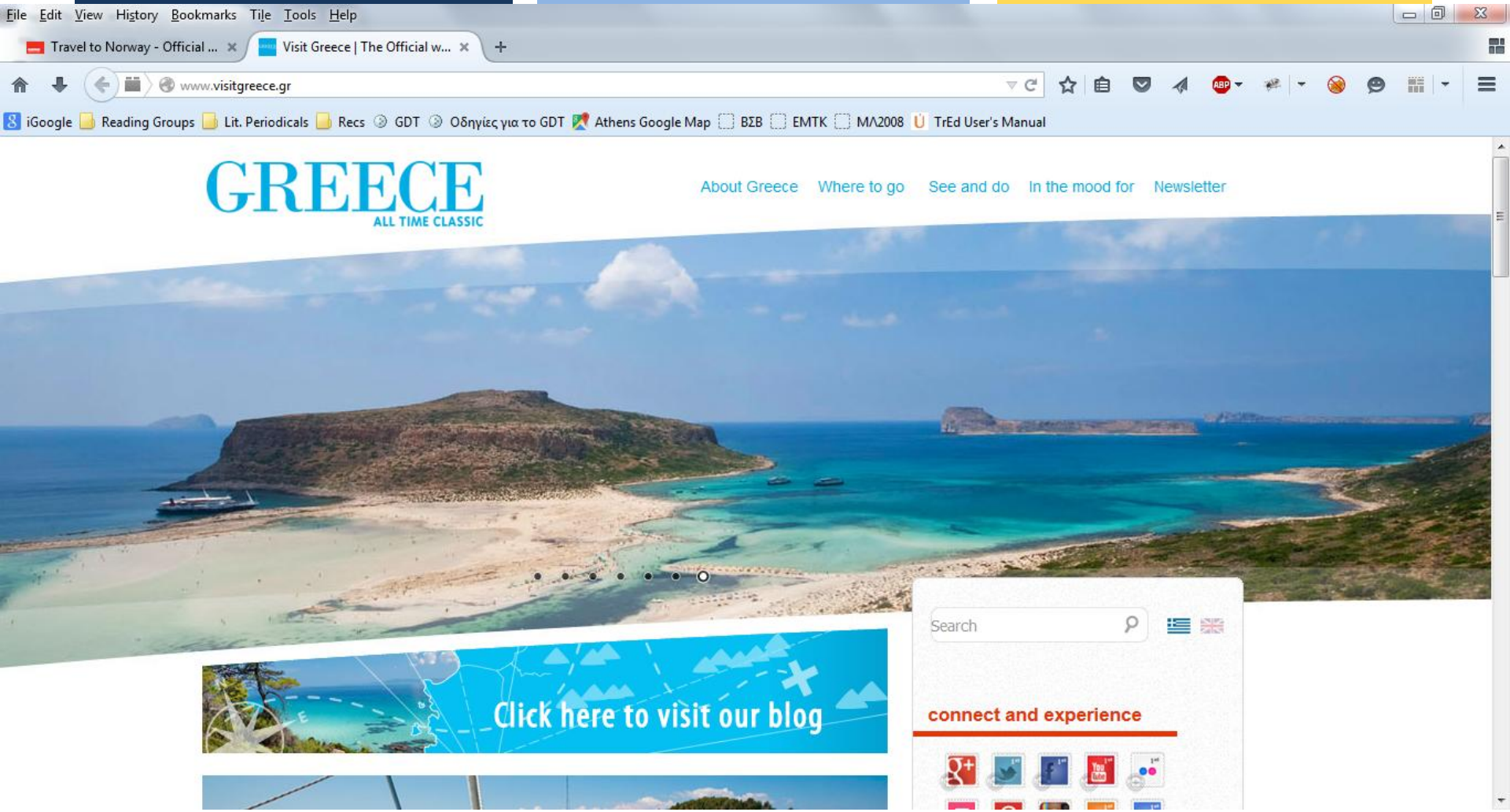
➢ **RAW DATA TO PROCESS**

Eine Sprachressourcen Pipeline

- Wie kann dieser Prozess automatisiert werden?
- Wir fangen an mit digital verfügbaren Daten
  - OCR kann für Sprachen eingesetzt die weniger in digitaler Form vertreten sind

European Language Resource Coordination
Connecting Europe Facility

Viele Seiten bieten Inhalt in mehreren Sprachen an



European Language Resource Coordination
Connecting Europe Facility

Wie kommen wir an diesen Inhalt?

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

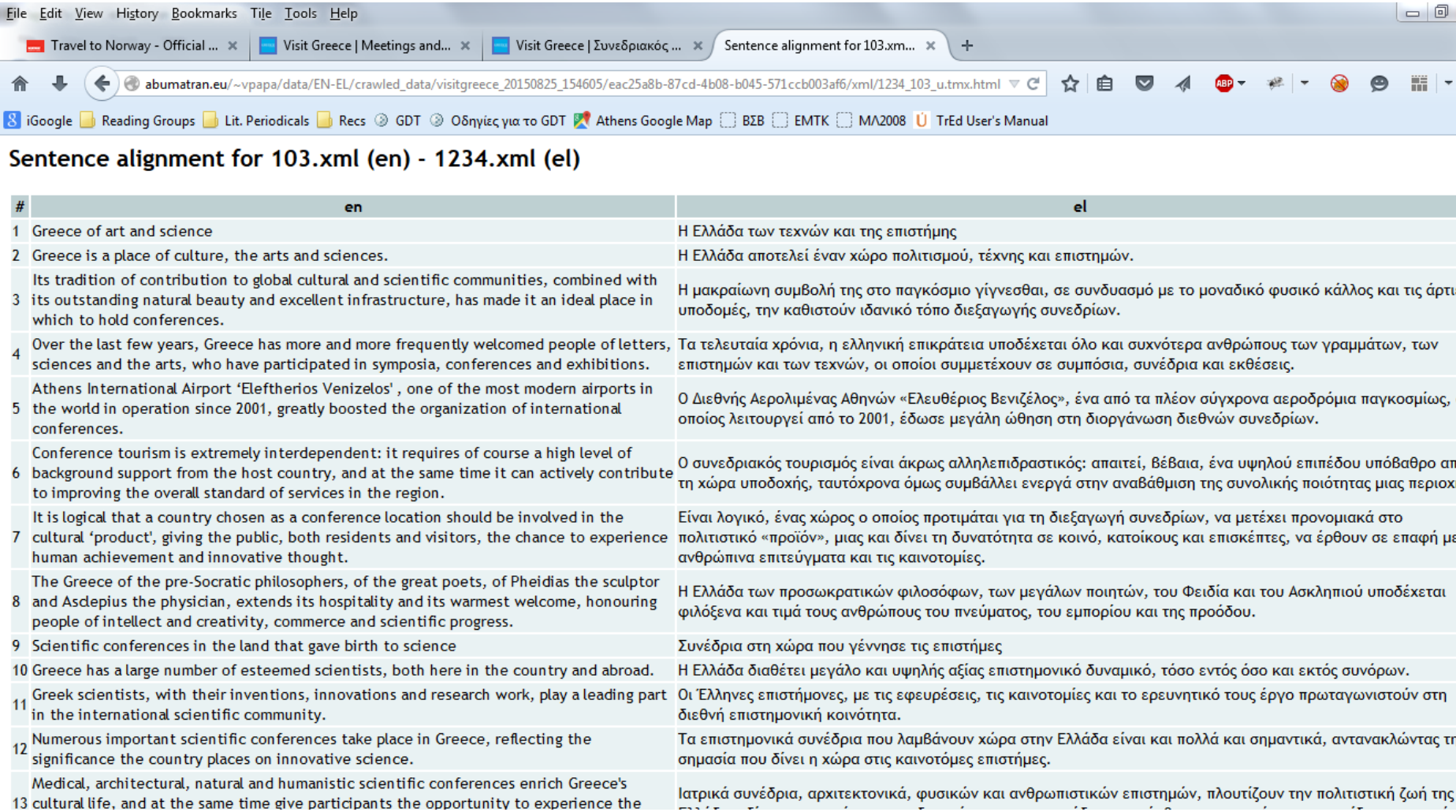
Η Ελλάδα αποτελεί έναν χώρο πολιτισμού, τέχνης και επιστημών. Η μακραίωνη συμβολή της στο παγκόσμιο γίγνεσθαι, σε συνδυασμό με το μοναδικό φυσικό κάλλος και τις **άρτιες υποδομές**, την καθιστούν ιδανικό τόπο διεξαγωγής συνεδρίων. Τα τελευταία χρόνια, η ελληνική επικράτεια υποδέχεται όλο και συχνότερα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

Und konvertieren in Daten für die SMÜ?



