European Language Resource Coordination
Connecting Europe Facility

What Data Is Needed? Why?

Dr. Khalid Choukri
(Evaluations and Language resource Distribution Agency)

What types of data? Translation

- Predominant approach of ***data-driven paradigm***
  - MT systems learn from existing data
  - Focus for ELRC: Data in all languages (EU/CEF)

- Language Resources are produced from:
  - Documents & data
  - Important that you help us with the data you have or you know about

What counts as Data for MT?

- Anything that contains “words”, preferences for “sentences”, even for sentences expressed in multiple languages, e.g.
  - Reports,
  - Speeches,
  - Contents on web pages,
  - Brochures, etc.

- Bags of “words”, “sentences”, multiple bags

What counts as data for MT?

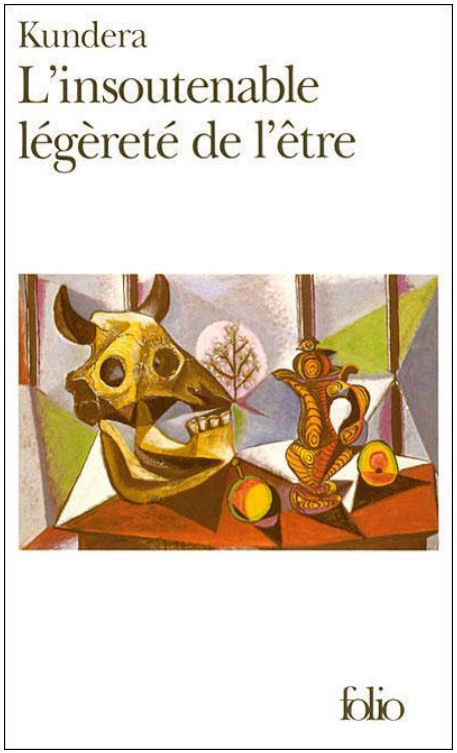
European Language Resource Coordination
Connecting Europe Facility

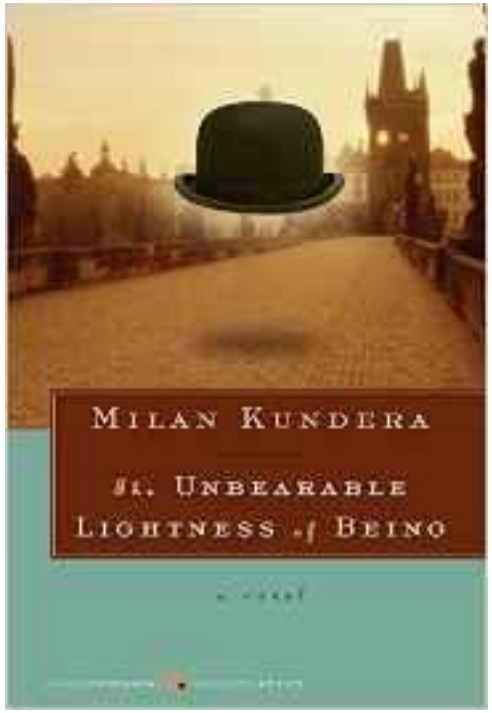
What types of data?
“Aligned” Translation

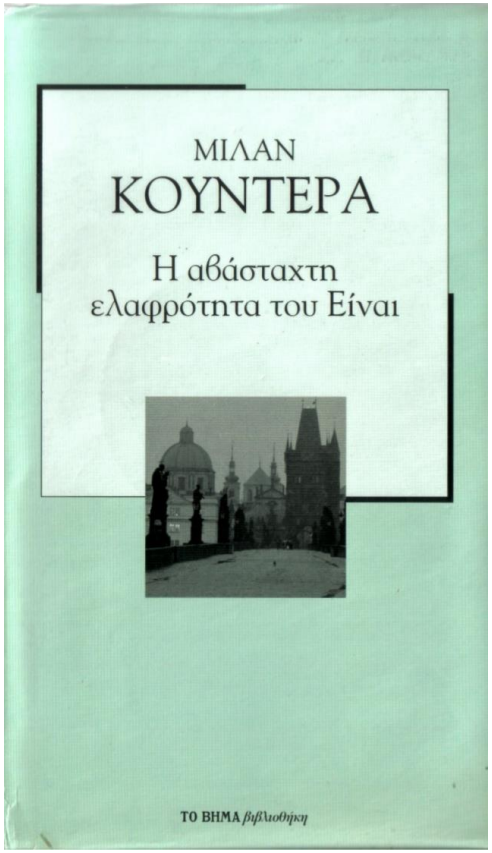
English

French

What types of data? Translations







Comparable Collections
2 …3 … Languages

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

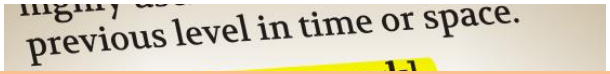
German page is slightly different but these are (never) translations of one source!!

