# Supporting Consistencies in Multi-Language Knowledge Sharing

Amit Pariyar

# Abstract

The goal from this thesis is to support the design of multi-language knowledge sharing system with a focus on consistency in content shared among communities. Though unprecedented growth in online collaboration has attracted diverse communities to participate in knowledge sharing, example among resource rich and resource poor communities, the possibility for inconsistency in content shared is increased. This is problematic for multi-language knowledge sharing system since it is not practical to state consistency rules in advance for content shared among communities. Consequently the design of multi-language knowledge sharing has to shift focus from consistency rules and pay attention to cases that cause inconsistency in the shared content. The cases such as content omitted or content updates not shared and the presence of conflicting content are expected to occur in collaboration and are the potential cause for inconsistency. Though the occurrence of such cases may seem trivial at first nonetheless the complexity is raised as each community participates in its own language and so inconsistent content is shared in several languages. Further such cases also have the potential to cause inconsistency at global and local scales leading to globally and locally shared inconsistent content. Regional discrepancies from inconsistent content shared with communities in several geographic regions are also equally anticipated in knowledge sharing.

Another problem is the constraint in content consistency due to divergent knowledge sharing goals of communities. This means where the goal is to leverage knowledge equally exact correspondences in shared content is preferred with a rigid consistency policy and where the goal is to customize

knowledge sharing there is a need to restrict sharing to specific languages and specific communities with a non-rigid consistency policy.

Grounding on the consequences from sharing inconsistent content and the constraint in content consistency that arises from disparate knowledge sharing goals of communities this thesis makes following contributions towards the design of multi-language knowledge sharing system.

1. Synchronization of User Editing Activities to Detect Inconsistency in Multilingual Content.

   The challenge in leveraging knowledge equally among communities is elevated from the participation in several languages. Inconsistency due to omitted content, updated content not shared and content conflict occur among languages which is undesirable to communities. Towards dealing with inconsistency in multilingual content, a process-based technique is proposed to detect missing content, updated facts or information and content conflict between languages. The proposed technique is based on the concept of synchronizing user editing activities which provides an alternative to content-based techniques. To realize this concept a state transition model is proposed to define states in multilingual content, set of actions and transition functions. Inconsistency detection rules are then designed using the combination of states in multilingual content. Experimental results from applying the proposed process-based technique to multilingual Wikipedia articles in English and Nepali languages showed satisfactory results with an average precision of 88% and a recall of 86% in detecting inconsistency. Since the proposed technique is not language specific it has an advantage over the content-based techniques by supporting variety of languages.

2. Guidelines on Consistency from Preferences in Sharing specific Content Categories.

   Given that several content categories are published in websites and shared among communities analysis based on propagation is proposed

to examine the influence of specific content categories on preferences in sharing. The approach is to qualitatively compare content in webpages and examine their propagation among country-specific websites first in website graph (inter-connecting the available websites) and then in website pairs. For this study 480 webpages from 80 websites representing 10 global brands (Nivea, 3M, Starbucks, Acer, Samsung, KPMG, HP, Nestle, Avon, John Deree) are analyzed. A total of 480 comparisons of webpages in website graph and 1680 comparisons in website pair are performed to determine the preferences in sharing specific content categories. From examining propagation in website graph we revealed that "Corporate Information" has tendency to be shared globally and "Customer Support Information" has tendency to be shared locally while "Product Information" tends to be locally and regionally suitable for sharing. Implication is the guidelines on content consistency needed for specific content, example global consistency required for 'corporate related information' while local consistency required for 'customer support related information'. From examining propagation in website pair coupling in websites is revealed with high coupling for 'corporate related information' which decreases as the content becomes local. Implication is the guidelines on setting priority where high coupling means higher priority for content consistency for example 'corporate related information' is of high priority in content consistency. Such guidelines are useful in dealing with global and local inconsistency in cross-site content.

3. Guidelines on Consistency from Preferences in Sharing within and beyond Geographic Regions.

   Country-specific websites that offer various content categories also represent geographic regions such as Europe, Asia Pacific, North America and so on which is important to consider as regional discrepancies in cross-site content are found to present in such websites. Analysis based on propagation is proposed to determine preferences among communities in sharing within or beyond specific geographic

regions. The proposed approach is to qualitatively compare content in webpages and examine their propagations in several geographic regions. For this study 80 websites from geographic regions North America, Asia Pacific, Europe and Middle East-Africa are analyzed. A total of 240 comparisons of webpages within region and 1440 comparisons among regions are performed to determine preferences in sharing for specific region. From examining propagation within geographic regions high coupling in websites among countries in Europe and low coupling in websites inside North America is revealed. Websites in Europe tend to be more dependent and prefer to share most content in comparison to websites in North America while websites inside North America tend to be autonomous and prefer to participate less in sharing. Implication is the guidelines that among all regions European region is more vulnerable to intra-regional discrepancies and have higher priorities for content consistency. From examining propagation among geographic regions the autonomous nature of websites in North America is further suggested. Guidelines on higher priories for content consistency are suggested among Asia Pacific, Europe and Middle-East Africa to avoid inter-regional discrepancies in cross-site content.

4. Deploying Pattern of Sharing to Propagate Content Updates.

   To support content consistency allowing community preferences in customizing knowledge sharing, a technique based solely on the concept of propagating content updates restricted to specific languages or specific community is proposed. Pattern of sharing (a) Internationalization (b) Regionalization and (c) Localization with rules for restricting the publication and description of content to specific languages or community is deployed in knowledge sharing. Community preferences specified with pattern of sharing is able to deal cross-site content inconsistency from scaling content specificity for global, regional or local communities and propagating content updates confined to specific communities. The advantage is its simplicity in applying

either automatically or executed manually as policies.

5. Support for Consistency without reliance on content processing.

   The problem surfacing limited support to resource poor languages is the dependence on content processing and necessity for massive linguistic corpuses in training systems which is unfortunately not available for resource deprived communities. To support content consistency in variety of languages including the resource poor languages techniques proposed in this thesis do not require content processing. The techniques are based on novel concept of synchronizing user editing action and restricting content updates with propagation which is not language specific and hence support community participation including the resource deprived ones.

From the techniques that are simple and applicable to variety of languages along with the guidelines for content consistency to deal with (a) inconsistency in multilingual content (b) global and local inconsistency as well as (c) regional discrepancies in cross-site content; this thesis contributed in the design of multi-language knowledge sharing system catered to knowledge sharing goals of communities both for leveraging knowledge equally and customization in knowledge sharing.

# Acknowledgments

First and foremost I would like to express my deepest gratitude to my advisor Professor Toru Ishida for the continuous support of my Ph.D study and related research, for his patience, motivation and immense knowledge. A teacher is a compass that activates the magnets of curiosity, knowledge, and wisdom in the pupils. I am extremely fortunate to have you as my teacher who inspired me to grow as a research scientist. This thesis is only possible because of your persistent guidance. I am also thankful for giving me this lifetime opportunity to acquaint with Japanese culture and tradition.

I also owe my sincere gratitude to my thesis committee members, Professor Katsumi Tanaka and Professor Katsuya Yamori for their insightful comments, encouragement and thought provoking questions, which stimulated creative thinking and polished my research.

I also like to express my gratitude to faculty members and alumni at Ishida and Matsubara Laboratory: Associate Professor Shigeo Matsubara, Associate Professor David Kinny, Associate Professor Hiromitsu Hattori, Assistant Professor Yuu Nakashima, Assistant Professor Arisa Ema, Lecturer Reiko Inaba, Researcher Masayuki Otani, Researcher Takao Nakaguichi. My sincere appreciation goes to Associate Professor Yohei Murakami and Assistant Professor Donghui Lin who always accompanied me with fruitful discussions, gave their practical advices and closely monitored my research progress. I thank them for co-authoring papers with me. I also greatly appreciate our coordinators, Ms. Hiroko Yamaguchi, Ms. Terumi Kosugi, Ms. Yoko Kubota, and Ms. Yoko Iwama, for their help in administrative affairs throughout the academic process.

Words are not enough to express my joy and excitement in sharing this achievement with my parents, father Dr. Madan Pariyar and mother Urmila Pariyar. Both are the pillars behind my success, they encouraged me and at times consoled me in pursuing this dream. I touch their feet (respectful salutation in Nepal) and thank them wholeheartedly. My family members, older brothers Bipin and Sachin, brother-in-law Dr. Alin, older sister Sapana and her children Anwesha and Saumya, all encouraged me with their wishes. I also offer my prayers to the holy temple, Pashupatinath.

Special thanks also go to all my lab mates: Ari Hautasaari, Julien Bourdon-Miyamoto, Andrew W. Vargo, Huan Jiang, Chunqi Shi, Mairidan Wushouer, Kemas Muslim Lhaksmana, Xin Zhou, Trang Mai Xuan, Shinsuke Goto, Hiroaki Kingetsu, Xun Cao, Wenya Wu, Nguyen Cao Hong Ngoc, Arbi Haza Nasution, Mondheera Pituxcoosuvarn, Victoria Abou Khalil, Shohei Hida, Akihiko Itoh, Hiromichi Cho, Kaori Kita, Daisuke Kitagawa, Jun Matsuno and many others. Thank you all for enriching my life in Japan. I will always cherish the beauty of Kyoto city and its people. We only part to meet again, Sayonara!

My stay in Kyoto University was supported by the Japanese Government (Monbukagakusho) Scholarships from April 2011 to March 2015. This research was partially supported by Service Science, Solutions and Foundation Integrated Research Program from JST RISTEX, and a Grant-in-Aid for Scientific Research (S) (24220002) from Japan Society for the Promotion of Science.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Overview

With recent advances in technological support for online community participation knowledge is no longer confined within specific languages or localities; rather knowledge is shared beyond linguistic, cultural and geographical barrier [Ardichvili et al., 2006, Fong Boh et al., 2013]. The ever growing online knowledge repository such as Wikipedia with content contributed in more than 200 languages or the adoption of wiki as project management, technical support tool in multinational project elucidate the increasing participation for knowledge sharing despite language differences. The prospect for online collaboration is further raised from the growing trend in promoting product and services world-wide via content offered with websites targeted for specific localities, mostly noticed in global brands. According to web globalization report [Yunker, 2014], leading global brands offer country-specific websites for more than 100 countries supporting more than 40 languages. The implication is massive content published and shared in several languages; a daunting and challenging task is how to manage inconsistency in shared content to support multi-language knowledge sharing among the communities. The term "multi-language" is meant to depict community participation in a broader context which includes knowledge sharing in common language as well as in different languages among communities.

The term "Inconsistency" has several definitions in database system and knowledge-based system. In database system, "Inconsistency typically means different copies of same data having different values [Sumathi and Esakkirajan, 2007]. Inconsistency in logic is said to occur from data with conflicting values, for example data (Obama's age = 60) and (Obama's age = 55) in multiple records cause logical inconsistency since a single person cannot have multiple ages at the same time. Such inconsistency in database system is usually avoided as relationship between data that must hold for logical consistency is explicitly expressed during the database design. This means integrity constraints such as entity, referential constraint is enforced to assure data consistency.

However in knowledge-based system, "Inconsistency" typically means contradiction in knowledge bases with a notion of strong and weak contradiction resulting in inconsistency in logic and fact about the world [Nguyen, 2007]. A contradiction is strong when fact $X$ about the world and its negation $\neg X$ both hold simultaneously which makes knowledge base logically inconsistent. For example in the world of Obama the knowledge bases (Obama plays tennis in weekend) and (Obama does not play tennis in weekend) is logically inconsistent. A contradiction is weak when fact $X$ about the world has overlapping in its descriptions. For example (Obama plays tennis in weekend) and (Obama plays tennis and soccer in weekend) is not logically inconsistent as a person can have multiple hobbies but are factually inconsistent due to overlapping in the description of the facts. Contradiction in logic-based knowledge bases is usually avoided from expressing relation as logical formulae $\{x \wedge y, x \vee \neg y, \neg x\}$ and checking their interpretation (truth or false) with an inference engine.

Previous definitions overlap in the following points (a) inconsistency in logic and facts about the world and (b) the importance of expressing relation in the form of consistency rules to resolve inconsistency [Easterbrrok and Nuseibeh, 1996, Sumathi and Esakkirajan, 2007, Nguyen, 2007, Nuseibeh et al., 2001]. Inconsistency in both logic and fact are also equally applicable in multi-language knowledge sharing system. However the problem of inconsistency is unique because it is impractical

to produce consistency rules for content shared in each language among community and that too in advance. Therefore the design of multi-language knowledge sharing system has to shift focus from consistency rules and highlight on cases that can possibly cause inconsistency in logic and fact as communities share content.

Past studies have hinted on poor quality translations, asymmetries, intransitivity in shared content due to back translation and cascaded translation as a potential source for inconsistency in multilingual collaboration [Yamashita and Ishida, 2006, Inaba et al., 2007]. Given the evolving nature of content, the cases such as content omitted or content updates not propagated and conflicts from content updated by several communities are likely to occur in collaboration for knowledge sharing with potential for causing inconsistency in logic and fact. Such cases are an important consideration when designing multi-language knowledge sharing system.

Another issue that has to be clarified in design of multi-language knowledge sharing is consideration for open world or closed world assumption. In knowledge-based system, a "closed world assumption (CWA) typically means that when a fact cannot be derived from the knowledge base then that fact must be false [Brodie and Mylopoulos, 2012]. This means that the knowledge base is assumed to be complete. For example if a knowledge base contains only one fact Play(Obama, Tennis) then under CWA the fact Play(Obama, Soccer) is false. However in an "open world assumption" (OWA) the fact Play(Obama, Soccer) is unknown (either true or false). It can only be verified if there are more information available on hobbies of Obama. The OWA assumes that the knowledge base is incomplete.

Similar concept applies in multi-language knowledge sharing system. Under CWA whatever content is published by community C1 also implies for community C2 or vice versa. Which means consistency in knowledge sharing is achieved only when content is shared between community C1 and C2 or vice versa, example content Play(Obama, Tennis) published by community C1 and shared with community C2. Inconsistency occurs when the content Play(Obama, Tennis) is missing in community C2. In OWA the content published by community C1 may not imply for community C2 and

so community C2 may chose not to use the content published by community C1. As in above example, Inconsistency do not occur when the content Play(Obama, Tennis) published by community C1 is missing in community C2. In fact both communities are free to publish their own content and chose to share or not to share. Consistency in knowledge sharing is achieved only from the consensus among the communities C1 and C2.

Considering the potential cause for inconsistency in logic and facts, the varying notion of inconsistency in open world and closed world assumptions along with the previous definitions in inconsistency management

**Definition.** Inconsistency in Multi-Language Knowledge Sharing System is defined for situation in which relation that must hold logically and factually consistent in the shared content under closed world assumption is violated from cases such as absence of content, lack of updates propagation and content conflict.

The complexity of inconsistency due to mentioned cases such as absence of content, lack of updates propagation and content conflict increases from the fact that the communities participate in their preferred language and such cases potentially reside in several languages. More severe is regional discrepancies which is also anticipated from the presence of inconsistent content shared possibly among country-specific websites managed in global brands which is a serious concern. Empirical studies have stressed that inaccurate, outdated information in websites have the potential to form poor perception of the brand leading to dissatisfaction in customer [Barnes and Vidgen, 2002, Palmer, 2002]. On this ground, inconsistency in shared content is problematic for knowledge sharing and an important concern in the design of multi-language knowledge sharing system.

Besides inconsistency that are plausible among communities in knowledge sharing, the content consistency constraint is also equally important when it comes to the specific needs of the communities [Hofstede and Hofstede, 2001, Hillier, 2003, Sun, 2001]. Constraint in con-

tent consistency means 'the diverging view on supporting content consistency' which is found reliant on the goals of knowledge sharing. Where the goal is to leverage knowledge equally among communities, content consistency that strictly enforces one-to-one correspondences in the shared content, in other words a 'rigid consistency' policy is appropriate. Documents such as technical manuals, software documentations that are produced with the intention to share same information in several language editions are the possible candidates for knowledge sharing that require exact correspondences in shared content. However, inconsistency is bound to occur in such documents as the content is reviewed and edited in multiple languages by several communities. The necessities for sharing consistent content in several languages for such documents also emerge from the growing proximity and cultural homogeneity among the communities.