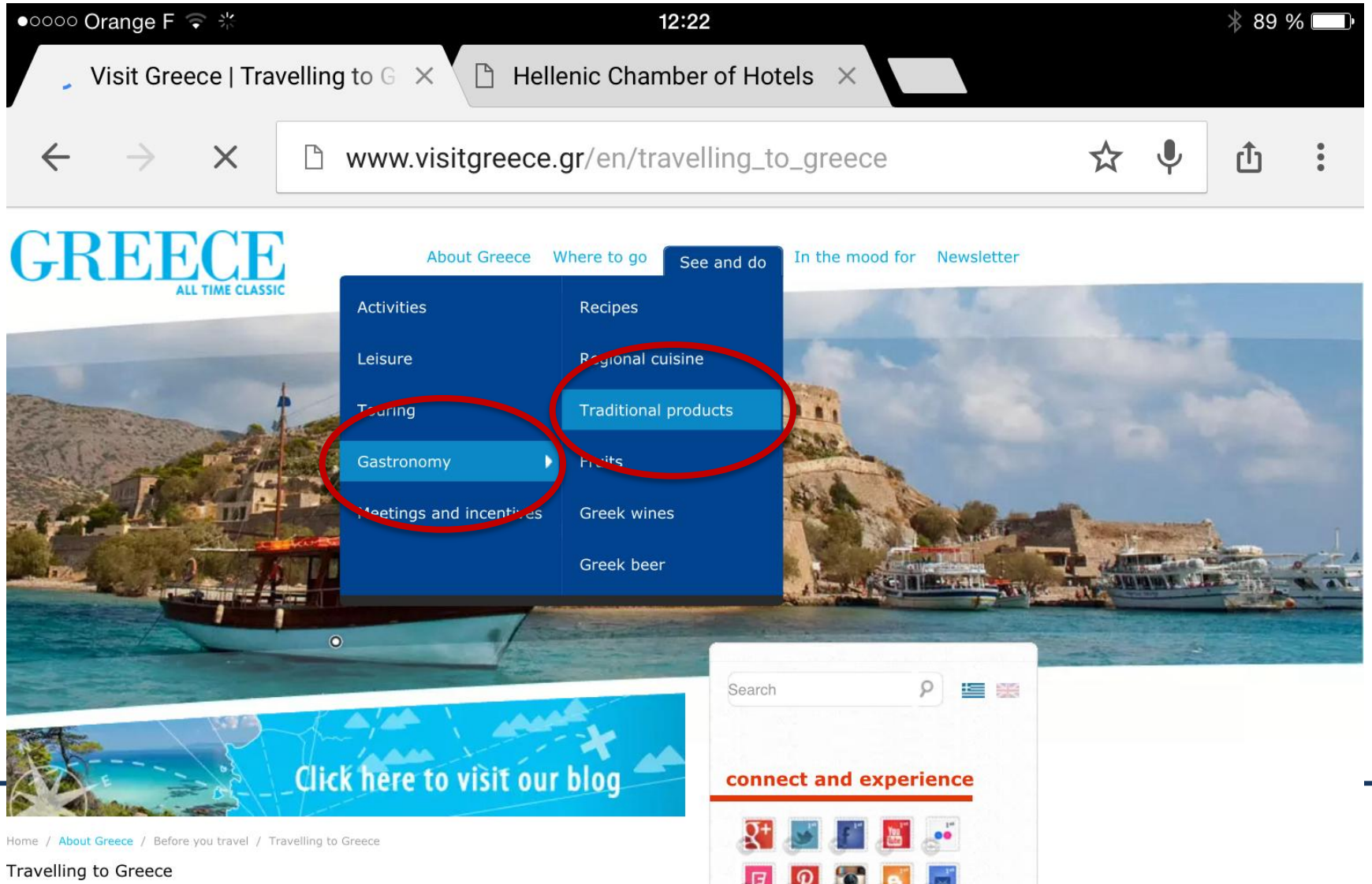
Sentence alignment for 103.xml (en) - 1234.xml (el)

#	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.
5	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.
6	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.
7	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.
8	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.
9	Scientific conferences in the land that gave birth to science	Συνέδρια στη χώρα που γέννησε τις επιστήμες
10	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.
11	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.
12	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.
13	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

Von der Webseite in die Pipeline

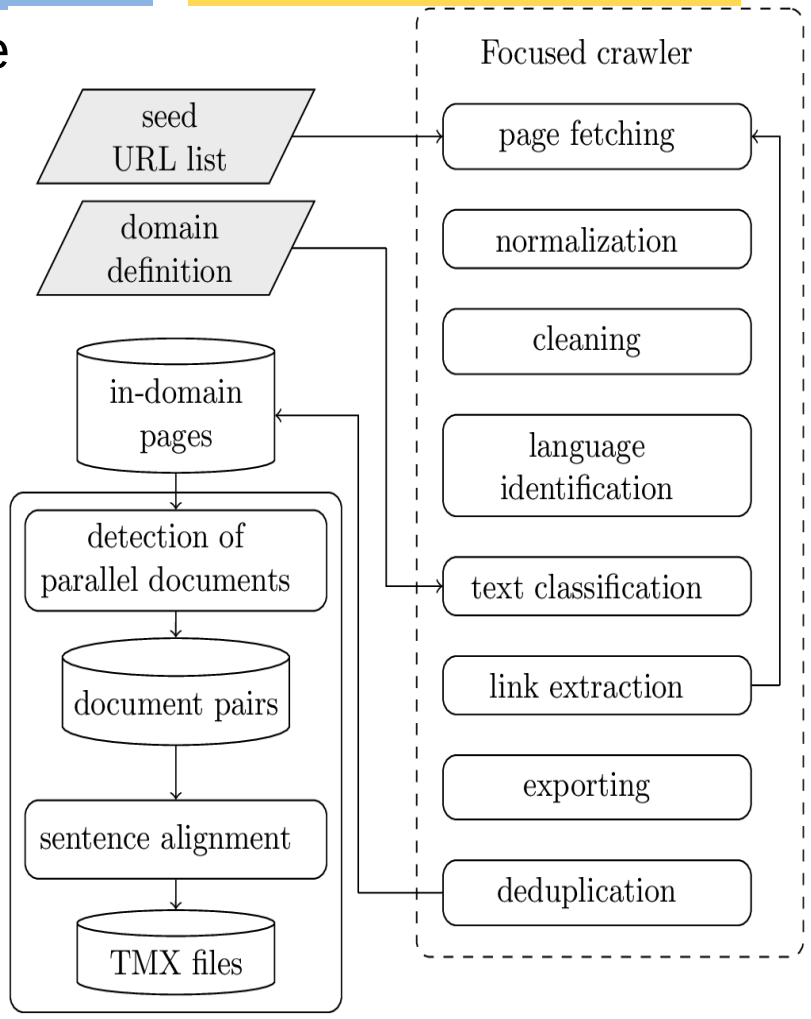
- Wir gehen vor indem wir:
  - Eine Quelle als nützlich identifizieren (gute multilinguale Daten)
  - **Alle weiteren Links abarbeiten**
    - **Analog zum anklicken eines jeden Links (Crawling)**
- Erhalt der Seiten mit der zugehörigen übersetzten Seiten
- Identifikation von Domain und Genre, wenn möglich
- Säuberung der Daten
- Alignierung (auf Dokument-, Abschnitt-, Satzebene)
- Überprüfung der Alignierung
- Use & Share

