

De quelles données avons-nous besoin?

Aspects techniques et pratiques

Khalid Choukri
(ELRA/ELDA)
De la part du consortium ELRC



- L'approche prédominante est un paradigme « apprentissage » à partir des données
 - systèmes de Traduction Automatique apprennent à partir des données existantes
 - le focus pour ELRC: Les données linguistiques dans toutes les langues (UE / CEF)
- Les Ressources Linguistiques (RLs) sont produites à partir de:
 - Documents et autres données linguistiques
 - Diverses sources comme le web
- Votre concours est important (avec les données que vous avez ou dont vous connaissez les détenteurs)



- Tout ce qui contient des « mots », préférences pour des « phrases », surtout des phrases exprimées en plusieurs langues, par exemple:
 - Rapports, documents,
 - Discours (transcriptions),
 - Contenus de pages web,
 - Brochures, etc.
- Sacs de « mots », « phrases », plusieurs sacs

Traductions « alignées »

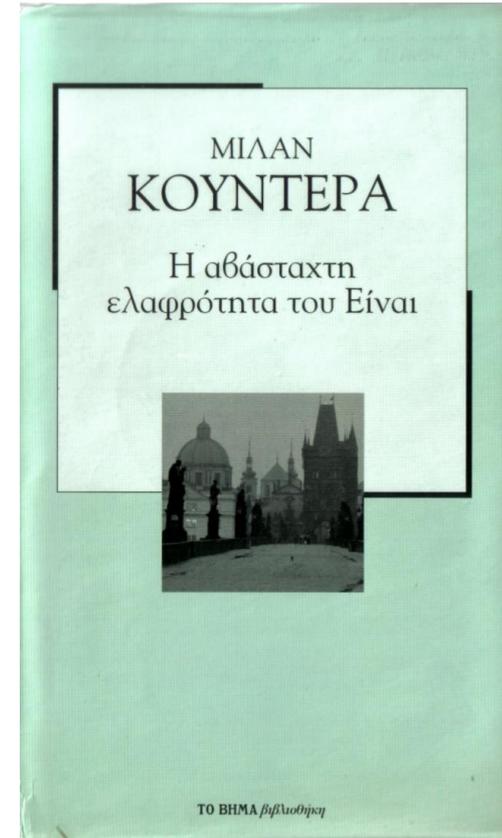
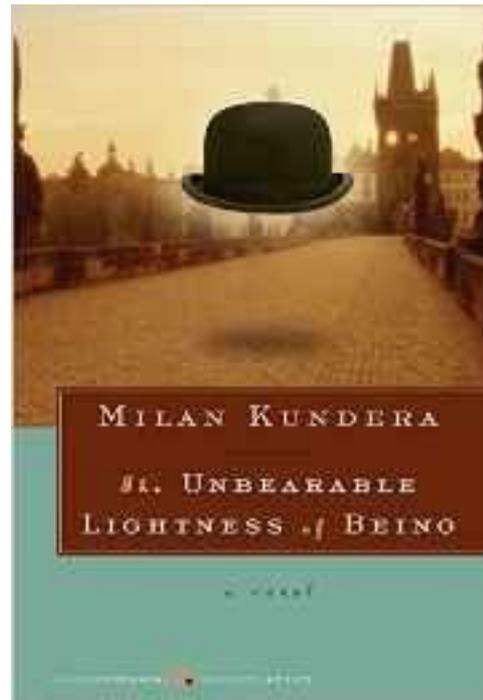
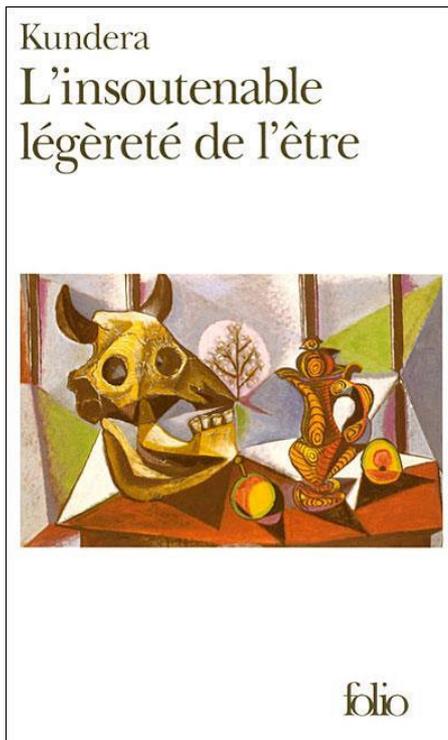


Anglais



Français

Exemple ... Illustrations





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.



