MT Bazaar: Translation Ecosystems in the Cloud

Sven Christian Andrä
Andrä AG
Torstraße 61
D-10119 Berlin
`sca@andrae-ag.de`

Jörg Schütz
bioloom group
Bahnhofstraße 12
D-66424 Homburg
`joerg@bioloom.de`

Abstract

In this paper, we introduce and discuss various elements of emerging translation ecosystems that facilitate new community-based, collaborative and crowdsourced approaches and business models based on cloud computing technology platforms, smart marketplaces with aggregation, recombination and hyperdistribution, and that make the online world and its user services exponentially better than the simple sum of their individual parts. Besides the elaboration of the methodological and computational frameworks, our discussion also includes two succinct case studies that show how a globally operating German automotive manufacturer and an SME in the field of serious online gaming will employ this new service infrastructure by extending the capabilities of an already deployed flexible Translation Management Framework within a three-dimensional translingual webservices-based information space.

1   Introduction

The language industry, and in particular the GILT[1] industry, is characterized by a slow innovation adoption policy although the main players and stakeholders see themselves as facilitators for enterprises that demand to "act global". This situation is especially true when it comes to technological innovations. Some technology vendors argue that it is up to the customers to demand or to push for the use of new technologies while customers argue the opposite.

With the ever growing volumes of dynamic, user-generated content enterprises are faced with serious difficulties to handle, manage and analyze information in many languages and across different cultural boundaries. Therefore, there is a need to employ more open and collaborative approaches based on recent Web technologies and the concept of utility computing to allow the existing language ecosystems to successfully evolve to the next generation of technology offerings and associated services. Recent innovation drivers in this scenario are cloud computing on the technology side and crowdsourcing on the social side.

In the following sections, we first introduce the concept of cloud computing, and what this computational service paradigm and infrastructure concept in tandem with crowdsourcing opens up for the future translation ecosystems which we call MT Bazaar to also acknowledge the social dimension. After a discussion of the methodological aspects and the baseline architecture, we present two succinct case studies on how a big globally operating automotive enterprise and an SME offering an online game will benefit from these ecosystems, and how they fit with the already existing translation infrastructures. The paper closes with concluding remarks and some recommendations for MT Bazaar users and MT-Bazaar service providers and developers.

2   MT Bazaar Overview

2.1   Cloud Computing

Cloud computing is a term that has been hyped up in many ways with no collectively understood definition, although there is nothing fundamentally new. Cloud computing simply is the continuous development of a variety of technologies that have come together to alter an enterprise's approach to build, maintain and leverage an IT infrastructure. The services themselves have long been referred to as Software-as-a-Service (SaaS).

---

[1]GILT stands for Globalization, Internationalization, Localization and Translation.

The data center hardware and software is what we will call a Cloud. When a Cloud is made available in a pay-as-you-go manner to the general public, we call it a Public Cloud; the service being sold is utility computing. In contrast, the term Private Cloud refers to internal data centers of a business organization, and is not made available to the general public. Thus, Cloud Computing is the sum of SaaS, utility computing and virtualization, SUV for short, and does not include Private Clouds.

For the language industry and the GILT service communities in particular, cloud computing, however, serves as the next generation globalization enabler because it opens new ways of building, offering and delivering translingual services and technologies that will further transmute into transcultural services.

In addition, it will also allow the masses to easier employ various language technologies on demand from everywhere and across time zones, and to contribute to further enhance and reshape services and the underlying technologies through crowdsourced activities. Utility sources can be machine translation services, terminology support and other language related services that range from simple lookup and management facilities to sophisticated exchange capabilities through agreed upon or even standardized interchange formats, in ways and with quality promises that were previously only available to industrial users. To bridge the gap between theory and practice the crowd would need tools and the knowledge of how to use them which certainly will be a lot easier for the generation of digital natives than the current generation of digital emigrants. Today, this phenomenon is already demonstrated by LSP practices when deploying Internet services such as the Google Translator's Toolkit in traditional translation projects.

