Framework Agreement for the sharing of translation memory data generated as the result of translations of Public Administration documents.

1. INTRODUCTION.

Since the adoption in 2003 of the first set of rules on the re-use of public sector information, the volume of data generated has increased exponentially worldwide, while new types of data are being generated and collected.

At the same time, we are witnessing a permanent evolution of language processing technologies for data analysis, exploitation and processing.

This rapid technological evolution allows the creation of new services and applications based on the use, aggregation or combination of data.

The massive amounts of data or big data collected through different technological tools or extracted from large collections of information in different formats are capable of generating new knowledge in the most diverse sectors, but, at the same time, pose problems both for data ownership and for its subsequent use.

Public Administrations (PA's) and the bodies and entities that make up the public sector are major producers of data which is of great use for (natural) language processing industries. Public Administrations generate enormous amounts of data in the undertaking of their duties. The use and exploitation of that data can be of great interest both to the PA itself and to the industry.

The use of these data by Public Administrations combined with data mining and big data techniques and natural language processing technologies can facilitate public decision-making and the effectiveness of public policies.

This Framework Agreement wishes to be a recommendation to Member States for the implementation of open licenses that remove possible legal obstacles in order to incorporate these types of data to the creation of new knowledge.

It has become a current practice for different governments, corporations and international organizations producing public data to rely on these licenses -in particular, the PPDL, GPL and Creative Commons (CC) modalities in which all rights to databases are waived- as an appropriate mechanism to favour the use and reuse of these large volumes of data.

2. Why a European Framework Agreement for Data Sharing management.

Reasons for the framework agreement for sharing data from translation memories.

2.1. General approach.

The FATDS (Framework **Agreement** on **TMX** Data **Sharing)** is neither coincidental nor purely episodic or punctual. On the contrary, it responds to a need that finds various motivations. European Public Administrations generate thousands of translations every day, and those translations generate databases whose release as data is already a legal mandate (Open Data Europe Directive).

There are two points of great relevance which reflect the lack of a basic consensus for adequate regulation and the management of the problems linked to the sharing of the data generated by translations of the texts coming from public administrations. This is **one of the sectors that generates the greatest amount of data** of this kind during their daily business. It is also the sector that invests the most in data storage infrastructures worldwide. We focus particularly on the data generated by the translation memories of documents owned by the Public Administration, given the needs of the administered, and the numerous languages not only within a territory or country, but in all the countries that make up the EU.

It is very important to comply with the recommendations for data reuse and interoperability in the Public Administration sector. Just as important is the laying out of the conditions for the reuse and interoperability.

Without a doubt, the optimal state would be for the data to be fully accessible and for its reuse not to require specific permits, as established by the European Directive on Open Data.

However, in order for accessibility and resue to happen, certain conditions are required, and this is why licenses are necessary - in this case translation memories should be as open as possible. It is our priority objective to establish a framework for action that favours its greater dissemination and capacity for reuse,

It is more than advisable not to generate more types of licences, but quite the opposite, simplifying the types as recommended by *the European Digital* Agenda (European Commission, 2010).

Currently, there are two main international alternatives that promote the use of licenses in the context of data and information, and that regulate and promote the free access and use of information. These licenses are Creative Commons (CC), Apache 2 and GPL and ODC. Creative Commons applies to both data and other documents, while the latter are only used in the data area. <u>The international trend indicates that to regulate the conditions of access and use of the data, open license models are being used, based mainly on CC, although with variations adapted to the characteristics of each data portal and each country where they are applied.</u>

All of this is to the benefit of an increasingly digital Administration that is capable **of measuring the impact of its actions in terms of social return on investment,** which is the aim of actions such as the National and European Central Translation Memory (NECTM) through this Framework Agreement for sharing data generated by translation memories from texts translated for Public Administrations.

Therefore, since there is currently no consensus on how to regulate the processing, the use and the sharing of data from translation memories generated as the result of the

Public Administration publication needs, either an abstentionist option could be proposed, referring to specific existing laws of a general nature in each Member State. The better alternative is a particular regulation for this issue, in the shape of a regulatory model herewith proposed and based on "soft standards", of a unilateral, non-binding nature, and a practical approach such as a FRAMEWORK AGREEMENT.

