Supporting Flexibility and Awareness in Localisation Workflows

David Lewis, Stephen Curran, Gavin Doherty, Kevin Feeney, Nikiforos Karamanis,
Saturnino Luz, John McAuley
Centre for Next Generation Localisation
Trinity College Dublin
www.cngl.ie
dave.lewis@cs.tcd.ie; stephen.curran@cs.tcd.ie; gavin.doherty@cs.tcd.ie; kevin.feeney@cs.tcd.ie;
nikiforos.karamanis@cs.tcd.ie; saturnino.luz@cs.tcd.ie; john.mcauley@cs.tcd.ie

Abstract

A key strategy for supporting users in distributed work systems is to help them maintain awareness of the state of the work system and of the work being done by others. At the same time, many knowledge intensive industries are embracing the technologies that have underpinned the Web 2.0 movement to allow open user generation, annotation and modification of content. These technologies can potentially provide a useful platform for supporting awareness and distributed teamwork. However, as distributed content generating activities become more valuable, organisations aim to optimise them, often by modelling and monitoring the workflows involved and augmenting them with software services. Currently, however, these two approaches do not integrate well and there is little system support that integrates the centralised monitoring and management of workflow with the open communications that is characteristic of web-based user content generation. In this paper we examine the use of both techniques in the localisation industry, and based on this analysis we propose a platform that combines the visibility and awareness support of open content generation between users with their involvement in a centrally managed workflow.

Keywords: *localisation workflow crowdsourcing service-oriented meta-data management*

1. Introduction

Situation awareness has long been recognised as critical for supporting effective and resilient performance in complex work systems (Endsley 2000). While the literature has to a large extent concentrated on situations such as process control, command and control etc., supporting situation awareness is an important factor in maintaining the ability of any complex organisation to effectively detect and respond to unforeseen and exceptional situations.

Our analysis focuses on the localisation industry, which translates textual content into different languages so that products and services can be marketed and used in different countries and regions around the world. We start by presenting an abstract description of the localisation workflow whereby such content is translated, based around the view commonly taken by the vendors of workflow products supporting this sector. We then present several examples of the recent trend for crowd-sourcing in localisation, whereby content is distributed for translation to a group of bilingual individuals engaged in a community around the product or service being localised. This is in contrast to the traditional localisation process of outsourcing translation to a professional translation agency.

To provide a more realistic picture of the information exchanged between actors in this domain we present an analysis of the results from a recent field study carried out within the localization industry. Preparatory work included offline study of the available tools, review of background materials on the localization industry (including previous issues of Localisation Focus, the proceedings of the LRC conference and archives of localisation fora such as www.localisationworld.com and www.multilingual.com), and review of the research literature pertaining to localisation including the descriptions of the localisation process by Esselink (2003) and by Wittner and Goldschmitt (2007). A series of 13 semi-structured interviews was then carried out, with interviews typically lasting between 45-minutes to an hour each. The interview subjects were employees of a large company and a multi-language service provider. The interviews were accompanied with observations of the employees using several tools to perform their tasks. The results of this study highlight the interactions that occur both within and external to the formal workflow and its support in the workflow management system. We then compare this analysis to observations made about the communication channels needed in crowdsourced localisation. From this analysis we propose a generic platform for managing workflows with control over the integration of open communication mechanisms related to quality issues. The aim of the platform is to provide greater visibility of the work system, supporting improved awareness, to integrate transient and ad-hoc communications in a more structured manner, supporting knowledge capture, and to support a more flexible and realistic conception of the workflow and work system.

