

From Server-based to Web-based Translation Memory Systems: Benchmarking, Testing and Implementation in STAR7

Enrico Gai¹, Nicola Poeta² and David Turnbull³

¹STAR7, S.p.A., Language Technologies Expert, Corso Orbassano 336, Torino, Italy

²STAR7, S.p.A., Global Content, Service Line Leader, Corso Orbassano 336, Torino, Italy

³STAR7, S.p.A., Language Lead, Corso Orbassano 336, Torino, Italy

Keywords: Corporate Language Management, TMS, Transit NXT, WebEdit.

Abstract: In the age of cloud computing and Software as a Service (SaaS), the need for web-based translation management solutions is on the rise. Connecting authoring tools and other Content Management Systems (CMSs) to Translation Management Systems (TMSs) is key to reaching a global audience quickly, effortlessly and efficiently. To this end, new web-based TMSs have been developed to automate the entire information lifecycle and allow cooperation among all stakeholders – from authoring to translation, review and publishing. In this paper we describe the features of a web-based TMSs, identify the contexts in which such a system is required and why, and look at their benefits. To do so, we will describe three different processes which use three different technologies based on varying automation and collaboration needs, from a server-based installation of Transit NXT to a web-based solution such as CLM WebEdit.

1 INTRODUCTION

Computer-Assisted Translation (CAT) tools have been a part of the translation industry since the late 1980s. For instance, STAR Group's STAR Transit is the second-oldest commercial-grade CAT software ever created and it is the oldest one still on the market.

In its essence, a CAT tool is made up of four main components:

1. A Translation Editor, where linguists can see the source texts and insert target translations;
2. A Translation Memory (TM) system which enables pre-translation¹ and fuzzy matching² between the source files and previous translations, as well as concordance searches;
3. A Termbase (TB) Management System in which specialised glossaries are created, shared and maintained;
4. A Quality Assurance (QA) system, which allows for advanced linguistic and structural conformity checks.

From the early days up until the late 2000s, a standard CAT tool would consist of a piece of software installed on the end user's personal computer, typically a Windows PC.

This approach was based on a software licencing model, in which users had to acquire a physical copy of the software or download it from the Internet; then, the software was activated with a licence number.

This model is still in use today. However, it has some disadvantages:

- The CAT tool's performance is strictly dependent on the local machine's computing power;
- Translation assets – typically language pairs (source text and target translations), TMs, TBs, and any other reference material – must be packed and transferred via e-mail³ to the linguists and back to the project managers;
- Real-time collaboration is not possible, as each linguist is tied to the files on their local machine;
- Project automations and connectors with authoring systems cannot be set up, meaning that files

¹ 'Pre-translation' is the 100% recovery of translations from previously translated texts, independent of source-file formats.

² A 'fuzzy match' is a TM suggestion deriving from similar already translated texts.

³ This can also cause data integrity and security issues, as discussed in subsequent paragraphs.

must be manually imported and delivered to clients;

- Client portals and project monitoring dashboards cannot be implemented in manual workflows.

A step forward in this approach is the use of server-based solutions. Many of the available CAT tools are also offered in a ‘server’ version, which can be installed and run from a Windows Server machine. This is beneficial to anyone in an organisation who has access to that server, for instance via a Virtual Machine (VM). In fact:

- CAT tools installed in a server environment can typically leverage higher computational power and better reliability.
- Translation assets can be transferred to a database and shared across the network.
- Real-time collaboration is possible for anyone who can access the VM.
- While running 24/7 on a dedicated server, integrations can be created and run as a service to allow for automated project creation and delivery via APIs or hotfolders⁴.
- Client portals and project monitoring dashboards can be implemented.

While this architecture is an improvement, adding a web-based layer to the system would allow for greater flexibility: the systems could be accessed directly via a web browser, without the need to physically install anything on the user system and without the need for setting up VMs and remote accesses. In this context, the only real requirement is a reliable Internet connection – everything else is taken care of remotely.

In this paper we will describe how STAR7’s CAT technologies have evolved from server-based to web-based, and what the benefits of this system are in terms of productivity by analysing actual implementations of these technologies in the production environment.