Contrary to this, the persistence of cultural differences among communities is also widely supported in the past studies which shifts the goal from leveraging knowledge equally to knowledge sharing that is relevant to specific communities. In such situations knowledge sharing is not uniform among communities and exact correspondence in shared content is not always preferred. The constraint in content consistency also arises from the need to restrict publication and description of content in specific languages which makes 'non-rigid consistency' policy a better choice. Since communities prefer to share content of interest or significance restricted to specific communities, it also becomes essential to keep reference of what content is shared and to whom, in order to impose exact correspondence in the shared content where required. Apparently the complexity is also raised as the preferences of communities for knowledge sharing vary and not a single solution for content consistency rather content consistency customized for specific communities is appropriate. Due to constraint in content consistency that arise from the varying preferences in sharing among communities, the design of multi-language knowledge sharing system for managing inconsistent content has to take into account the knowledge sharing goals of the communities which mean either leveraging knowledge equally or customizing knowledge for specific communities.

## 1.2 Objective

Given the content consistency constraint in multi-language knowledge sharing where correspondences in shared content is either strictly enforced or customized due to the underlying preferences of specific communities, it is required to have an understanding of knowledge sharing that occur in real world cases such as Wikipedia or websites. Towards the design of multi-language knowledge sharing system, the objectives in this thesis are as following.

- **To determine influence of specific content categories on preferences in sharing among communities.**

  As several categories of content is shared either in websites or in wiki, the possibility for constraint in content consistency arising from the preference that vary for sharing specific kind of content cannot go unnoticed. The diverging perspective on cultural influences in the design of websites is also relatable to the cultural influences in sharing. Where the view on cultural homogeneity favors the standardization of product and services across the globe; the standardization in knowledge sharing is yet to be investigated. Similarly where majority of researches have lenience towards Hofstede typology of cultural differences, knowledge sharing is also expected to be influenced from such differences. Of past researches where the design and content features are shown to vary with cultural groups, industry, product types and so on the scale in sharing for specific content categories is not investigated which largely determines the content consistency requirement for that specific content.

- **To determine influence of geographic region on preferences in sharing among communities.** Cultural differences are also found to be prevalent among geographic regions and are raised as a concern in past studies for localizing websites to specific locale. The depiction of western societies as individualistic low-context culture and eastern

societies as collectivistic high-context culture in their preferences for the use of instant messaging among North America and Asia have implications in knowledge sharing goals that cater to specific geographic region. Geographic influence is also shown to vary the perception of customer towards marketing stimuli and website effectiveness from specific region such as North America and Europe. With a concern for regional discrepancies in shared content, it is indeed required to understand the preference for sharing specific content categories restricted to specific geographic region.

From expanding our understanding on the preferences in sharing among communities, the content consistency constraint is explored which are useful in the design of multi-language knowledge sharing system. However from the technical standpoint there are also additional objectives that are to be met in the design of multi-language knowledge sharing system.

- **To enable participation of communities with inadequate language resources in knowledge sharing.**

  Towards bridging knowledge gap, past researches have tackled inconsistent multilingual content with techniques from language generation to language processing. However the support for content consistency to resource rich languages mostly European languages is predominant. Such techniques are also reliant on huge linguistic corpuses for training system in finding overlapping and differential content which is unfortunately limited for resource poor languages. Due to language dependency, replicating such techniques for content consistency in resource poor languages is also not practical. From inadequate support for consistency in shared content to resource poor languages, the essence of multi-language knowledge sharing is not truly achieved as communities with limited language resources are not involved in knowledge sharing. To encourage the participation of such communities in knowledge sharing, the techniques for content consistency that minimizes the need for language processing and support variety of languages including resource poor languages is appropriate.

- **To support content consistency applicable for both monolingual and multilingual cases.**

  As already highlighted, the term "multi-language" in this study associates the notion of community participation for sharing in preferred language which can be either in common language or different languages, hence collaboration in both monolingual and multilingual settings. The techniques for dealing inconsistency in shared content are also required to be equally supportive of both situations.

Following these objectives, the design of multi-language knowledge sharing system is an important undertaking in this thesis with the focus on consistency in shared content among communities while considering the knowledge sharing goals of the communities.

## 1.3 Issues and Approach

Though the necessity for content consistency in multi-language knowledge sharing among communities has gathered enormous attention due to increasing collaboration, several issues are apparent in the design of such system.

1. **Consequences from content inconsistency in knowledge sharing.** Due to content shared in several languages, inconsistency from cases that may seem trivial such as omitted content, updated content not shared, content modified in multiple languages is also painstakingly difficult to locate from enormous content. Further the source content that serves as originating source of information keeps changing between languages as contributions are made by communities complicating content consistency when translation needs to refer the source content. Globally and locally shared inconsistent content also emerge from such cases. Also severe are inter-regional and intra-regional discrepancies due to inconsistent content shared, most likely in websites

targeted for specific countries in global brands, for example inconsistent content in product specification for customer in Asia Pacific and North America. Such inconsistency in websites is not tolerable as they can form a poor perception of the brand.

2. **Preferences determining content consistency constraint are unknown.** From the view of fact that the goals of knowledge sharing vary from leveraging knowledge equally to customizing knowledge for specific communities, the challenge for imposing content consistency in multi-language knowledge sharing is two-fold. A rigid consistency policy for exact correspondences in shared content is to be enforced for knowledge sharing that requires content consistency among all communities whereas non-rigid policy is favorable for customized knowledge sharing that imposes restrictions in sharing of content among specific communities and in specific languages. On this part, one of the issues is the factors which set apart the content consistency constraints for knowledge sharing is not known.

3. **Inadequate support for resource poor languages in knowledge sharing.** Another issue for multi-language knowledge sharing arises from the minimal support for content consistency in resource poor language which deters the participation of communities with limited language resources. Though majority of world knowledge repositories are available in resource rich languages which at first glance seem an opportunity for resource deprived communities, sooner inconsistent content appear in resource poor languages in multi-language knowledge sharing due to dearth of techniques supporting consistency for such languages. The problem is how to support content consistency for communities with limited language resources for participation in multi-language knowledge sharing.

4. **Diminishing role of translation activity in promoting content consistency.** Though translation is seen as an important activity towards an attempt to bridge knowledge gap among communities that collaborate in dif-

ferent languages, the suitability of translation practices also raises concern. The conventional translation practices are primarily inappropriate for multi-language knowledge sharing due to their inability to propagate changes; lack of support for content reuse and reliance on pivot language usually English for all major changes. The role of translation is even diminished as inconsistency in shared content among communities is likely to occur not only among distinct languages but also in same language.

Content inconsistency in same language is possible in content shared among websites representing countries that share their official languages; example French is an official language common to countries in Europe and Africa. Inconsistency in shared content of global significance for example 'Outbreak of Ebola Virus' is likely from the contribution of content globally by communities where some share a common language. The problem is to how to shift focus solely from translation activity in handling content consistency to a broader context such as activities that support translation activity for content consistency in multi-language knowledge sharing.

Given the issues that highlight the content consistency constraints, the consequences from content inconsistency and the concern over encouraging participation of resource deprived communities, this thesis undertakes studies to promote consistency in shared content among communities in multi-language knowledge sharing. The goal of this thesis is to propose techniques for content consistency that cater to the knowledge sharing needs of the communities both for leveraging knowledge equally and to customize knowledge sharing for specific communities. Also equally of concern are the techniques that are not bound to specific languages, in other words applicable even for resource poor languages. To address the issues, following approaches are presented in this thesis as shown in Fig. 1.1.

1. **Synchronization of User Editing Activities to Detect Inconsistency in Multilingual Content.**

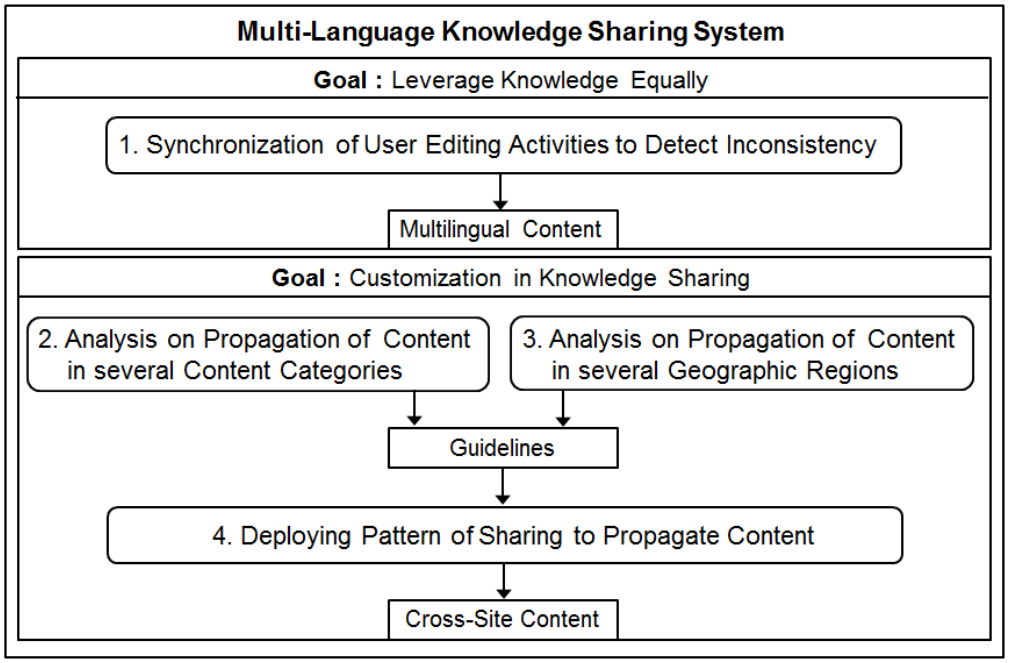



Figure 1.1: Outline on Issues and Approaches.

Primarily for knowledge sharing where exact correspondences in shared content is important the concept that is based on synchronizing user edit actions in reviewing and updating content in several languages is employed to detect the presence of inconsistent content. To realize this concept a state transition model is proposed to define the states of multilingual content, set of action performed on the content and set of transition functions that describe the state transition of the content. Based on this model, inconsistency detection rules are presented to specify the inconsistent states of the multilingual content. The proposed process-based technique is supportive to variety of languages including the resource poor languages due to no prerequisites for content processing, thus enabling the participation of resource deprived communities in multi-language knowledge sharing. From the experiment with multilingual Wikipedia articles in English and Nepali

language, the proposed technique is found to have an average precision of 88% and a recall of 86% in detecting inconsistency which is satisfactory given that the technique is based only on user actions. Such a technique integrated to the NLP based approaches can also simplify earlier phases in content synchronization before processing content.

For knowledge sharing where exact correspondences in shared content is not a compulsion, this thesis also posed to determine factors that give rise to preferences in sharing among communities. Realizing content consistency both in same and different languages, the state transition model cannot be directly applied when content is shared in same language, hence the techniques that rely solely on propagating content updates is proposed in later stages. Analytical studies are undertaken to determine the preferences in sharing among communities where knowledge sharing is customized for specific communities.

2. **Analysis on Propagation of content in several content categories to determine their preferences in knowledge sharing.**

   Qualitatively comparing the propagation of content in webpages among country-specific websites managed in global brand, the analytical study attempts to determine the preferences in sharing specific content categories. Examining propagation in a website graph that links available country-specific websites, traits such as scales in sharing that vary for specific content categories are revealed. It is found from the analysis that corporate related information are preferred to be shared globally and customer support related information are preferred to shared locally while product related information are preferred for sharing either locally and regionally. The varying preferences of sharing specific content categories give an indication on the content consistency constraints that exists while sharing such content in knowledge sharing. Further analyzing propagation in website pair revealed coupling in sharing for specific content categories. High

coupling in websites is found in sharing content related to corporate information which is useful as guidelines for manager in enforcing strict content consistency only for specific content.

3. **Analysis on Propagation of content among communities in several geographic regions to determine their preferences in knowledge sharing.**

   Qualitatively comparing the propagation of content in webpages among country-specific websites that represent several geographic regions, the analytical study attempts to determine preferences in sharing for specific region to account for regional discrepancies. Examining propagation within geographic region high coupling in websites among countries inside European region while low coupling in websites in North America are revealed which raised an important concern that content shared inside European region are more vulnerable to intra-regional discrepancies. From examining propagation among geographic region, websites from Asia Pacific, Europe and Middle-East Africa are found to participate mostly in sharing among themselves; hence more vulnerable to inter-regional discrepancies from sharing inconsistent content with customer in these region. Further inspection also revealed that websites in North America have higher preferences to generate specialized product related information which are not shared with other region; while customer support related information are specialized inside all regions and not shared among each other. Such an understanding of how communities that represent several geographic regions respond when it comes to sharing content from specific content categories among each other is useful as guidelines for web manager willing to promote consistency in knowledge sharing.

4. **Deploying Pattern of sharing to propagate content among communities.**

   Primarily for knowledge sharing where customization in content

shared is essential the concept that is based on propagation restricted to specific language or communities due to specific pattern is employed to share up-to-date knowledge. Rules are associated with such pattern that restricts the publication and description of content in specific language and hence restricts the propagation of content updates. Pattern of sharing such as Internationalization, Regionalization, Localization and their combinations are proposed for consistent knowledge sharing which delivers consistent content by propagating content updates restricted to global, regional or local communities. The advantage of such pattern of sharing is that it can be deployed either automatically once specific content categories are identified or manually as policies.

From the techniques proposed for content consistency and guidelines compiled from determining preferences in sharing among communities, this thesis caters to the knowledge sharing goals of leveraging knowledge equally and customization for specific communities and makes an important contribution in the design of multi-language knowledge sharing system towards promoting consistency in shared content.

## 1.4 Thesis Outline

This thesis consists of seven chapters including Chapter 1. The content of each of the remaining chapters are summarized next.

Chapter 2 introduces the background on knowledge sharing which includes the discussion of previous works that are focused along the spectrum of leveraging knowledge equally among communities at one end and customizing knowledge sharing for specific communities at the other end. In doing so, the techniques and tools that cater to both knowledge sharing goals are studied to have an understanding of their support to communities in general, including the resource deprived ones. Further the shortcomings of the conventional translation practices along with the essential features and difficulties of collaborative authoring and translation is also presented for an

overview of design requirements in multi-language knowledge sharing system.

Towards addressing the knowledge sharing goals of communities in leveraging knowledge equally, Chapter 3 proposes a process-based technique to detect the presence of inconsistent content shared among communities. From the real world example of collaboratively generated multilingual Wikipedia article between resource rich and resource poor language, this chapter depicted inconsistency in content shared in multiple language editions and raised the need to promote knowledge sharing equally among communities, including the resource deprived ones. To promote content consistency, this chapter introduces a state transition model and inconsistency detection rules which is based on user editing actions and supports variety of languages.

Chapter 4 shifts the focus on knowledge sharing goals of communities from leveraging knowledge equally to sharing customized for specific communities. In doing so, the technique for content consistency that embodies collaboration among communities in both same and different languages is appropriate. Moreover, the preference among communities in sharing is an important issue and a deciding factor for imposing content consistency constraints. From the real world example of content shared among country-specific websites, this chapter depicted inconsistency in content shared among communities inside global brand, especially in cross-site content. With an intention to determine the underlying preferences in sharing analysis based on propagation is proposed in this chapter to qualitatively compare content in website and examine their propagation in website graph and website pairs. The tendency in sharing specific categories that vary in scales and coupling is revealed which is important to prepare guidelines on content consistency in multi-language knowledge sharing system.