2. Conventional Centralised Localisation Workflow

Figure 1 depicts a generalized localization workflow. This process is based on one of the author's experience (Stephen Curran) working as a software engineer in localisation. The chain of activities in the workflow is as follows:

- **Extraction:** The process of extracting translatable text from the source documents. Documents coming from desktop publishing packages often contain a lot of structural information that is not needed for the translation process. Textual content is separated from structural content.
- **Segmentation:** The process of dividing up the source document into translatable units of text. Normally these are sentences but they don't have to be. They could for instance be paragraph headings or diagram captions.
- **Creation of Project TM:** The segmented documents are analysed against the central translation memory to produce a project translation memory. The project translation memory contains all relevant translations from the central translation memory.
- **Pre-translation:** The target of every exact match in the project translation memory is inserted into the translation placeholder of the corresponding segment in the document. This is an optional step in the workflow. Sometimes it is preferable to have the translator manually insert the exact match from the project translation memory into the document themselves so that the translation unit is getting a certain amount of review.
- **Machine Translation (MT):** Sometimes a machine translation system is used to generate translations for segments in the documents that do not have any match in the translation memory.
- **Generation of Translation Kit:** All files needed to perform the translation are zipped up and sent to the translator. This includes the documents to translate and the project translation memory. It may also include a glossary containing any relevant terms and their translations and any reference material required to give context for the translation.
- **Manual Translation:** The translation kit is downloaded and unzipped by the translator. The translator opens the documents, translation memory and glossary in their translation environment and iterates through and provides a translation for each document segment. When the translator opens a segment in the environment any match in the TM or glossary is presented to the translator for insertion into the target segment along with a notification of the match value.
- **Review and Editing:** The translated documents go through a cycle of review and editing. This includes both linguistic review and functional testing.
- **Translation Memory Update:** Once the documents have been signed-off from review the translation memory is updated with the translations from the documents. The updating process includes inserting any new translations and updating any previous translations.
- **Creation of Target Documents:** The final target version of the documents are created. The translated segments are combined with the document structural information to produce the final version of the documents.

3. Crowd-sourced Localisation

The localisation industry is increasingly turning to crowd-sourcing to address the scalability problem of current processes. In localisation crowd-sourcing, the translation job traditionally done by a professional translator is done in a more informal fashion by a group of volunteers. These volunteers are usually engaged in a community around the product being translated. Some notable organisations have adopted in varying degrees the crowd-sourcing approach to localisation :

Facebook

The social networking site Facebook crowdsources the translation of their user interface (Facebook Translations). Users, through a Facebook application, can submit translations for strings in the user interface. A translation memory is incrementally built and made use of by the community. Quality control is achieved through the community commenting and rating each other's translations.