European Language Resource Coordination
Connecting Europe Facility

What types of data?
"Aligned" Translation

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.	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.
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."	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.
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.	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.

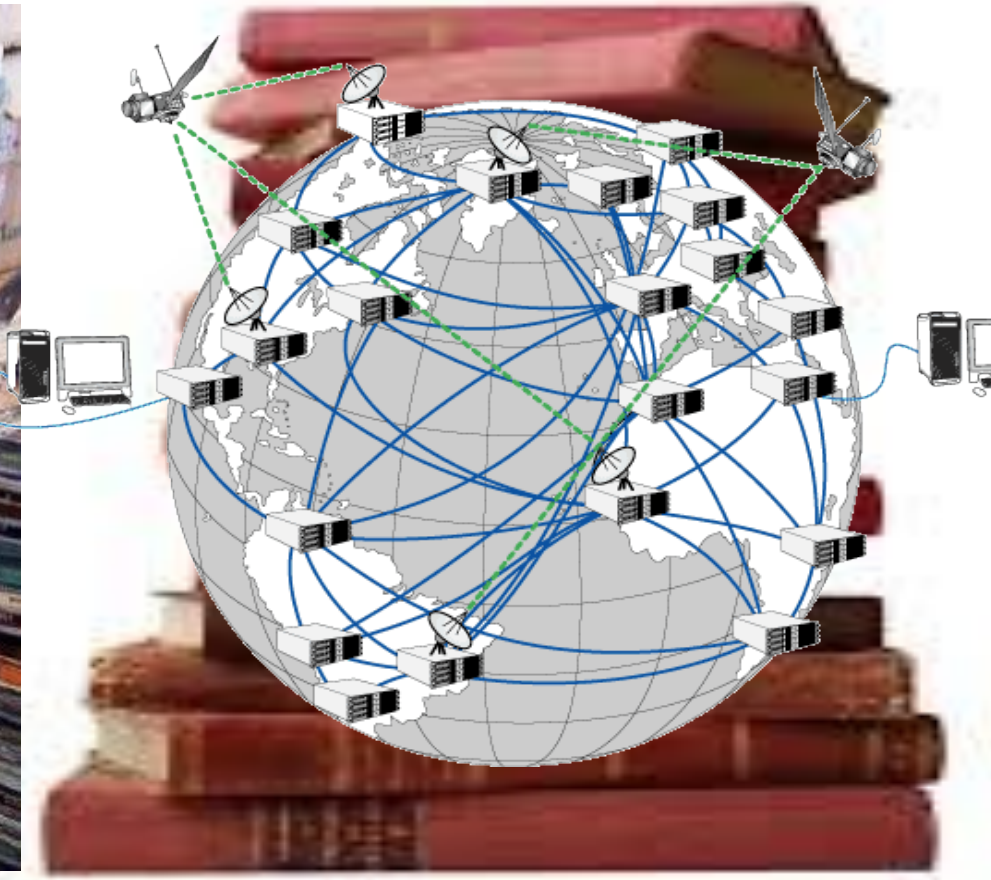
Dictionaries / Terminologies /Ontologies



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

What types of data?
“Aligned” Translation

English



French

European Language Resource Coordination
Connecting Europe Facility

What format is needed? Digital textual data

Documented Data (Meta-data)

- On a case by case we can define what is ESSENTIAL:
  - Sources of data (trustability, quality, etc.)
  - Domain specific
  - Languages
  - Rights if not public