Further knowledge sharing among communities that represent several geographic regions is studied in Chapter 5 to determine their preferences in sharing. Again from the real world example of content shared in webpages from websites representing several geographic regions, this chapter illustrates the prospects for inter-regional and intra-regional discrepancies from

inconsistent shared content. Analysis based on propagation is undertaken to examine the managerial preferences in sharing content by qualitatively comparing propagation within and beyond geographic region as depicted in the country-specific websites managed inside global brands. Traits such as coupling and scales in sharing that vary for specific geographic region and specific content categories are revealed which serves as a guideline for manager in the design of content consistency policy customized for specific region.

In Chapter 6 grounding on the results of the analytical studies from former chapters 4 and 5 for determining preferences in sharing, a simple technique that enables propagation of consistent content among communities is presented. With a concept on restricting propagation of content updates to specific languages or specific communities, pattern of sharing such as Internationalization, Regionalization and Localization with rules for restricting the description and publication of content is proposed for consistent knowledge sharing. Besides the simplicity of technique, the advantage is its suitability for both monolingual and multilingual cases including the support for automation as well as policy specified manually in sharing.

Finally Chapter 7 concludes this thesis by discussing the summary of contributions made for supporting content consistencies in multi-language knowledge sharing and also suggesting possible future directions.

# Chapter 2

# Understanding Knowledge Sharing Goals among Communities

Globalization has fostered from the rapid technological advancement facilitating upsurge in the communication and collaboration among diverse communities that are geographically, culturally and linguistically distinct. Wikipedia reflects one such community that collaboratively creates and maintains online knowledge repository. Where knowledge is pulled from communities in Wikipedia [OLeary, 2009b, Kussmaul and Jack, 2009]; the presence of websites that are representative of global organization is geared toward pushing knowledge to specific community [O'Leary, 2008]. Notable differences exist in knowledge sharing goals among such communities that have different orientation, either towards language or location [Yunker, 2002]. Wikipedia is primarily language-oriented with vast amount of knowledge dispersed in several languages for consumption to global communities while websites are location-oriented with knowledge sharing customized for specific country, its culture and official languages in mind. Bulk of past researches have addressed the knowledge sharing goals along the spectrum with research communities at one end focused on leveraging knowledge equally across the languages for bridging the knowledge gap;

while research communities at other end focused on the strategies in customization with the cultural, linguistic and business context of target audiences for sharing knowledge, for example web globalization.

## 2.1 Leveraging Knowledge Equally

Though Web has opened channel for the multilingual communities to contribute knowledge it has resulted in knowledge that is scattered in several languages widening the knowledge gap. Wikipedia which embodies the repository of world knowledge diversity with several language editions is an illustration of such knowledge gap among communities. Such diversity in knowledge offers a unique opportunity for the researchers to investigate for the potential exploitation at a global scale. The availability of high quality information resources maintained only from the volunteer editorial services is also attractive to social science researchers.

To address the knowledge gap past researches have employed language generation to language translation with automated techniques as well as those that rely upon having "human in the loop". From centralized representation of multilingual correspondences to creation and maintenance of multilingual content collaboratively in a decentralized context has also been proposed. The researches converge towards the tool implementation that support range of users from technical writers to novice along with features for leveraging content from one language to several languages in bridging the knowledge gap.

### 2.1.1 Language Generation Technique

Language Generation techniques were popular with the task for generating natural languages from the machine representation system such as knowledge base or a logical form. Such representation opened opportunities for automating most of the repetitive tasks for multilingual authoring by automatically generating multilingual instructions from the underlying formal models. Alternative to the contemporary translations from source docu-

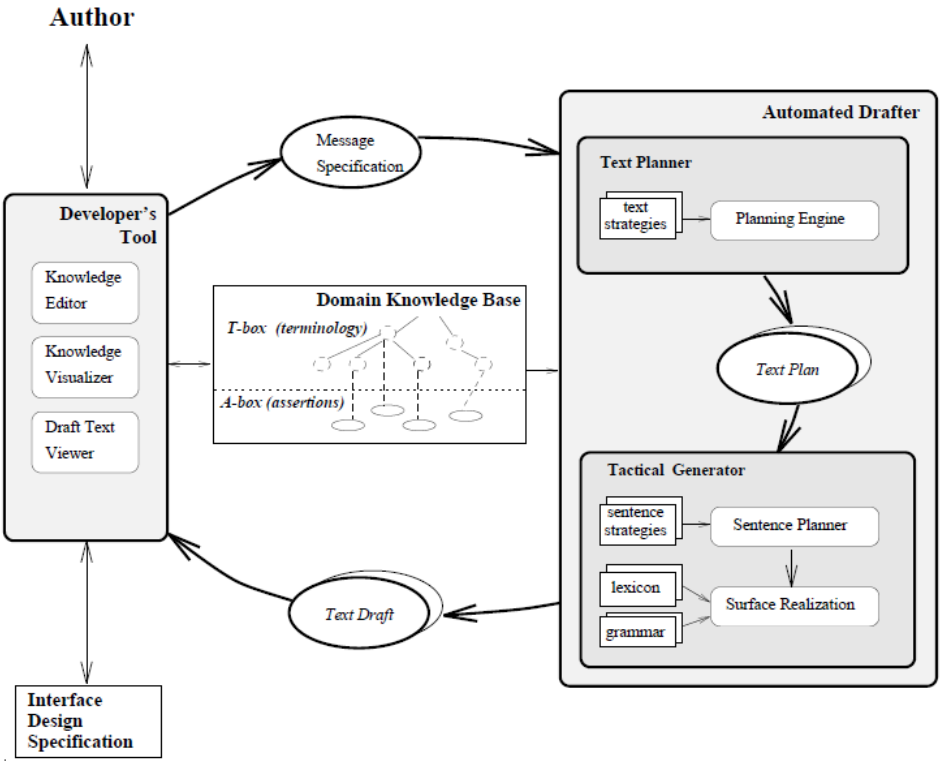



Figure 2.1: Architecture for Language Generation in Drafter [Hartley and Paris, 1997].

ment, language generation technologies eliminated "source language biases" from a language neutral representation of master source in generating document in several languages independently and automatically.

Based on language generation techniques, symbolic authoring is employed in [Scott and Evans, 1998] to represent symbolic representation of the content implemented as LOOM knowledge base with a vision towards multilingual document management without translation. The automated generation of multilingual instructions for software manuals from such semantic knowledge base is also attempted with an interactive drafting tool called Drafter [Hartley and Paris, 1997, Paris and Vander Linden, 1996]. Fig. 2.1 presents the architecture of Drafter with two main supporting tools, the developers tool that allows technical communicators to specify formally

the procedures for certain task and drafting tool to generate the text automatically with specific styles from the domain knowledge base.

Equally applicable to pharmaceutical industries, the representation of the knowledge base as master model for the generation of medical documentation in several languages is implemented as PILLS (Patient Information Language Localization System) project [Bouayad-Agha et al., 2002]. The support for accurate and consistent terminology along with the propagation of content updates throughout document and languages are achieved from such tool. The consistency of content among the linguistic versions of the documents is achieved with modifications only occurring at a single place i.e. the underlying knowledge base.

However in such language generation tools the challenge is to find a convenient way of creating and maintaining the content model. Much followed is WYSIWYM editing with the user interface that enables the author in constructing and modifying the knowledge base graphically. Multilingual documents authored in controlled languages using direct manipulation interface of the knowledge base is also presented in [Power et al., 2003] with corresponding text generation taking longer. Unlike other collaborative tools which are discussed below, Drafter is primarily meant for technical writers and requires expertise knowledge in constructing the knowledge base which limits its usefulness to the online community of volunteer translators. In addition to this modifications to the content are made by modifying the underlying language neutral representation such as knowledge base which complicates the task for non-experts.

### 2.1.2 Language Processing Technique

With an inclination towards wiki systems, several techniques employ language processing using resources such as machine translations, dictionaries and so on utilizing the linguistic corpuses for leveraging knowledge from one language to several languages automatically or with human assistance. Tracking changes in the linguistic versions and highlighting discrepancies among multilingual knowledge repositories also form the majority of re-

search contributions. The restructuring of multilingual correspondences in a unified format is also practiced to leverage knowledge sharing.

**(a) Maintaining Multilingual Correspondences**

The differential growth rates in the linguistic content provide a unique opportunity for leveraging the articles in one or more languages to improve the content in another. The researches in [Adar et al., 2009] are focused on the notion of information arbitrage across Wikipedia as a mechanism to exploit linguistic differentials in detecting missing, old or incorrect information in one language's corpus to fix the data in another.

[Adar et al., 2009] introduced Ziggurat, an automated system to align info boxes in several languages for matching the field values pairs to fill missing information or detect discrepancies. Though the system minimizes the dependence on machine translation; self-supervised learning techniques are employed to build classifiers requiring the accumulation of huge data sets which may not be available for resource poor languages. The complexity also increases with the increase in the languages for comparing field-values in each language pairs; currently it is limited to four large language domains mostly European languages. Where Ziggurat is technically focused, the design requirements are the main contributions of LizzyWiki [Désilets et al., 2006].

In addition to the feature of the wiki engine, LizzyWiki largely incorporated design requirements from the user-centered design experiment for identifying the user roles. The separation of user roles into visitor, content author and content translation and the need for switching between the roles is addressed with LizzyWiki. The support limited to bilingual sites and inflexibility in the workflow for the content author to always bringing the page first up-to-date in their linguistic version with respect to its counterpart before making any subsequent modifications are the major drawback of the tool. Unlike Ziggurat, [Bronner et al., 2012] highlighted on the inconsistencies in content from the diversity of contributions in different languages and proposed content synchronization with a framework CoSyne. Employing

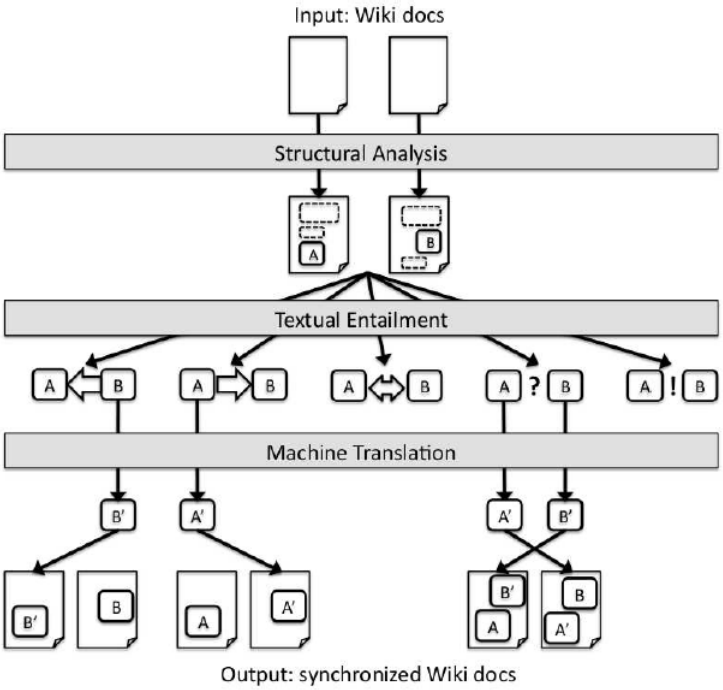



Figure 2.2: Content Synchronization Framework in CoSyne [Bronner et al., 2012].

the state of the art machine translations, natural language processing techniques, multilingual concept networks and cross-lingual entailments; synchronization of content is however limited to European languages. Several components in Cosyne for identifying semantically coherent segments in multiple language and the entailments for uni/bidirectional relation between segments to achieve content synchronization is depicted in Fig. 2.2. While Ziggurat focused on field value pairs, CoSyne focused on the body of the articles pinpointing the topically related pieces of information in different languages; identifying the information that is missing or less detailed in any of the language; translating them in appropriate language in the appropriate place. Clearly the direction is towards the automation with well-resourced language in mind.

Systems like Wikibhasha and Google translator toolkit are also

employed in leveraging knowledge among multilingual communities [Kumaran et al., 2008]. The aim is towards achieving partial automation with the use of machine translation to aid human in propagating information from one Wikipedia typically English to another language edition with a small number of articles. Such collaborative framework enables leveraging fairly stable content from source language to create rough initial content in target language which is then collaborative corrected by the target communities. Despite content reuse across language communities is simplified, it is also crucial to focus on the correspondences between multilingual content as the multilingual editing cannot be controlled. Creation of multilingual content from the collaborative participation of communities is promising towards bridging the knowledge gap.

**(b) Highlighting Knowledge Discrepancies**

The vast majority of differences in the knowledge represented for specific concept in several language editions of Wikipedia are also highlighted with an interactive visualization techniques in Omnipedia [Bao et al., 2012]. From 25 language editions of Wikipedia, the coverage of articles in various language domains is simultaneously accessed identifying the most commonly and globally discussed aspects of a concept. Such visually rich features are appealing for the communities to gather specific knowledge from referring to the multiple language communities. Though design premise for Omnipedia is language neutral way enabling the user to switch the interface language to any of the supported languages, the mechanism itself involves complicated information organization strategies.

Bridging knowledge gap from the collaborative authoring and translation of content in multiple languages, architecture for cross lingual wiki engine is proposed by [Huberdeau et al., 2008]. The use of abstract change tokens independent of language and textual content is employed to track changes in multiple languages. Tokens for each edit are generated and added to an edit set when updating the particular linguistic version of a page. The missing edit tokens in the pages of remaining languages are sufficient to sig-

nal the updates needing propagation in the corresponding pages. However the part of text in the page that needs actually to be propagated or translated to other language is not known.

The limitation in the accuracy of the up-to-datedness measure is also highlighted as it is based on the counting the number of insertion and deletion of characters. In natural language changing from singular to plural even with the addition of single character can have major change in the overall context of the sentence and score highly for the up-to-datedness measure. The distinction in the kind of edits such as major or minor edits which is normal in wiki system is also beneficial to the architecture proposed in cross-lingual wiki engine.

**(c) Restructuring Multilingual Correspondences**

The correspondence between the linguistic versions is also highlighted in [Al Assimi and Boitet, 2001] to deal with non-centralized evolution of multilingual documents from the modifications in several languages. [Hajlaoui and Boitet, 2005] presented methodology for managing multilingual correspondences between segments in parallel multilingual document with alignment tools centralizing the collection of modified segments, organizing them to recreate modified document in source language and translating them back to subsequent language versions. Such centralized representation of multilingual correspondences is inappropriate for volatile environment where the originating source of information continually changes among languages as it delays the propagation of content updates. To exploit the collaborative and open editing functionalities on the web, the methodology is extended with TransBey [Bey et al., 2006] tool to annotate the sources. XML based architecture DITA is also used in [Traicu and Prostean, 2012] as a model for multilingual document management integrating automated translation management components and machine translation algorithm.

Even in multilingual websites the problem of content inconsistencies between languages is severe and demands for consistency management tech-

```
<HTML>
<HEAD>
    <TITLE>
        <ML lang="en"> Cups of coffee consumed </ML>
        <ML lang="it"> Tazze di caffe' consumate </ML>
    </TITLE>
</HEAD>
<BODY>
    <TABLE border="1">
    <TR>
        <TH>
            <ML lang="en"> Name </ML>
            <ML lang="it"> Nome </ML>
        </TH>
        <TH>
            <ML lang="en"> Cups </ML>
            <ML lang="it"> Tazze </ML>
        </TH>
        <TH>
            <ML lang="en"> Type of sugar </ML>
            <ML lang="it"> Tipo di zucchero </ML>
        </TH>
    </TR>
    </TABLE>
</BODY>
</HTML>
```

Figure 2.3: Restructuring Page to MLHTML [Tonella et al., 2002].

niques. Not only the available information are to be displayed in same format; also intra and inter language hyperlinks should be consistent with the overall site organization. [Tonella et al., 2002, Tonella et al., 2006] proposed restructuring process as the single target representation MHTML (Multilingual XHTML) for centralizing the language dependent variants of a web page as shown in Fig. 2.3 thus guaranteeing the propagation of content updates across several pages. The techniques are reliant on natural language processing employing language identifications, page alignments to identify corresponding parts of the pages in representing a unified structure.