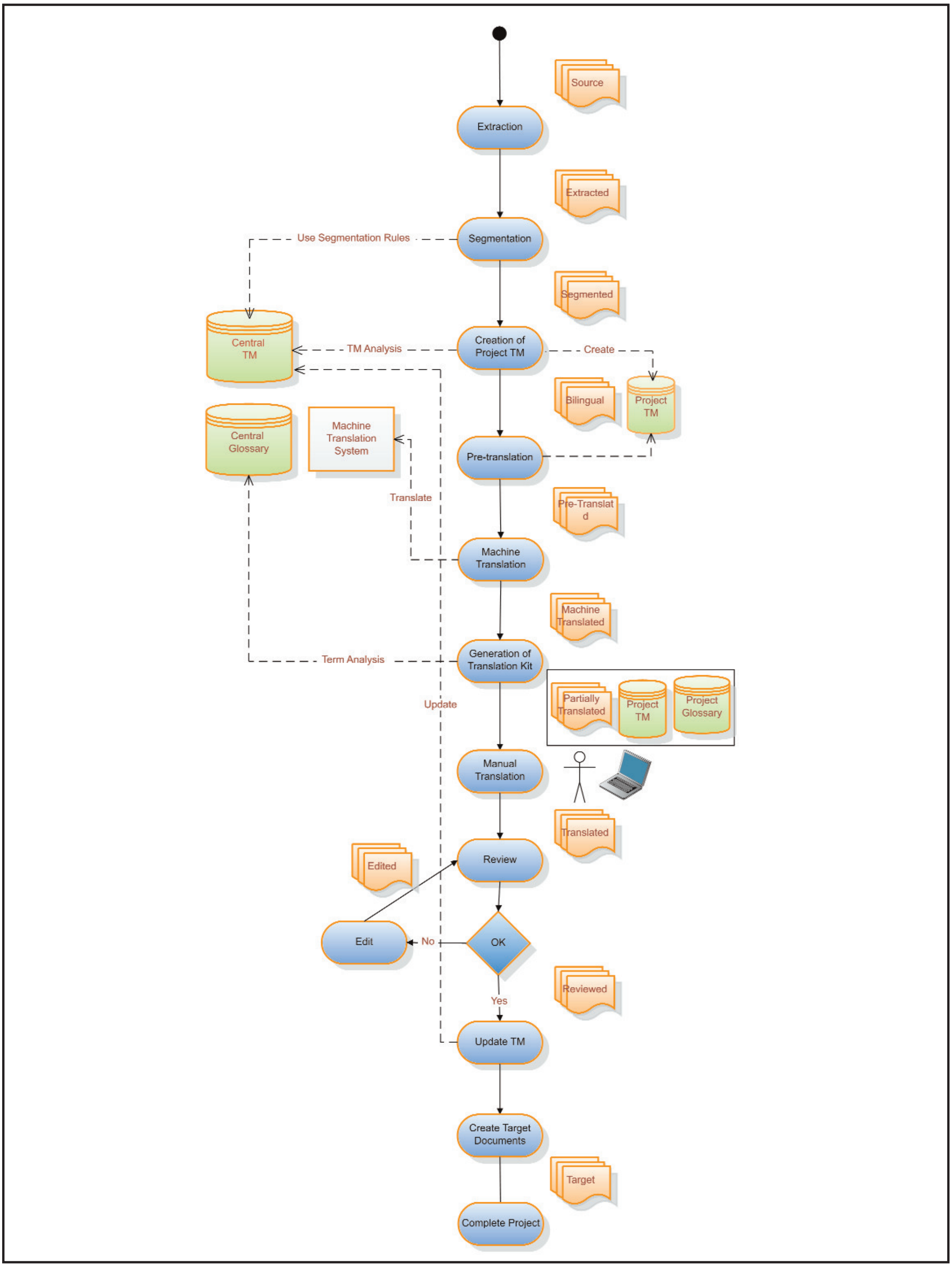


Figure 1: Generalized Localization Workflow

We use this model as the basis for conducting a study of the actual interactions that are part of the current practice of the interviewed practitioners. These interactions include informal communications as well as those that are part of formal, business processes supported by workflow tools.

For each target language, the community translation process goes through the following steps:

- The core terminology is translated by the community to build a bilingual glossary.
- The user interface strings are translated by the community making use of the glossary.
- The contributed translations are rated by the community and winning translations are determined.
- The translations pass through an internal review stage before being approved.