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.

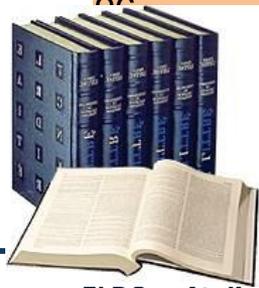
Por metonimia, también se denomina telecomunicación (o telecomunicaciones, indistintamente) a la disciplina que estudia, diseña, desarrolla y explota aquellos sistemas que permiten dichas comunicaciones; de forma análoga, la ingeniería de telecomunicaciones resuelve los problemas técnicos asociados a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

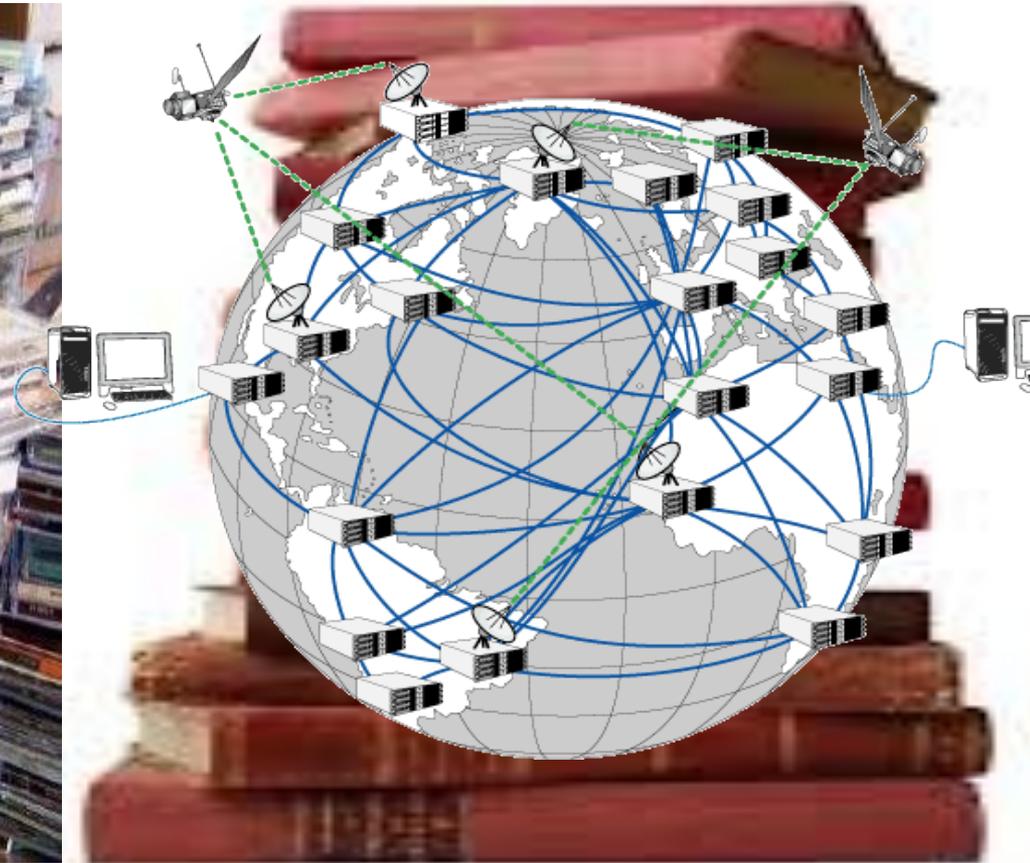
Source: Premières phrases de l'article « Télécommunications » dans le Wikipédia anglais, grec et espagnol.

