Accord-cadre sur le partage de données de mémoires de traduction générées par la traduction de documents des administrations publiques.

1. INTRODUCTION.

Depuis l'adoption en 2003 du premier ensemble de règles sur la réutilisation des informations du secteur public, le volume des données générées augmente de manière exponentielle dans le monde entier, tandis que de nouveaux types de données générées et collectées apparaissent.

Dans le même temps, nous assistons à une évolution permanente des technologies de traitement du langage servant à l'analyse, à l'exploitation et au traitement des données.

Cette évolution technologique rapide permet de créer de nouveaux services et de nouvelles applications sur la base de l'utilisation, de l'agrégation ou de la combinaison de données.

Les quantités massives de données ou les mégadonnées collectées par différents outils technologiques ou extraites de grands ensembles d'informations de différents formats peuvent générer de nouvelles connaissances dans les secteurs les plus divers. Cependant, elles posent également des problèmes, liés autant à la propriété des données qu'à leur utilisation ultérieure.

Les administrations publiques et autres organismes et entités composant le secteur public sont de grands producteurs de données qui sont très utiles aux secteurs du traitement du langage (naturel). En exerçant leurs fonctions, les administrations publiques génèrent d'énormes quantités de données. L'utilisation et l'exploitation de ces données peuvent être d'un grand intérêt, tant pour les administrations publiques elles-mêmes que pour lesdits secteurs.

Utilisées par les administrations publiques, avec des techniques pour les mégadonnées et l'exploration de données, ainsi que des technologies de traitement du langage naturel, ces données peuvent faciliter la prise de décision publique et améliorer l'efficacité des politiques publiques.

Le présent accord-cadre se veut une recommandation aux États membres pour la mise en œuvre de licences ouvertes qui suppriment les éventuels obstacles juridiques afin d'intégrer ces types de données à la création de nouvelles connaissances.

Pour les divers gouvernements, grandes entreprises et organisations internationales produisant des données publiques, il est désormais une pratique courante de s'appuyer sur ces licences en tant que mécanismes appropriés pour favoriser l'utilisation et la réutilisation de ces grands volumes de données. Ils se servent particulièrement des modalités PPDL, GPL et Creative Commons (CC), où tous les droits sur les bases de données sont abandonnés.

2. Pourquoi un accord-cadre européen pour la gestion du partage de données.

Les motifs de l'accord-cadre pour le partage de données de mémoires de traduction.

2.1. Approche générale.

L'ACPDT (**Accord**-cadre sur le **partage** de données **TMX**) n'est ni fortuit, ni purement épisodique ou ponctuel. Au contraire, il répond à un besoin qui trouve diverses motivations. Les administrations publiques européennes génèrent des milliers de traductions chaque jour, et ces traductions génèrent des bases de données dont la diffusion en tant que données est déjà requise par la législation (directive européenne sur les données ouvertes).

Il existe deux éléments de grande importance qui reflètent l'absence d'un consensus de base pour une réglementation adéquate et pour la gestion des problèmes liés au

partage des données générées par la traduction de textes provenant des administrations publiques. Il s'agit de **l'un des secteurs qui génèrent les plus grandes quantités de données** de ce type dans le cadre de leurs activités quotidiennes. C'est également le secteur qui investit le plus dans les infrastructures de stockage de données à l'échelle mondiale. Nous nous concentrons particulièrement sur les données générées par les mémoires de traduction des documents appartenant à l'administration publique, étant donné les besoins des administrés et les nombreuses langues, non seulement sur un même territoire ou un pays, mais dans tous les pays qui composent l'Union européenne.

Il est très important de se conformer aux recommandations en matière de réutilisation des données et d'interopérabilité dans le secteur de l'administration publique. Il est tout aussi important de définir des conditions pour la réutilisation et l'interopérabilité.

Sans aucun doute, la meilleure situation serait que les données soient pleinement accessibles et que leur réutilisation ne nécessite pas de permis spécifique, comme l'établit la directive européenne sur les données ouvertes.