2 RELATED WORK

The benefits of cloud computing and web-based CAT technologies in general were extensively described by Muegge back in 2012. In his article, the translation

technologies expert explains the benefits of a centralised Translation Memory System (TMS):

- No need to install any application
- Software is always updated
- Cross-platform compatibility
- Ease of access and enhanced collaboration
- Low cost

Nonetheless, some critical drawbacks should be considered as well:

- Constant, good quality Internet connectivity is essential
- Privacy and confidentiality issues
- Control over and ownership of linguistic assets

Muegge concludes that the advantages far outweigh the drawbacks – and this is particularly true considering the advances of the last ten years. According to data from Eurostat, as of 2021, 90% of EU households now have access to a broadband Internet connection, a 17% improvement compared with data from 2012. As for data protection and privacy concerns, the General Data Protection Regulation (GDPR) established in 2016 has helped in ensuring the protection of personal data and privacy in the European Union and in the transfer of such data outside the EU.

The inefficiency of the ‘traditional desktop paradigm’ compared to SaaS solutions has also been explained by Zydrón. In his article, the focus is set on the benefits of a centralised SaaS solution not only for Language Service Providers (LSPs), but also for all the stakeholders in a translation project. In addition, in such a solution, the CAT tool would only be a part of a much wider system of integrated components aimed at automating tasks. As he states, “A proper SaaS TMS/CAT solution will allow you to integrate a customizable web-based ordering and payment system into your own website so that customers can upload, get quotes, pay online and initiate the workflow for translation jobs.”

The translation industry in general is also moving towards a cloud-based approach. According to a survey run by Nimdzi in 2020, the majority of TMS installations are now on the cloud – be it private or public.

In this paper we will move forward from this and discuss how the implementation of systems with varying degrees of centralisation and process automation has been beneficial in practical terms.

⁴ A hotfolder is an FTP folder that is constantly monitored by a service to detect changes in its content and trigger automated actions.

3 METHODOLOGY

As a member of the STAR Group Network – the world’s thirteenth largest Language Services Provider (LSP) in terms of revenue in 2022 – STAR7 has been using STAR Group’s technologies since its foundation in 2000. Therefore, the technology selection benchmarking has been conducted internally and on proprietary technologies.

The results described in this paper outline STAR7’s evolution in terms of translation technologies and how this has been beneficial in terms of workflow automation and scalability. In particular, the paper reports how the implementation of the Corporate Language Management (CLM) system⁵ has helped STAR7 to effortlessly manage millions of words per year, and how the WebEdit Web-based component is helping to move the bar even further.

This said, the three models presented in the results are still all in use at STAR7: each automation level has its own pros and cons, and the technology selection for each client is evaluated based on volumes, automation potential and turnaround requirements.

4 RESULTS

4.1 Model 1 – Server-based Installation: Transit NXT

In this scenario, Transit NXT (STAR Group’s proprietary CAT tool) is installed on a Windows Server machine that can be accessed by all project managers and any internal linguists or QA specialists. The server also connects to the Terminology and Translation Memory SQL Server databases, to centralise both terminology and translation memories.

While this approach is beneficial as it allows for a higher level of centralisation, the resources can only be accessed internally within the organisation. The ‘thin’ clients⁶ can access the Terminal Server installation and TM and TB assets only via VMs, so access to the local area network is required. For instance, external linguists are excluded from this architecture.

Translations are sent in the form of Project Package Files (PPFs) containing language pairs, TM extracts containing only reference segments useful for the project, TBs, and any other additional supporting

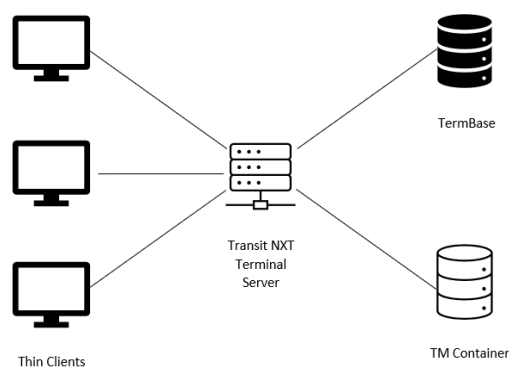


Figure 1: Terminal Server Architecture.

material. This package is typically shared with linguists via e-mail, who can then unpack the project files to their local Transit NXT installation.

Linguists can return their Translation Package Files (TPFs) to project managers, who can receive them, export the completed files, and update the centralised TMs with the finished translations.

4.2 Model 2 – Server-based Installation with FTPS File Transfer: CLM and WebTransit

A second level of automation has been implemented with the introduction of the CLM system. This is an advanced Translation Management System that can be accessed via web interface and lets clients create new translation requests or even connect to third-party Content Management Systems (CMSs) or hot-folders to automate translation requests.