Le texte espagnol est légèrement différent mais il ne s'agit jamais de traductions de la même 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
22	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

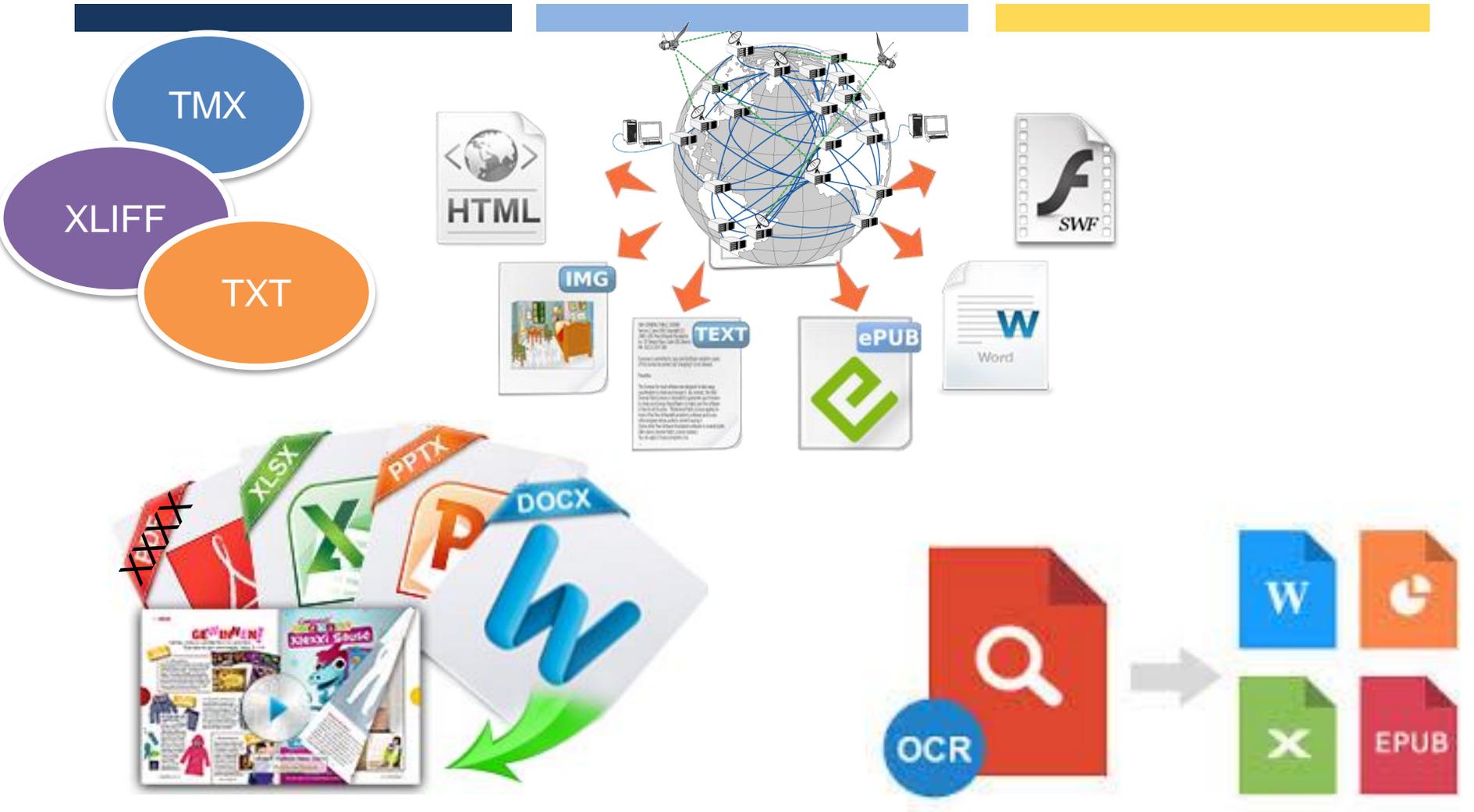




more languages

Quel est le bon format ?

Texte numérique & manipulable





- On peut définir ce qui est essentiel au cas par cas
- Sources de données (fiabilité, qualité, etc.)
 - Domaines spécifiques
 - Langues
 - Droits (si données non publiques)

Les éléments descriptifs (metadata) du Dublin Core

1. Titre // Title
2. Créateur // Creator
3. Sujet // Subject
4. Description // Description
5. Éditeur // Publisher
6. Contributeur // Contributor
7. Date // Date
8. Type // Type
9. Format // Format
10. Identifiant // Identifier
11. Source // Source
12. Langue // Language
13. Relation // Relation
14. Couverture // Coverage
15. Droits // Rights



- Des données brutes (« raw data ») comme des pages html avec tableaux, images, etc.) peuvent être converties
 - Découvrir et identifier les sources (e.g.: URL)
 - Clarifier les aspects juridiques (propriété intellectuelle, licence)
 - Obtenir les données (réception/envoi, téléchargement « crawling »)
 - Nettoyer les données (par exemple détecter et supprimer les « boilerplate », « modèles », des images, des balises html, etc., convertir le format)
 - Documenter les données (à la « Dublin Core » ou selon notre meta-data)
 - Aligner les traductions lorsque identifiées et segmentation en « unités » de traduction
 - Calculer un score de fiabilité de l'alignement
 - Partager

➔ Exemple de moissonnage de données



Documents Word provenant de <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



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).