The maintenance related to updating pages is performed on the ML-HTML representation to ensure consistent propagation of changes to all site versions in different languages. However determining the originating source

of information in a particular language is still problematic with centralized representation and similar to [Al Assimi and Boitet, 2001] approach this requires management overhead in first bringing the linguistic version of content to a unified representation Though researches discussed here are motivated towards leveraging knowledge equally among communities for the benefit of global communities; customization due to culture, language, business context or geography and so on in knowledge sharing is also prospering. The next section presents the strategies and the implication of content features in their customization for sharing knowledge.

## 2.2 Customizing Knowledge Sharing

With growing proximity among communities from globalization, the persistence of cultural differences is quite debatable; researchers view cultural homogeneity among countries arising from the dominance of western culture favoring standardization in the product and services across the globe [Hall, 1997, Main, 2001]. Contrary with this, the popular Hofstedes typology of culture emphasizes cultural differences among countries implying the significance of product and service localization to the target market [Hofstede and Hofstede, 2001]. Confounding the homogenizing effect of international culture, previous research in particular web globalization has suggested strategies in employing globalization with customization in knowledge sharing via websites [Kale, 1991, Singh et al., 2005].

### 2.2.1 Strategies in Sharing Content

Incorporating opposing cultural views; globalization strategies in organization are widely studied in websites. Following strategies for knowledge sharing in relation to content shared among in-country offices from managing websites are compiled [Singh, 2011, LionBridge, 2009].

**(a) Centralization**

Strategies for globalization are handled centrally in most organization especially by headquarter office centralizing activities related to knowledge sharing occurring unidirectional only from headquarter office to in-country offices. In managing websites, the content for corporate websites are managed centrally such as what content is to be published for certain market, what is to be translated, what not to be translated and so on are decided from the central authority. Though such effort toward centralization streamlines the activities with business goal ensuring brand preservation, they fail to involve knowledge from the regional view and prospect adequately from the participation of in-country offices. Simply translating the corporate website in different languages for specific markets does not resonate with the target audiences in sharing knowledge from centralized effort.

**(b) Decentralization**

Country-specific websites that are managed by their respective in-country offices offer knowledge sharing with decentralization strategies in globalization. Though the customization in design and content are achieved for each target market, the possibilities for inconsistent branding, fragmented localization, inappropriate content occurring in the absence of well-defined guidelines are problematic for knowledge sharing.

The better approach for globalization strategies is a hybrid model, for example collaboration that brings together central corporate marketing department and regional marketing team to ensure brand preservation while developing local programs that complies with corporate goal and standards. Tools supporting knowledge sharing among teams scattered in global organization for managing websites should support hybrid model to promote global consistency and local flexibility in knowledge sharing where required.

### 2.2.2 Features in Sharing Content

Content and design features in the websites are also investigated in past researches to study their influences in sharing among various cultural groups. The content component addresses what is included in the website and identifies the various type of information which is customizable for sharing. Given this organization are able to shape and define their image from information rich websites. Whether such content and design features in websites are globally standardized or customized provides clues in knowledge sharing customized for specific categories of content.

[Robbins and Stylianou, 2003] proposed content features which include: corporate information, communication/ customer support financial information and so on as shown in Fig. 2.4 and identified that the majority of content features are significantly different across the websites in various cultural groups. Such finding is useful as it clearly illustrates knowledge sharing that is not standardized; rather localized for specific communities. This has implication to researches focused on bridging knowledge gap in considering customization of content features while sharing. Similar study without cultural influences, by [Huizingh, 2000] also illustrated content and design features related with the size of websites; with large websites incorporating most of the content and design features. Integration of content features in sharing provides customization techniques to target market for knowledge sharing.

[He, 2001] also studied the adaption of content features in websites that are meant for local and global communities. As described by Hopkin certain content are written directly in each language for the local market and is a reflection of local culture while some content are translated to many languages for worldwide use and is relatively insensitive to national or cultural differences. Such claims are interesting as it supports the need for customization of knowledge sharing while the specific categories of content that exhibit such features are unknown. [Yunker, 2002] suggested a global content model to incorporate features with scoping content for corporate, regional and local to support content creation and management from author-

ing, reviewing to approvals. Though content features and content scopes are suggested from past researches; the relation of content and their scope in publication are crucial for knowledge sharing.

In a broader context, the cultural differences among geographic region in previous researches have also raised concerns in content features targeted for specific region. The depiction of western societies as individualistic low-context culture and eastern societies as collectivistic high-context culture in their preferences for the use of instant messaging among North America and Asia also have implications for sharing content that cater to specific geographic region [Kayan et al., 2006]. The differences in the perception of customer to marketing stimuli and website effectiveness from specific region such as North America and Europe also support geographic considerations in shared content [Chakraborty et al., 2005]. Catering to specific region, the location specificity in knowledge sharing among certain geographic region also put forth that relevance of knowledge is confined within specific region and transferring the same knowledge to other region is a futile practice [Ambos and Ambos, 2009].

Several other features such as the number of languages for publishing content; navigation to locate locale-specific content; global consistency with global design template across locales while local relevancy of website with user's culture and country in [LionBridge, 2009] are also crucial for implementing knowledge sharing among multilingual country-specific websites.

### 2.2.3 Categories in Sharing Content

Several categories of websites are proposed in researches emphasizing the extent of localization effort in the content and design; a clue in the level of managerial support needed for managing knowledge sharing. [Tixier, 2005] classified websites into three categories: global websites offering no cultural adaptation, glocal websites offering cultural adaptation to appeal to local markets but not truly localized and local websites fully localized for local cultures and customs.

Such categories imposes restriction in the sharing of content among the

| Content features | Design features |
|---|---|
| Corporate information | Presentation |
| Biographical sketches | Animation |
| History | Frames |
| Message from CEO | Graphics |
| Mission statement | Sound |
| Organizational charts | Video |
| Press releases | Navigation |
| Vision statement | Hyperlinks to other sites |
| Communication/customer support | Protected contents |
| Corporate phone number | Search engine |
| E-mail opportunity | Site/map/index |
| Frequently asked questions | Security |
| Headquarters address | Secure access |
| On-line chat with an expert | Speed |
| Currency | Download time of home page |
| Current content | Download time between pages |
| Last updated indicator | Tracking |
| Financial information | Use of cookies |
| Annual report | |
| Financial highlights | |
| Employment opportunities | |
| Employment overview | |
| Job openings | |
| Social issues | |
| Cookie disclosure | |
| Cultural sensitivity | |
| Language translation | |
| Privacy issues | |
| Social responsibility | |

Figure 2.4: Content and Design Features [Robbins and Stylianou, 2003, Huizingh, 2000].

country-specific websites, for example what content is to shared globally or what content is to localized for specific markets, what content to be targeted for specific regions and so on. [Singh and Pereira, 2005, Singh et al., 2009]

further extended the categories enabling more restrictions that can be imposed in sharing knowledge. Singh introduced localized variables such as language translation, navigation structure and so on to identify five categories of websites: standardized, proactive, global, localized and highly localized website. The adoption of country-specific websites presented in [Daryanto et al., 2013] is also influenced from the presence of web categories. Such web categories presumably either depicts standardization in knowledge sharing with same content offered in entire country-specific websites or localization with localized content offered for target country and not shared between the country-specific websites.

With respect to the diversity of languages offered in the websites, [Esselink, 2000] viewed processes in globalization referring them as: internationalization for generalizing product and services in handling multiple languages and localization for targeting locale (country/regional and language). Such processes also emphasize the restriction in knowledge sharing for global communities in several languages or to regional to local communities only in specific languages. The interaction of internationalization and localization from integrating both global and local cultural characteristics in [Maynard and Tian, 2004] also introduces glocalization for sharing to both global and local communities. Next, the motivation that drives online community to collaborate for knowledge sharing is summarized.

Though this chapter broadly categorized past researches in knowledge sharing among communities as leveraging knowledge equally and customized for specific communities; overcoming language barrier is also interesting to researchers. Over the next sections, the limitation and feature of translation practices are discussed for their prospects to achieve knowledge sharing in the collaborative environment.

## 2.3 Collaboration Prospect in Knowledge Sharing

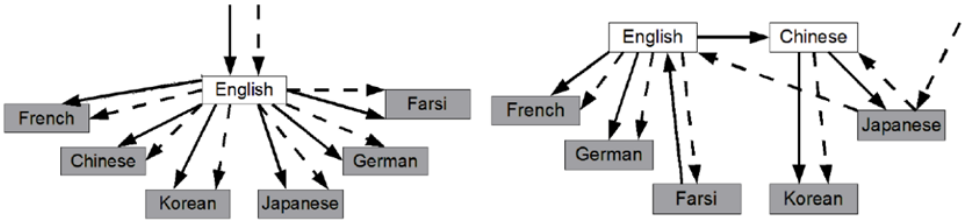
The philosophy and sociology of sharing with collaborative web authoring has changed from strict editorial control to democratic peer review with the advent of web. Though conventional translation processes are still widely practiced in organization, these do not work well with the realities in the world of internet. Emergence of online communities of volunteer ranging from novice to experts with aptitude of language skills has lifted off editorial control as they collectively author and translate important documents online in several languages.

### 2.3.1 Discrete Goals of Online Translator Communities

Online translator communities are broadly categorized as a) mission oriented translation communities which are strongly connected group of volunteers involved in translating clearly defined set of document like technical documentation of projects such as Linux, Mozilla documentation or b) subject oriented translator communities who have no prior orientation but share similar opinions about events [Bey et al., 2006]. Such communities of translator have separate goals with respect to sharing where mission oriented translation communities are committed toward enforcing equality in shared knowledge with the multilingual documents, the subject oriented translator are motivated towards sharing specific knowledge that would require specialist in a field.

### 2.3.2 Limitations of Conventional Translation

The conventional translation processes has specific limitation with their suitability for such online community of volunteer translator. The primary difference of conventional translation processes with the collaborative environment lies in the editorial control. The limitations are compiled below [Désilets et al., 2006, Huberdeau et al., 2008].




(a) Conventional Translation (b) Collaborative Translation

Figure 2.5: Content Flow in Translation Processes [Désilets et al., 2006].

**(a) Enforces Strict Editorial Control.** The central control is practically non-existent in online community of volunteer translators. The authoring and translation of content in collaboration is characterized from the irregular patterns in the flow of content occurring across the languages. In conventional translation, the content flow is unidirectional from the master language usually English. As depicted in Fig. 2.5(a) with page creation (full arrow) and subsequent edits (dotted arrow) are first done in master language and propagated to other languages Fig. 2.5(b). In case of collaborative environment page creation and subsequent edits can occur in any language. With no strict editorial control, it therefore becomes necessary to keep track of the language where the source text originates from such irregular patterns.

**(b) Difficulty Managing Volatile Content.** Sequential translation is particularly not appropriate for online communities as the contribution from the volunteer translator are ever changing the content without guaranteeing the stable final version in a specific language. Though flexible workflow is supported regarding changes made possible even after translation with incremental just in translation; the assumption of single master language usually English is not reasonable in an online context where volunteers can contribute directly in their native languages.

**(c) Inadequate Support for Content Reuse.** Parallel authoring results

in parallel communities working from scratch generating content on the same topic. In an online context, the shortage of domain experts in a target language can lead to communities without having access to the knowledge that resides in another language. Parallel authoring in Wikipedia also technically do no promote content reuse among languages which heavily depends upon the volunteers to leverage and translate content from article in one language to another.

Due to such limitations, the tool and workflow supporting conventional translation processes are not suitable for online communities motivated for knowledge sharing among communities from collaboration either for leveraging knowledge or customizing knowledge.

### 2.3.3 Difficulties in Integrating Translation

Though online communities use machine translation to overcome language barrier it is essential to have an understanding of how translation affects sharing. Past researches highlighted the problems of common grounding in multilingual collaboration mostly in the context of conversational communication between speaker (writer) and listener (reader) in their native languages. Knowledge sharing among multilingual communities is also impacted in the absence such mutual understanding even though the communities are able to translate contents in their native language. Therefore the contribution of research communities in MT mediated collaboration is also useful for multi-language knowledge sharing in the design of the tools supporting collaboration for bridging knowledge gap. The problems from integrating machine translation in collaboration for knowledge sharing are compiled below.

- Typographical errors are a big source of translation errors that hinder mutual understanding [Climent et al., 2003]. In knowledge sharing, the typographical errors in one language would not be successful in conveying the consistent knowledge in another language.

- Natural refereeing behavior is not supported with translation mediated collaboration. This was depicted from the experiments among triad and pairs by [Yamashita and Ishida, 2006] where the collaborating parties refrained from changing their written texts over repetitive task such as shortening referring expressions (lexical entrainment) due to asymmetries in machine translation. In knowledge sharing the paraphrasing of text in language is one of such several examples that should be taken as surficial changes and not as major changes that requires translation to other languages in avoiding the consequences from asymmetries.

- Cascading several translation services are problematic in communication. The drifting in word meaning from inconsistency, asymmetry and intransitivity during translation among languages are highlighted in [Tanaka et al., 2011]. Inconsistencies from the translation of same word varying among different sentences; Asymmetries from back translated word different from source word and Intransitives from drifting of word sense as translation progress along several languages are three specific cases that are highlighted by [Tanaka et al., 2011] when the context of communication is not taken into consideration. These are also severe problems in knowledge sharing when using machine translations and this would result in discrepancies of shared knowledge among multilingual communities.

The awareness of such existing problems when integrating machine translation for knowledge sharing is useful in the design of the collaborative tools that supports authoring and translation in multi-language knowledge sharing.

### 2.3.4 Essential Features in Collaborative Tool

To foster the online collaboration for the communities of volunteer translators, the collaborative tool has to integrate following features to support democratic peer review editorial participation in knowledge sharing.

(i) *Abolition of Master Language.* The support for a single master language usually English as a source in generating multilingual content should be lifted as the volunteer authors may not be fluent enough to write quality content in that language.

(ii) *Avoidance of Controlled Language pairs.* This is closely related to the use of master language which limits the generation of content only in specific language that can be translated from the master language. The online collaboration should be able to support the translation in any pairs even to under-resourced languages.

(iii) *Avoidance of Edit Freeze.* The adaptation of continual changes to the content in the source text or in any other languages should be supported for the online communities.

(iv) *Adhere to End Users.* The online volunteer of translator communities is limited in their exposure to the tool, processes and linguistic training so the collaborative tools must be simple to use and cater to their needs.

(iv) *Support for Coordination.* The activities of the online volunteer communities are not coordinated due to the absence of central authority. The tools must provide coordination in the form of cues that signal what translation work needs to be done, which linguistic version is the latest and so on.

(v) *Switching User roles.* The role of a user as content author and content translator is not segregated in an online environment. Collaborative tools must support the transition of roles for online communities of volunteer.

The above listed features are worthy of considerations when designing and developing tools for supporting collaborative authoring and translation in multi-language knowledge sharing system. Multilingual service platforms such as Language Grid offers language resources [Ishida, 2011]

which can be deployed in knowledge sharing. Such systems are also deemed important in [OLeary, 2009a] with growing requirements in multilingual knowledge management.

## 2.4 Summary

With growing multilingual presence, the majority of world knowledge repositories are scattered in several languages from resource rich to resource poor languages. Such diversities offer unique opportunity in sharing knowledge among multilingual communities. This chapter expanded our understanding on the current state of art in knowledge sharing goals along the spectrum with one end that focuses on leveraging knowledge equally across the languages and other end that focuses on customization in knowledge sharing due to cultural, linguistic and business context of target community. For bridging knowledge gap, the techniques and tools exploiting multilingual corpuses in detecting missing, outdated or incorrect information among several languages, tracking changes in linguistic versions, highlighting knowledge discrepancies and content synchronization are studied. Toward knowledge sharing customized for specific communities, the strategies in achieving global consistency while promoting local flexibility in the websites are studied. To foster online collaboration, the limitations in translation practices along with the features and difficulties of collaborative authoring and translation in knowledge sharing are also studied.