FATDS takes sides for both sides involved in the process, the PA and the vendors. Firstly, it makes it clear that the translation memory data sharing problem exists, that it is real for the organizations involved and that it is necessary to intervene, while rejecting excesses and abuses. Secondly, FATDS proposes a complementary and specific framework for regulation, as well as a systematic management, which is also flexible, adaptable for all Administrations, private entities and Member States.

To this end, it is highly recommended and necessary in the interest of a society that is increasingly concerned with the value of the data it generates that the data be made available to the responsible contractor or a central body within each Member State, with each of the translation service contracts that are carried out, making available not only the text or documents of the translation in question but also the translation memory generated as a result of the process.

Likewise, thanks to the implementation of the NEC TM database, conditions are therefore established to facilitate the reuse of these translation memories thanks to their accessibility, and above all that they will not be subject to technical or legal restrictions that limit or hinder such reuse.

3. Multidimensionality of the problem of data sharing of translation memories generated by Public Administrations: protection offered by FATDS.

In the strict sense of its legal nature, this Framework Agreement is intended to be an Agreement whose implementation will take place in accordance with the practices and procedures of each Member State experience and taking into account practices that are already common and not unknown to vendors in the translation industry.

On the basis of the above considerations, it seems clear that it is not possible to recognise the Framework Agreements with a direct legal effectiveness as if it were just another piece of Community legislation, in this case from an autonomous source, but it is also not possible to underestimate its effectiveness. The agreement is of purely voluntary compliance, strictly dependent on the power of the signatory subjects to bind the organizations.

Thus, FATDS understood as an Autonomous Agreement aims to be rather a practical instrument, that is to say, a set of guiding guidelines that specify effective rules for the management, centralisation and sharing of data, the Framework Agreement being strictly complementary to Community and national legislation.

In this way, a non-normative Agreement, such as FATDS, can constitute, and hereby constitutes, a useful instrument of concretion and clarifying interpretation of the sharing of the data generated by the translation memories created as a result of the work carried out by the bidders/vendors.

The function of FATDS is therefore to clarify and specify the right of Public Administrations to request all the data generated in translation contracts, their ownership of the original texts and translation as a service contracted and derived from an original, reflecting the consensus of the actors involved, understanding as those involved on the one hand the bidding companies (the vendors) and on the other the administrative body from each Member State and finally the central body of the EU that will be determined for this purpose.

In this sense, it should be remembered that the most accredited configuration of this model or paradigm of juridicity illustrates the existence of instruments that, without conforming to traditional types of legal norm, cannot be excluded from the world of law either, and therefore from the obligatory nature, in order to express a firm commitment of the signatory subjects who carry out the contract of services in the exercise of their competences under a legally foreseen procedure, putting into practice an instrument endowed with a minimum or relative juridicity.

The fundamental content of FATDS focuses more on the establishment of models and guidelines, forming a typical instrument of Reflective Law.

Ultimately, this paradigm of regulation materialized by FATDS is framed flat within the new approaches and new methods of regulation and action currently proposed, as well as by the EU, which relies more on instruments that transform the legal rule into practice rather than establishing more rules, in our case defining the access, use and possible reuse of the information contained in the databases regulated by law, as well as through contracts and/or licenses.

From the EEC, the publication of the communication on open data and the European Commission's proposal for amendment form a policy of opening up data and promoting an information market that has its central point in the approval of Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information.

To this we must add the revision of the Directive on the Reuse of Public Sector Information, introduced at the end of 2011, which forms part of the European Digital Agenda, an initiative undertaken by the European Commission (2010), this Agenda being aimed at favouring online services within the Union, and having as a priority the opening up of public data for reuse, the simplification of the licensing system for the exchange of content and the implementation of interoperability standards.

5. FATDS' regulatory option

Public data held by public bodies in the European Union are subject to specific treatment under the 2003 and 2013 Public Sector Information Reuse Directives, which provide for

PA documents to be made available to the public for reuse "any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording) kept by public sector bodies for commercial or non-commercial purposes". Therefore, the general principle is the free availability of such content, if any tariff is applied, it "shall be limited to the marginal costs incurred for its reproduction, making available and dissemination" (art. 6.1).