From there, projects are created automatically using standardised project templates, and the system engages a Transit NXT command line session to automate the project creation, file import, package creation and delivery to linguists.

In this automated workflow, linguists receive a system notification when a new translation is requested. Instead of sharing them via e-mail, packages are sent via a secure FTPS connection and received via WebTransit – Transit NXT’s integrated component to receive and distribute translation jobs created in CLM.

Translated jobs can be uploaded to the system via WebTransit itself, and the CLM system handles both the update of the TMs and the final file delivery to the clients, who can retrieve their finished jobs either via their CLM portal or in SFTP folders.

⁵ For further information, see: <https://www.star-group.net/en/downloads/star-clm.html>

⁶ A ‘thin’ client is any PC that accesses a main server to perform the actual computational operations. In this

case, the PCs access a server where Transit NXT is installed, and all operations are carried out in that environment.

The benefits of this approach are the higher level of automation, centralisation and scalability the system offers. Nonetheless, the system still requires linguists to have Transit NXT installed on their local machines.

4.3 Model 3 – Web-based Installation: CLM and Webedit

In the third evolution phase, the CLM system is improved upon by offering a fully fledged online Translation Editor in the form of the WebEdit module.

In this framework, the capabilities of CLM in terms of workflow automation are combined with the benefits of online translation, in that translators and reviewers are no longer tied to software residing on a local PC. This is valuable in other ways, too:

- Data security is ensured, as translation files are no longer saved to local machines but on a centralised server.
- Data integrity is also secured, as the risk in terms of data loss is lower and a Disaster Recovery Plan (DRP) can be implemented in case of disaster.
- Linguists can start working on one machine and resume working on another one without the need to transfer data.
- Linguists using an operating system other than Windows can still be involved in translation projects, as the web-based system is OS-independent.

In other words, CLM WebEdit preserves all the benefits of CLM described in Model 2, while adding online translation and review functionalities. In this framework, the TMS can also be useful on the client side, as the review step can be handled internally without the need for additional software installs.

As the system is accessible not only to linguists – who are used to working in a TMS environment – but also to non-experienced users, a great deal of effort has been made to provide a simplified, user-friendly experience while preserving all the core functionalities of a TMS. Features like a quick access toolbar (1), the language pairs in column form (2), the terminology and QA windows (2, 3) as well as the Fuzzy matching and Concordance search windows (5, 6) are all present and easily accessible, as shown in the figure 2.

User training and feedback have shown that the learning curve for the use of WebEdit is particularly favourable: both linguists and clients have reported ease of use and clarity among the most appreciated features.

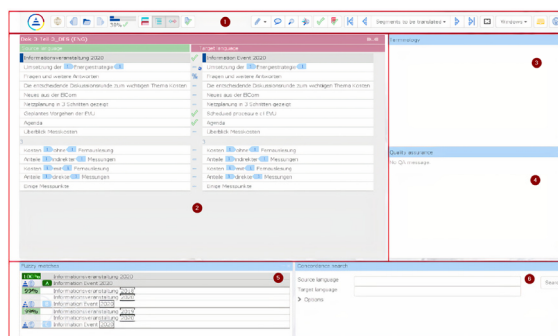


Figure 2: WebEdit’s Translation Editor.

4.4 Model 4 – Cloud-based Installation: Future Developments

While the approach in Model 3 is an improvement in relation to the first two models, it can still be improved upon. At the moment, a server infrastructure is needed for the installation of the system components, which may be an issue in terms of costs and security concerns. Work is underway to make CLM and WebEdit deployable as cloud-based systems, available both in services like Azure and AWS. This would mean outsourcing the computing power to commercial-grade IT infrastructures for even higher reliability and performance.

4.5 Implementation

The server-based solution has been implemented since 2014 to automate the translation workflow of a top-tier client in the Truck & Bus sector. In this scenario, the client’s authoring system saves XML files for translation in a shared FTPS hotfolder, which is constantly monitored by the CLM FTP component. New files in the hotfolder trigger the creation of new projects by using project templates, which are selected based on codes in each file name. Translation and review jobs are sent to selected translators based on a ranking logic, and the completed files are saved back to the FTPS folder, ready to be handled by the client’s authoring system.