Automatisches Verfolgen aller referenzierter Links



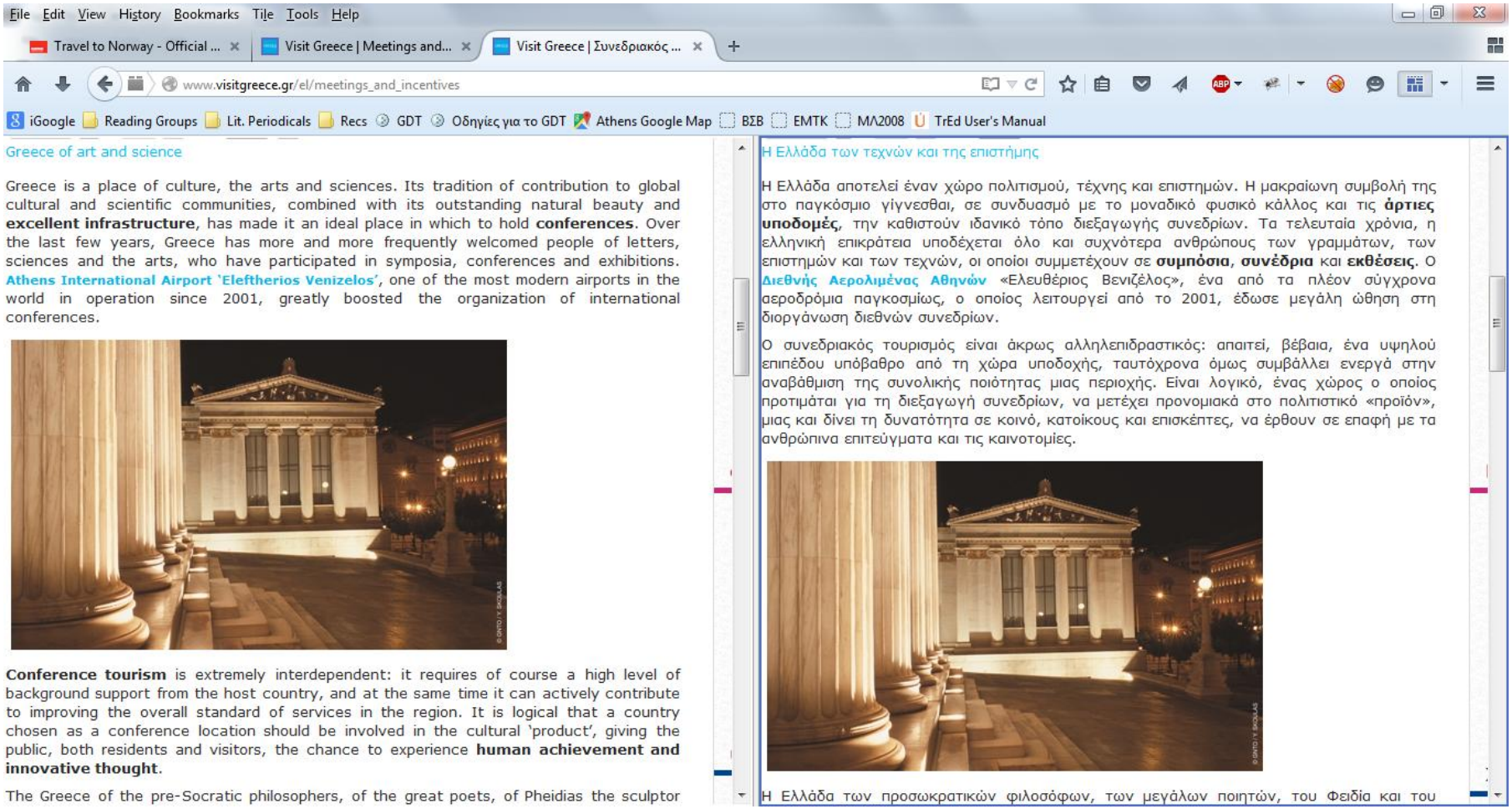
ILSP Focused Crawler

- Prototyp zum Erwerb von generellen, ode Domain-spezifischen ein- oder mehrsprachigen Korpora
- **Eingabe**:
  - Definition von Domänen (Termini)
  - Seed URL
- Module(open source libraries/toolkits):
  - Page Fetching/Text Extraction
  - Normalization and Metadata Extraction
  - Boilerplate Detection (Boilerpipe)
  - Spracherkennung (> 50 Sprachen)
  - Textklassifikation
  - Entfernen von Duplikaten
  - Erkennen der Dokumentenpaare
  - Satzalignierung (Hunalign and others)
- **Ausgabe**
  - Dokumentenpaare
  - Segmentpaare in TMX Format



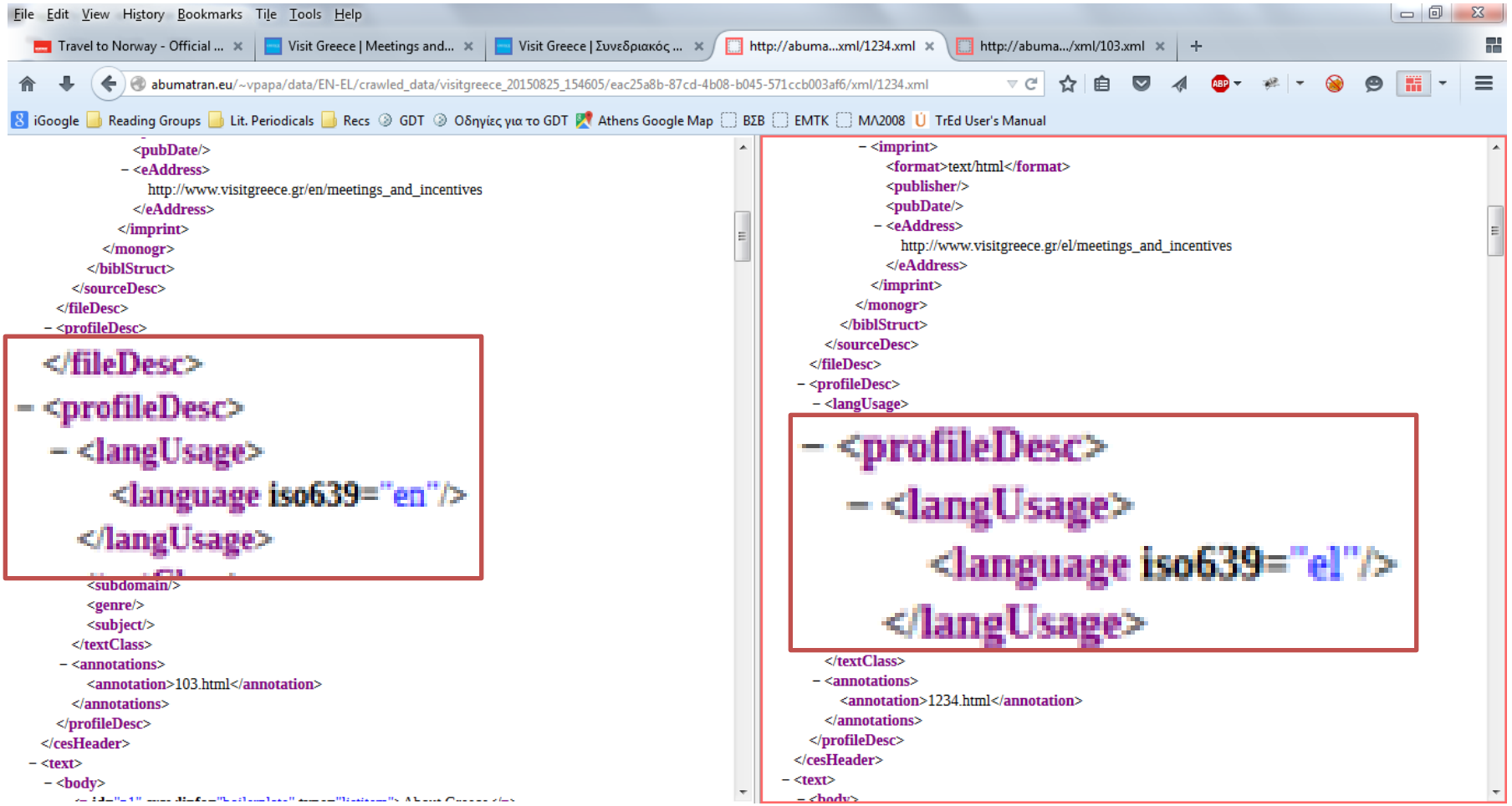
European Language Resource Coordination
Connecting Europe Facility