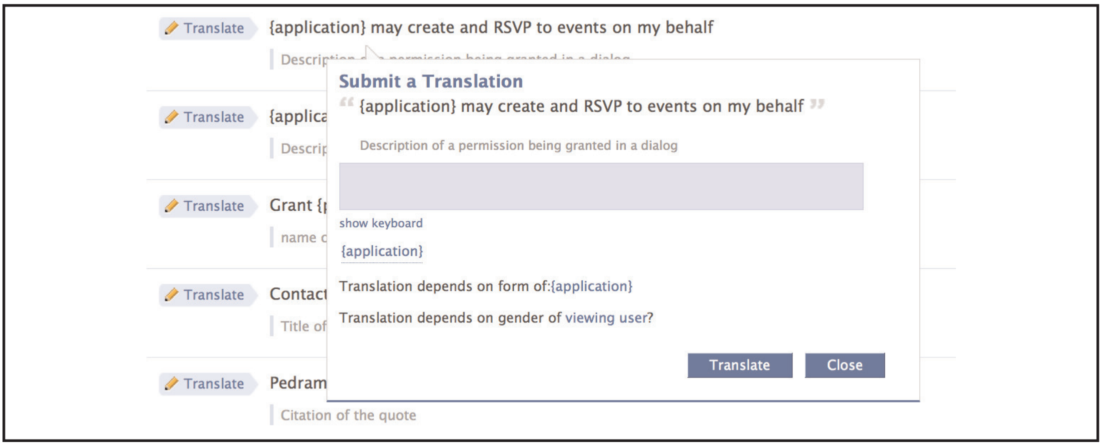
Microsoft

Microsoft has a forum on its MSDN website that allows the community to contribute and rate translations of terminology in Microsoft products (Microsoft's Terminology Community Forum). Better translation of product-specific terminology can often be forthcoming from the product user community than from a professional translator who does not use the product. Whereas Facebook crowdsources the translation of sentences, Microsoft crowdsources only the translation of core terms. The community are not involved in the translation process other than providing suggested translations for these core terms.

TED Translations

TED is a non-profit organisation that runs conferences on topics in technology, entertainment and design. Recently, in an attempt to reach out to a wider audience, they have published their conference presentations in languages other than English. They have used a crowd-sourcing approach to translate the presentations to these languages (TED Translations).

TED, like Facebook, get the community to translate sentences rather than just terminology. However, unlike Facebook, the translator tends to work on the entire text rather than on small segments. There is no concept of on-going rating or feedback of other parallel translation taking place within the text. The translator tends to work independently on the entire text.



Translate {application} may create and RSVP to events on my behalf

Description o

Submit a Translation

“ {application} may create and RSVP to events on my behalf ”

Description of a permission being granted in a dialog

show keyboard

{application}

Translation depends on form of:{application}

Translation depends on gender of viewing user?

Translate {applica

Descrip

Translate Grant {

name c

Translate Contact

Title of

Translate Pedram

Citation of the quote

Figure 2: Facebook translation interface

	Community involvement	Interaction in the community	Resources that the community interacts with
Facebook	High. Full involvement in the translation of glossary and the translation and review of product.	High. The community discuss and rate translation.	Translation memory and glossary.
Microsoft	Low. Input into the translation of glossary. No involvement in the translation and review process.	High. The community discuss and rate translation.	Just glossary.
TED	High. The community translates all of transcript.	Low. A transcript normally translated by one individual.	No resources used. Translation memory and glossary not maintained.

Table 1: Analysis of Translation Crowd-sourcing Systems

Unlike Facebook, no translation memory or term-base is maintained. The language used in the talks is not controlled and the topics of the talks are varied, so its not clear that there is much benefit in using a translation memory or terminology manager.

Summary of approaches

Table 1 summarises these approaches and their relative involvement with the communities concerned. A key distinction from conventional localisation workflows is the emphasis on supporting peer interaction within the community. Supporting the development of a sense of community by enabling different translators to communicate freely contributes to a sense of shared endeavour that serves to motivate translators in the absence of direct financial reward for their efforts.

From a systems point of view, crowd-sourcing platforms place an emphasis on promoting communication between those involved. This contrasts with workflow management systems which only concern themselves with task related control and data flow between participants, and therefore tend to ignore communication and information sharing that occurs outside of the workflow.

However, as has often been observed, workers involved in workflow often encounter problems that are not fully modelled in the prescribed procedure to which they are expected to work. They therefore often seek solutions by using informal communication with other workers. In a shared office environment such communication can take place readily, but in workflows where human knowledge based activities within a process are undertaken in different organisations and by geographically distributed staff, such informal communication may be less easy to initiate as workers are less likely to be personally acquainted. The ability of web based communication technologies to support and encourage open communication between unacquainted crowd-sourcing workers may therefore offer some benefits to commercial workflow systems. To understand better what potential channels of communication could be beneficial to existing commercial localisation workflows, we look at the current working practices and problems revealed in the study. We examine those aspects that are not addressed in current localisation workflows and are therefore not implemented in the systems that support them.