In short, FATDS is basically revealed as a "Recommendation" based on a minimum or relative juridicity, based on existing service contracts and data (translation memories) generated ex profeso. This "Recommendation" is the practical specification of the duty to share data generated in the form of translation memories originating from translation services contracted by Public Administrations, whatever their medium or format.

6. Contents of the Framework Agreement: the management system for the transfer of data from the translation memories promoted by the agreement.

A translation memory is a language database that continuously stores parallel translations generated by professionals in order to be able to use them in the future with a view to obtaining greater terminological and style consistency, as well as savings due to total or partial matches between new texts and old translations that have already been completed.

Translation memories are digital repositories composed of lines of text from the content in the original language aligned with its translation in other languages. These texts can also be efficiently aligned by translation units. The translation units that are stored together with their equivalents are defined in different ways (by phrase, paragraph, word or group of words, etc.) the segmentation after a punctuation mark marking the end of the sentence or a paragraph return being typically the most frequent separator by default in the environment of computer-assisted translation systems.

The main function of translation memories (hereinafter TMs) is to extract total or partial suggestions or matches from a sentence and concordances for terms. During translation, segments of the source language are searched in the TM database. If the TM has a segment in the source language that matches exactly or partially, this segment will be suggested to the translator, along with the translation retrieved from that database, and any additional information that has been saved with the segment in the database. Computer-Assisted Translation tools (CAT tools) show degrees of similarity (known in the industry as *fuzzy* matches). In the translation industry, these matches can useful for linguists from a range starting at 50%-65% up, and price scales are also established for the effort involved in translating completely new phrases or others for which there are similar suggestions and percentage matches, within the same context. A 100 % degree of similarity is considered to be a complete match between a requested sentence and an identical sentence in the database.

The free format for the exchange of translation memories is TMX (Translation Memory eXchange), generally in version 1.4b. This is an XML standard of the type DTD *(document type* definition). It was created by the OSCAR committee (Open Standards for Container/Content Allowing Re-use).

Through the application of the TMX format, it is more feasible for people or companies to collaborate in translation projects. The TMX format also makes it easier to migrate from one Computer-Assisted Translation system to another, which favours competitiveness between the technologies offered and their constant development in order to make a difference with respect to their competitors. Like other open standards, this format was developed with a view to reducing compatibility issues, driving reuse of linguistic resources, simplifying the exchange of data and thus stimulate technological innovation.

This Framework Agreement aims to establish, as we have been defining well, a "framework of good practice" for obtaining data generated in translation service contracts by European Public Administrations, its organisation through the implementation of the NEC TM software system and the data centralisation initiatives designated by the European Commission with a view to the adoption of a Protocol for the centralisation of bilingual data by Public Administrations, its use and benefit to society in general and the Member States themselves, assistance in the creation of a corpus of national Big Data and the sharing of the parts that each national Administration deems relevant at European level.

To this end, the centralization of data hosted on national computer infrastructures will be managed by the competent body of the State in question - for example, the specific State Secretariat for Digital Advancement - or the competent Ministry in the Member State in question. These data may be shared, in case of accession to this Framework Agreement and if so determined by the competent authority, at a higher level with the relevant body of the European Commission (e.g. ELRC-Share).

Having said this, the first of the steps will be the centralisation of the data from the translation memories generated by the Public Administrations at a national level through the adoption of this Agreement, and after that, and in accordance with it, their sharing with a pan-European body for the benefit of the PAs of the other EU Member States. The use of the "National and European Central Translation Memory" (NEC TM) software will provide the input/output (I/O) connection for administrations, who will be able to privately store their translation memories, hosted on a central server and share them with those who decide to make their work more efficient and cost-saving at national level (typically in-house translators or external vendors), create national Big Data and select the data they wish with the central server where the European central translation memory is housed.