Verfolgt Links auf der Webseite…



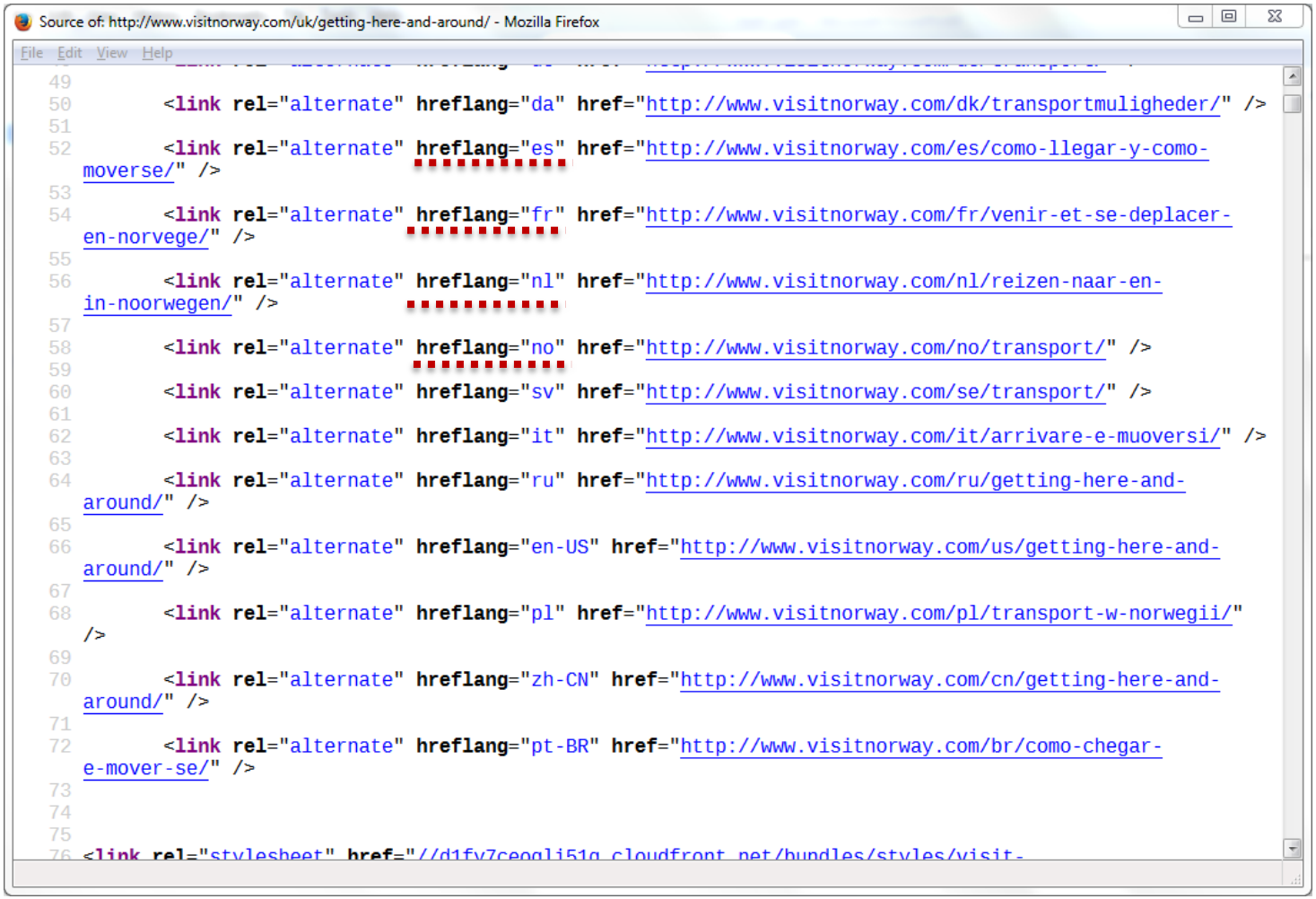
European Language Resource Coordination
Connecting Europe Facility

