

Acordo-Quadro para a partilha de dados de memórias de tradução gerados a partir de traduções de documentos da Administração Pública.

1. INTRODUÇÃO.

Desde a adoção do primeiro conjunto de regras sobre a reutilização de informações do setor público, em 2003, o volume de dados gerados tem aumentado de forma exponencial em todo o mundo, enquanto surgem e se recolhem novos tipos de dados.

Simultaneamente, assistimos a uma evolução contínua das tecnologias de processamento da linguagem para a análise, a exploração e o tratamento de dados.

Esta rápida evolução tecnológica possibilita a criação de novos serviços e novas aplicações com base na utilização, agregação ou combinação de dados.

As quantidades elevadíssimas de dados ou de Big Data recolhidas através de diferentes ferramentas tecnológicas ou extraídas de grandes coleções de informações em formatos diferentes têm a capacidade de criar novos conhecimentos nos setores mais diversificados. Contudo, também constituem problemas ao nível da propriedade de dados e da utilização subsequente dos mesmos.

As Administrações Públicas, os órgãos e as entidades que formam o setor público são importantes produtores de dados, que são grandemente utilizados pelas indústrias do processamento da linguagem (natural). Em particular, as Administrações Públicas geram enormes quantidades de dados no exercício dos seus deveres. A utilização e a exploração desses dados podem ser de um elevado nível de interesse para as Administrações Públicas em si, mas também para a indústria.

A utilização destes dados pelas Administrações Públicas, em conjunto com as técnicas de Big Data e prospeção de dados e as tecnologias de processamento da linguagem natural, pode facilitar a tomada de decisão pública e a eficácia das políticas públicas.

O presente Acordo-Quadro pretende estabelecer-se como uma recomendação para os Estados-Membros no que respeita à implementação de licenças livres que suprimem possíveis obstáculos legais, de forma a incorporar estes tipos de dados na criação de novos conhecimentos.

Para diferentes governos, empresas e organizações internacionais que geram dados públicos, tornou-se prática corrente depender destas licenças (especialmente, das modalidades de PDDL, GPL e CC (Creative Commons), nas quais todos os direitos a bases de dados são renunciados), como um mecanismo adequado para favorecer a utilização e a reutilização destes grandes volumes de dados.

2. Fundamentação do Acordo-Quadro europeu para a gestão da partilha de dados.

Motivos para a criação do Acordo-Quadro para a partilha de dados de memórias de tradução.

2.1. Abordagem geral.

O surgimento do FATDS (**Acordo-Quadro para a Partilha de Dados em TMX**) não ocorreu por coincidência nem é episódico ou pontual. Pelo contrário, é a resposta a uma necessidade, que tem diversas motivações na sua base. Todos os dias, as Administrações Públicas europeias produzem milhares de traduções que, por sua vez, geram bases de dados, cuja divulgação de dados já é um requisito legal (Diretiva UE relativa aos dados abertos).

Existem dois aspetos de grande relevância que refletem a falta de consenso de base no que concerne à regulamentação adequada e à gestão dos problemas associados à

partilha dos dados gerados pelas traduções dos textos provenientes de Administrações Públicas. Por um lado, este é **um dos setores que gera a maior quantidade de dados** deste tipo nas suas atividades diárias. Por outro, também é o setor que investe mais nas infraestruturas de armazenamento de dados em todo o mundo. Centramos as nossas atenções, particularmente, nos dados gerados pelas memórias de tradução de documentos que são propriedade da Administração Pública, dadas as necessidades de administração e as diversas línguas, não só num dado território ou país, como também em todos os países que constituem a União Europeia.

É muito importante cumprir as recomendações de reutilização de dados e interoperabilidade no setor da Administração Pública. Igualmente importante é o estabelecimento de condições para essa reutilização e interoperabilidade.

Indubitavelmente, o estado ideal seria a acessibilidade total aos dados e a respetiva reutilização sem serem necessárias autorizações específicas, conforme disposto pela Diretiva europeia relativa a dados abertos.