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 regarding food insecurity and malnutrition and is related to the achievement of the first Millennium Development Goal (MDG 1) (reduction of extreme poverty and hunger by half by 2015).

With almost 870 million chronically undernourished people in 2010-2012, the number of hungry people in the world remains unacceptably high.

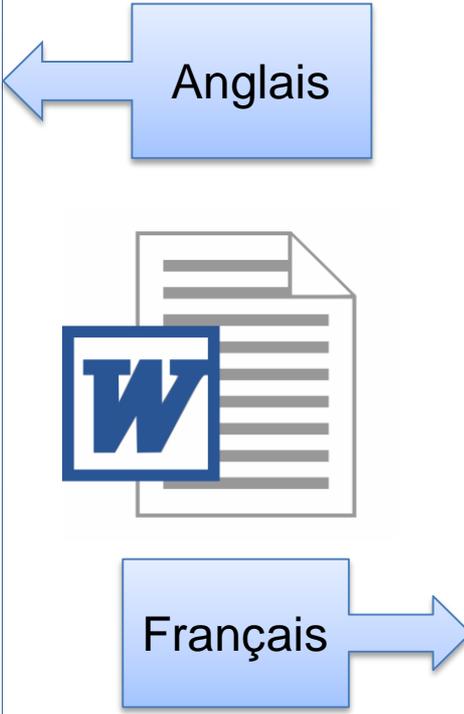
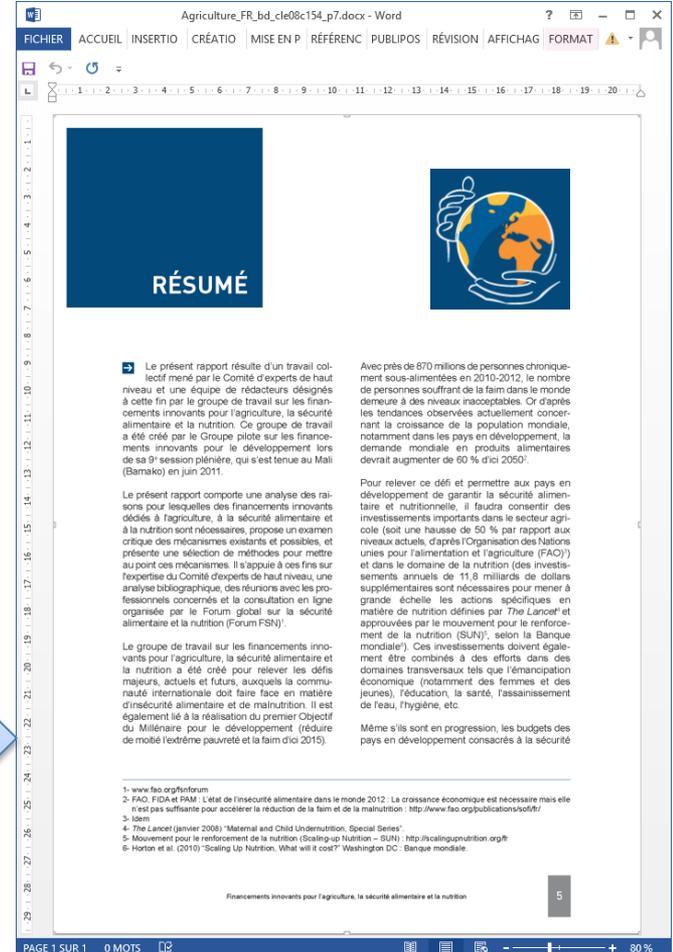
Given the current trends in world population growth, particularly in developing countries, the global demand for food is expected to increase by 60 percent by 2050¹.