<u>Relationship with private entities and assignment of TMX</u>

The translation work carried out by bidding companies (vendors) will entail the supply of the translation memories generated by them as a result of the provision of their service. To this end, in the notice appearing in the Official Gazette of the Member State, region or municipality, i.e. in the contractor's portal, a clause should be included referring to the NEC TM FATDS, including the CPV CODE, which will determine THAT THE DATA OF THE TRANSLATION MEMORY WILL BE SUPPLIE TO THE SERVER OF THE COMPETENT PUBLIC AUTHORITY for the purposes that in the future, and if so deemed by that national authority could be shared with the body of the European Commission to be determined (ELRC-Share or similar initiative, for example).

For this purpose, the bidding companies (vendors) will deliver the translation memories generated together with their work, applying the principle of non-retroactivity, meaning that this Framework Agreement must not have backwards effects in time, and that only this obligatory nature will operate from the adhesion and signature of the clause by which they undertake to deliver, together with the work carried out, the parallel data generated by their translations (TMX or similar compatible format).

7. INTELLECTUAL PROPERTY AND DATA BASE CONDITIONS OF USE

With regard to intellectual property, there can be no doubt that the data generated in the performance of the work in compliance with the contract, for which the translation in question has been awarded, is entirely and exclusively that of the Public Administration which is the author or manager of the original text that is the subject of the translation.

The fact cannot be ignored that to the extent that most of the time the information originates in the Public Administrations, the regulation of the reuse of public information and the opening of public data must also be borne in mind.

Thus, the successful bidder (vendor) will assign exclusively and without any time or territorial limits, the rights of any type of documentation or data generated independently of their support or format, understanding distribution and transformation in this transfer of rights of reproduction.

Consequently, the data generated by such translation memories may not be used by the successful tenderer for any lucrative purpose and may only be used to provide information if this is the case for the work carried out.

Intellectual property and databases conditions of use.

The DGT-TM database is the exclusive property of the European Commission. The Commission grants, free of charge and on a worldwide basis, throughout the period of protection of these rights, its non-exclusive rights to re-users for all types of uses meeting the conditions set out in the Commission Decision of 12 December 2011 on the re-use of Commission documents, published in the Official Journal of the European Union L 330 of 14 December 2011, pages 39 to 42.

Any reuse of the database or of the structured elements contained therein must be identified by the reuser, who is obliged to indicate the source of the documents used: the website address, the date of the last update and the fact that the European Commission retains ownership of the data.

This database is therefore optimal for initially populating the NEC TM version of a Member State.

8. DATA PROTECTION. APPLICATION OF REGULATION EU 2016/679 AND REGULATION 2018/1725

In relation to personal data that have to be processed on the basis of the fulfilment of the contract by the Public Administration and the successful bidder, both parties shall be obliged to comply with the below General Regulations on Data Protection.

Regulation (EU) 2016/679 of the European Parliament and of the Council, of 27 April 2016, as well as compliance with current state legislation on data protection in each of the Member States that adopt this Framework Agreement, as well as the Regulation of the European Parliament and of the Council of 23 October 2018 on the protection of individuals with regard to the processing of personal data by the institutions, bodies, offices and agencies of the Union and on the circulation of such data.

FINAL CONCLUSIONS.

The Public Administration as promoter of the language technologies industry, with the creation of common platforms for natural language processing and automatic translation and the development of resources for the Reuse of Public Sector Information (RISP), is obliged to develop data sharing policies and to lay the foundations for this sharing to be real and effective among all the entities that participate in the process, also ensuring that participants in the process outside the Administration acquire the commitment to sign the clause of acceptance of their tenders, to undertake to deliver along with their work, the data generated by their translations (translation memories) forming all of them a set of linguistic resources of invaluable value.

REFERENCES. Ferrer-Sapena and others (2011) in their article on access to public data; Ferrer-Sapena and Peset (2012) on the reuse of cultural data or Ramos Simón and others (2012) in their study on European data portals.

Licensing Open Data: A Practical Guide for the Higher Education Funding Council for England (Korn and Oppenheim,2011). Also noteworthy are the *Guide to Open Data Licensing* (Open Knowledge Foundation, n.d.) and the guidelines emanating from the *European Digital Agenda* (European Commission, 2010).