2.2 Translation Ecosystems

In this context, we talk about language ecosystems and more specifically about translation ecosystems because these terms describe adequately the evolutionary character of the emerging online landscape with multi-faceted collaborative and community-driven niche markets and their specific crowdsourced demands and requirements besides the industrial main stream developments.

In general, the introduction of new elements or agents – technologies, methods and approaches – into an ecosystem usually tend to have a disruptive effect. In some cases, this can lead to an ecological collapse and the death of many species within the ecosystem. Of course, we have seen this with many MT systems and MT approaches in the past decades.

Under this deterministic vision, the abstract notion of ecological health attempts to measure the robustness and recovery capacity of an ecosystem in terms of how far the ecosystem is away from its steady state. Often, however, ecosystems have the ability to rebound from a disruptive element or agent. The difference between collapse or a gentle rebound is determined by two factors: the toxicity of the introduced element, and the resiliency of the original ecosystem. You might compare this also with the up and down of MT development as well as MT deployment.

Our envisioned translation ecosystem – the MT Bazaar – comprises all translation related processes and the necessary tools and resources and even policies, which as cloud applications leverage the Internet in the underlying system architecture, and break the existing sequential process chains that are traditionally associated with the translation process as well as the proprietary business models, and the lack of access to multiple language resources by introducing stigmergic collaboration of humans and machines, and emergence into the overall architecture and its internal information flows. In general, stigmergy deals with the social behavior and the group dynamics in ecosystems, and today it is also termed crowdsourcing.

These characteristics in particular make the approach distinct from other approaches, even those that already deploy SaaS and virtualization technology in SOA[2] environments, because it allows for effective improvements of quality and performance through adaptation, correction and (self-) learning.

The approach of the MT Bazaar is economically viable because of its apparent overall cost effectiveness as a cloud computing application, and being potentially based on business models that could extend to satisfying the "long tail" markets by opening the cloud applications to the mass markets, for example, as a translation-on-demand service with attached quality tags for the general public.

---

[2]SOA stands for Service-Oriented Architecture.

A successful cloud services architecture in terms of application re-architecture requires disruption to the status quo, and is not simply a matter of deploying new technology and building service interfaces to existing applications; it requires redesign of the application portfolio. And it requires a massive shift in the way IT operates. You might compare this situation with the introduction of SOA because the small group of organizations that has seen real gains from SOA did so by treating it as an agent of transformation. In each of these success stories, SOA was just one aspect of the transformation effort. And this is the solution to success: a cloud services architecture needs to be part of something bigger. If it is not, then you need to ask yourself why you have been doing it. The latest shiny new technology will not make things better. Incremental integration projects will not lead to significantly reduced costs and increased agility. If one wants real visible gains, then one needs to make a thorough commitment to change. Here, crowdsourcing enters the stage that with its adaptability contributes to making the process pervasive and powerful, and to being a phenomenon of creative destruction in near real time.

2.3 Economic Challenge

Although MT and LT services and systems are cost effectively deployed by very large, globally operating companies, the masses of small- and medium-size companies are handicapped by the accompanying financial investments and the necessary human resources for such systems.

Therefore, the MT Bazaar as an Internet service ecosystem allows in particular small and medium companies to benefit similarly and efficiently from various automated language tasks, and to enter successfully new markets, which would not have opened because of the existing language barriers.

The MT Bazaar is a timely contribution to the current critical economic situation, and it is an important challenge to demonstrate the advances in developing real computational intelligence that can learn and adapt to altering demands and needs. The compensation of costs and the continuous improvement of services quality through advanced leverage of sophisticated technologies together with new business models that also include open source software models and the collaboration between cloud computing service providers

and customers are the ingredients of the economy's future driving forces.

Besides reducing the setup and management cost of an application that is associated with cloud computing resources, there are other advantages. For example, when a company separates itself from its resources by the Internet, it does not really matter where those resources reside. They could be, for example, in a location that offers appropriate terminological mining know-how in a certain domain and therefore minimizes application usage.