No entanto, de forma a concretizar a acessibilidade e a reutilização referidas, são necessárias determinadas condições, motivo pelo qual também são necessárias licenças; neste caso, as memórias de tradução deveriam ser tão livres quanto possível. O nosso objetivo prioritário é definir um plano de ação que favoreça a maior divulgação e capacidade de reutilização dos dados.

Recomenda-se vivamente que não se criem mais tipos de licenças; pelo contrário, dever-se-á simplificar os tipos existentes, conforme recomendado pela *Agenda Digital Europeia* (Comissão Europeia, 2010).

Atualmente, existem algumas alternativas internacionais principais que fomentam a utilização de licenças no contexto dos dados e das informações e que regulam e promovem o acesso e a utilização livres das informações. Estas licenças são: CC, Apache 2, GPL e ODC. A CC, em particular, aplica-se a dados e outros documentos, enquanto a ODC apenas é utilizada na área de dados. A tendência internacional indica que estão a ser utilizados modelos de licenças livres para regulamentar as condições de acesso e de utilização dos dados, principalmente, com base na CC, embora com

variações adaptadas às características de cada portal de dados e cada país em que se aplicam.

O objetivo é beneficiar as Administrações cada vez mais digitais, capazes de **mensurar o impacto das respetivas ações em termos de retorno social do investimento**, que é a meta a atingir por ações como a NEC TM (National and European Central Translation Memory) através deste Acordo-Quadro para a partilha de dados gerados por memórias de tradução de textos traduzidos pelas Administrações Públicas.

Por conseguinte, uma vez que não existe, atualmente, consenso sobre a forma de regulamentação do tratamento, da utilização e da partilha de dados a partir de memórias de tradução provenientes das necessidades de publicação da Administração Pública, poderia ser proposta uma opção abstencionista, fazendo referência à legislação específica existente de natureza geral em cada Estado-Membro. A melhor alternativa seria um regulamento particular a esta questão, sob a forma de um modelo regulamentar proposto com o presente documento e com base em “normas leves”, de natureza unilateral e não-vinculativa, e uma abordagem prática, como um ACORDO-QUADRO.

O FATDS visa ajudar ambas as partes envolvidas no processo: a Administração Pública e os fornecedores. Em primeiro lugar, clarifica que existe o problema da partilha de dados de memórias de tradução, que esta é uma realidade para as organizações envolvidas e que é necessário intervir, rejeitando excessos e abusos. Em segundo lugar, o FATDS propõe um plano de regulamentação complementar e específico, bem como uma gestão sistemática, que também é flexível e adaptável a todas as Administrações, entidades privadas e Estados-Membros.

Para este fim, recomenda-se vivamente, pois tal é necessário no interesse de uma sociedade que está cada vez mais preocupada com o valor dos dados que gera, que os dados sejam disponibilizados ao contratante responsável ou a um órgão central no território de cada Estado-Membro, com cada um dos contratos de prestação de

serviços de tradução realizados, ficando disponível não só o texto ou os documentos da tradução em questão, mas também a memória de tradução gerada neste processo.

Da mesma forma, graças à implementação da base de dados da NEC TM, são, assim, definidas condições para facilitar a reutilização destas memórias de tradução devido à sua acessibilidade. Acima de tudo, não estarão sujeitas a restrições técnicas ou legais que limitem ou impeçam a referida utilização.

3. A multidimensionalidade do problema da partilha de dados de memórias de tradução geradas por Administrações Públicas: a proteção oferecida pelo FATDS.

No sentido estrito da sua natureza jurídica, o presente Acordo-Quadro pretende funcionar como um acordo cuja implementação ocorrerá de acordo com as práticas e os procedimentos da experiência de cada Estado-Membro e tendo em consideração práticas que já são comuns e do conhecimento dos fornecedores na indústria da tradução.