Meeting this challenge and allowing developing countries to ensure food security and nutrition imply an important investment in the agricultural sector (i.e. 50 percent more than the current level according to FAO²) and in the field of nutrition (i.e. additional annual investment of USD 11.8 billion is needed to implement at scale the priority nutrition-specific interventions identified by the Lancet series³ and endorsed by the Scaling Up Nutrition (SUN) movement⁴, according to the World Bank⁵). These investments need to be as well combined with efforts across cross-cutting sectors such as economic empowerment (especially women and young people), education, health, water sanitation and hygiene, etc.

Although they are progressing, budgets for food security, including agriculture and nutrition components, in developing countries are severely constrained. Regarding the agriculture side, the dedicated Official Development Assistance (ODA) has increased in recent years, with higher amounts allocated to emerging middle income countries compared to Sub-Saharan Africa and less developed countries. The challenge is all the more important in Sub-Saharan Africa, the continent where population growth will be the highest, where yields have tended to stagnate

1- FSN Forum : <http://www.fao.org/fsnforum/>
 2- FAO, IFAD and WFP: 'The State of food insecurity in the world (SOFI) 2012. Economic growth is necessary but not sufficient to accelerate reduction of hunger and malnutrition.' <http://www.fao.org/publications/sofi/>
 3- Idem
 4- The Lancet (janvier 2008) 'Maternal and Child Undernutrition, Special Series'
 5- SUN movement: <http://scalingupnutrition.org/>
 6- Horton and al. (2010) 'Scaling Up Nutrition. What will it cost?' Washington DC: World Bank

international

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 9^e 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).

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 internationale doit faire face en matière d'insécurité alimentaire et de malnutrition. Il est également lié à la réalisation du premier Objectif de Millénaire pour le développement (réduire de moitié l'extrême pauvreté et la faim d'ici 2015).

Avec près de 870 millions de personnes chroniquement sous-alimentées en 2010-2012, le nombre de personnes souffrant de la faim dans le monde demeure à des niveaux inacceptables. Or, d'après les tendances observées actuellement concernant la croissance de la population mondiale, notamment dans les pays en développement, la demande mondiale en produits alimentaires devrait augmenter de 60 % d'ici 2050¹.

Pour relever ce défi et permettre aux pays en développement de garantir la sécurité alimentaire et nutritionnelle, il faudra consentir des investissements importants dans le secteur agricole (soit une hausse de 50 % par rapport aux niveaux actuels, d'après l'Organisation des Nations unies pour l'alimentation et l'agriculture (FAO)) et dans le domaine de la nutrition (des investissements annuels de 11,8 milliards de dollars supplémentaires sont nécessaires pour mener à grande échelle les actions spécifiques en matière de nutrition définies par The Lancet³ et approuvées par le mouvement pour le renforcement de la nutrition (SUN)⁴, selon la Banque mondiale⁵). Ces investissements doivent également être combinés à des efforts dans des domaines transversaux tels que l'émanipation économique (notamment des femmes et des jeunes), l'éducation, la santé, l'assainissement de l'eau, l'hygiène, etc.

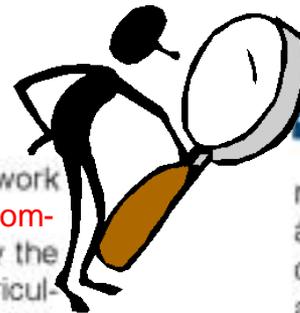
Même s'ils sont en progression, les budgets des pays en développement consacrés à la sécurité

1- www.fao.org/fsnforum/
 2- FAO, IFAD et PAM: 'L'état de l'insécurité alimentaire dans le monde 2012. La croissance économique est nécessaire mais elle n'est pas suffisante pour accélérer la réduction de la faim et de la malnutrition.' <http://www.fao.org/publications/sofi/>
 3- Idem
 4- The Lancet (janvier 2008) 'Maternal and Child Undernutrition, Special Series'
 5- Mouvement pour le renforcement de la nutrition (Scaling-up Nutrition - SUN) : <http://scalingupnutrition.org/>
 6- Horton et al. (2010) 'Scaling Up Nutrition. What will it cost?' Washington DC : Banque mondiale.

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



→ 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

La version anglaise – Données Brutes

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
[...]

La version française – Données Brutes

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
[...]

Alignement des deux versions

S1. Executive Summary

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.

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. 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 [...]

S1. Résumé

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.