…identifiziert die Sprache…



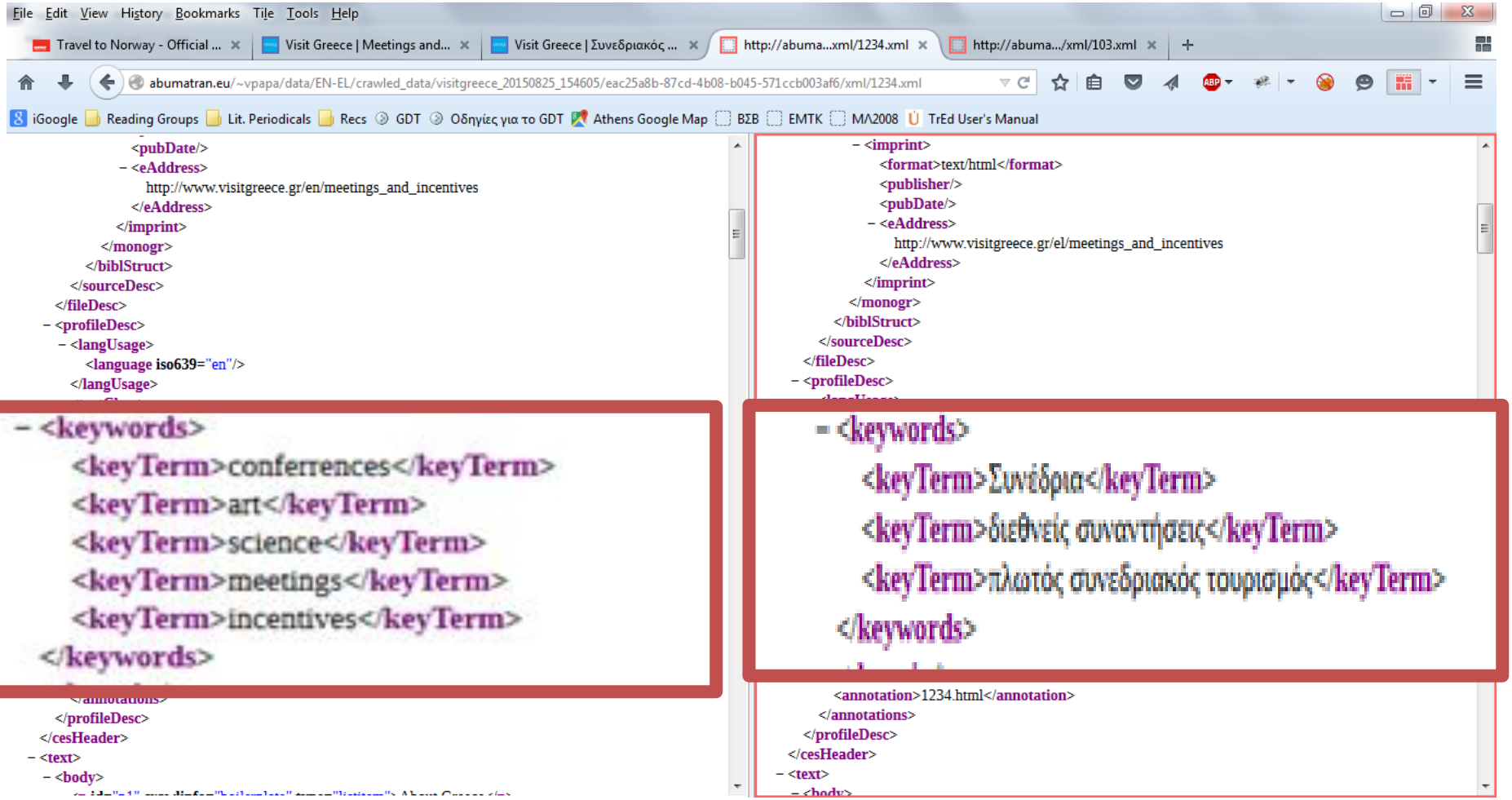
European Language Resource Coordination
Connecting Europe Facility

… auf jeder einzelnen Seite…



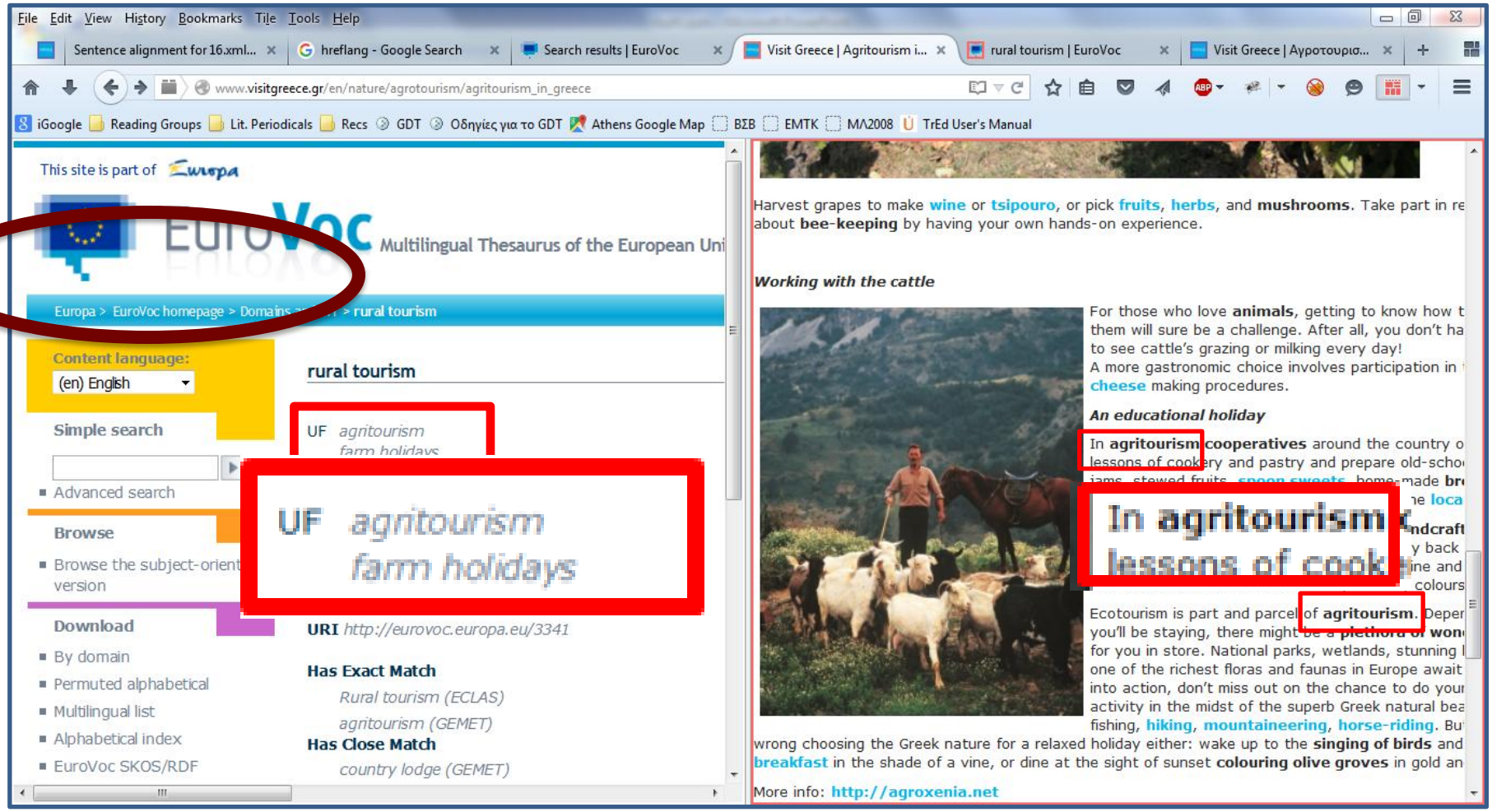
European Language Resource Coordination
Connecting Europe Facility