Com base nas considerações supra, parece claro que não é possível reconhecer os Acordos-Quadro com uma eficácia jurídica direta como se fossem outro constituinte da legislação comunitária, neste caso, de uma fonte autónoma. Simultaneamente, também não é possível subestimar a sua eficácia. O acordo prende-se, puramente, com conformidade voluntária, estando estritamente dependente do poder dos signatários para vincularem as organizações.

Assim, o FATDS, visto como um acordo autónomo, tem em vista constituir um instrumento prático, isto é, um conjunto de diretrizes orientadoras que especificam regras efetivas para a gestão, a centralização e a partilha de dados, sendo o Acordo-Quadro estritamente complementar para a legislação nacional e comunitária.

Desta forma, um acordo não-normativo, como o FATDS, pode constituir (e constitui, efetivamente, por meio do presente) um instrumento útil de consolidação e interpretação esclarecedora da partilha dos dados gerados pelas memórias de tradução criadas em resultado do trabalho desenvolvido pelos proponentes/fornecedores.

A função do FATDS é, portanto, clarificar e especificar o direito das Administrações Públicas de pedirem todos os dados gerados em contratos de tradução e a respetiva propriedade dos textos originais e das traduções como um serviço contratado e derivado do original, refletindo o consenso das partes envolvidas, que se entende serem, por um lado, as empresas proponentes (os fornecedores) e, por outro lado, o órgão administrativo de cada Estado-Membro e, por fim, o órgão central da UE, que será determinado para este fim.

Neste sentido, deve ter-se em mente que a configuração mais reconhecida deste modelo ou paradigma de juridicidade ilustra a existência de instrumentos que, sem estarem conformes aos tipos tradicionais de normas jurídicas, não podem ser excluídos do mundo do Direito e, por isso, da natureza de obrigatoriedade, de modo a expressar um compromisso seguro dos signatários que realizam o contrato de prestação de serviços, no exercício das suas competências, ao abrigo de um procedimento previsto na lei, colocando em prática um instrumento provido de juridicidade mínima ou relativa.

O conteúdo fundamental do FATDS centra-se mais na determinação de modelos e diretrizes, constituindo um instrumento típico de Direito refletivo.

Em última análise, este paradigma de regulamentação materializado pelo FATDS é harmonizado com as novas abordagens e os novos métodos de regulamentação e ação atualmente propostos, incluindo pela UE, que se baseia mais em instrumentos que transformam a norma jurídica em prática, em vez de estabelecer mais normas, no nosso caso, definindo o acesso, a utilização e a possível reutilização das informações contidas nas bases de dados reguladas por lei, assim como através de contratos e/ou licenças.

A publicação da comunicação sobre dados abertos da CEE e a proposta da Comissão Europeia para correção compõem uma política de abertura de dados e promoção de um mercado de informações que tem como ponto central a aprovação da Diretiva 2003/98/CE do Parlamento Europeu e do Conselho, de 17 de novembro de 2003, sobre a reutilização de informações do setor público.

Para tal, temos de acrescentar a revisão da Diretiva sobre a reutilização de informações do setor público, introduzida no fim de 2011, que faz parte da Agenda Digital Europeia, uma iniciativa por parte da Comissão Europeia (2010). Esta Agenda destina-se favorecer os serviços online dentro da União e tem como prioridade a abertura de dados públicos para reutilização, a simplificação do sistema de licenças para a troca de conteúdos e a implementação de normas de interoperabilidade.

5. A opção regulamentar do FATDS

Os dados públicos detidos por organismos públicos na União Europeia estão sujeitos a um tratamento específico ao abrigo das Diretivas de 2003 e 2013 relativas à reutilização de informações do setor público, que estipulam a disponibilização dos documentos da Administração Pública ao público para reutilização de “qualquer conteúdo, seja qual for o seu suporte (documento escrito em suporte papel ou eletrónico, registo sonoro, visual ou audiovisual), na posse dos organismos do setor público para fins comerciais e não-comerciais”. Por conseguinte, o princípio geral é o da disponibilidade gratuita desse conteúdo e, no caso de serem aplicadas tarifas, estas “devem limitar-se aos custos marginais incorridos na sua reprodução, disponibilização e divulgação” (Art.º 6.º, alínea 1).