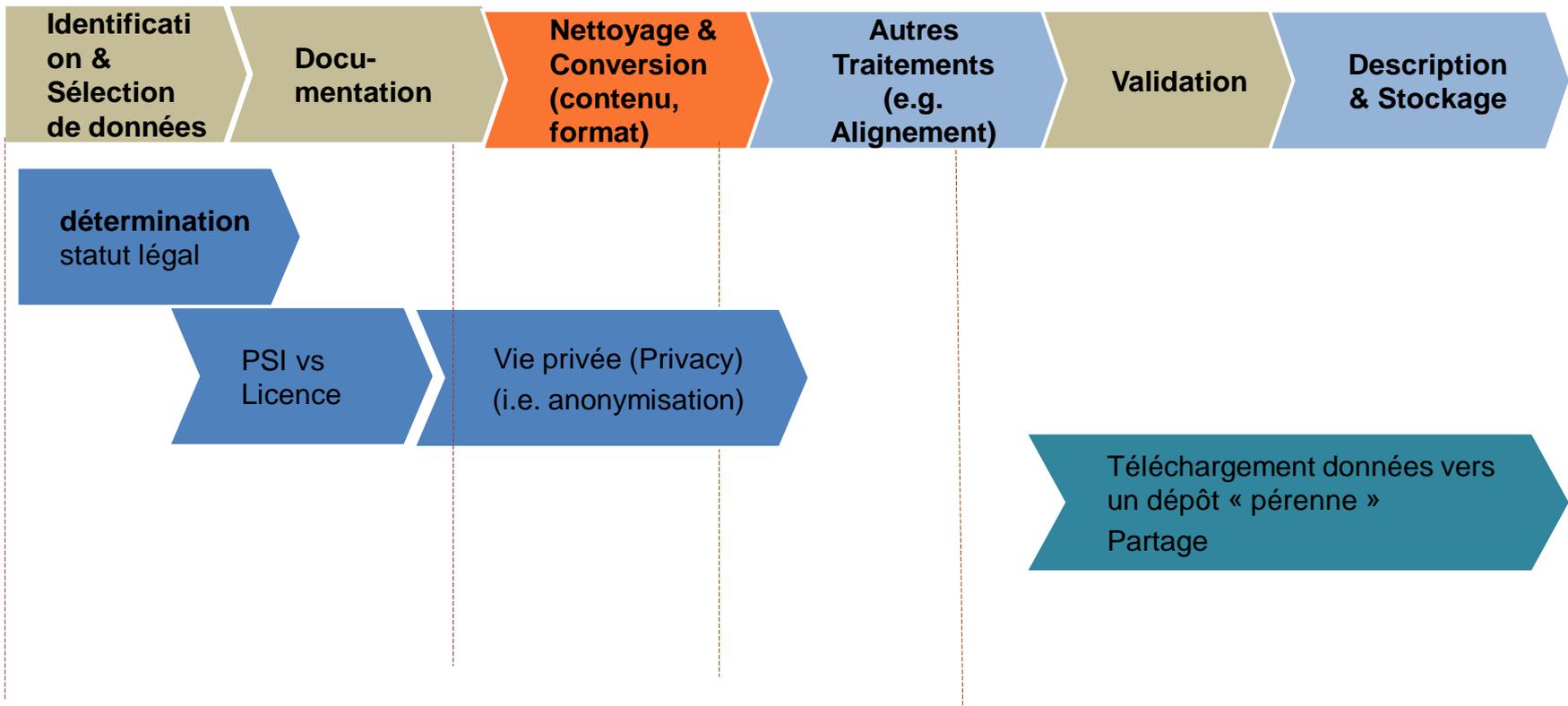
S4. Le présent rapport comporte une analyse des raisons pour lesquelles des financements innovants dédiés à l'agriculture, à la sécurité alimentaire et à la nutrition sont nécessaires, propose un examen critique des mécanismes existants et possibles, et présente une sélection de méthodes pour mettre 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.

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 [...]