Toutefois, pour que l'accessibilité et la réutilisation soient possibles, certaines conditions sont requises, et c'est pourquoi des licences sont nécessaires. Dans ce cas, les mémoires de traduction doivent être aussi ouvertes que possible. Notre objectif prioritaire est d'établir un cadre d'action qui favorise une plus grande diffusion et une plus grande capacité de réutilisation.

Il est fortement conseillé de ne pas créer davantage de types de licences, mais au contraire, de simplifier les types de licences, comme le recommande la *Stratégie numérique pour l'Europe* (Commission européenne, 2010).

Il existe actuellement plusieurs alternatives internationales principales qui encouragent l'utilisation de licences dans le domaine des données et des informations, et qui réglementent et promeuvent le libre accès aux informations et leur utilisation libre. Il s'agit des licences Creative Commons (CC), Apache 2, GPL et ODC. Les licences Creative Commons s'appliquent à la fois aux données et aux autres documents, alors que les autres licences mentionnées sont utilisées uniquement dans le domaine des données. La tendance internationale en ce qui concerne la réglementation des

conditions d'accès et d'utilisation des données est le recours à des modèles de licences ouvertes, basés principalement sur CC, mais avec des variations adaptées aux caractéristiques de chaque portail de données et de chaque pays où ils sont appliqués.

Tout cela bénéficie à une administration de plus en plus numérique, capable **de mesurer l'impact de ses actions en termes de retour sur investissement social.** C'est d'ailleurs l'objectif d'initiatives telles que la mémoire centrale de traduction nationale et européenne (NEC TM), par le biais du présent accord-cadre pour le partage des données générées par les mémoires de traduction de textes traduits pour les administrations publiques.

Par conséquent, étant donné qu'il n'existe actuellement pas de consensus sur la manière de réglementer le traitement, l'utilisation et le partage des données de mémoires de traduction générées à la suite des besoins de publication des administrations publiques, on pourrait choisir l'option abstentionniste, en se référant simplement aux lois spécifiques existantes de nature générale dans chaque État membre. Cependant, la meilleure option serait une réglementation propre à cette question, sous la forme du modèle réglementaire proposé par les présentes, basé sur des « normes douces », de nature unilatérale et non contraignante, ainsi qu'une approche pratique telle qu'un ACCORD-CADRE.

L'ACPDT tient compte des deux parties impliquées dans le processus, les administrations publiques et les fournisseurs. D'abord, il indique clairement que le problème du partage des données de mémoires de traduction existe, qu'il est réel pour les organisations concernées, et qu'il est nécessaire d'intervenir, tout en rejetant les excès et les abus. Ensuite, l'ACPDT propose un cadre réglementaire complémentaire et spécifique, ainsi qu'une gestion systématique, mais aussi flexible, adaptable à toutes les administrations, entités privées et États membres.

À cette fin, et dans l'intérêt d'une société de plus en plus soucieuse de la valeur des données qu'elle génère, il est fortement recommandé, et même nécessaire, que ces données soient mises à la disposition du contractant responsable ou d'un organisme

central au sein de chaque État membre, pour chaque contrat de services de traduction exécuté, en fournissant non seulement le texte ou les documents de la traduction en question, mais aussi la mémoire de traduction générée à la suite du processus.

De même, grâce à la mise en œuvre de la base de données NEC TM, certaines conditions sont établies pour faciliter la réutilisation de ces mémoires de traduction du fait de leur accessibilité, la condition la plus importante étant qu'elles ne soient pas soumises à des restrictions techniques ou juridiques qui limitent ou entravent cette réutilisation.

3. Multidimensionnalité du problème du partage des données de mémoires de traduction générées par les administrations publiques : la protection offerte par l'ACPDT.

Au sens strict de sa nature juridique, le présent accord-cadre est conçu comme un accord dont la mise en œuvre se déroule conformément aux pratiques et procédures découlant de l'expérience de chaque État membre, et en tenant compte des pratiques qui sont déjà courantes et qui sont connues des fournisseurs du secteur de la traduction.