Em suma, o FATDS demonstra ser, basicamente, uma “recomendação” fundamentada numa juridicidade mínima ou relativa, com base nos contratos de prestação de serviços existentes e nos dados (memórias de tradução) gerados *ex professo*. Esta “recomendação” é a especificação prática do dever de partilhar dados gerados sob a

forma de memórias de tradução que resultem de serviços de tradução contratados pelas Administrações Públicas, independentemente do suporte ou formato.

6. Conteúdos do Acordo-Quadro: o sistema de gestão para a transferência de dados a partir das memórias de tradução promovido pelo acordo.

Uma memória de tradução é uma base de dados linguística que armazena continuamente traduções paralelas geradas por profissionais, de forma a poderem utilizá-las no futuro, com vista a obterem uma maior consistência terminológica e estilística, tal como poupanças devido às correspondências totais ou parciais entre novos textos e traduções antigas, que já foram concluídas.

Ademais, as memórias de tradução são repositórios digitais compostos de linhas de texto dos conteúdos na língua de partida alinhados com as respetivas traduções noutras línguas. Estes textos também podem ser alinhados de forma eficiente por meio de unidades de tradução. As unidades de tradução que são armazenadas em conjunto com os respetivos equivalentes definem-se de diferentes formas (por expressão, parágrafo, palavra ou grupo de palavras, etc.). A segmentação após um sinal de pontuação que assinala o fim da frase ou quebra de parágrafo é, geralmente, o separador mais frequente por predefinição no ambiente dos sistemas de tradução assistida por computador.

As principais funções das memórias de tradução (doravante denominadas TM) são extrair sugestões ou correspondências totais ou parciais de uma frase e fazer concordâncias de termos. Durante o processo de tradução, os segmentos da língua de partida são pesquisados na base de dados da TM. Se a TM contiver um segmento na língua de partida que corresponde exata ou parcialmente ao texto de partida, este segmento será sugerido ao tradutor, juntamente com a tradução obtida dessa base de dados e quaisquer informações adicionais que tenham sido guardadas com o segmento na base de dados. As ferramentas de tradução assistida por computador (conhecidas como *CAT tools*) apresentam graus de semelhança (na indústria, estas

correspondências são chamadas de *fuzzy matches*). Na indústria da tradução, estas correspondências podem ser úteis para os linguistas a partir dos 50-65 %. Além disso, são estabelecidas tabelas de preços de acordo com o esforço aplicado na tradução de frases completamente novas e de outras para as quais existem sugestões semelhantes e correspondências em percentagem, no mesmo contexto. Considera-se que um grau de semelhança de 100 % é uma correspondência completa entre uma frase solicitada e uma frase idêntica presente na base de dados.

O formato gratuito para a troca de memórias de tradução é o TMX (Translation Memory eXchange), geralmente, na versão 1.4b. Este é um modelo de XML do tipo DTD (*document type definition*). Foi criado pela comissão da OSCAR (Open Standards for Container/Content Allowing Re-use).

Através da aplicação do formato TMX, é mais praticável as pessoas e as empresas colaborarem em projetos de tradução. O formato TMX também facilita a migração de um sistema de tradução assistida por computador para outro, o que favorece a competitividade entre tecnologias oferecidas e o seu constante desenvolvimento, de modo a distinguirem-se da concorrência. À semelhança de outros modelos livres, este formato foi desenvolvido tendo em vista a redução de problemas de compatibilidade, o incentivo da reutilização de recursos linguísticos, a simplificação da troca de dados e, assim, o estímulo da inovação tecnológica.