In the current economic crisis, the language industry, in particular the GILT market, is one of the exceptions for which analysts still forecast an increasing growth by a minimum of approximately 5 % per year within the next 5 to 10 years. However, such a healthy growth can only be achieved if a balance between language services demands and actual costs can be given so that through the investments in language services not yet another critical factor is introduced. Apparently community efforts and new, innovative technologies besides the already existing ones are needed to allow industries - small, medium and large - to master the multilingual language threat.

New dynamic technologies in a virtual translation automation space then is the main idea of the MT Bazaar which comprises several horizontal and vertically organized translation automation related services on an industrial scale with associated different language resources that effectively support and efficiently facilitate these services. In addition, the services are also enablers of translingual and transcultural communication as well as of knowledge management systems because they add essential value to business intelligence and predictive analytics solutions.

In summary, the MT Bazaar is the consequent further development of already existing SaaS offerings in the area of translation automation process and workflow management that most often integrate translation memory and terminology service capabilities. These applications easily extent to offerings with affordable direct access to MT and other language technology services with value-adding built-in quality assessment and quality assurance functions at different levels for the industry and the crowd.

Because the MT Bazaar creates new questions about privacy, security, trust and quality in general, we also have to focus on these aspects, and

what the given flexibility that allows different dynamics means for the traditional system of "create, patent and license" that has dominated intellectual property for the last few decades. We discuss these questions in the forthcoming ASLIB paper [Andrä, S. and J. Schütz, 2009].

3 Methodology

3.1 MT-Bazaar Baseline

The primary service and technology ingredients of the envisioned translation ecosystem, which resembles a great babbling bazaar of differing agendas and approaches, are:

- Content creation with negative translatability indicators

- Horizontal and vertical terminology governance

- Linguistic and cultural governance

- Translingual assets management including information and data sharing across domain and applications

- Process and workflow management

- MT integration management and workflow

- MT output monitoring and revision in post-editing environments with feedback coordination and management

- Human and machine feedback lifecycle management

Today's translingual business activities comprise manifold processes in and around the proper translation workflow that need to be supported by software tools and services in a transparent, coherent and efficient manner to ensure effective quality and process management, and to efficiently enable additional business intelligence and predictive analytics applications as well as the integration of consumers as active contributors and collaborators. At least three dimensions are mostly important to accomplish the vision of a complete translation ecosystem.

The first dimension represents the entire workflow which also includes pre- and post-translation tasks that deal with processes such as content creation and optimization, glossary handling, proofreading, legal approval, and specific market adaptations in terms of content, form and function. The second dimension depicts the different information sources that deal with a specific subject matter in various publishing and presentation formats and linguistic realizations such as marketing brochures, web pages, training documents, technical descriptions, workshop manuals, and the huge amounts of valuable user-generated content. The third dimension reflects the management of the information's life cycle in terms of its evolution in creation, translation and adaptation.

In addition, today's global communication processes demand for an easy deployment of such tools and services, and their worldwide accessibility as well as a seamless interaction with other systems either through connecting existing techniques together, combining processes that are based on different techniques into an overall architecture, or extending and augmenting core techniques in various ways. These demands require the effective support of standard exchange and interchange formats, and the provision of secure, open and powerful interfaces. Although these demands and requirements are not directly related to the most often quoted business goals for the translation process, namely time, cost and quality, our experience proves that a strict support of the three dimensions plus the fulfillment of the additional demands and requirements are enablers to save time and money, and to assure and continuously enhance translation quality.

In the MT Bazaar, the offered services and their various combinations are fully customizable according to the specific demands and requirements of a service user including the leverage of the user's feedback, and there is no need to follow any pre-defined process or integration flow. Each service is independently deployable and offers appropriate interfaces so that these services might also work in tandem to build more complex system incarnations that would offer capabilities beyond the summation of the single service capabilities because a complex ecosystem is not just the sum of competences and performances of its components and agents. Imagine, for example, the emergent power of a collaborative system that integrates MT, post-editing and translation memory services with feedback cycles to assess and improve the service output quality in a certain domain.