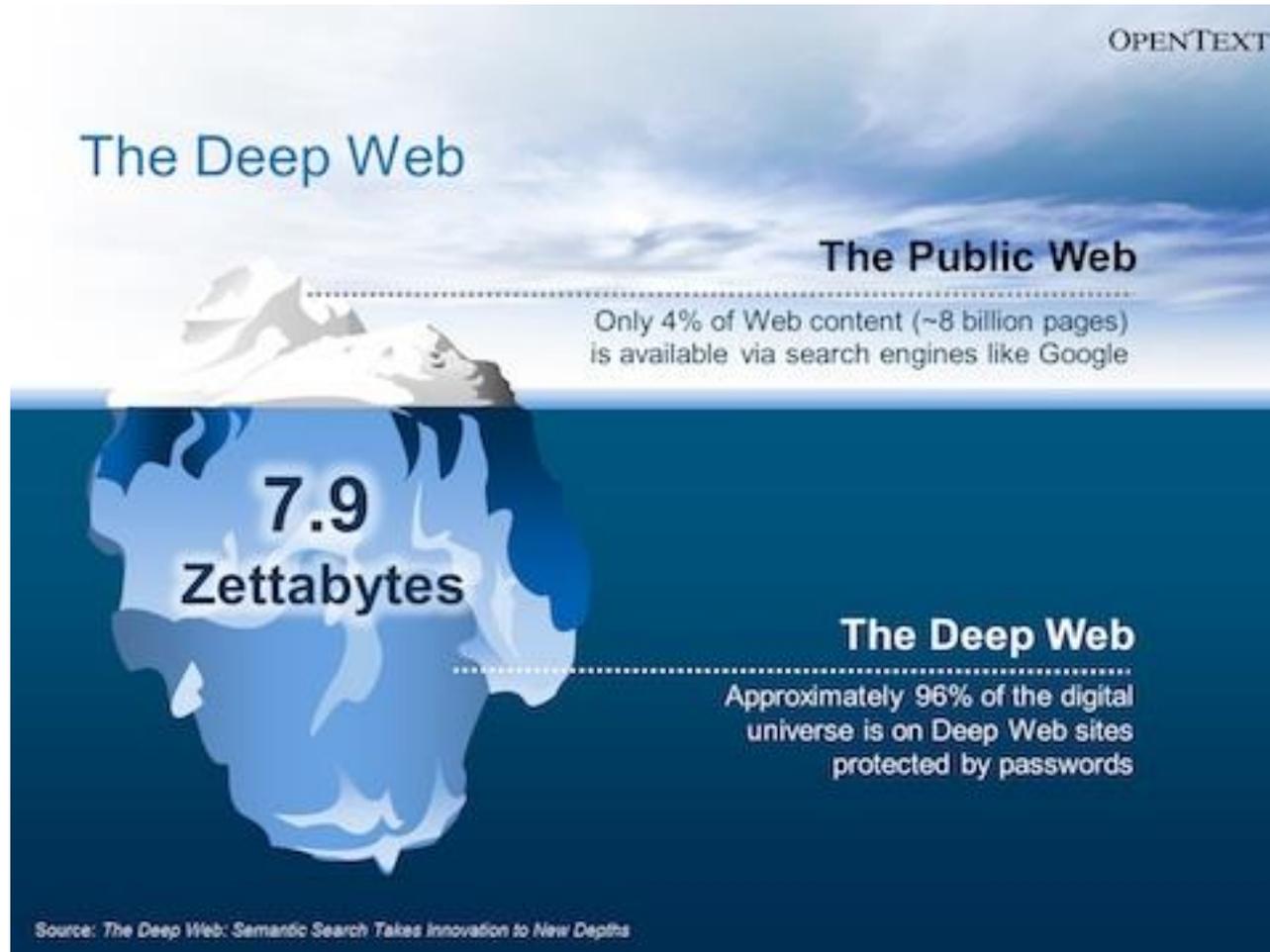
Données → Ressources Linguistiques

La chaîne de valeur





- On ne peut obtenir que la partie « visible » du web
- Il y a beaucoup plus dans les organisations publiques
- Nous avons besoin de votre aide pour identifier ces sources
- Ce processus peut aboutir à une « usine » de production de RLs
 - Automatisation de la procédure avec votre support