… extrahiert Typen von Datenbeschreibung (Metadaten)…



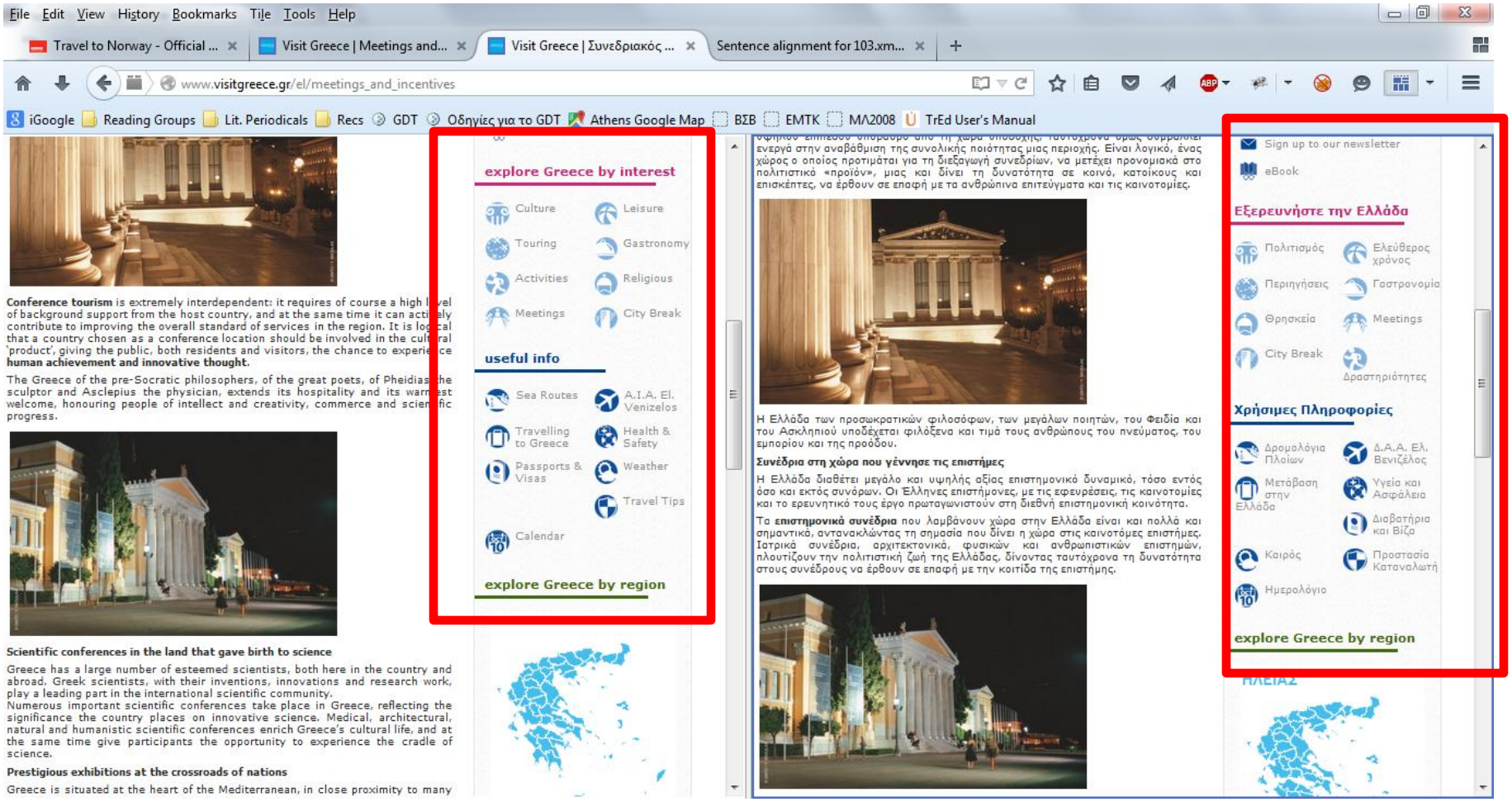
European Language Resource Coordination
Connecting Europe Facility

… und klassifiziert optionell auf Relevanz, gegeben der Vorgaben des Benutzers…



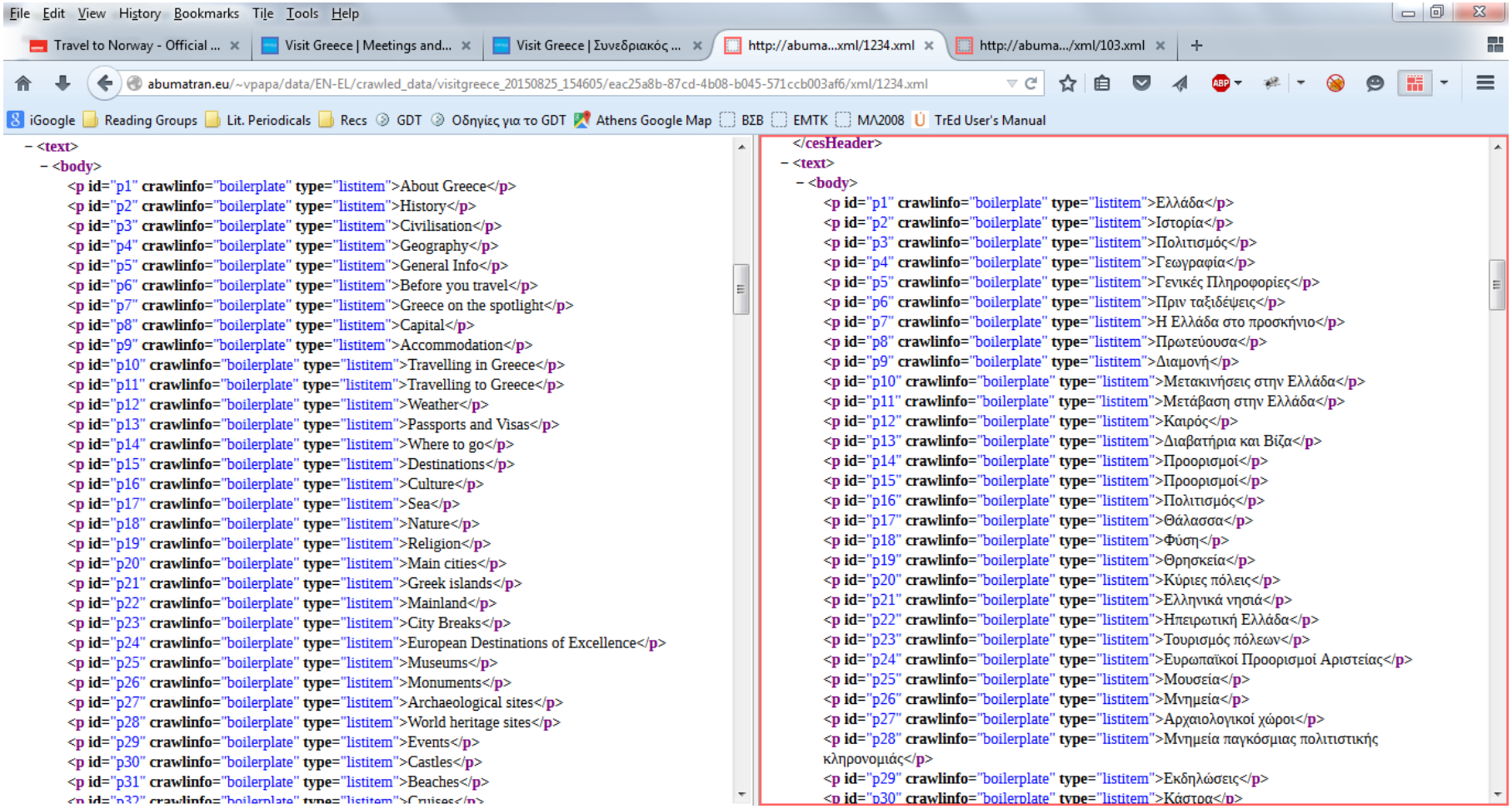
European Language Resource Coordination
Connecting Europe Facility

...es erkennt s.g. *Boilerplate* Text...