4. Role Interactions in Localisation Workflow

Across the centralised, commercial localisation workflow and crowd-sourced localisation the following abstract roles can be identified:

- Content author: produces the source text
- Terminology Manager: manages a consistent set of terms and expressions for a project in the source language.
- Linguist: a language specific specialist responsible for maintaining translations of terms, translation manuals and the guidelines used in assessing quality.
- Project Manager: overviews the translation process on behalf of the clients. In a commercial setting this is sometime the activity of third party LSPs, while in a crowd-sourced setting this may take the form of an online community moderation role.
- Post-editors/Translators: manually translate source text to target language text. Post editing involves reviewing existing translations, whether produced by human or automated translators and involves selecting from alternatives, modifying a translation or providing a new one.

The following table illustrates the points of interaction between these identified roles. The interactions marked "I" are informal ones that are poorly supported in current localisation processes while the interactions marked "W" are those that are already supported in existing localisation workflow systems. Those marked "N" are ones that arose through discussion as being of potential value but which are not currently common practice. What is clear from this analysis is that current workflow systems and processes focus on communication that follows the workflow process from one role to the next. What is not well supported is upstream communication from the roles operating at later portions of the workflow to those involved at the earlier parts. Although such interactions may help improve the overall process, they lie outside of the main flow of communication. Since the workflow model is seen as the primary route to value generation and therefore the basis for contractual arrangements, this means little relative value is attached to these communications.

To/From	**Terminology Manager**	**Content Author**	**Linguist**	**Project Manager**	**Translator / Posteditors**
Terminology Manager	Share term bases and techniques for achieving high compliance to controlled language guidelines (N)	Detail problems encountered with applying the term base and controlled language guidelines (I)	Relate language-specific problems in translating specific terms from term base (I); propose changes to controlled language guidelines to improve efficiency of translation to a specific language (I)	Relate problems with conformance to controlled language and missing terminology (via linguist) (I)	Relate problems with conformance to controlled language and missing terminology (via project manager) (I)
Content Author	Term-base and controlled language guidelines (W)	Share notes about complying with controlled language guidelines and appropriate terminology	Specify the job target level for controlled language compliance (I)	Relate translation problems with provided content and its context (I)	Errors in source content and missing contextual information (I)
Linguist	Response to stated translation problems with specific terms (I)	not applicable	Share problems in translating term base to different languages (I)	Quote for job (W); Relate problems with using terminology translations (I); Problems with use of translation guidelines (I)	Suggest different terminology translations (N)
Project Manager	Term base and its context (W)	Content and its context (W)	Translation dictionary and guidelines (W)	Share problems with translating jobs into parallel languages, handling specific content, performance of specific TM, MT and translators (I)	Progress in translation job (W); Problems in translating content, erroneous content, terminology or term translation (I)
Translator	Term base and its context (via project manager) (W)	Content and its context (via project manager) (W)	Translation dictionary and guidelines (via project manager) (W); Responses to translation problems with terminology (I)	Job allocation and quality targets (W)	Share problems with use of terminology translations and lack of terminology definition/translation; queries to more experienced translators; Feedback on quality of TM and MT translations (I)

Table 2 : Summary of Role Interaction in Localisation Workflow

5. System Support for Localisation Workflow Awareness

It seems clear that the future of localisation will involve some element of crowd-sourcing. However, to reach its full potential of crowd-sourcing to optimally complement commercial localisation activities, such next-generation localisation will need new integrated platforms that integrate crowd-sourcing technologies and workflow management technologies together seamlessly. The study shows that current commercial workflow could benefit from improved communication that goes beyond that dictated by the major value flow of the workflows. However, the very open communication characteristics of the crowd-sourcing environment are not always appropriate for workflows operating within the constraints of commercial contracts. A client may make use of several Localisation Service Providers (LSPs), perhaps with different financial arrangements, and may therefore be sensitive to free communication between them. Equally, LSPs compete for clients and may not be willing to expose the less formal, but valuable interactions that have built up with a particular client.