O presente Acordo-Quadro pretende estabelecer, conforme temos vindo a definir, um “plano de boas práticas” para a obtenção de dados gerados em contratos de prestação de serviços de tradução por Administrações Públicas europeias, a respetiva organização através da implementação do sistema de software da NEC TM e as iniciativas de centralização de dados designadas pela Comissão Europeia com vista à adoção de um protocolo para a centralização de dados bilingues pelas Administrações Públicas, a respetiva utilização e as vantagens para a sociedade em geral, bem como para os próprios Estados-Membros, a assistência na criação de um *corpus* de Big Data nacional e a partilha das partes que cada Administração nacional considera relevante a nível europeu.

Para este fim, a centralização dos dados alojados em infraestruturas informáticas nacionais será gerida pelo órgão competente do Estado em questão (por exemplo, a Secretaria de Estado para os Avanços Tecnológicos em concreto) ou o Ministério competente no Estado-Membro em questão. Estes dados poderão ser partilhados, caso se adira a este Acordo-Quadro e se assim for determinado pela autoridade competente, num nível mais elevado com o órgão relevante da Comissão Europeia (p. ex., o ELRC-Share).

Dado o supra-exposto, o primeiro passo a dar será a centralização dos dados das memórias de tradução geradas pelas Administrações Públicas a nível nacional através da adoção do presente Acordo; posteriormente, de acordo com o mesmo, outro passo será a respetiva partilha com um órgão pan-europeu em benefício das Administrações Públicas de outros Estados-Membros da UE. A utilização do software da NEC TM fornecerá uma ligação de entrada/saída (E/S) às Administrações, que poderão armazenar, de forma privada, as respetivas memórias de tradução, alojadas num servidor central, e partilhá-las com quem decidir tornar o trabalho mais eficiente e rentável a nível nacional (geralmente, tradutores internos ou fornecedores externos), criar Big Data nacional e selecionar os dados que quiserem com o servidor central no qual a memória de tradução central europeia está alojada.

Relação com entidades privadas e atribuição de TMX

O trabalho de tradução desenvolvido por empresas proponentes (fornecedores) implicará o fornecimento das memórias de tradução geradas por elas em resultado da prestação dos serviços. Para tal, no aviso publicado no Diário Oficial do Estado-Membro, da região ou do município, isto é, o portal do contratante, deve ser incluída uma cláusula que mencione o FATDS da NEC TM, incluindo o CÓDIGO CPV, que determinará que OS DADOS DA MEMÓRIA DE TRADUÇÃO SERÃO FORNECIDOS AO SERVIDOR DA AUTORIDADE PÚBLICA COMPETENTE para o fim de, no futuro e se assim for entendido por essa autoridade nacional, poderem ser partilhados com o órgão da Comissão Europeia a ser determinado (o ELRC-Share ou iniciativa semelhante, por exemplo).

Com este fim em vista, as empresas proponentes (fornecedores) entregarão as memórias de tradução geradas juntamente com o respetivo trabalho, aplicando o princípio da não-retroatividade, o que significa que o presente Acordo-Quadro não pode ter efeitos retroativos e que apenas esta natureza de obrigatoriedade advirá da adesão e da assinatura da cláusula por meio da qual se comprometem a entregar, juntamente com o trabalho realizado, os dados paralelos gerados pelas respetivas traduções (em TMX ou formato compatível semelhante).

7. PROPRIEDADE INTELECTUAL E CONDIÇÕES DE UTILIZAÇÃO DA BASE DE DADOS

Relativamente à propriedade intelectual, não podem subsistir dúvidas de que os dados gerados na realização do trabalho de acordo com o contrato que atribuiu a tradução em questão são, inteira e exclusivamente, da Administração Pública que é autora ou gestora do texto original sujeito a tradução.

Na medida em que, na maioria das vezes, as informações têm origem nas Administrações Públicas, não se pode ignorar o facto de que a regulamentação da reutilização das informações públicas e a abertura de dados públicos também têm de ser tidas em consideração.