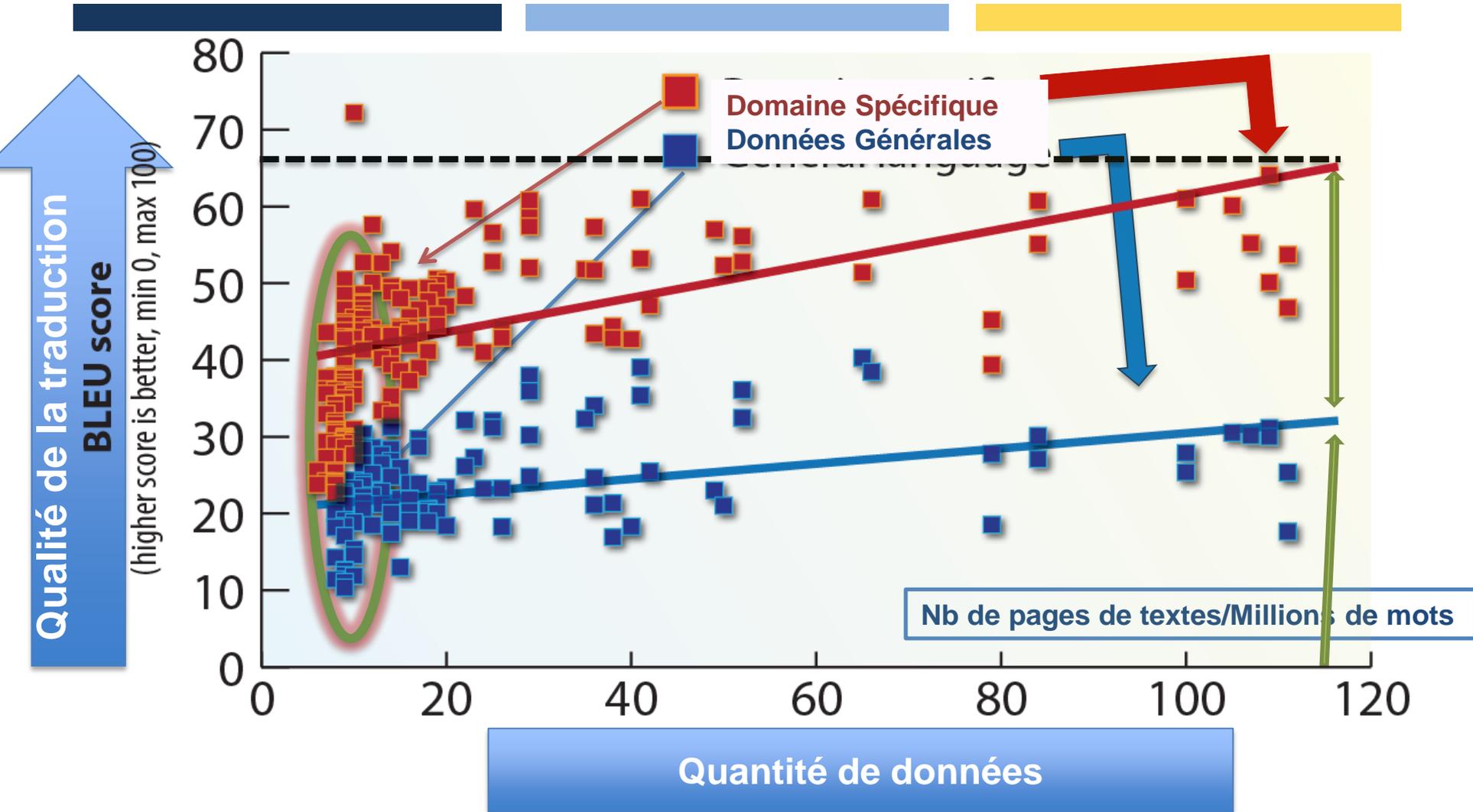




- Ces documents existent déjà:
 - Dans les nombreux centres de documentation (Rapports, Brochures, Discours transcrits, fiches, des Mémoires de Traduction, Termino, ...)
 - Auprès de vos prestataires de services le.g. linguistiques (sous-traitants des travaux de traduction)
- **Nous avons besoin de votre assistance pour les identifier**

➤ De quelles données avons-nous besoin?

➤ **De combien de données avons-nous besoin ?**



- Comment produit-on les données : essentiellement par des données existantes (développées pour d'autres finalités, réorientation/requalification, etc.)
- L'importance des données : Paradigme d'apprentissage (Data Driven Paradigm), encore plus exigeant avec les modèles de réseaux de neurones
- *Dans ce contexte, la valeur de vos données est inestimable*
- *Comment pouvez-vous contribuer à cet effort collectif?*
 - *Contribuer et bénéficier du CEF.AT*