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

How Language Resources (LR) are produced from data

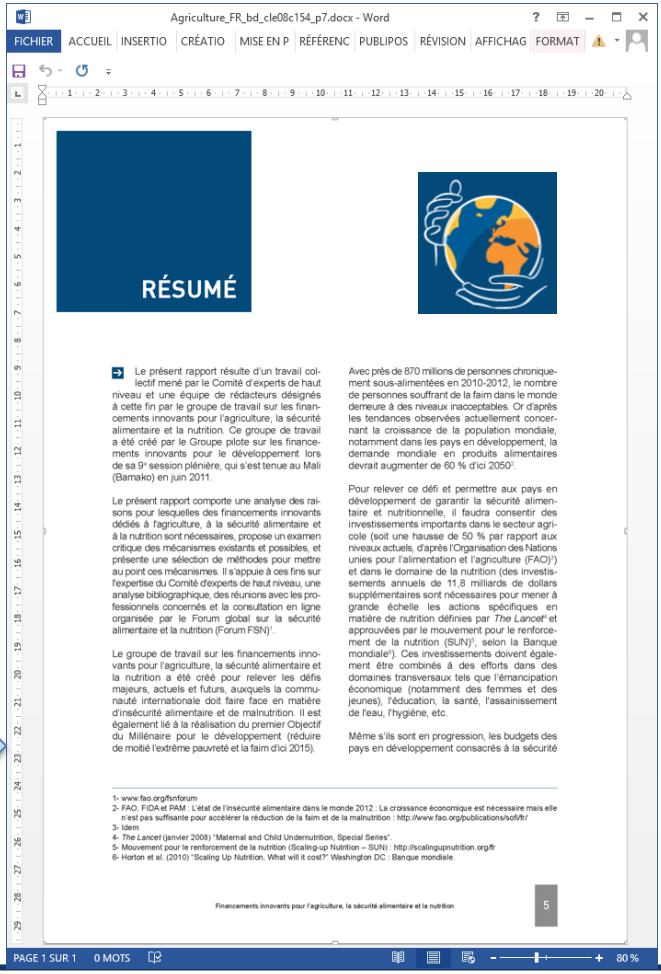
- Let us see some examples of raw data (html with tables, pictures, etc.) and how they become LRs
  - Discover & identify sources
  - Clear IPR and Get the data (Receive, Download, harvest, crawl, …)
  - Clean the data (e.g. detect and remove the “boilerplate”, “templates”, pictures, html tags, etc.,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

Management of Bilingual Data Example (1/4)

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

English version

French version

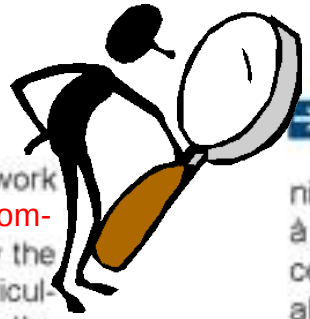


Management of Bilingual Data Example (2/4)

EXECUTIVE SUMMARY

This report is the result of a collective work carried out by the high-level expert Committee and a writing team commissioned by the Task Force on Innovative Financing for agriculture, food security and nutrition created by the Leading Group on Innovative Financing for Development at its 9th plenary session in Mali (Bamako) in June 2011.

The report includes an analysis of the need for innovating financing dedicated to the agricultural, food security and nutrition sector, a critical review of existing and possible mechanisms and a proposed selection of avenues for the development of such mechanisms on the basis of the



RÉSUMÉ

Le présent rapport résulte d'un travail collectif mené par le Comité d'experts de haut niveau et une équipe de rédacteurs désignés à cette fin par le groupe de travail sur les financements innovants pour l'agriculture, la sécurité alimentaire et la nutrition. Ce groupe de travail a été créé par le Groupe pilote sur les financements innovants pour le développement lors de sa 9e session plénière, qui s'est tenue au Mali (Bamako) en juin 2011.

Le présent rapport comporte une analyse des raisons pour lesquelles des financements innovants dédiés à l'agriculture, à la sécurité alimentaire et à la nutrition sont nécessaires, propose un examen critique des mécanismes existants et possibles, et

Management of Bilingual Data Example (3/4)

English version – Raw text

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

French version – Raw text

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

Management of Bilingual Data Example (4/4)