Though several benefits are achieved with past researches for knowledge sharing, the growing collaboration among multilingual communities poses challenges that are partly inaccessible. The problems ranged from necessity of expert online communities to support for limited languages. The goal of sharing knowledge to global communities are also partly met with previous techniques as replicating them to resource poor language are problematic primarily from (a) dependence on content processing (b) necessity for massive linguistic corpuses in training systems. In previous techniques, the support for global consistency while local flexibility are not met in knowl-

edge sharing due to the focus only on content inconsistency ignoring the restrictions in the publication and description of content to specific locales. Collaborative tools in place are also limited in knowledge sharing due to dependence on specific language, inflexible workflow, incompatible translation practices and grounding difficulties in translation. Overcoming some of the challenges the design of multi-language knowledge sharing system has to cater to knowledge sharing goals of the communities from leveraging knowledge both equally and customized for specific communities along with encouraging participation of communities with limited language resources.

# Chapter 3

# Supporting Consistency in Leveraging Knowledge Equally

The challenges in leveraging knowledge equally among communities are elevated from the participation of communities with distinct language preferences. This chapter focuses on knowledge discrepancies caused from the cases such as omitted content, updated content not shared and content conflict that may seem trivial but have a profound effect when existent in several languages. Such inconsistences are undesirable among communities that prefer to share knowledge equally in diverse languages. Towards enabling content consistency in multi-language knowledge sharing this chapter focuses specifically on multilingual communities and proposes process-based technique to detect inconsistency in multilingual content. The proposed technique is based on synchronizing user editing actions and is an alternative to content-based technique. Since the proposed technique is not language specific it is able to support content consistency in variety of languages including the resource poor languages.

## 3.1 Background

With Wiki system gaining popularity as the platform for co-creating knowledge inconsistencies in multilingual content generated from collab-

oration among multilingual communities is an impediment for sharing consistent knowledge [Sousa et al., 2010, Wagner, 2004, OLeary, 2009b, Cormican and Dooley, 2007]. Inconsistencies in multilingual content shared among communities emerge from cases (a) content omitted in several languages while sharing (b) updated content in one language not propagated to remaining languages and (c) same content updated in several languages resulting in content conflicts. Such inconsistencies are undesirable in multilingual documents such as technical manual, software documentation, product catalogue and so on which are produced with an aim to circulate consistent information globally in several language editions. Implications of such inconsistent content shared among communities is the misconception due to knowledge bias with one language favored over another and most importantly widening the knowledge gap.

**Definition.** Content is said to be Inconsistent between languages $l_1$ and $l_2$ for situation when cases (a) missing information in $l_1$ or $l_2$ (b) updates not propagated from $l_1$ to $l_2$ or vice versa and (c) conflicting information between languages $l_1$ and $l_2$ occur in knowledge sharing.

Previous attempts on managing multilingual content have considered centralized representation such as MLHTML to represent multilingual correspondences which do not keep references of the source content where the information originates in a particular language [Tonella et al., 2002, Al Assimi and Boitet, 2001]. The source content that serves as originating source of information is a concern while propagating content updates from one language to another. Approaching language generation, conceptual model in propagating change consistently have also been considered in multilingual authoring tools such as DRAFTER, PILLS [Hartley and Paris, 1997, Bouayad-Agha et al., 2002]. The problem with such tools is the expertise needed in building and making necessary changes to knowledge models which clearly excludes novice users and is impractical.

Though language processing have opened the state of art machine

translation researches towards sharing content from one language to another, the constrains in conventional translation practices have put forth the need to revisit the need in the context of collaboration. Some of the constraints in translation practices for content reuse are explored in LizzyWiki and WikiBABEL which provides insight into the usefulness of tool support in detecting inconsistencies in community collaboration [Kumaran et al., 2008, Adar et al., 2009]. However inconsistencies in content between the languages are not dealt. The automation in content synchronization with NLP and machine translation often favor resource rich language that have adequate language resources such as bilingual dictionary, high quality machine translations [Bronner et al., 2012]. With the abundance of content in resource poor languages, it is also need to support such communities to share content consistently with resource poor language. The collaborative wiki-style translation in [Huberdeau et al., 2008] is also limited from highlighting specific inconsistent cases that is useful in applying content consistency accordingly. Grounding on the limitations of previous approaches, particularly inadequate support for resource poor languages, content consistency that caters to communities with several language backgrounds is appropriate for the design of multi-language knowledge sharing system.

## 3.2 Inconsistency in Multilingual Content

To investigate the causes of inconsistency in multilingual content, we use Wikipedia articles as it is available in multiple languages with content revised and managed by the collaborative effort of the communities. We inspect an event based article titled "2013 ICC World Cricket League Division Three" (referred as Article 1)* in English and Nepali language, a resource rich and resource poor language respectively. The article is managed by the communities contributing content or translating content in either English or Nepali language and is available as parallel content. We identify following

*http://en.wikipedia.org/wiki/2013_ICC_World_Cricket_League_Division_Three

cases causing inconsistency between the language versions of the article.

### 3.2.1 Content Omitted and Not Propagated

From the table of content in each of the language version of Article 1, it is evident that the multilingual content is not equally available between the languages. As in Fig. 3.1 comparing the content in English and Nepali language, it is found that not all content is translated and therefore information in one language is largely missing in another language. The article in Nepali language appears to have more information than the English article. Also as Article 1 is created with the collaborative effort of multilingual communities, the changes made to the content by the editor in one language needs to be propagated to another languages for the consistency of information. As shown in Fig. 3.2 inconsistency from the lack of propagation of content updates between languages is found as the latest information related to 'the score point of the player' that is available in the revised version of English


*Source*: http://en.wikipedia.org/wiki/2013_ICC_World_Cricket_League_Division_Three.

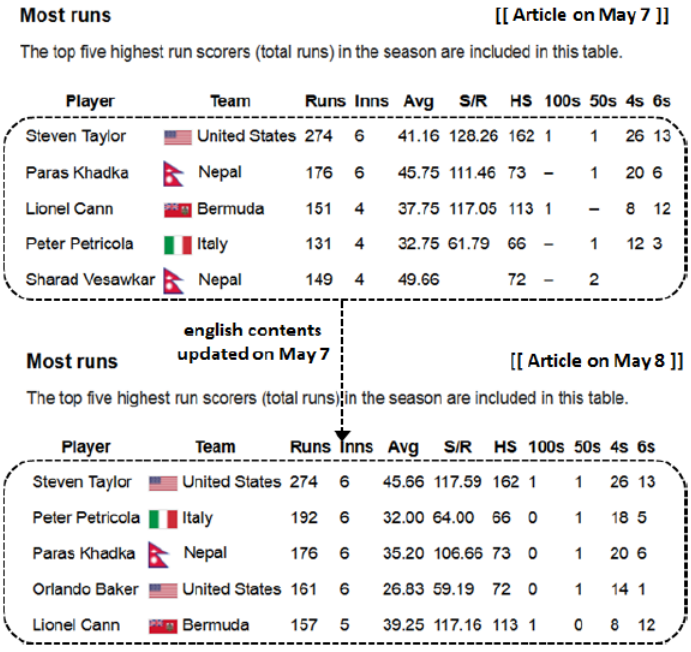
Figure 3.1: Missing Content.

article on May 8 is not available in the Nepali language.

### 3.2.2 Content Conflict

Such situation arises when parallel content is modified independently resulting in content that is no longer translation of each other. The conflicts in content related to 'the score point for maximum wicket taken by the player' appear between the English and Nepali language version of the article as shown in Fig. 3.3.

With the depiction of content inconsistency from the cases such as omitted content, updates not propagated and content conflicts in the collaboratively created multilingual Wikipedia article, the need to support content consistency in multilingual collaboration is emphasized from this example.



**Most runs** [[ Article on May 7 ]]

The top five highest run scorers (total runs) in the season are included in this table.

| Player | Team | Runs | Inns | Avg | S/R | HS | 100s | 50s | 4s | 6s |
|---|---|---|---|---|---|---|---|---|---|---|
| Steven Taylor | United States | 274 | 6 | 41.16 | 128.26 | 162 | 1 | 1 | 26 | 13 |
| Paras Khadka | Nepal | 176 | 6 | 45.75 | 111.46 | 73 | – | 1 | 20 | 6 |
| Lionel Cann | Bermuda | 151 | 4 | 37.75 | 117.05 | 113 | 1 | – | 8 | 12 |
| Peter Petricola | Italy | 131 | 4 | 32.75 | 61.79 | 66 | – | 1 | 12 | 3 |
| Sharad Vesawkar | Nepal | 149 | 4 | 49.66 | | 72 | – | 2 | | |

english contents updated on May 7

**Most runs** [[ Article on May 8 ]]

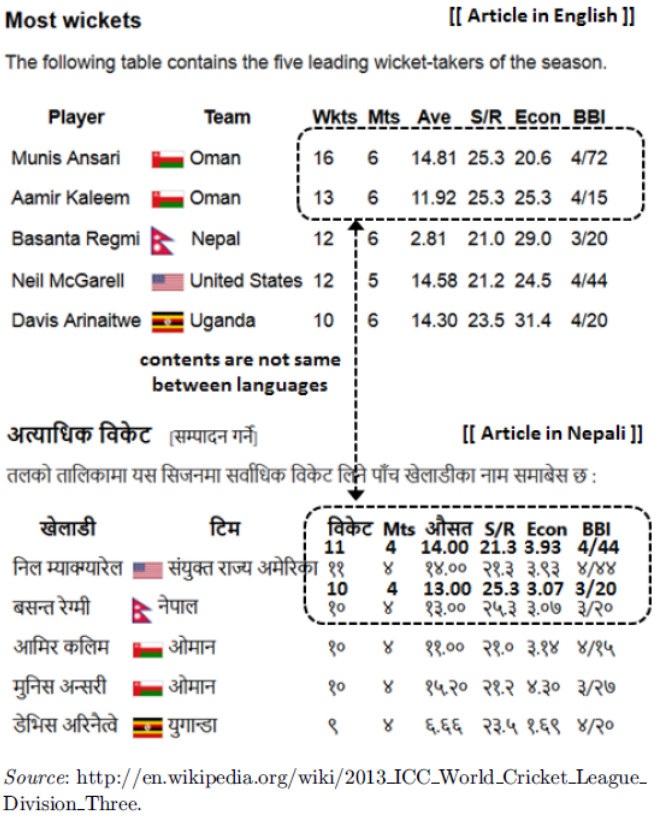
The top five highest run scorers (total runs) in the season are included in this table.

| Player | Team | Runs | Inns | Avg | S/R | HS | 100s | 50s | 4s | 6s |
|---|---|---|---|---|---|---|---|---|---|---|
| Steven Taylor | United States | 274 | 6 | 45.66 | 117.59 | 162 | 1 | 1 | 26 | 13 |
| Peter Petricola | Italy | 192 | 6 | 32.00 | 64.00 | 66 | 0 | 1 | 18 | 5 |
| Paras Khadka | Nepal | 176 | 6 | 35.20 | 106.66 | 73 | 0 | 1 | 20 | 6 |
| Orlando Baker | United States | 161 | 6 | 26.83 | 59.19 | 72 | 0 | 1 | 14 | 1 |
| Lionel Cann | Bermuda | 157 | 5 | 39.25 | 117.16 | 113 | 1 | 0 | 8 | 12 |

*Source*: http://en.wikipedia.org/wiki/2013_ICC_World_Cricket_League_Division_Three.

Figure 3.2: Updated Content in English Language from May 7 to May 8.

In addition to this, the collaboration between communities that are representative of resource rich language ‘English’ and resource poor language ‘Nepali’ in this example also illustrated that the lack of support for content consistency to resource poor language can lead to majority of local information in Nepali language that is of global interest not shared with English communities. Inconsistency detection mechanism are thus needed that enable multilingual editors including the resource deprive communities in sharing consistent content in multilingual documents.



**Most wickets** **[[ Article in English ]]**

The following table contains the five leading wicket-takers of the season.

| Player | Team | Wkts | Mts | Ave | S/R | Econ | BBI |
|---|---|---|---|---|---|---|---|
| Munis Ansari | Oman | 16 | 6 | 14.81 | 25.3 | 20.6 | 4/72 |
| Aamir Kaleem | Oman | 13 | 6 | 11.92 | 25.3 | 25.3 | 4/15 |
| Basanta Regmi | Nepal | 12 | 6 | 2.81 | 21.0 | 29.0 | 3/20 |
| Neil McGarell | United States | 12 | 5 | 14.58 | 21.2 | 24.5 | 4/44 |
| Davis Arinaitwe | Uganda | 10 | 6 | 14.30 | 23.5 | 31.4 | 4/20 |

**contents are not same between languages**

**अत्याधिक विकेट** [सम्पादन गर्ने] **[[ Article in Nepali ]]**

तलको तालिकामा यस सिजनमा सर्वाधिक विकेट लिने पाँच खेलाडीका नाम समाबेस छ :

| खेलाडी | टिम | विकेट | Mts | औसत | S/R | Econ | BBI |
|---|---|---|---|---|---|---|---|
| | | 11 | 4 | 14.00 | 21.3 | 3.93 | 4/44 |
| निल म्याक्यारेल | संयुक्त राज्य अमेरिका | ११ | ४ | १४.०० | २१.३ | ३.९३ | ४/४४ |
| | | 10 | 4 | 13.00 | 25.3 | 3.07 | 3/20 |
| बसन्त रेग्मी | नेपाल | १० | ४ | १३.०० | २५.३ | ३.०७ | ३/२० |
| आमिर कलिम | ओमान | १० | ४ | ११.०० | २१.० | ३.१४ | ४/१५ |
| मुनिस अन्सरी | ओमान | १० | ४ | १५.२० | २१.२ | ४.३० | ३/२७ |
| डेभिस अरिनैत्वे | युगान्डा | ९ | ४ | ६.६६ | २३.५ | १.६९ | ४/२० |

*Source*: http://en.wikipedia.org/wiki/2013_ICC_World_Cricket_League_Division_Three.

Figure 3.3: Content Conflict between English and Nepali languages.

## 3.3 Synchronizing Editing Activities to Detect Inconsistency

In this section, we consider inconsistency detection for the multilingual document that is to be shared among multilingual communities. We augment the parallel content in [Al Assimi and Boitet, 2001] with the information about the states of the parallel content and employ inconsistency detection rules to identify the cases of inconsistency arising from the state of the parallel content.

First we will illustrate the state transition model to define the states, actions and the state transition of the sentences during content modifications. We then define inconsistency detection rules to check for inconsistency in multilingual content. The notation used throughout is illustrated next.

**Notation.** A Monolingual Document $d^l$ is the document with the content available in language $l$. A sentence $e_i^l$ in the document $d^l$ is the $i^{th}$ sentence in language $l$. Content in monolingual document are organized into a collection of sentences $d^l = \{e_i^l \mid 1 \leq i \leq n\}$. If $L$ is the set of languages used in the multilingual document then parallel multilingual document is the collection of several monolingual documents $D^L = \{d^l \mid l \in L\}$.

With this granularity of the document, we will focus on the alignment i.e. consistency of multilingual content at the sentence level with parallel aligned sentences in multilingual documents. We refer to [Hopcroft, 1979] for basic concepts in automata theory.

### 3.3.1 State Transition Model

The state transition model can be described as a tuple: $M = (S, \Sigma, \delta, S_0)$ where

(1) $S = \{\text{Q,NQ,T}\}$ is the set of states of the sentences corresponding to Qualified, Non-Qualified and Translated states respectively. $S_0 = \{\text{Q,NQ}\}$ is the set of initial states.