Sur la base des considérations qui précèdent, il semble évident qu'il n'est pas possible de concéder à cet accord-cadre, provenant d'une source autonome, des effets juridiques directs au même titre que les autres éléments de la législation communautaire, mais il n'est pas non plus question de sous-estimer ses effets. L'adhésion à cet accord est purement volontaire et dépend strictement du pouvoir des signataires de contraindre les organisations.

Ainsi, l'ACPDT, en tant qu'un accord autonome, se veut plutôt un instrument pratique, c'est-à-dire un ensemble de lignes directrices qui précisent des règles efficaces en ce qui concerne la gestion, la centralisation et le partage des données, l'accord-cadre étant strictement complémentaire de la législation communautaire et nationale.

De cette manière, un accord non normatif tel que l'ACPDT peut constituer, et constitue par les présentes, un instrument utile de concrétisation et de clarification de l'interprétation du partage des données générées par les mémoires de traduction créées à la suite de travaux effectués par les soumissionnaires/fournisseurs.

La fonction de l'ACPDT est donc de clarifier et de préciser non seulement le droit des administrations publiques à demander toutes les données générées dans le cadre des contrats de traduction, mais aussi la propriété de ces administrations sur les textes originaux et leur traduction en tant que service contracté et dérivé d'un original, en reflétant le consensus des acteurs impliqués. Ces acteurs comprennent, d'une part, les entreprises soumissionnaires (les fournisseurs), d'autre part, l'organisme administratif de chaque État membre et, enfin, l'organisme central de l'Union européenne qui est désigné à cette fin.

En ce sens, il convient de rappeler que la configuration la plus accréditée de ce modèle ou paradigme de juridicité illustre l'existence d'instruments qui, bien qu'ils ne se conforment pas aux types de normes juridiques traditionnels, ne peuvent pas non plus être exclus du monde du droit, et dont le caractère obligatoire ne peut pas être exclu. Ainsi, il est possible d'exprimer l'engagement ferme des signataires qui exécutent le contrat de services dans l'exercice de leurs compétences selon une procédure légalement prévue, en mettant en pratique un instrument doté d'une juridicité minimale ou relative.

Le contenu fondamental de l'ACPDT concerne surtout l'établissement de modèles et de lignes directrices, ce qui en fait un instrument typique du droit réfléchi.

Finalement, ce paradigme de réglementation matérialisé par l'ACPDT est entièrement cadré par les nouvelles approches et les nouvelles méthodes de réglementation et d'action actuellement proposées, mais aussi par l'Union européenne. Cette dernière s'appuie davantage sur des instruments transformant la règle juridique dans la pratique que sur l'établissement de règles supplémentaires qui, dans le cas présent, définiraient

l'accès, l'utilisation et la réutilisation éventuelle des informations contenues dans les bases de données réglementées par la loi, les contrats et/ou les licences.

Depuis la CEE, la publication de la Communication sur l'ouverture des données publiques et la proposition d'amendement par la Commission européenne constituent une politique d'ouverture des données et de promotion d'un marché des informations, politique dont l'élément central a été l'approbation de la directive 2003/98/CE du Parlement européen et du Conseil du 17 novembre 2003 concernant la réutilisation des informations du secteur public.

Il faut ajouter à cela la révision de la directive concernant la réutilisation des informations du secteur public, introduite fin 2011, qui s'inscrit dans le cadre de la Stratégie numérique pour l'Europe. Initiative de la Commission européenne (2010), cette stratégie vise à encourager la mise en place de services en ligne au sein de l'Union et accorde une importance capitale à l'ouverture des données publiques pour leur réutilisation, à la simplification du système de licences pour l'échange de contenus, ainsi qu'à la mise en place de normes d'interopérabilité.