Alignement of English and French versions

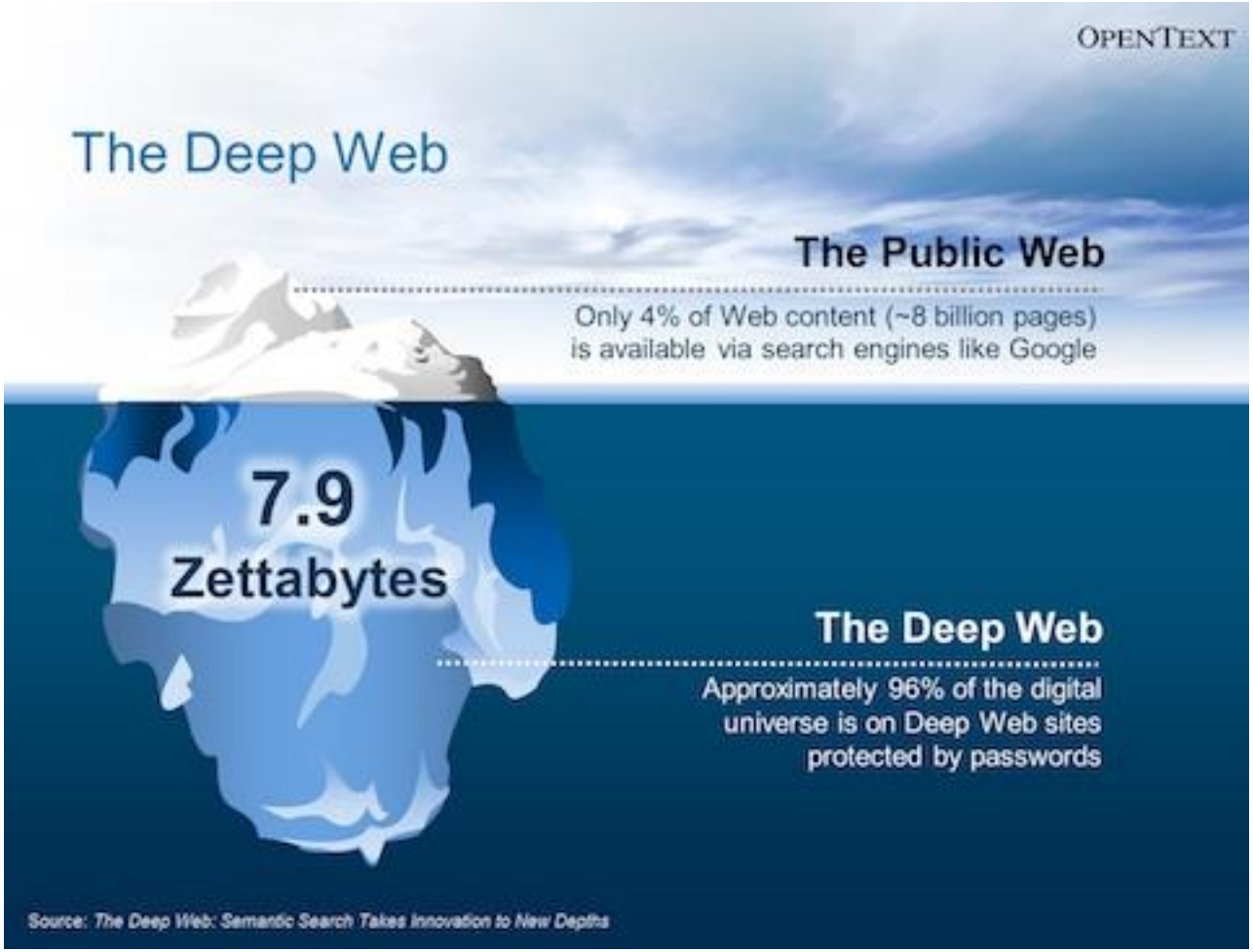
S1. Executive Summary	**S1.** Résumé
S2. This report is the result of a collective work carried out by the high-level expert Committee and a writing team commissioned by the Task Force on Innovative Financing for agriculture, food security and nutrition created by the Leading Group on Innovative Financing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

European Language Resource Coordination
Connecting Europe Facility

A better alternative

- What we can get is only the “visible” part, there are many more in your organizations
- Help us identify sources of data
- This process can be turned into a factory of LR production (Automation of the Procedure) with your support (Collect all your documents, reports, files, etc.)

Visible data versus existing ones



Our contributions … Deep web

European Language Resource Coordination
Connecting Europe Facility

A better alternative

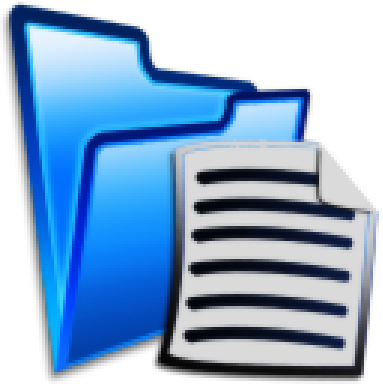
- Such documents exist already:
  - At the various documentation centers (translated reports, leaflets, brochures, speeches, web pages, etc.)
  - At the Language Service Providers (LSP), to whom translation works are subcontracted
- Help us identify and liaise with both sources
  - (see next Panel interactions)