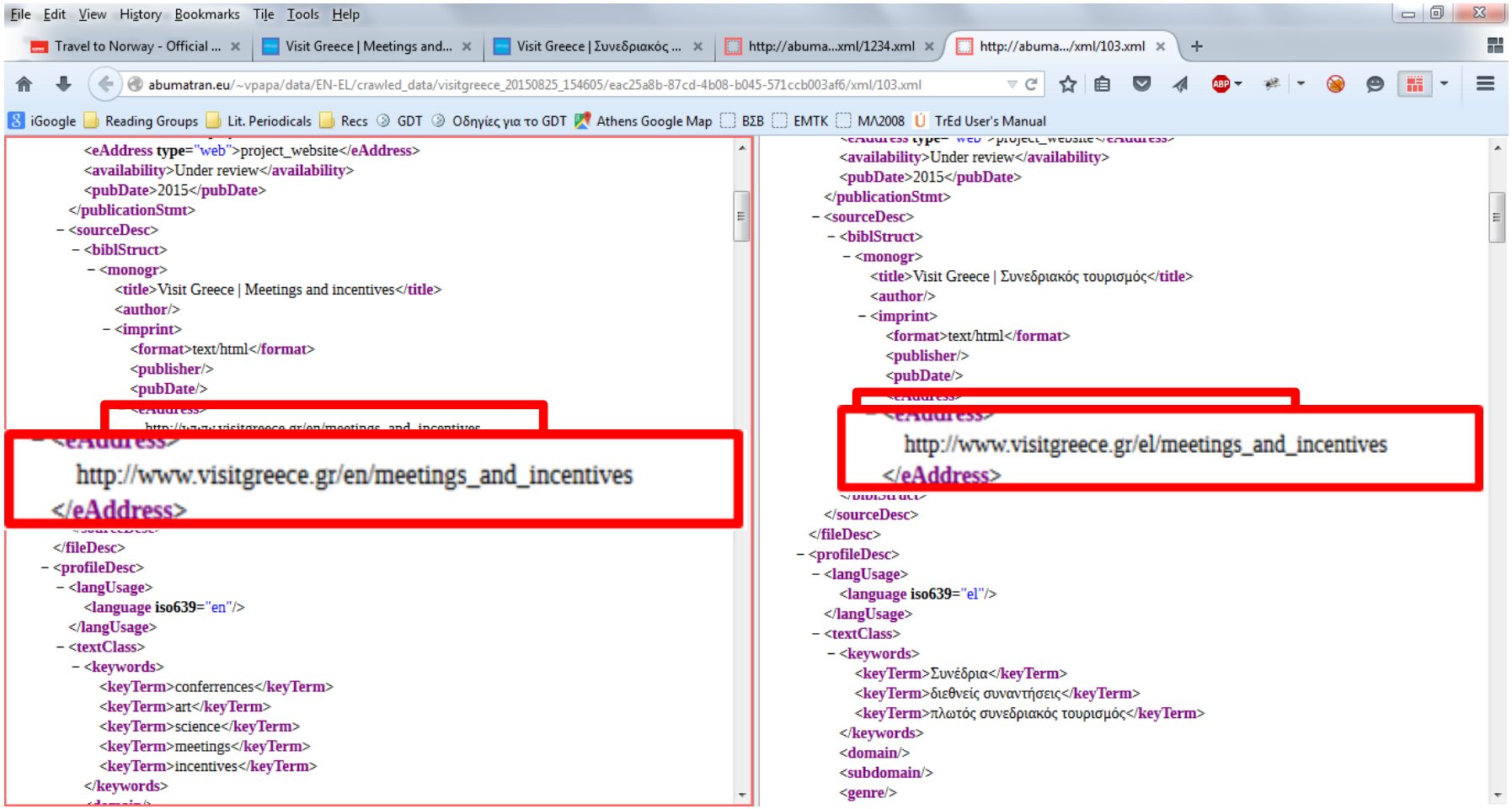


European Language Resource Coordination
Connecting Europe Facility

… und ermöglich den Fachleuten schnelles Ausfiltern…



… nutz HTML Struktur und URL Ähnlichkeiten zur Erkennung von Dokumentenpaaren…



European Language Resource Coordination
Connecting Europe Facility

…und aligniert auf Satzebene.

#	en	de
1	Reliable partnership based on trust	Vertrauensvolle und verlässliche Partnerschaft
2	"The stage has been set for further intensive cooperation," declared Chancellor Angela Merkel after her meeting with Colombian President Juan Manuel Santos.	"Die Weichen für eine weitere intensive Zusammenarbeit sind gestellt." Das hat Bundeskanzlerin Merkel nach dem Treffen mit dem kolumbianischen Präsidenten Santos erklärt.
3	The Chancellor intends to support the peace process in Colombia, partly through cooperation in the fields of research, education and climate change mitigation.	Den Friedensprozess in Kolumbien will Merkel unter anderem durch weitere Kooperationen in den Bereichen Forschung, Bildung und Klimaschutz unterstützen.
4	Colombia's President seeks support in the peace process with FARC rebels Photo:	Kolumbiens Präsident wirbt für Unterstützung beim Friedensprozess mit den FARC-Rebellen. Foto:
5	Bundesregierung/Denzel	Bundesregierung/Denzel
6	"We have a cordial and reliable partnership based on trust," declared Chancellor Angela Merkel after her meeting with Colombian President Juan Manuel Santos.	"Wir sind in einer vertrauensvollen, freundschaftlichen und verlässlichen Partnerschaft", erklärte Bundeskanzlerin Angela Merkel nach dem Treffen mit dem kolumbianischen Präsidenten Juan Manuel Santos.
7	Their talks focused on Colombia's peace process, bilateral relations and economic and regional issues.	Im Mittelpunkt des Gesprächs standen der kolumbianische Friedensprozess, die bilateralen Beziehungen sowie wirtschaftliche und regionalpolitische Themen.
8	Fostering the peace process	Friedensprozess fördern
9	"The situation today in Colombia is marked by the courageous peace process initiated by the President, which is currently in a crucial phase," said the Chancellor following talks with Colombia's President Juan Manuel Santos.	"Die aktuelle Situation in Kolumbien ist dadurch gekennzeichnet, dass der Präsident einen mutigen Friedensprozess initiiert hat, der im Augenblick in einer entscheidenden Phase ist", erläuterte die Kanzlerin nach dem Treffen mit dem kolumbischen Präsidenten Juan Manuel Santos.
10	She reported that she had pledged Germany's full support in this process.	Hierfür habe sie die volle deutsche Unterstützung zugesagt, sagte Merkel.
11	In Columbia a conflict has been smouldering for decades between right-wing paramilitary groups, left-wing guerrillas and the Colombian army.	In Kolumbien schwelt seit Jahrzehnten ein Konflikt zwischen rechtsgerichteten Paramilitärs, linksgerichteten Guerillatruppen und der kolumbianischen Armee.