5. L'option réglementaire de l'ACPDT

Les données publiques détenues par des organismes publics dans l'Union européenne font l'objet d'un traitement spécifique en vertu des directives de 2003 et 2013 sur la réutilisation des informations du secteur public, qui prévoient la mise à disposition du public des documents des administrations publiques pour réutiliser « tout contenu, quel que soit son support (écrit sur papier, stocké sous forme électronique ou sous forme d'enregistrement sonore, visuel ou audiovisuel) conservé par des organismes du secteur public à des fins commerciales ou non commerciales ». Par conséquent, le principe général est la libre disponibilité de ce contenu. En cas d'application d'un tarif ou de redevances, « lesdites redevances sont limitées aux coûts marginaux de reproduction, de mise à disposition et de diffusion » (article 6.1).

En résumé, l'ACPDT se présente essentiellement comme une « recommandation » fondée sur une juridicité minimale ou relative, sur la base des contrats de service existants et des données (mémoires de traduction) générées dans un cadre professionnel. Cette « recommandation » est la spécification pratique du devoir de partager les données générées sous forme de mémoires de traduction provenant des services de traduction contractés par les administrations publiques, quel que soit leur support ou leur format.

6. Contenu de l'accord-cadre : le système de gestion pour le transfert des données des mémoires de traduction dont cet accord fait la promotion.

Une mémoire de traduction est une base de données linguistiques qui stocke de manière continue les traductions parallèles produites par les professionnels afin de pouvoir les utiliser plus tard. Elle permet de faire preuve d'une plus grande cohérence terminologique et stylistique, mais aussi de faire des économies grâce à des correspondances totales ou partielles entre les nouveaux textes et les anciennes traductions déjà réalisées.

Les mémoires de traduction sont des dépôts numériques composés de lignes de texte issu du contenu dans la langue d'origine et aligné avec sa traduction dans d'autres langues. Ces textes peuvent également être alignés efficacement par unités de traduction. Les unités de traduction, stockées avec leur équivalent, sont définies de différentes manières : par phrase, par paragraphe, par mot ou groupe de mots, etc. Dans l'environnement des systèmes de traduction assistée par ordinateur, la segmentation s'effectue le plus souvent par défaut après un séparateur tel qu'un signe de ponctuation en fin de phrase ou un retour à la ligne en fin de paragraphe.

La fonction principale des mémoires de traduction (ci-après « TM ») est d'extraire des suggestions ou des correspondances totales ou partielles de phrases, et des options de concordance de termes. Lors de la traduction, des segments de la langue source sont recherchés dans la base de données de la TM. Si la TM comporte un segment dans la

langue source qui correspond exactement ou partiellement au segment sur lequel se trouve le traducteur, ce segment lui sera suggéré, ainsi que la traduction correspondante extraite de cette base de données, et toute information supplémentaire qui a été enregistrée avec le segment dans la base de données. Les outils de traduction assistée par ordinateur (outils de TAO) indiquent le degré de similarité (dans le jargon, on parle correspondance partielle ou *fuzzy matches*). Dans le secteur de la traduction, ces correspondances peuvent être utiles aux linguistes lorsque la similarité est égale ou supérieure à 50 %-65 %. Des barèmes de prix sont également établis pour refléter l'effort nécessaire à la traduction de phrases complètement nouvelles ou bien d'autres phrases pour lesquelles il existe des suggestions similaires et des pourcentages de correspondance dans le même contexte. Un degré de similarité de 100 % est considéré comme une correspondance totale entre la phrase demandée et la phrase identique dans la base de données.

Le format libre pour l'échange de mémoires de traduction est TMX (Translation Memory eXchange), généralement dans la version 1.4b. Cette norme XML est une DTD (*document type definition*). Elle a été créée par le comité OSCAR (Open Standards for Container/Content Allowing Re-use).