Your involvement is essential so please let us work together

RESOURCES

BRING YOUR OWN LANGUAGE RESOURCES

Your involvement is essential so please let us work together



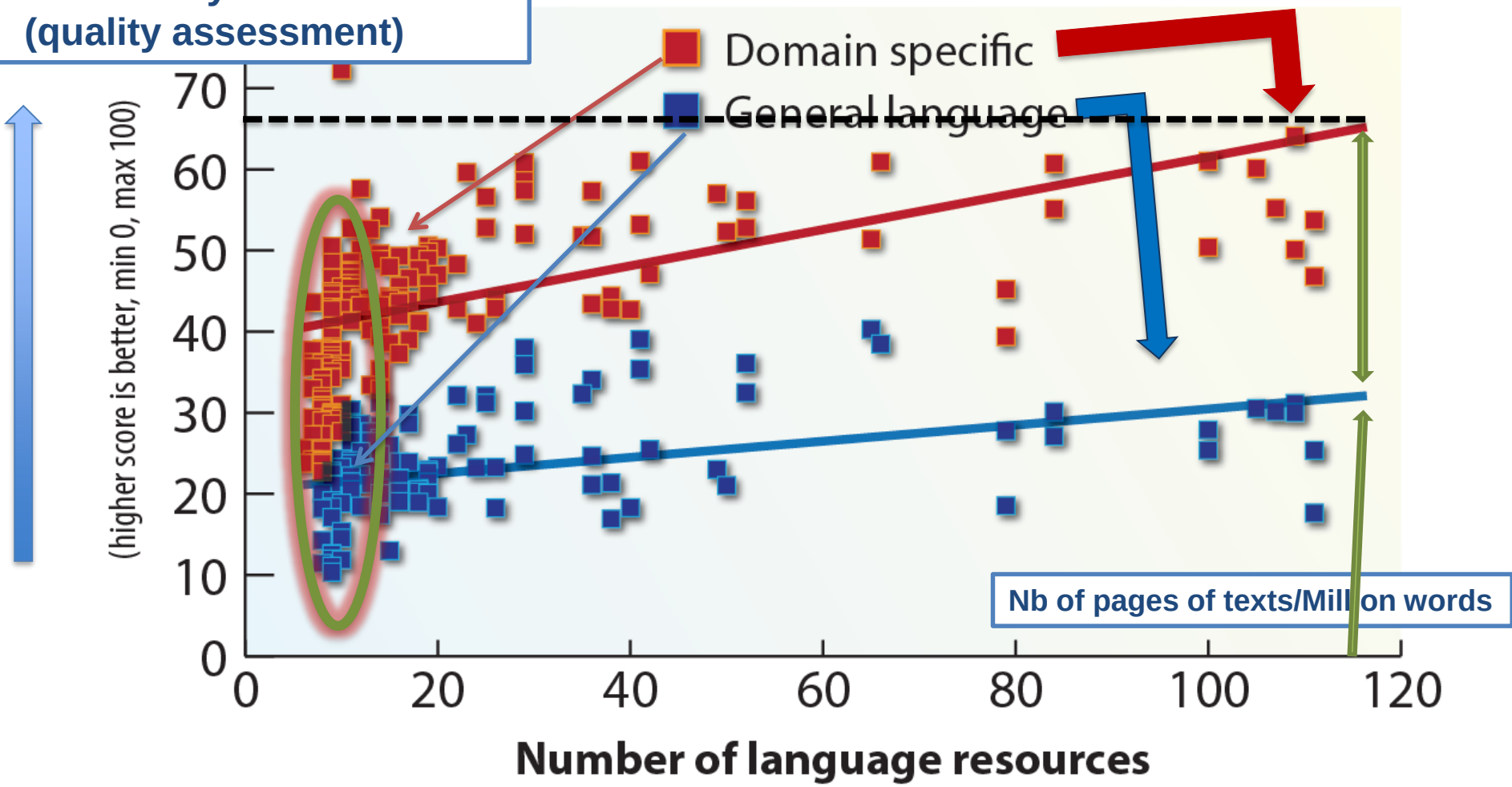
- How to help upload the data
- See the information on the REPOSITORY set-up for this

- **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 measure (quality assessment)



European Language Resource Coordination
Connecting Europe Facility

CONCLUSIONS

- How data is produced: **repurposing and repackaging existing data**
- Why is important: the data driven paradigm is very efficient

- *Let us not under estimate the value of our resources*
- *How can you contribute and benefit from CEF.AT*
  - *(next sessions)*