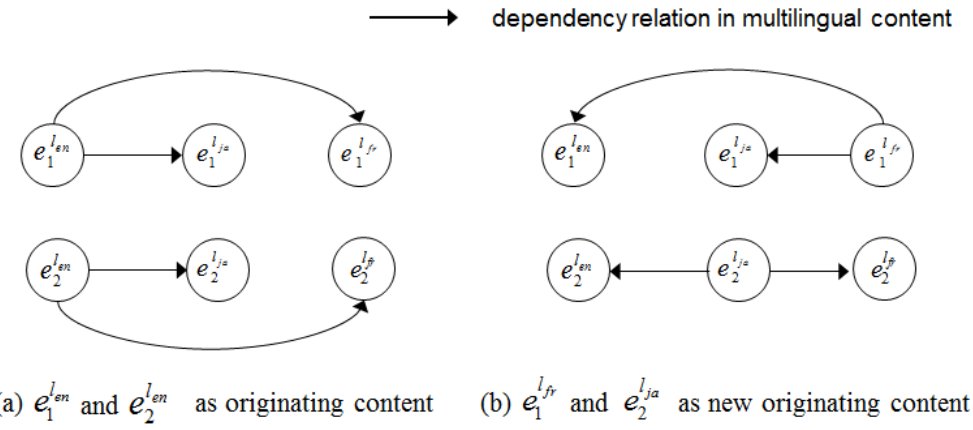



Figure 3.4: Dependency Relation between Parallel Aligned Sentences.

(2) $\Sigma$ = {modify,qualify,translate} is the set of actions performed on the sentence.

(3) $\delta$ is the state transition function given by $\delta : S \times \Sigma \rightarrow S$.

**(a) States.** To define the states for the sentence, we consider two aspects in the parallel aligned sentences: i) relation of content originating in one language with the content derived from translation in another language and ii) modification to the content that change overall context of the sentence (addition or deletion of facts or information) or preserve the meaning of the sentence. (e.g. paraphrasing the text) [Faigley and Witte, 1981, Jones, 2008, Sommers, 1980]

Table 3.1: State Transition Table

| **State S** | **Action $\Sigma$** | | |
|---|---|---|---|
| | **modify** | **qualify** | **translate** |
| **Qualified (Q)** | NQ | - | T |
| **Non-Qualified (NQ)** | NQ | Q | - |
| **Translated (T)** | NQ | - | - |

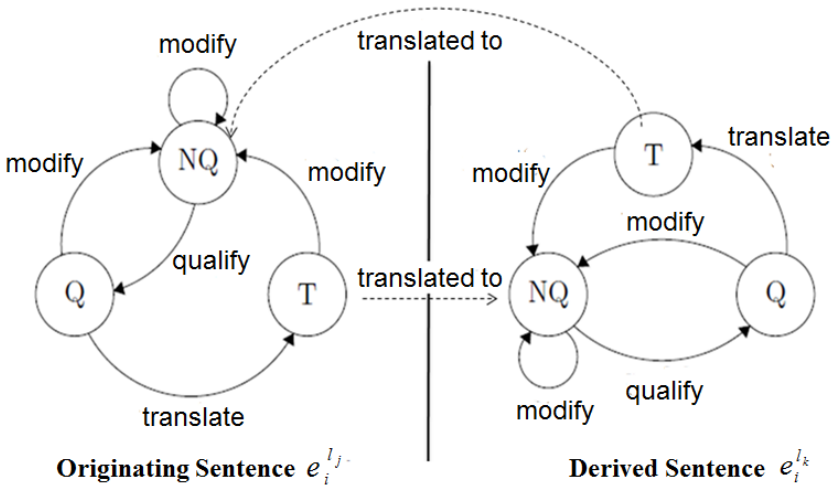



Figure 3.5: State Transition Diagram of Parallel Aligned Sentences $e_i^{l_j}, e_i^{l_k}$

Fig. 3.4 depicts dependency relation in languages showing the source and its translation among parallel aligned sentences in multilingual document. As shown in Fig. 3.4(a) the sentences $(e_1^{l_{en}}, e_2^{l_{en}})$ in the English document is the originating content which is translated to produce aligned sentences $(e_1^{l_{ja}}, e_2^{l_{ja}})$ in the Japanese document and $(e_1^{l_{fr}}, e_2^{l_{fr}})$ in the French document respectively. Under this illustration, the sentences $(e_1^{l_{en}}, e_2^{l_{en}})$ hold the qualified content that is used for sharing in other language. As the content is modified in either of the documents, the sentences $e_1^{l_{fr}}$ in the French document and $e_2^{l_{ja}}$ in the Japanese document is modified with updated information in Fig. 3.4(b) the originating source of information changes. Such modification can also be enriched with the information about whether the updated content qualifies or do not qualify for translation in another language. In other words, if content in knowledge resources is updated with additional facts or information which is considered useful to share among communities. The states of sentences in multilingual document are defined as follows:

**(i) Qualified:** A sentence $e_i^l$ in the multilingual document is said to be in

Qualified state (Q) if the contents of the sentences are eligible for translation in another language. By eligible, we mean the originating content or updated facts that qualify for translation. For example the information about 'the score points of the player' in the article used in the previous section is additional information and qualified content for translation.

**(ii) Non-Qualified:** A sentence $e_i^l$ in the multilingual document is said to be in Non-Qualified state (NQ) if the content is not eligible for translation in another language. For example the sentence modified by paraphrasing or improving grammar that do not change the overall meaning is not necessarily retranslated. A sentence that holds the derived contents also corresponds to Non-Qualified state.

**(iii) Translated:**A sentence $e_i^l$ in the multilingual document is said to be in Translated state (T) if the content of the sentence is translated into another language.

We use Qualified, Non-Qualified and Translated states to model the parallel aligned sentences so that during modification, qualified content is used as the originating source for translation in another language.

**(b) Transition Function.** The state transition of the sentences when actions are performed for modification is presented in Table 3.1. The transition of the states of the sentences is described with state transition diagram of a parallel aligned sentence $e_i^{l_j}$ as originating sentence in language $l_j$ and $e_i^{l_k}$ as derived in language $l_k$ in Fig. 3.5. The translate action results the qualified content $e_i^{l_j}$ to translate to $e_i^{l_k}$ . The documents $d^{l_j}$ and $d^{l_k}$ are comprised of several such parallel aligned sentences each with the states.

### 3.3.2 Inconsistency Detection Rules

We use the states of the multilingual content for generating inconsistency detection rules to be used in the multilingual documents. Inconsistencies in multilingual content to be detected are from the cases a) missing information

Table 3.2: Category of State Combinations

| Category | State of aligned sentence pair ($e_i^{l_j}$, $e_i^{l_k}$) |
|---|---|
| 1. Presence of both Qualified States | (Q,Q) |
| 2. Presence of single Qualified State | (Q,NQ) (Q,T) (NQ,Q) (T,Q) |
| 3. Presence of both Non-Qualified States | (NQ,NQ) |
| 4. Presence of all Translated State | (T,T) |
| 5. Presence of Translation and Non-Qualified State | (T,NQ) (NQ,T) |

or part of document not translated b) content modified in one language not propagated or translated to other languages c) content modified in multiple language independently such that content is no longer translations of each other. We want to detect such inconsistencies in the multilingual content for producing consistent multilingual documents. The content in multilingual knowledge resources is consistent when the content is translated from the originating language.

**(a) Design.** Inorder to design the specification rules, we will consider a pair of monolingual documents, $d^{l_j}$ and $d^{l_k}$ from parallel multilingual documents $D^L$. The documents $d^{l_j}$ and $d^{l_k}$ are in language $l_j$ and $l_k$ with set of sentences $d^{l_j} = \{e_i^{l_j} \mid 1 \leq i \leq n\}$ and $d^{l_k} = \{e_i^{l_k} \mid 1 \leq i \leq n\}$. Inconsistencies detected in parallel contents of any document pair ($|L| = 2$) in parallel multilingual document $D^L$ leads to inconsistencies in ($|L| \geq 2$) documents. Therefore, the rules presented here are naturally extended in the case of parallel multilingual documents.

**(b) Formulation.** Next, we will use the editing activities in Table 3.3

Table 3.3: Editing Activity in Multilingual Document

| **Editing Activity** | | **Parallel Multilingual Content** | | **Remark** |
|---|---|---|---|---|
| | | **State of $e_i^{l_j}$** | **State of $e_i^{l_k}$** | |
| *(i)* | *isCreated* ($e_i^{l_j}$) | **Q** | - | **Inconsistent:** A new content is created but not translated |
| *(ii)* | *isTranslated* ($e_i^{l_j}$) | Q → **T** | **NQ** | Content in $l_j$ is translated to $l_k$ |
| *(iii)* | *isModified* ($e_i^{l_j}$) | T → **NQ** | **NQ** | Content in $l_j$ is modified |
| *(iv)* | *isQualified* ($e_i^{l_j}$) | NQ → **Q** | **NQ** | **Inconsistent:** Qualified content in $l_j$ is not translated to $l_k$ |
| *(v)* | *isTranslated* ($e_i^{l_j}$) | Q → **T** | NQ → **NQ** | Qualified content in $l_j$ translated |
| *(vi)* | *isModified* ($e_i^{l_j}$) | **T** | NQ → **NQ** | Content in $l_k$ is modified but not qualified |
| *(vii)* | *isQualified* ($e_i^{l_j}$) | **T** | NQ → **Q** | **Inconsistent :**Qualified content in $l_k$ is not translated to $l_j$ |
| *(viii)* | *isTranslated* ($e_i^{l_j}$) | T → **NQ** | Q → **T** | Qualified content in $l_k$ is translated |
| *(ix)* | *isModified* ($e_i^{l_j}$) *and isModified* ($e_i^{l_j}$) | NQ → **NQ** | T → **NQ** | Contents in both $l_j$ and $l_k$ is modified |
| *(x)* | *isQualified* ($e_i^{l_j}$) *and isQualified* ($e_i^{l_j}$) | NQ → **Q** | NQ → **Q** | **Inconsistent:** Both modified contents $l_j$ and $l_k$ are qualified. |

to generate the possible state transitions to reach the combination of states categorized in Table 3.2. The current state of the sentence is highlighted in

the Table 3.3. As shown in Table 3.3, the first editing activity (i) refers to the creation of new content $e_i^{l_j}$ which causes inconsistency as contents are not translated resulting in missing contents in language $l_k$.

**Rule 1:** Missing information or part of document not translated.

$$\forall e_i^{l_j}, e_i^{l_k}: isCreated(e_i^{l_j}) \Rightarrow isInconsistent(e_i^{l_j}, e_i^{l_k})$$

The creation of newly created content $e_i^{l_j}$ causes inconsistency when the contents are not translated resulting in missing contents in language $l_k$.

The editing activity in Table 3.3 from (iii) to (viii) corresponds to the contents modified in single language either in $e_i^{l_j}$ or $e_i^{l_k}$ . Inconsistencies occurring at (iv) & (vii) are the combination of states (category 2 in Table 3.2) in the language pair that results in qualified content not translated.

**Rule 2:** Changes not propagated due to contents modified in one language not translated in another language.

$$\forall e_i^{l_j}, e_i^{l_k} : isStateOf(e_i^{l_j}, \mathrm{Q}) \wedge (isStateOf(e_i^{l_k}, \mathrm{NQ}) \vee isStateOf(e_i^{l_k}, \mathrm{T})) \Rightarrow isInconsistent(e_i^{l_k}, e_i^{l_j})$$

The editing activity (ix),(x) corresponds to the contents modified in both sentence $e_i^{l_j}$ and $e_i^{l_k}$. Inconsistencies occurring at (x) are the combination of states (category 1 in Table 2) in the language pair that results in both qualified contents in language $l_j$ and $l_k$ which are not translations of each other.

**Rule 3:** Change not propagated due to contents modified in multiple languages independently such that content is not translations of each other.

$$\forall e_i^{l_j}, e_i^{l_k} : isStateOf(e_i^{l_j}, \mathrm{Q}) \wedge isStateOf(e_i^{l_k}, \mathrm{Q}) \Rightarrow isInconsistent(e_i^{l_j}, e_i^{l_k})$$

For category 3, the presence of both Non-Qualified states in (iii), (ix) in Table 3.3 is the case when the qualified content for translation is unknown. The combination of states for $e_i^{l_j}$ and $e_i^{l_k}$ in category 4, with both Translated state is not possible as the qualified contents after translation causes the derived content to be in Non-Qualified state.

The category 5, represents the combination of states for the consistency between the multilingual contents $e_i^{l_j}$ and $e_i^{l_k}$ . The presence of combination of Translated and Non-Qualified state pair in (ii),(v),(vi),(viii) in (Table 3.3) is the case of multilingual contents that are translations pairs conveying same contents. Note that in (vi) the modified contents has not been qualified for translation therefore it is not a context changing modifications. The qualified contents after translation always results in the combination of states as state of $e_i^{l_j}$ is Translated and state of $e_i^{l_k}$ is Non-Qualified.

**Example.** We present an example to model the states of the parallel contents and apply inconsistency detection rules to identify inconsistencies in multilingual documents. Table 3.4 represents the editing activity for English and Japanese document. The English sentence (*"Today is hot"*) created in editing activity (i) is the originating qualified content for translation, so it is in Qualified state Q. From Rule 1, as the originating content is not translated, this result in inconsistencies from the case of missing content in Japanese language.

In editing activity (ii) missing content is translated (*"EN: Today it's hot"*) to become available in Japanese language. With translation the derived content in Japanese sentence is set as Non-Qualified state NQ. The content in Japanese sentence is modified and qualified (*"EN: Today it is the hot day of the year"*) as eligible for translation in editing activity (iii). The state combination of the English and Japanese sentence is Translated T and Qualified State Q respectively. From Rule 2, inconsistency from content modified in one language not propagated in another language is detected from this combination. The modified content in Japanese sentence is translated to English language in editing activity (iv). With translation,

Table 3.4: Example of Inconsistency Detection

| | Parallel Multilingual Content | | | | |
|---|---|---|---|---|---|
| **Editing Activity** | **English** | **State** | **Japanese** | **State** | **Inconsistent** |
| (i) Create English sentence | Today is hot. | Q | | - | **Rule 1:** Missing content |
| (ii) Translated English sentence | Today is hot. | T | Kyou wa atsui.<br><br>*EN: Today it's hot.* | NQ | - |
| (iii) Updated contents in Japanese sentence is qualified. | Today is hot. | T | Kyou dewa, kotoshi no atsui hi desu.<br><br>*EN: Today it is the hot day of the year.* | Q | **Rule 2:** Modified contents not translated |
| (iv)Translating contents in Japanese sentence | Today it is the hot day of the year. | NQ | Kyou dewa, kotoshi no atsui hi desu.<br><br>*EN: Today it is the hot day of the year.* | T | - |
| (v) Updated contents in both sentences are qualified | It is usually hot in August. | Q | Sore wa 8 gatsu ni natsuyasumi desu.<br><br>*EN: It is a summer vacation in August* | Q | **Rule 3:** Contents not translations of each other |

the derived content in English sentence is set as Non-Qualified state NQ. In editing activity (v) both content in English and Japanese are modified and qualified ("It is usually hot is August") and (*"EN: It is a summer vacation in August"*) as context changing modifications eligible for translation. The state of both English and Japanese sentence is Qualified state Q and incon-

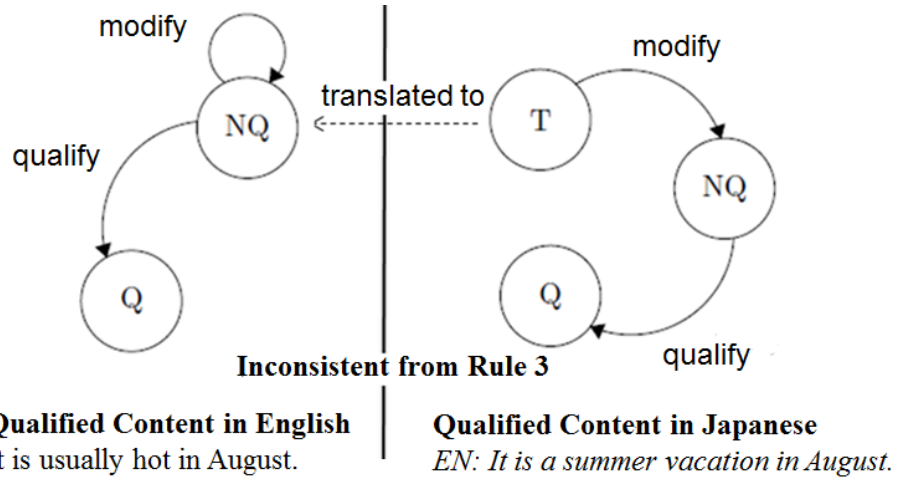



Figure 3.6: Inconsistency in Multilingual Content due to Multiple Source.

sistency is detected from Rule 3, as the content is not translation of each other. Fig.4. represents inconsistency detected for the editing activity (v) in which states of both English and Japanese sentence is Qualified state Q. From this example we illustrated inconsistencies occurred in editing activity (i), (iii) and (v) detected as the cases of missing content, changes not propagated and content no longer translation of each other.