The introduced primary ingredients of the MT Bazaar fuel at least the following capabilities to name just a few:

- Ability to effectively distribute and manage centralized and decentralized resources of GILT related applications horizontally and vertically to ensure a semantic footprint across different quality levels

- Fostering teamwork between humans and machines based on collaborative community platforms and on collective intelligence and emergence computational models to accelerate the sharing of resources across domains and to collaborate with users and partners to fulfill specific goals in the language product lifecycle, which across all industries including the GILT industry itself traditionally have been separated, such as: content creation including (source) language proofing as an initial linguistic quality assurance, glossary setup with horizontal and vertical sharing, translation proper including translation memories and MT deployment, proofreading including MT post-editing, and linguistic quality assurance based on, for example, industry standards such as J2450, as well as information assimilation and dissemination with inbound and outbound MT

- Beating competition in resources and technologies with emergence and stigmergy based services

- Improving products and services through automated learning – unsupervised and supervised – and self-repairing methods as well as crowdsourced approaches

- Accessing services on demand, at anytime and from anywhere

- Providing transparent measures for validating service quality in terms of performance and competence

These capabilities can be adapted vertically to efficiently support specific application domains and industrial settings, and they provide horizontal services across-applications by

- Offering access statistics and predictive analytics to answer questions such as

"What is missing?" to facilitate effective and efficient service arrangements

- Acting as an early warning and alert system to allow the tracking of, for example, questions such as "Where and why is quality endangered?". This also includes certain error or quality failure prevention activities

3.2 Further Components and Services

A very important area of the MT Bazaar is the broad field of terminology which comprises several sub-services that are necessary for an effective employment of domain-dependent vocabularies in an MT application scenario, although it is often claimed that terminology is just a matter of "plug and play". These services include:

- Discovery and extraction of terminology

- Management of vocabularies including quality validation

- Import and export from external and internal resources with quality assessment

- Role-based access from other processes and services

- Domain and context dependent usage control

- Effective marketplace with sharing, bartering and crowdsourced capabilities based on well-defined or agreed upon interchange and quality rules

Translation proper is also more than a single service application; it is a series of several services that collectively facilitate inbound and outbound translation tasks. These services include:

- Automated pre-editing that eliminates certain errors in the source language information

- Integrated translation asset management including the selection of the most appropriate MT engine for a specific translation request based on meta-information, as well as the gathering of further additional training data for automated translation trainers if SMT and HMT is employed

- Automated post-editing that corrects certain machine translation errors

- Marketplace for crowdsourced translation assistance with a challenging incentive's model

Given these application scenarios, the MT Bazaar initially may distinguish three main employment roles:

- Provider: The provider is the owner and operator of the infrastructure

- User: The user is the consumer and the active actor of the MT Bazaar services with particular demands for competences, privacy and quality

- Vendor: The vendor sells specific products and services that facilitate the delivery, adoption and use of MT Bazaar related applications such as MT, language quality proofing, machine learning, etc.

A further area for a healthy evolution of the MT Bazaar are standards that, on the one hand, consist of a number of already existing, typically lightweight, open standards that facilitate the execution of the services and apply to areas such as application, client, infrastructure, platform, service, storage, and so forth, and on the other hand, resemble language and translation related standards such as DITA, XSL, XLIFF, TMX, TBX, and others.