Deste modo, o proponente (fornecedor) bem-sucedido atribuirá, exclusivamente e sem prazo ou limites territoriais, os direitos sobre qualquer tipo de documentação ou dados gerados, independentemente do respetivo suporte ou formato, sendo abrangidas nesta transferência de direitos de reprodução a distribuição e a transformação.

Por conseguinte, os dados gerados por essas memórias de tradução não poderão ser utilizados pelo adjudicatário para fins lucrativos e apenas poderão ser utilizadas para fornecer informações para o trabalho desenvolvido, se este for o caso.

Propriedade intelectual e condições de utilização da base de dados.

A base de dados da DGT-TM (Memória de tradução da Direção-Geral da Tradução) é propriedade exclusiva da Comissão Europeia. A Comissão concede aos reutilizadores, gratuitamente e a nível mundial, os respetivos direitos não-exclusivos, pelo período de proteção dos mesmos, sobre todos os tipos de utilizações que cumpram as condições estabelecidas na Decisão da Comissão Europeia, de 12 de dezembro de 2011, relativamente à reutilização dos documentos da Comissão, publicada no Jornal Oficial da União Europeia, L 330, de 14 de dezembro de 2011, páginas 39 a 42.

Qualquer reutilização da base de dados ou dos elementos estruturados contidos na mesma terá de ser identificada pelo reutilizador, que é obrigado a indicar a fonte dos documentos utilizados: o endereço do Web site, a data da última atualização e o facto de a Comissão Europeia preservar a propriedade dos dados.

Portanto, esta base de dados é ideal para povoar a versão da NEC TM de um Estado-Membro numa fase inicial.

8. PROTEÇÃO DE DADOS. APLICAÇÃO DO REGULAMENTO (UE) 2016/679 E DO REGULAMENTO 2018/1725

No que diz respeito aos dados pessoais que têm de ser tratados no âmbito do cumprimento do contrato pela Administração Pública e o proponente bem-sucedido, ambas as partes são obrigadas a cumprir os seguintes regulamentos gerais em matéria de proteção de dados:

O Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho, de 27 de abril de 2016, além da conformidade com a legislação estatal atual em matéria de proteção de dados em cada um dos Estados-Membros que adotem o presente Acordo-Quadro, tal como o Regulamento do Parlamento Europeu e do Conselho, de 23 de outubro de 2018, sobre a proteção dos indivíduos relativamente ao tratamento de dados pessoais por instituições, órgãos e organismos da União Europeia e à circulação dos referidos dados.

CONCLUSÕES.

Na qualidade de promotora da indústria das tecnologias da linguagem, com a criação de plataformas comuns para o processamento da linguagem natural e a tradução automática e com o desenvolvimento de recursos para a reutilização de informações do setor público, a Administração Pública é obrigada a desenvolver políticas sobre a partilha de dados e a definir as bases para esta partilha se tornar real e efetiva entre todas as entidades que participam no processo, assegurando, também, que os participantes no processo fora da Administração assumem o compromisso de assinar a cláusula de aceitação das respetivas propostas e se comprometem a entregar, juntamente com o respetivo trabalho, os dados gerados pelas suas traduções (memórias de tradução), criando, tudo isto, um conjunto de recursos linguísticos de valor incalculável.

REFERÊNCIAS. Ferrer-Sapena *et al.* (2011) no seu artigo sobre o acesso a dados públicos; Ferrer-Sapena e Peset (2012) sobre a reutilização de dados culturais ou Ramos Simón *et al.* (2012) no seu estudo sobre os portais de dados europeus.

Licensing Open Data: A Practical Guide, para o Higher Education Funding Council for England (Korn e Oppenheim, 2011). Também são dignos de nota o *Guide to Open Data Licensing* (Open Knowledge Foundation, n.d.) e as diretrizes da *Agenda Digital Europeia* (Comissão Europeia, 2010).