Grâce à l'application du format TMX, il est plus facile pour les personnes ou les entreprises de travailler ensemble à des projets de traduction. Le format TMX facilite également la migration d'un système de traduction assistée par ordinateur à un autre, ce qui favorise la compétitivité des technologies proposées et leur développement constant afin de se différencier de la concurrence. Comme dans le cas d'autres normes ouvertes, ce format a été développé en vue de réduire les problèmes de compatibilité, de favoriser la réutilisation des ressources linguistiques et de simplifier l'échange de données, stimulant ainsi l'innovation technologique.

Comme nous l'avons clairement défini, cet accord-cadre vise à établir un « cadre de bonnes pratiques » concernant l'obtention des données générées dans le cadre des contrats de services de traduction par les administrations publiques européennes, ainsi que l'organisation de cette obtention par la mise en œuvre du système logiciel NEC TM

et des initiatives de centralisation des données déterminées par la Commission européenne. Cet accord-cadre est proposé en vue de l'adoption d'un protocole concernant la centralisation des données bilingues par les administrations publiques, l'utilisation de ces données et leur apport à la société en général et aux États membres eux-mêmes, l'aide à la création d'un corpus de mégadonnées nationales et le partage des éléments que chaque administration nationale juge pertinents à l'échelle européenne.

À cette fin, la centralisation des données hébergées sur les infrastructures informatiques nationales sera gérée par l'organisme compétent de l'État en question, par exemple, le Secrétariat d'État au numérique, ou le ministère compétent dans l'État membre en question. Ces données peuvent être partagées, en cas d'adhésion au présent accord-cadre et si l'autorité compétente le décide, à un niveau supérieur avec l'organisme compétent de la Commission européenne (par exemple, ELRC-Share).

Cela étant dit, la première des étapes sera la centralisation des données des mémoires de traduction générées par les administrations publiques à l'échelle nationale grâce à l'adoption du présent accord, puis, conformément au présent accord, le partage de ces données avec un organisme paneuropéen au profit des administrations publiques des autres États membres de l'Union européenne. L'utilisation du logiciel de « mémoire centrale de traduction nationale et européenne » (NEC TM) fournira une connexion d'entrée/de sortie (E/S) aux administrations. Elles pourront ainsi stocker de manière privée leurs mémoires de traduction, hébergées sur un serveur central, les partager avec les acteurs souhaitant rendre leur travail plus efficace et plus économique à échelle nationale (généralement des traducteurs internes ou des fournisseurs externes), créer des mégadonnées nationales et sélectionner les données souhaitées avec le serveur central où est hébergée la mémoire centrale de traduction européenne.

Relation avec les entités privées et cession de TMX

Les travaux de traduction effectués par les entreprises soumissionnaires (les fournisseurs) impliquent la fourniture des mémoires de traduction qu'elles ont générées à la suite de la prestation de leur service. C'est pourquoi l'avis figurant au Journal officiel de l'État membre, de la région ou de la municipalité, c'est-à-dire sur le portail du

contractant, doit comporter une clause faisant référence à l'ACPDT pour NEC TM, incluant le CODE CPV, qui spécifie QUE LES DONNÉES DE LA MÉMOIRE DE TRADUCTION SONT FOURNIES AU SERVEUR DE L'AUTORITÉ PUBLIQUE COMPÉTENTE afin qu'à l'avenir, si cette autorité nationale le juge opportun, ces données puissent être partagées avec un organisme de la Commission européenne à déterminer (ELRC-Share ou une initiative similaire, par exemple).

À cette fin, les entreprises soumissionnaires (les fournisseurs) livrent les mémoires de traduction générées conjointement à leur travail, le principe de non-rétroactivité s'appliquant ; c'est-à-dire que les effets du présent accord-cadre ne doivent pas s'exercer sur une période antérieure, et que cette obligation s'applique seulement à partir de l'adhésion et de la signature de la clause dans laquelle les entreprises soumissionnaires s'engagent à livrer, en même temps que le travail effectué, les données parallèles générées par leurs traductions (TMX ou format compatible similaire).