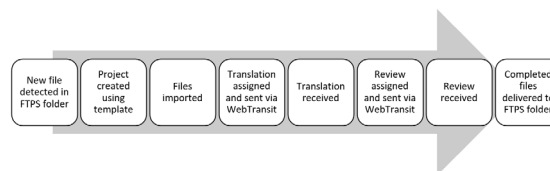


Figure 3: Server-based workflow automation with FTPS hotfolder.

The number of words processed for the client has been rising ever since implementation in 2014 and has now reached approximately 520 million words per year. This volume would not be feasible in a traditional environment, both in terms of workforce and in terms of system capabilities: for instance, the main TM for the customer now contains 80 million unique segments in more than 35 languages. The result is an exceptionally high level of leverage in terms of recovery from existing translations.

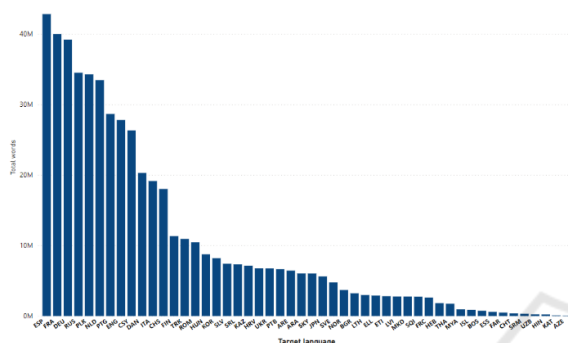


Figure 4: Total words managed for the client in STAR CLM, by target language (2021).

The system has also been used to manage the translation requests of clients ranging from the Agriculture sector to Automotive, as well as clients in the Home Appliances and Fashion sectors, proving that the system is well suited to all kinds of workflows. In situations where no connection between CMS and TMS needs to be established, clients can also use CLM's Client Portal to upload translation requests and trigger the same automated workflows.

With the release of CLM WebEdit in 2020, STAR7 has been hard at work migrating existing clients from the previous server-based system to the new web-based solution. In addition, new clients have been migrated to WebEdit, as the online translation and review module is also attractive in terms of in-house client review. In that respect, a client in the field of Sports & Fitness has successfully implemented the CLM WebEdit solution to request InDesign catalogue translations from STAR7 while managing internal market reviews by using WebEdit. Previously, the final step was performed using comments in PDFs, in which the client would report corrections that STAR7 had to make in both the target files and in the TM. Using WebEdit has drastically improved productivity, as corrections are now directly implemented in the working files and in the TM.

5 CONCLUSION

In this paper we set out to describe different automation models and workflows using STAR7's translation technologies. As translation processes require higher automation levels and translation volumes grow higher, the need for reliable, structured and scalable solutions grows consequently. This is why STAR7 decided to adopt the server-based model first, and the web-based model later, to 'future-proof' translation workflows. As new technologies and IT architectures are developed, research activities are constantly pushed forward to optimise translation workflows and attract existing or prospective clients with additional features and processes aimed at simplifying tasks that could otherwise be automated. An area that is currently under development is that of Machine Translation (MT) and Post-Editing Machine Translation (PEMT) workflows, which have been successfully implemented in CLM using STAR MT technology as well as commercial MT engines.

Potential future developments can be made – as already mentioned – in cloud computing and in implementing Artificial Intelligence models to improve upon existing processes that are still human-driven, to assist the many actors in the translation industry.

REFERENCES

- Eurostat Data on Households with Broadband Access*, <https://ec.europa.eu/eurostat/data-browser/view/tin00073/default/bar?lang=en>, last accessed 2022/08/04.
- Muegge, U. (2012). The silent revolution: Cloud-based translation management systems. In *tcworld*, July 2012, 17–21.
- Nimdzi 100: The 2022 ranking of the largest language services providers in the world*, <https://www.nimdzi.com/nimdzi-100-top-lsp/>, last accessed 2022/08/04
- Nimdzi History of TMS Technology*, <https://www.nimdzi.com/the-history-of-tms-technology-from-the-80s-to-2010>, last accessed 2022/08/04.
- Nimdzi Language Technology Atlas*: <https://www.nimdzi.com/nimdzi-language-technology-atlas-2020>, last accessed 2022/08/04.
- Official Journal of the European Union, Regulation (EU) 2016/679*, <https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN>, last accessed 2022/08/04.
- Zydroń, A. (2012): Perspectives: Cloud computing, SaaS and translation tools. In *MultiLingual*, January 2012, pp. 20–21.