A last but not least area comprises MT systems that will evolve to the next generation MT systems based on machine learning techniques and computational intelligence to facilitate

- Automated knowledge discovery tasks such as the clustering and classification of information entities from feedback cycles

- Collaborative behavior to ensure and to assure fixed and volatile quality measures used for validation and evaluation purposes

- Effective self-learning and adaptable machine translation

4 Results and Analysis

4.1 Translation Management Framework

ontram – the Online Translation Management Framework of Andrä AG – has been designed and implemented with some of these fundamentals as its leading architectural and development guidelines, and has given evidence in several large-scale industrial projects over the past seven years that the concept of an entire web browser based translation process management system that also can be deployed as a cloud application has exceptional advantages in achieving the translation related business objectives of different companies.

ontram's web browser based, open approach fully scales with the various business needs and demands of translation projects, and therefore medium and large enterprises and even small organizations benefit from its employment in the translingual value chain. Because the downgrading of requirements can be accomplished much easier than the effective upgrading of requirements, we demonstrate two best practice cases with a translation business analysis of an automotive enterprise that is among the worldwide 10 largest companies, and an Online Game Developer SME along the introduced three basic dimensions of an translation ecosystem, and show how this framework is able to evolve into an MT Bazaar environment.

4.2 Automotive Manufacturer

Firstly, this big auto manufacturer required an easy and intuitive, worldwide access to the Translation Management Framework (TMF), which certainly is directly connected to our first dimension, the entire workflow. All people involved in the process chain work with one secure online platform that is centrally controlled and managed through a scalable and flexible workflow engine and complete, consequent logging capabilities of any changes in content, form and function of the language materials. This makes the entire value chain and its resources most transparently, and even non-professional users may easily access the system and retrieve essential business information for their work. Feedback cycles and response times in the approval processes are drastically reduced, and because no direct installation activities besides the setup of a web browser are required, training efforts are minimized and new users or roles can be added in quasi no time on the fly. To accomplish optimal acceptability within the TMF user community, the system provides powerful interfaces to different editing systems as well as translation memory systems to allow users to work in their familiar environments. The system also manages its own translation memory and

glossary sources to allow for optimized re-usability, and it keeps full control of a customer's intellectual property rights.

Secondly, to optimize synergies between different data sources, media and documents was another key requirement as well as to fully leverage the content of translation memories across media formats to gain significant savings but still maintaining the overall quality. ontram solves this demand in the second dimension by keeping the basic applications independent of the used media and by providing preview functions for the most common media types. This enables the system user to view her current working text version in the target layout and media. This design decision means that the standard user interface within the web browser looks the same for e.g. a webpage and a marketing brochure but by activating the preview function the user can view her text either as a webpage in a web browser window or as an PDF with its target layout. As soon as translation and adaptation processes have been finalized in their approval steps, the results are directly accessible by follow-up projects thus allowing for instant leverage of the translation memory content. Yet another very important aspect for the second dimension is the need to communicate efficiently with other systems such as Content Management Systems, Learning Management Systems, and even other TMFs to exchange content data as well as translation memory information without loss of coherence and reference.

Thirdly, several translation project analyses of this auto manufacturer have shown that 50-80 percent of all new translation projects are in most cases updates of already existing translingual information. Therefore, this fact must be taken into account seriously because it is an important resource for saving time and for ensuring a continuous quality. To gain maximum benefit it is not only the translation memory match that contributes to the reuse success, it is of crucial importance to take the life cycle information of each text element as well as of entire documents – the third dimension within our translation ecosystem – as an essential factor. This action requires to store as much meta-information as possible within the TMF, and to use that information to reduce the need for manual (re)work by the user within the process management.

In addition, globally operating enterprises also require high volume, high speed and accurate sift-ing of unstructured digital information to gain market insights, to identify appropriate strategies and to extract and derive knowledge for their field operations. This information often comes in from a variety of languages, and first must be transformed into a common language – sort of pivot language – which is supported by the employed applications for Business Intelligence and predictive analysis, to ensure an effective analysis of the information as well as optimized knowledge extraction from the information. For example, the automotive company is interested in conducting market research, competitive intelligence, or measuring the opinion of the company at a global level. When multilingual information is involved, it is critical to have automated translation as part of the solution set to enable timely access to relevant information and inferred knowledge. Therefore, machine translation and crowdsourced approaches are now entering the TMF environment as a consequent next step in its successful deployment.