7. PROPRIÉTÉ INTELLECTUELLE ET CONDITIONS D'UTILISATION DE BASES DE DONNÉES

En ce qui concerne la propriété intellectuelle, il n'y a aucun doute que les données générées au cours de l'exécution du travail et conformément au contrat, attribué pour la traduction en question, est entièrement et exclusivement celle de l'administration publique qui est auteur ou gestionnaire du texte original faisant l'objet de la traduction.

On ne peut pas ignorer le fait que, dans la mesure où les informations proviennent la plupart du temps des administrations publiques, il faut également garder à l'esprit la réglementation de la réutilisation des informations publiques et de l'ouverture des données publiques.

Ainsi, le soumissionnaire retenu (fournisseur) cède exclusivement, sans limite temporelle ni territoriale, les droits de tout type de documentation ou de données générées, indépendamment de leur support ou format, ce transfert de droits comprenant la reproduction, la distribution et la transformation.

Par conséquent, les données générées par ces mémoires de traduction ne peuvent pas être utilisées par le soumissionnaire retenu à des fins lucratives et ne peuvent être utilisées que pour fournir des informations, si les travaux réalisés le requièrent.

Propriété intellectuelle et conditions d'utilisation de bases de données.

La base de données de DGT-TM est la propriété exclusive de la Commission européenne. La Commission accorde, gratuitement et sur une base mondiale, pendant toute la période de protection de ces droits, ses droits non exclusifs aux réutilisateurs pour tous les types d'utilisation répondant aux conditions énoncées dans la décision de la Commission du 12 décembre 2011 relative à la réutilisation des documents de la Commission, publiée au Journal officiel de l'Union européenne, L 330 du 14 décembre 2011, pages 39 à 42.

Toute réutilisation de la base de données ou des éléments structurés qu'elle contient doit être identifiée par le réutilisateur, qui est tenu d'indiquer la source des documents utilisés : l'adresse du site web, la date de la dernière mise à jour et le fait que la Commission européenne reste propriétaire des données.

Cette base de données est donc optimale pour l'alimentation initiale de la version de NEC TM d'un État membre.

8. PROTECTION DES DONNÉES. APPLICATION DU RÈGLEMENT UE 2016/679 ET DU RÈGLEMENT 2018/1725

En ce qui concerne les données à caractère personnel qui doivent être traitées sur la base de l'exécution du contrat par l'administration publique et le soumissionnaire retenu, les deux parties sont tenues de respecter les règlements généraux sur la protection des données ci-dessous.

Le règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, ainsi que la législation nationale en vigueur en matière de protection des données dans chacun des États membres qui adoptent le présent accord-cadre, et le règlement du Parlement européen et du Conseil du 23 octobre 2018 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel par les institutions, organes et organismes de l'Union et à la libre circulation de ces données.

CONCLUSIONS FINALES.

En tant que promoteur du secteur des technologies linguistiques, avec la création de plateformes communes pour le traitement du langage naturel et la traduction automatique, et le développement de ressources pour la réutilisation des informations du secteur public (RISP), l'administration publique est tenue de développer des politiques de partage des données et de poser les fondations nécessaires afin que le partage entre toutes les entités qui participent au processus soit réel et efficace. Elle doit également veiller à ce que les participants au processus ne faisant pas partie de l'administration s'engagent à signer la clause d'acceptation de leurs offres et à effectuer la livraison, avec leur travail, des données générées par leurs traductions (mémoires de traduction), constituant un ensemble de ressources linguistiques d'une valeur inestimable.

RÉFÉRENCES. Ferrer-Sapena et al. (2011) dans leur article sur l'accès aux données publiques ; Ferrer-Sapena et Peset (2012) sur la réutilisation des données culturelles ; Ramos Simón et al. (2012) dans leur étude sur les portails de données de l'Union européenne.

Licensing Open Data: A Practical Guide à destination du Higher Education Funding Council for England (Korn et Oppenheim, 2011). À noter également le *Guide to Open Data Licensing* (Open Knowledge Foundation, s.d.) et les lignes directrices fournies par la *Stratégie numérique pour l'Europe* (Commission européenne, 2010).