Furthermore, any common platform for supporting both commercial and crowdsourced localisation would need to support the evolution and migration of workflows between the two for both companies and individuals, as many will be involved in both, to varying degrees at different times.

The challenge in developing a common model for integrating commercial and crowdsourced localisation is to support a variety of levels of control over the interactions that can occur with a localisation process, ranging from tight central control to loose, highly devolved control.

The Centre for Next Generation Localisation (CNGL) is an integrated research project bringing together academic researchers and industrial partners to construct a framework to support the next generation of localisation systems that precisely encompasses both commercial and crowd-sourcing localisation in order to support a growing variety of new business models that can deploy both. It remains unclear whether the integration of features designed specifically for managing human tasks directly into the workflow management system is desirable as such an approach increases the complexity of the workflow specification, yet will significantly constrain the expressiveness that can be applied to human task management.

Below, we outline an integrated software architecture that is being assembled within the CNGL project to support standards based web service integration and workflow execution with flexibility in how the human communities involved in such workflows can communicate with each other.

Central Meta-data Repository

It was identified earlier in this paper that many of the transient communications in current localisation workflows are being lost because the workflow technology does not support them. In this section we introduce a potential solution to the problem.

On the internet there has been a move towards adding structured information to content so that it can be automatically reasoned over. With the semantic web initiative content publishers are being encouraged to attach additional meta-data to the resources that they publish. The Resource Description Framework (RDF) is a meta-data data model that is being used to support this initiative (RDF 2004). RDF allows for the representation of meta-data in the form of 3-place relations, subject, predicate and object. The data model can therefore support any arbitrary data schema. This is necessary as it is not known a priori the range of meta-data that content publishers will want to express.

We propose using an RDF repository to store communications in and across organisational boundaries. A flexible meta-data schema is necessary since it is also not known the range of communication that might need to be represented in a localisation workflow. More precisely, each resource in the localisation workflow : TMs, glossaries, controlled-language rules etc. would have a set of communications associated with it that are stored in the repository. Storing the information centrally means that it can be easily aggregated and reasoned over. Since potentially many organisations could be involved in a localisation workflow and could contribute to this store, we propose using our Community Based Policy Management technology (see below) as a means of managing this store in a decentralised way.

Community Based Policy Management

Given the cross-organisational nature of localisation workflow, the communication meta-data repository would consist of communications within and between organisations with no one organisation owning all the communications. Therefore, each organisation should be able to manage the communications that it owns and control which other individuals and organisations should have access to it. Our Community Based Policy Management (CBPM) technology (Feeney et al 2004) can be used to allow the repository to be managed in such a decentralised fashion.

The CBPM is a policy based management system which allows for the sub-division of organisations into smaller groups. Each of these groups has its own set of policies applied to it, meaning that each group has a certain set of rights over the resources owned by the organisation. CBPM also supports the delegation of management rights to various sub-groups or federated groups, meaning that management can be decentralised.

Community Management Framework

We have taken the Drupal Content Management System (CMS) (Drupal 2009) and integrated it with CBPM. Drupal is a web content framework with a pluggable architecture and a collection of add-on modules contributed by a development community. These modules include such things as forums, messaging, blogs and other social networking and communication technologies.

The integration of Drupal and CBPM allows online communities to control the distribution of management authority over content in the CMS across the community. Since Drupal comes with these communication technologies, the Community Management Framework (CMF) can be used in a localisation crowd-sourcing scenario to help manage communication between volunteer translators in a fine grained manner. This allows community management decision makers (who may vary from professional community moderators to a democratic function of the whole community) to balance the benefits of completely open communication with those of more restricted, team based communication and to move easily from different models as the focus and activity level of the crowd-sourcing community shifts over time.

Business Process Execution Language and Human Tasks

A Service Oriented Architecture approach has been taken in development of the CNGL project demonstrators. Business Process Execution Language (BPEL) is used as a platform for creating and executing localisation processes. BPEL automates business processes through the orchestration of web services. Linguistic processing software components, performing functions such as Machine Translation or Text Analytics, are packaged as web services and used by BPEL processes. BPEL is good for task automation but the central standard does not support human tasks. In localisation processes, some tasks are manual: professional translation/post-editing, crowdsourced translation. We need a way to support the inclusion of such tasks.