## 3.4 Experimental Evaluation

To measure the effectiveness of proposed mechanism for detecting inconsistency in multilingual content, we collected edit histories of multilingual articles titled "2013 ICC World Cricket League Division Three" (will be referred as Article 1) and "2014 ICC World Twenty20" (will be referred as Article 2)[†] both are available in English and Nepali languages. The edit histories of multilingual article reflects on how the multilingual content has evolved over a time and is suitable for applying the proposed mechanism to detect inconsistent content shared between the languages. The step for this

[†]http://en.wikipedia.org/wiki/2014_ICC_World_Twenty20

experiment is as follows:

### 3.4.1 Data Collection

Wikipedia API is used to extract the revision history of the selected articles. Taking into account the duration of tournament and assuming more editing activities in multilingual content during the period, we collected revision histories from May 4 to May 9 for Article 1 and for Article 2 we collected revision histories from March 16 to April 6. The XML format of the data is cleaned to correct date and exported into Excel formats. We extracted 71 parallel content from Article 1 and 72 parallel content from Article 2 for this experiment. Articles in Nepali languages are created using English content as a reference either by creating the equivalent content in Nepali manually or by translating using Google Translate. Using Hindi as a pivot language to create content in Nepali language from English is also practiced. Most of the content is also found to be directly copied and appear as English text in Nepali articles which we ignored and focused only on parallel content between English and Nepali languages.

| Qualifying Modification(12) | Non-Qualifying Modification(18) |
|---|---|
| Addition of Category, Addition of Comment, Addition of different language version, Addition of external links, Addition of links, Addition of text, Creation of new article, Disambiguation, Merge, Recategorization, Removal of text, Reference | Addition or rephrasing of a short text, Alphabetization, Capitalization Cleanup, Copyedit, Correction, Formatting, Grammar, Headers Interwiki, Manual of Style, Move, Punctuation, Redirect, Revert to previous edit, Spelling, Tweaks, Typo |

**Source:**
https://en.wikipedia.org/wiki/Wikipedia:Edit_summary_legend/Quick_reference

Figure 3.7: Mapping Qualifying and Non-Qualifying Modifications.

### 3.4.2 Action Mapping

Referring to Taxonomy of Revisions in [Faigley and Witte, 1981] we mapped 30 modification actions as listed in Wikipedia Edit Summary Legend[‡]; out of which 12 actions change the context of sentences and are labeled as qualify action eligible for translation while remaining actions are minor changes Fig. 3.7. We map modify action to edits made in the article and translate action to represent parallel content generated either manually or using translation. The parallel multilingual content along with the actions tagged in each revision of the articles form the data samples in modeling the states of the multilingual content and applying inconsistency detection rules.

### 3.4.3 Evaluation

Comparing inconsistencies detected applying the proposed technique with inconsistencies identified from manual inspection of the selected articles, we compute the precision and recall as following.

$$\text{Precision} = \frac{\text{total no. of correctly detected inconsistencies}}{\text{total no. of inconsitencies detected with proposed method}}$$

$$\text{Recall} = \frac{\text{total no. of correctly detected inconsistencies}}{\text{total no. of inconsitencies detected manually}}$$

Table 3.5 quantitatively showed that for Article 1 overall precision of 94% and recall of 85% is achieved which suggested that the proposed technique is promising for detecting inconsistent content in collaboratively generated multilingual article. Further examination showed that the precision is roughly consistent in the initial period due to higher occurrences of missing content in Nepali article which is detected as inconsistent. As the missing content from English article is added to Nepali article in later period, it is also seen that there are fewer edits made to content already existing in the article. Fig. 3.8 show the missing content ("matches between Nepal-America,

[‡]http://en.wikipedia.org/wiki/Wikipedia:Edit_summary_legend

Table 3.5: Precision and Recall Measure (Article 1)

| Date | Inconsistencies Identified by Algorithm (Correctly Identified) | Inconsistencies Identified Manually | Precision | Recall |
|---|---|---|---|---|
| May 4 | 50 (50) | 56 | (50/50) = 1 | (50/56) = 0.892 |
| May 5 | 46 (46) | 50 | (46/46) = 1 | (46/50) = 0.92 |
| May 6 | 25 (24) | 29 | (24/25) = 0.96 | (24/29) = 0.827 |
| May 7 | 16 (14) | 18 | (14/16) = 0.875 | (14/18) = 0.77 |
| May 8 | 27 (24) | 29 | (24/27) = 0.888 | (24/29) = 0.827 |
| May 9 | 27 (24) | 28 | (24/27) = 0.888 | (24/28) = 0.857 |
| | | | **0.94** | **0.85** |

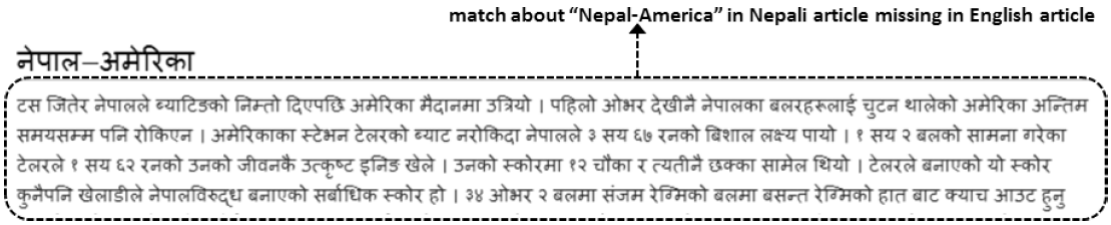



Figure 3.8: Detection of Missing Content (Article 1)

Nepal-Uganda, Nepal-Oman" in Nepali article, Revision Id: 337549) detected as missing content in English article. With the detection of large number of missing content, knowledge from one language can be leveraged into another language.

In case for Article 2 as shown in Table 3.6 the overall precision achieved is 82% and the recall is 87% which also suggested that most of inconsistencies in multilingual content are detected. Inconsistencies in Article 2 were mostly from cases in which content updated in articles either in English or Nepali language are not propagated to respective language. Such a case between content "Round 1 Group B" in English Article (Revision Id: 600298773) and Nepali article (Revision Id: 384275) is detected as updated content "match entries to Netherland vs Zimbabwe" is not propagated to Nepali article Fig. 3.9. With the detection of such inconsistencies the content updates made in one language can be notified or directly translated to

Table 3.6: Precision and Recall Measure (Article 2)

| Date | Inconsistencies Identified by Algorithm (Correctly Identified) | Inconsistencies Identified Manually | Precision | Recall |
|---|---|---|---|---|
| March 16 | 20 (20) | 31 | (20/20) = 1 | (20/31) = 0.645 |
| March 17 | 20 (25) | 23 | (20/25) = 0.8 | (20/23) = 0.8695 |
| March 18 | 18 (26) | 24 | (18/26) = 0.692 | (18/24) = 0.75 |
| March 19 | 23 (28) | 26 | (23/28) = 0.821 | (23/26) = 0.884 |
| March 20 | 21 (26) | 26 | (21/26) = 0.81 | (21/26)=0.81 |
| March 21 | 20 (27) | 26 | (20/27) = 0.74 | (20/26)=0.77 |
| March 22 | 30 (37) | 30 | (30/37) = 0.81 | (30/30)=1 |
| March 23 | 25 (34) | 28 | (25/34) = 0.74 | (25/28)=0.89 |
| March 24 | 22 (31) | 22 | (22/31) = 0.71 | (22/22)=1 |
| March 25 | 30 (37) | 38 | (30/37) = 0.81 | (30/38) = 0.789 |
| March 26 | 30 (36) | 38 | (30/36) = 0.83 | (30/38) = 0.789 |
| March 27 | 35 (40) | 38 | (35/40) = 0.9 | (35/38) = 0.921 |
| March 28 | 38 (43) | 40 | (38/ 43) = 0.88 | (38/40) = 0.95 |
| March 29 | 43 (47) | 45 | (43/47) = 0.91 | (43/45) = 0.96 |
| March 30 | 37 (51) | 39 | (37/51) = 0.73 | (37/39) = 0.948 |
| March 31 | 37 (50) | 38 | (37/50) = 0.74 | (37/38)= 0.97 |
| April 1 | 44 (49) | 47 | (44/49) = 0.9 | (44/47)= 0.977 |
| April 2 | 44 (48) | 45 | (44/48) = 0.917 | (44/45)= 0.860 |
| April 3 | 37 (46) | 43 | (37/46) = 0.80 | (37/43)= 0.85 |
| April 4 | 34 (46) | 40 | 34/46) = 0.74 | (34/40)= 0.857 |
| April 5 | 36 (46) | 42 | (36/46) = 0.78 | (36/42)= 0.857 |
| April 6 | 38 (46) | 40 | (38/46) = 0.83 | (38/40)= 0.95 |
| | | | **0.82** | **0.87** |

other language. However, the decrease in precision for Article 2 accounts from the absence of content processing involved in checking semantic relatedness in parallel content. It is found that the content “Round 1 Group A” in English article (Revision Id: 600298773) is wrongly detected as inconsistent content with Nepali content (Revision Id: 384275). Content conflict due to updating same content in both languages is also detected. As seen in Fig. 3.10 the content “entries for score points for Nepal” in English ar-


Figure 3.9: Detection of Updated Content Not Propagation (Article 2).

ticle (Revision Id: 600444676) and Nepali article (Revision Id: 384304) is detected as conflicting content and hence inconsistent. With the proposed technique for detecting inconsistency in the selected articles, we find an average precision of 88% and recall of 86% which is satisfactory in detecting inconsistency given that only user editing action is used. Though the recall estimates that most of inconsistent content in the articles were detected, some content in Article 1 such as ("Nepal 1st in 2012 ICC World Cricket League Division Four") in English language and (*"EN: Nepal 1st in 2011 ICC World Cricket League Division Four"*) in Nepali language which differ in only dates are not detected. The reason for this is not including semantics while conforming content consistencies.

The experimental results are convincing to illustrate the suitability of the proposed technique to detect inconsistency in multilingual content from cases (a) missing information or part of document not translated (b) changes made to contents not propagated or translated to other language and (c) contents that are no longer translations of each other. The experiment further suggested the applicability of proposed technique in a collaborative context such as Wiki in detecting inconsistency among multilingual communities.

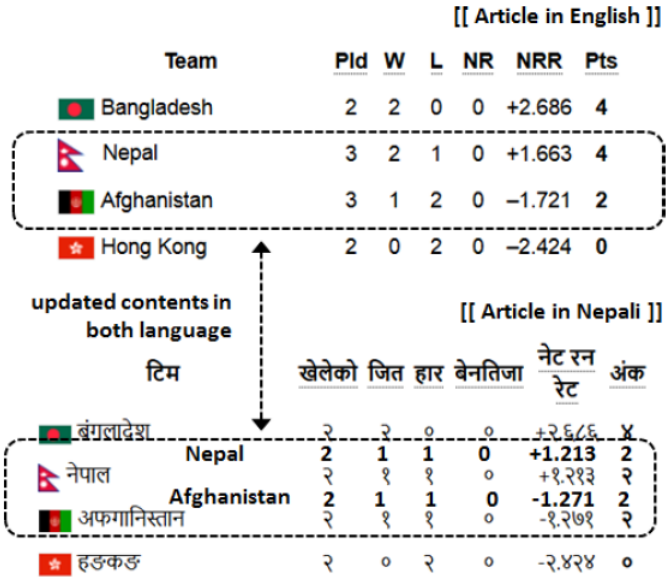



Figure 3.10: Detection of Content Conflict Between Languages.

## 3.5 Summary

The prime focus in this chapter has been to leverage knowledge equally among communities and in doing so promote content consistency even for communities with limited language resources. In this chapter, inconsistency in knowledge resources is depicted with multilingual content collaboratively created as multilingual Wikipedia articles. In an attempt toward supporting content consistency in knowledge sharing, this chapter contributes from the proposal of process-based technique to detect the presence of new content, updated facts or information and content conflict between languages. In addition to this, the proposed technique also does not require content processing making it eligible to support variety of languages; mostly of interest are the resource poor languages.

The proposed process-based technique is based on the concept of synchronizing user editing activities to detect the presence of inconsistency in shared content. To realize this concept, a state transition model is proposed

which is used to model multilingual content with states, action performed on them and the set of translation functions. Inconsistency detection rules for several cases such as content omitted, content conflict and so on are then designed which when applied to states in multilingual content detect inconsistency. From applying the proposed technique to the test set of revision histories in multilingual Wikipedia articles, we achieved satisfactory results with an average precision of 88% and a recall of 86% in detecting inconsistency. With the solution for detecting inconsistent content shared among communities even for the communities with limited resources, this chapter made an important contribution in the design of multi-language knowledge sharing system catered to leveraging knowledge equally.

# Chapter 4

# Determining Preferences in Sharing with Content Categories

The preference in sharing among communities is an important consideration when it comes to the design of multi-language knowledge sharing system to support customization in knowledge sharing. However due to the fact that exact correspondence in shared content is not mandatory for content consistency in such cases, it becomes essential to discover the factors that give rise to preferences in sharing, in other words the need to restrict content consistency in specific languages or for specific communities. Given that several content categories is published and shared among communities via websites targeted for specific locality such as country-specific websites in global brands; this chapter undertakes analysis based on propagation to examine the influence of specific content categories on preferences in sharing among communities. In doing so, the propagation in website graph interconnecting country-specific websites and propagation in website pair are extensively studied in the chapter. Traits such as coupling and scales in sharing are revealed that vary for specific content categories indicating the reason for preferences in sharing and an interesting finding to associate content consistency constraint with specific content categories.

## 4.1 Background

Global web presence is undoubtedly a strategic response of multinationals willing to promote their business internationally across the region. In doing so, country-specific websites that offer content and design targeted to specific communities is a trend seen among the global brands. Despite the advantage of offering content in multiple languages; the managerial challenge is raised from the difficulty in propagating content updates between the websites which results in omitted content and conflicting content shared among the communities. In addition to inconsistent content, propagating content updates also fails to notice the restriction in the publication and description of content giving rise to globally and locally inconsistent content shared among communities.

**Definition.** Cross-Site Content is said to be Inconsistent between websites $w_1$ and $w_2$ if $Semantic(Content, w_1) \neq Semantic(Content, w_2)$ due to (a) missing information in one of websites either $w_1$ or $w_2$ (b) updates not propagated from website $w_1$ to $w_2$ or vice versa and (c) conflicting information published between websites $w_1$ and $w_2$.