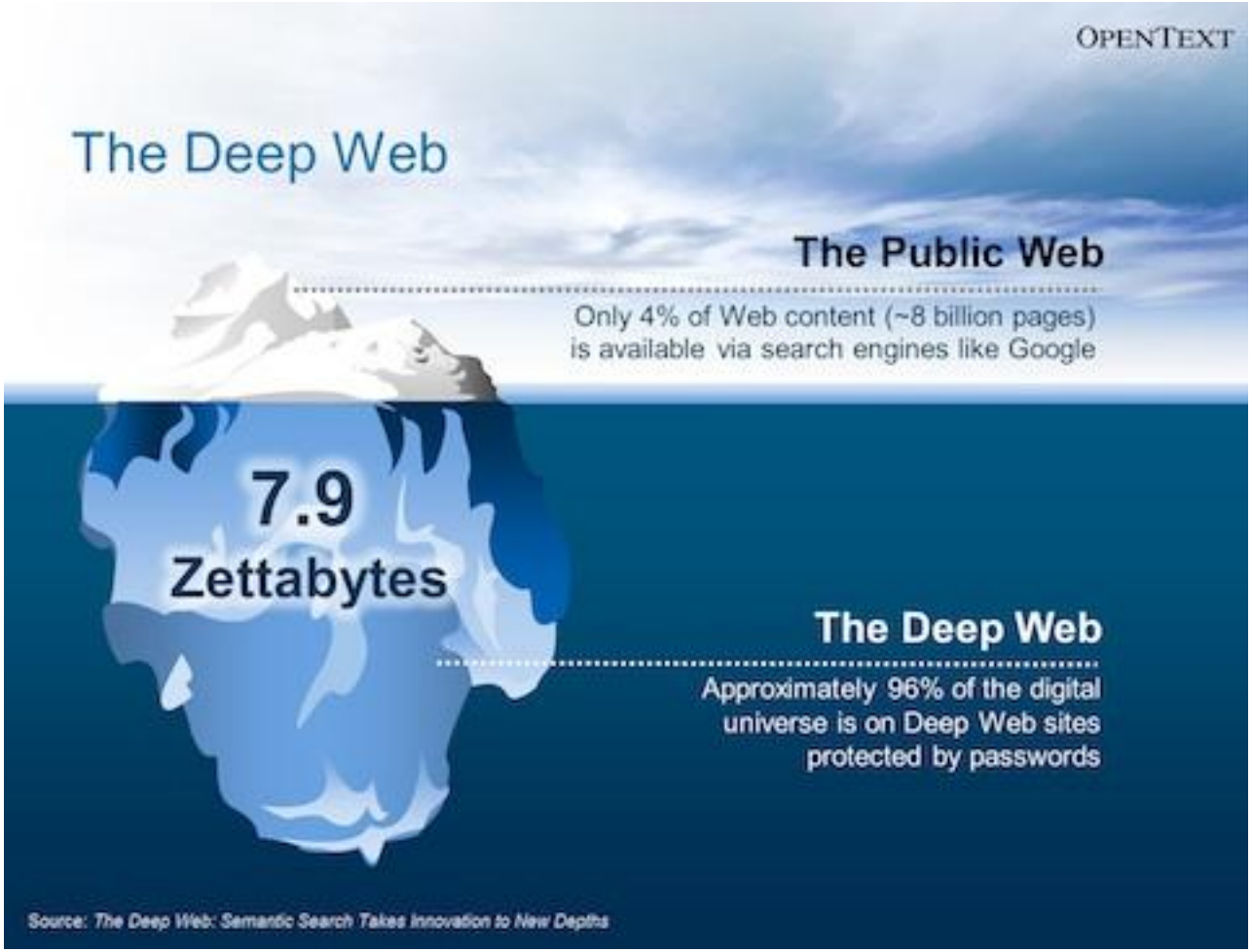
European Language Resource Coordination
Connecting Europe Facility

Weitere Daten...

- Wir können den nach außen sichtbaren Teil der Daten finden, aber es gibt noch viel mehr in Ihren Organisationen
- Helfen Sie uns diese Daten zu bekommen und zu nutzen!
- Dieser Prozess kann mit Ihrer Hilfe zu einer automatisierten Pipeline zur Herstellung von parallelen Sprachresourcen werden (sammeln Sie Ihre Dokumente, Berichte, Dateien, etc.)

European Language Resource Coordination
Connecting Europe Facility

Verhältnis von sichtbaren zu existierenden Daten



European Language Resource Coordination
Connecting Europe Facility

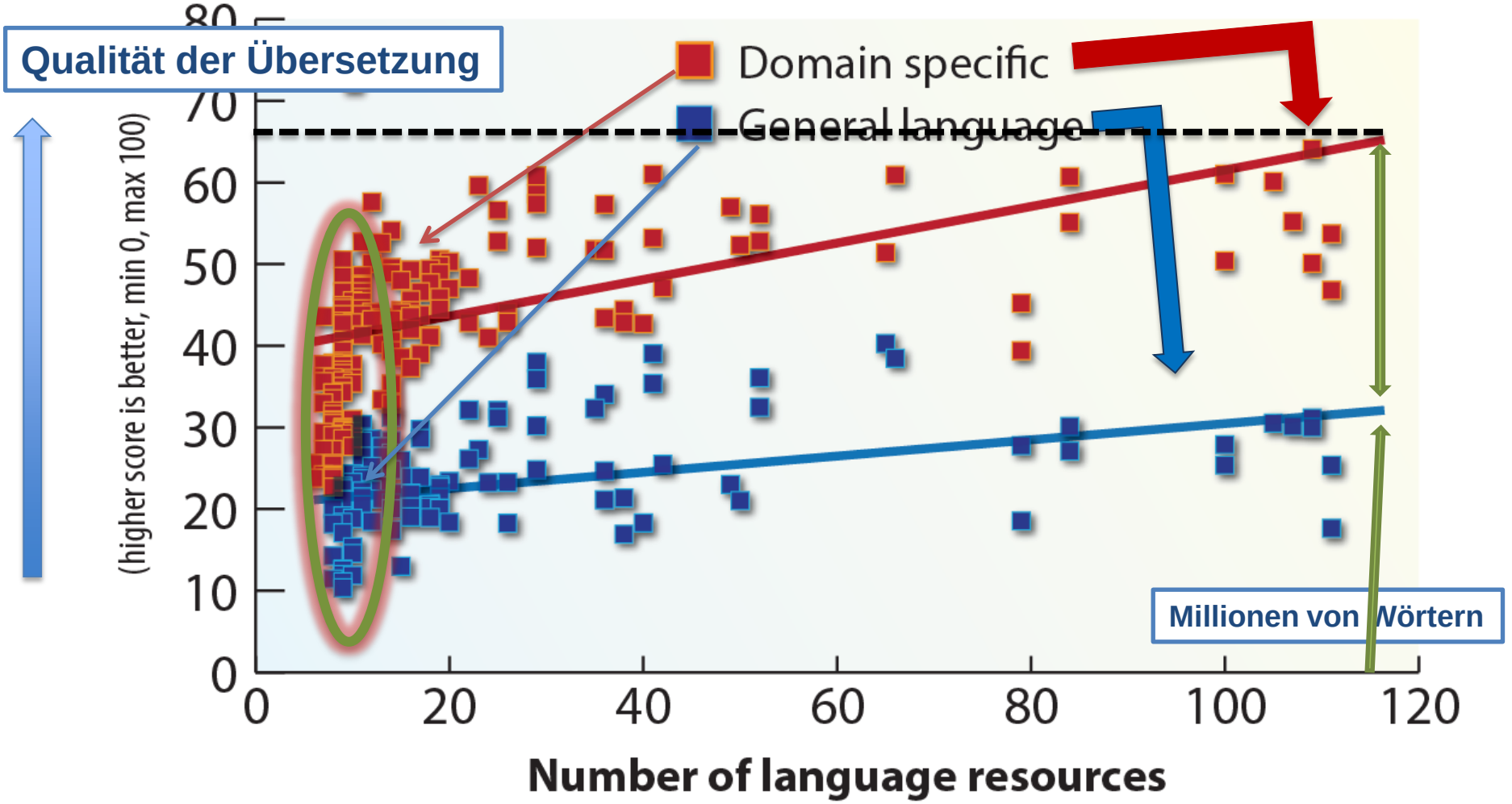
Ihr Beitrag – Deep Web

Weitere Daten

- Solche Dokumente gibt es in:
  - Dokumentationszentren (übersetzte Berichte, Broschüren, Vorträge, (interne) Webseiten, etc.)
  - Bei den Language Service Providers (LSP), welche Übersetzungsaufträge erhalten
- Helfen Sie uns diese Quellen zu identifizieren und mit diesen zu kooperieren

European Language Resource Coordination
Connecting Europe Facility

Impact of number of language resources on Statistical MT quality



Schlussfolgerung

- Verwertung von bereits vorhandenen Daten
- Effiziente Identifikation von Quellen ist
- Der Wert der vorhandenen Resourcen darf nicht unterschätzt werden
- Der Startpunkt der Suche ist oft das Schwierigste

Zusammenarbeit in der

Auffindung der Quellen ist essentiell