4.3   Online Game Developer

Meanwhile, the offered online game has more than 4 million players world-wide and supports over 30 languages. Most of the translation efforts are facilitated either by professional translators or by translations of the player community. It was not just saving money that has motivated the company to take the path of crowdsourced translation, and partly even crowdsourced development, but the community itself gave further good reasons to follow this path.

The challenge, however, is to keep a harmonized balance between using the collective intelligence and traditional processes. Integrating both in one TMF platform that controls the underlying resources, the translation memories, and assures the continuous quality is the key factor for the longterm success of a crowdsourced approach.

Besides all of the demands of the automotive company, this company has the following particular requirements that need to be managed within the TMF:

- Strong collaboration support in all communication and translation processes

- Leverage of community assets

- Crowdsourced translation

- Quality as an entire user experience

- Continuous development cycles

- Direct link to the gaming environment for effective previews

These requirements are based on a strict long-term business vision combined with organizational challenges of availability, redundancy and reliability. The internal processing steps and their collaborative behavior are modeled in the same fashion as other business services are modeled in the TMF environment which allows to employ the same monitoring, supervision and optimization functions that are used for ordinary business processes, and thereby guaranteeing the effective assessment and evaluation of the quality of the crowdsourced tasks.

Machine translation services that are provided through the Internet can be seamlessly integrated into these internal process flows. However, an actual feedback mechanism to the external services remains an open issue, and should be taken into account seriously by the machine translation system providers. To solve this aspect, we consider to adapt and optimize the translation results on the basis of various external information elements to identify possible survivable translation results.

5   Conclusions and Recommendations

An integrated, easy to use web browser based and SaaS deployable TMF such as the ontram ASP Edition provides many advantages for the whole translation and adaptation value chain across different media and all uses of language data, and presentation and publishing formats. If the involved parties all work on one online platform, this system, on the one hand, leads to optimized processes by means of cost, time and quality, and on the other hand, gives total transparency to the processes and their content.

A fully fledged translation ecosystem can only be achieved if the system is capable and consequently supports an open interchange with other systems and information sources and targets. ontram demonstrates in a consistent and industrial proven way how these findings have been realized in a complex but still easy to use web browser based application. In its next evolution stages, the system must further prove how it effectively integrates MT capabilities that work together with the already existing components, modules and services in an adaptable and emergent fashion and in an MT Bazaar deployment scenario, thus al-

lowing companies for additional savings, revenues and even new market gains.

Within the next 3 to 5 years, we will see more and more developments towards the MT Bazaar vision because

- Cloud computing further matures

- Sensitivity for crowdsourcing and language data sharing increases

- New markets and market opportunities emerge

- MT developers join in and MT systems evolve to the next generation due to either the disruptive effect of evolution, or the toxicity of the introduced new elements and the resiliency of the original ecosystem

This future will bring us more distributed, crowdsourced markets in which social negotiation and collaboration between humans and machines are stigmergically mediated by computational intelligence and Internet-based technologies

Our recommendation for future MT Bazaar users is:

- Actively collaborate and bring in your specific expertise and knowledge to the emerging communities and marketplaces

- Be open minded towards crowdsourcing and its evolving power.

For future MT Bazaar service provides and developers our recommendation is:

- Grow communities around products

- Do not rely on standards but provide effective support to continuously enhance quality as a user experience

- Develop MT with emergent behavior and provide MT that learns and adapts, e.g. by employing computational intelligence

References

Andrä, Sven Christian and Jörg Schütz, 2009. Cloud Computing in GILT Ecosystems and Evolution. In: Proceedings of ASLIB 2009, London, UK.

Schütz, J., 2008. Artificial Cognitive Post-Editing Intelligence. In: 8th AMTA Conference, Waikiki, Hawai'i, USA, pp. 448-453