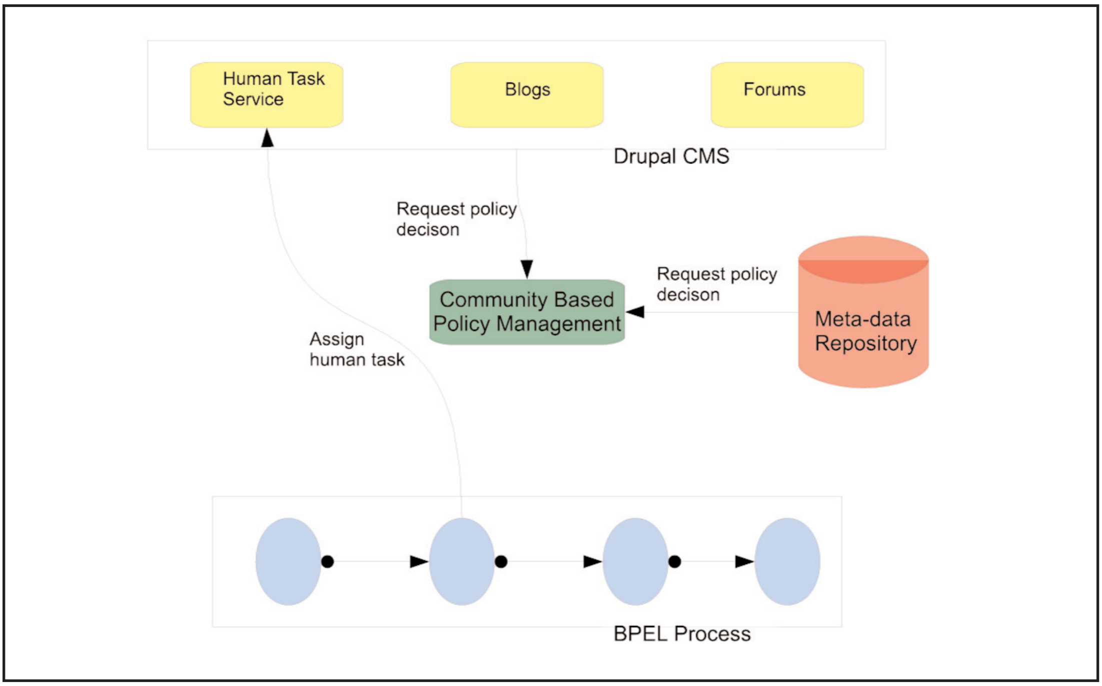


Figure 3: Integrated architecture for Flexible Interactions between Localisation Workflow Actors

Currently, human tasks are integrated into BPEL-based workflows through a human task web service that is included in the workflow as a partner service. The central BPEL process sends the task to this service and the task is completed by people interacting with the service. The web service implements such functionality as task assignment, task workflow and task lists. BPEL execution engines normally come with such a human task web service.

There is also an OASIS standard, BPEL4People, that aims to include support for human tasks directly in BPEL. The advantage of this is that humans and task assignments can be modelled directly in BPEL. However, there is very little support for this standard in BPEL execution engines.

We aim to create a human task web service and integrate it into the CMF so that through the application of CBPM rules, the tasks will be routed to the appropriate individuals within the organisation. This may be useful in the localisation crowd-sourcing scenario where appropriate translators can be selected based on policies in CBPM, taking into account such criteria as source and target language, the domain of translation, and the reputation of the translator.

Architecture for System Support for Flexible Interaction between Localisation Workflows Actors

Hence our integration framework includes a custom service interface designed to support the expression of a wide range of the management requirements that organisations typically wish to apply to human tasks. This human task service is distinguished from other implementations in that it provides native support for task delegation and decomposition within groups, rather than requiring the workflow designer to specify the intimate details of how each task should be performed and monitored. It is based on the CBPM, and allows tasks to be allocated to groups or communities in addition to individuals. These groups can then use the CBPM system to further break down tasks, allocate them to individuals and gain fine-grained control over the process. This ability to delegate portions of the workflow is crucial in supporting large, complex workflows that span organisations without requiring the workflow designer to have a complete understanding of every detail of the process.

6. Discussion and Further Work

In this paper we have discussed a number of issues (identified through fieldwork) relating to traditional workflow management systems, including the opaqueness and coarse granularity of the work, the lack of support for the continuous and out-of-band communications which are a feature of effective teamwork, as well as departures from the formal workflow needed in order to deal with changing circumstances or exceptional cases. We have analysed these issues with respect to the localisation process, by examining a number of role interactions, and the degree to which they are supported in the workflow. These interactions are not just between individuals in a role, but between individuals and groups, between groups and between different subsets of groups and the rest - especially when groups (or individuals with them) play more than one role.

We have examined the recent trend towards distributed content generation and management, which presents an opportunity to leverage the same technical infrastructure across a range of systems, ranging from centrally controlled to fully distributed. Localisation provides an excellent example, as there is an opportunity to achieve integration between crowdsourced and enterprise localisation technologies.

The solution proposed here is based on adding metadata and additional links to existing artefacts beyond the major transactions of the workflow. Within the context of localisation, a range of artefacts (terminology, TM entries, source text) that are exploited across the process can provide a vehicle and a set of interface concepts for achieving this integration. This approach will help to make work, effort and resources more visible, which will increase awareness throughout the workflow, and will also allow the many exceptions to be dealt within the formal structures of the system. We are in the process of prototyping such a system and will aim to trial it in a real localisation scenario in the future. While dynamic communication features are becoming a common feature within workflow systems, these should be integrated in a way that facilitates knowledge capture, so that they become a resource for the rest of the team.

Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under the Centre for Next Generation Localisation (www.cngl.ie).

References

Allee, V. (2002) The Future of Knowledge: Increasing Prosperity through Value Networks, Burlington: Butterworth-Heinemann

Drupal Content Management System (2001) [online], available: http://www.drupal.org [accessed 6th Feb 2009].

Endsley, M.R. (2000) 'Theoretical Underpinnings of Situation Awareness: A Critical Review', in Endsley M.R and Garland D.J. (eds.) Situation awareness: analysis and measurement, New Jersey: Lawrence Erlbaum Associantes, Inc., 3-33

Esselink B. (2003) 'Localisation and Translation', in Sommers H., ed., Computers and Translation, Amsterdam: John Benjamins, 67-87.

Facebook Translations Application (2008) [online], available: http://www.facebook.com/apps/application.php?id=4329892722 [accessed 25 Jan 2010]

Feeney, K., Lewis, D., Wade, V. (2004) 'Policy Based Management for Internet Communities', in Proc. of IEEE 5th Int'l Workshop on Policies for Distributed Systems and Networks, New York, 7-9 June, IEEE Computer Society Press, 23-34.

Microsoft Terminology Community Forum (2008) [online], available: http://www.microsoft.com/language/mtcf/mtcf_default.aspx [accessed 25 Jan 2010]

OASIS, WS-BPEL Extension for People (BPEL4People) (2009) [online], available: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=bpel4people [accessed 25 Jan 2010]

Resource Description Framework (RDF) (2004) [online], available: http://www.w3.org/RDF/ [accessed 24th July 2009]

TED Translations (2009) [online], available : http://www.ted.com/translate [accessed 25 Jan 2010]

Web Services Business Process Execution Language (BPEL) Version 2.0 [online], (2007), available: http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-0S.html [accessed 6 Feb 2009]

Wittner J. and Goldschmidt D. (2007) 'Technical Challenges and Localisation Tools', Multilingual #91, Localisation Guide: Getting Started.