In the realities of world when countries and their official languages are associated, the restriction occurring at global and local level becomes obvious. The presence of multiple official languages within a country supports sharing limited to its official languages whenever something of local significance is shared. However the ground truth on such restriction in knowledge sharing via country specific website is not known. Given that restriction in propagation is unknown, issues such as delivery of knowledge confined within specific country or several countries; restrictions in the publication and description of content in specific languages becomes predominant while propagating content updates for knowledge sharing via country-specific websites. Further difficulty in propagation is also raised as the restriction that is suited for specific categories of content is not known.

On this ground it is worth determining preference in sharing among

communities to address content inconsistencies. In the area of ubiquitous computing, preference in sharing has received much attention in the design of privacy policy for sharing personal or contextual information [Olson et al., 2005, Wiese et al., 2011]. The influence of social graph and interpersonal relation are explored to determine preferences as willingness to share information with family, colleagues and so on. Basing on past researches in this study,

**Definition.** Preferences in Knowledge Sharing among communities is defined as willingness in sharing non-personal content restricted to specific community and their languages primarily with an aim to design content consistency policy.

Country-Specific websites that are managed within global brands is an example of communities that share non-personal content and is worth for understanding preferences in sharing. The presence of inconsistencies in content shared among communities in their country-specific websites is also an indication of problem with knowledge sharing in the absence of information on preferences in sharing.

## 4.2 Inconsistency in Cross-Site Content

To investigate the presence of inconsistent content in knowledge sharing where customization is a norm we refer to content published in country-specific websites of global brands i.e. cross-site content. Fig. 4.1 is the screenshot of cross-site content "3M at a glance" offered in country-specific websites for United Kingdom (UK), Switzerland (CHE) and Canada (CA) managed by global brand '3M'. It can also be noted that 'English' is a common language between UK and Canada while Canada and Switzerland both share a common language (French) and also unshared language (Deutsch) respectively which means that inconsistences are bound to occur both in same and different language in knowledge sharing. Following cases of in-

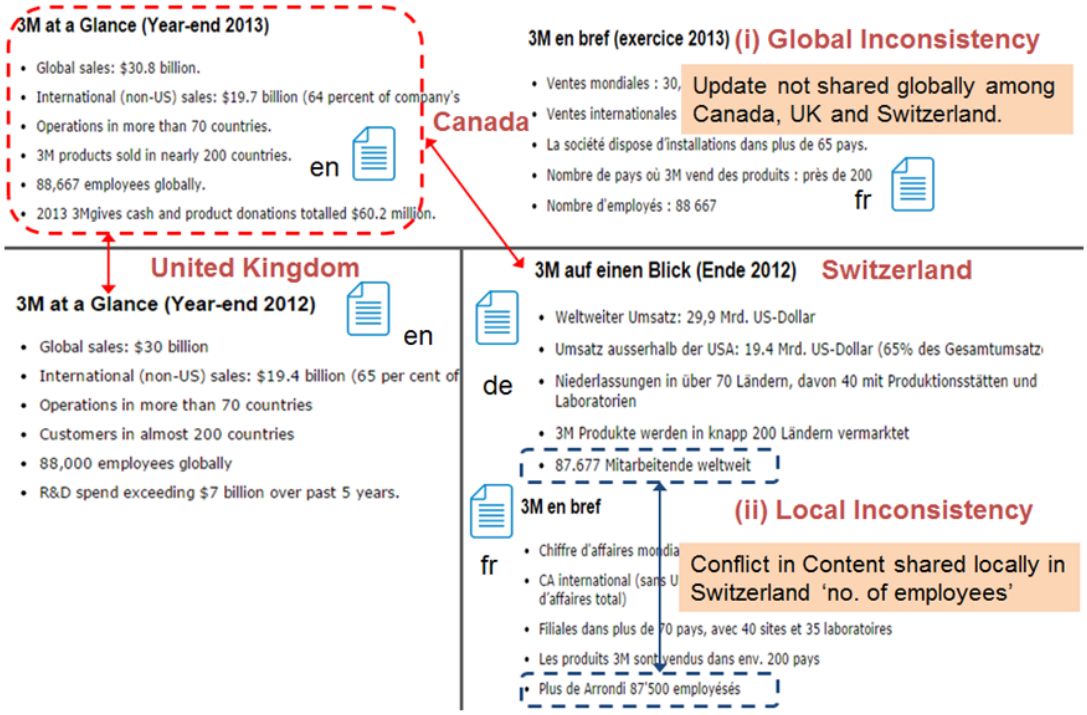



Figure 4.1: Global and Local Inconsistency in Cross-Site Content.

consistencies are compiled.

### 4.2.1 Global Inconsistency

As shown in Fig. 4.1 the content updates for the year 2013 in country- specific website for Canada is not shared with UK even though they share a common language English and with Switzerland neither in Deutsch (different language) and French (common language). The absence of propagating content updates to multiple countries gives rise to globally inconsistent content shared among communities. The case (i) in Fig. 4.1 illustrates inconsistent content shared due to the lack of propagating content updates in specific languages among the country-specific websites.

### 4.2.2 Local Inconsistency

Within a country-specific website for Switzerland, content conflict between the languages (Deutsch and French) for information on "statistics for the number of employees" is observed. Such a case (ii) in Fig. 4.1 highlights inconsistencies occurring locally within a community due to absence of propagating content updates locally to limited languages. In addition to this, the content in a common language (French) offered at country-specific websites for Canada and Switzerland also show possibilities for conflicts occurring among countries due to absence of propagating content updates.

Inconsistencies in cross-site content i.e. content shared among country-specific websites are indications on the difficulty in propagating content updates for knowledge sharing among communities. The depiction of globally and locally inconsistent content also shows that propagating updates is alone not sufficient. It is crucial to determine the restriction in propagation that is suited for specific content categories; an important factor in the design of multi-language knowledge sharing where the need is to customize knowledge sharing.

## 4.3 Hypothesis

Relating to cultural influence in sharing past researches have diverse perspectives on the impact of cultures in websites. The view on cultural homogeneity have stressed standardization in product and services across the globe [Hall, 1997, Main, 2001] suggesting that sharing of knowledge (content offered in webpage of websites) is standardized with same content published among websites. Whereas majority of researches have lenience towards Hofstede typology of cultural differences [Hofstede and Hofstede, 2001, Kale, 1991] among countries which suggest that sharing of knowledge (content offered in webpage of websites) is localized for specific country which mean different content is published among websites. Refuting the homogenizing effect of "international" culture, previous research in web globalization have also empha-

sized on varying globalization efforts placed for the web presence. Several categories of websites from standardized to highly-localized are suggested [Singh et al., 2009, Tixier, 2005] such as having global design template versus design adapted to local culture. Similar website categories in [Maynard and Tian, 2004] also depict the level of cultural adaptation catered to global appeal and local touch as glocal websites.

Bringing this notion into country-specific websites that are managed by global brands, the content shared among the websites presumably either depict standardization with same content offered for both domestic and international users or localization with content localized for international users and not shared among the country-specific websites. However, the presence of content in varying proportion among the websites indicates that sharing of content is restricted to specific websites. As several categories of content are published in websites, the restriction in their sharing is presumed to occur due to varied suitability among countries. Also previous researches have supported the differences in the content and design features among cultural groups which seem to be a possibility in shared content [Robbins and Stylianou, 2003, Huizingh, 2000, Okazaki and Alonso Rivas, 2002]. The clue for such differences in sharing content from specific content categories is obtained when their propagation among country-specific websites are examined. To shed light on the preferences in sharing for specific content categories, we set the following hypothesis.

**Hypothesis 1:** Propagations among County-Specific Websites are constrained by Content Categories: Corporate Information, Product Information, and Customer Support Information.

The goal from the stated hypotheses is to uncover the traits from examining propagation of content shared among country-specific websites, in other words the restriction in sharing specific content categories. In doing so, the contribution will be towards determining preferences in sharing specific content categories among the communities in the consideration for customization in knowledge sharing.

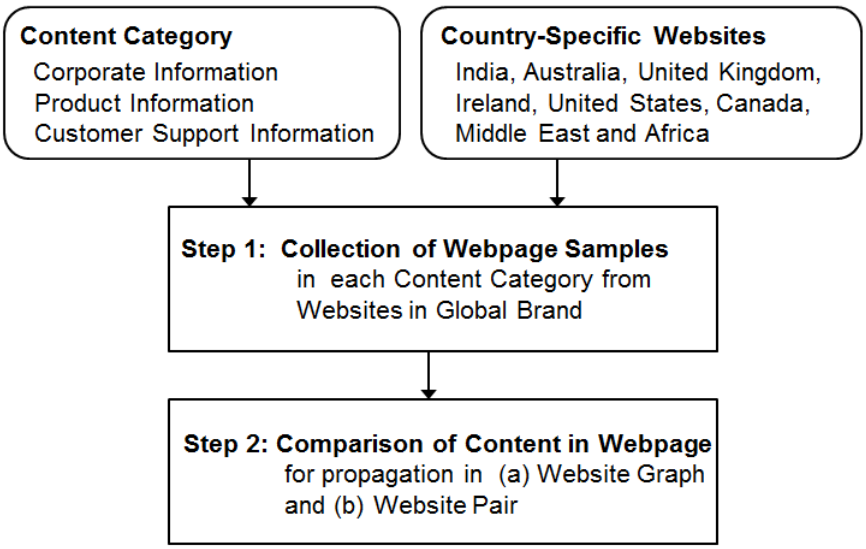



Figure 4.2: Outline on Examining Content Categories.

## 4.4 Outline on Methodology

An outline on examining sharing among countries from sampling webpages to comparing webpages among country-specific websites in global brands is shown in Fig. 4.2.

### 4.4.1 Websites and Content Categories

Websites from 10 global brands that are ranked highly in the web globalization report card [Yunker, 2014] is selected for this study. Each of the chosen global brand offer worldwide product and services with webpage published in more than 40 country-specific websites and more than 20 languages as in Table 4.1. For this study, webpage offered in shared language (English) among country-specific websites are selected. A sample of 8 country-specific websites in each global brand representing countries : India (IN) , Australia (AU), United Kingdom (UK), Ireland (IE), United States (US), Canada (CA), Middle East (ME) and South Africa (ZA) from several

Table 4.1: Statistics on Country-Specific Websites of Global Brand

| Global Brand | Industry | Country-Specific Websites | Sampled Websites* |
|---|---|---|---|
| Nivea | Skin and body care | 70 | 8 |
| 3M | Conglomerate | 100 | 8 |
| Starbucks | Coffee shop | 41 | 8 |
| Acer | Computer | 60 | 8 |
| Samsung | Conglomerate | 143 | 8 |
| KPMG | Professional Services | 143 | 8 |
| HP | Computer | 88 | 8 |
| Nestle | Food Processing | 75 | 8 |
| Avon | Personal Care | 74 | 8 |
| John Deeree | Heavy Equipment | 63 | 8 |

*Country-Specific Websites selected from various geographic regions

geographical regions: Asia Pacific, North America, Europe and Middle East - Africa are selected. A total of 80 country-specific websites are collected as the source for webpages to be used for comparison as in Table 4.2. From 8 country-specific websites, there are 28 possible websites pairs representing content sharing in pair of country for each global brand. For example, sharing among India and remaining countries occur in website pairs as: (IN, AU), (IN, UK), (IN, US) and so on.

Previous researches [Robbins and Stylianou, 2003][Huizingh, 2000] presented content features with categories that provide general company information, financial information, support and employment information to the customer and so on. Such features are associated with the design and cultural adaptation in the corporate websites. We also used the content categories in sampling webpages from each global brand which are (a) Corporate Information: in sampling webpages that provide background information of a company such as mission statements, history and its people

(b) Product Information: in sampling webpages on description, usage, and specification of product and (c) Customer Support information: in sampling webpages on ways to contact company or find answer to queries.

### 4.4.2 Comparison in Website Graph and Website Pair

We then manually analyzed webpages from each global brand and labelled them to specific content categories: "Corporate Information", "Product Information" and "Customer Support Information" respectively. From each global brand we collected 48 webpage samples making a total of 480 webpage samples Table 4.2.

From webpage samples, the content in webpages is qualitatively compared to determine whether propagation occurs or do not occur among the Websites. A paragraph of text in a webpage of a selected Website is used as a threshold to check for its presence among the remaining websites. Propagation is said to occur among the websites upon the presence of exactly same paragraph or comparable paragraph in their webpages. Comparable paragraph are paraphrased text that provide same information in the webpages of corresponding websites. Similarly, no propagation among websites is assigned when content is not same in the webpages of the corresponding website or webpage do not exist.

**Definition.** Propagation is said to occur between websites $w_1$ and $w_2$ managed in a global brand if webpages $p_a \in w_1$ and $p_b \in w_2$ have exactly same or comparable content. Since comparable content has to be checked between webpages manual effort is needed to examine their propagation among websites which means existing text-based method cannot be applied. Propagation among country-specific websites is examined in a website graph and in website pairs which are explained below.

Table 4.2: Statistics on Websites and Content Categories.

| Brand | Website from each Brand | | Content Category | Webpage from each brand | Total | |
|---|---|---|---|---|---|---|
| | | | | | Website | Webpage |
| | Individual | Pair | | | | |
| 10 | 8 | 28 | 3 | 48 | 10 * 8 = 80 | 10 * 48 = 480 |

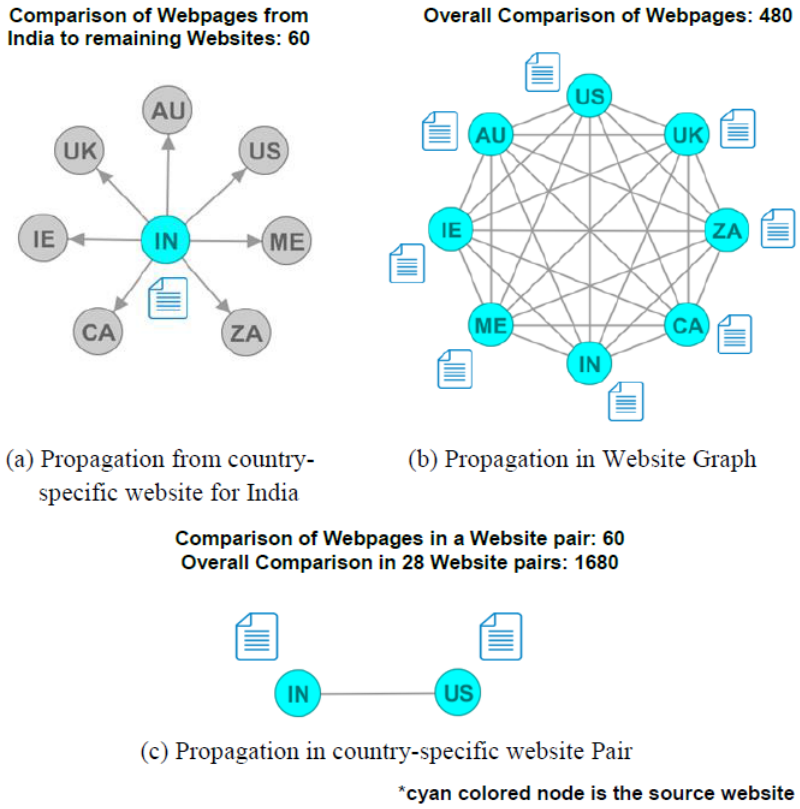
**(a) Website Graph**

Due to the absence of publicly accessible information on specific website where content originates in a webpage; each country-specific website is considered as a potential source for publishing content in a webpage and sharing with the remaining websites. Fig. 4.3(a) is an example depicting country-specific website for India chosen as a potential source for publishing content in a specific webpage. The presence of same or comparable content in corresponding webpages residing in the remaining country-specific websites such as Australia, United States, Canada and so on illustrates the propagation from India to remaining countries i.e. propagation occurs in (1: 7) country-specific websites.

We analyzed propagation in (1: 7) country-specific websites from 8 potential sources. The comparison of webpages performed from all potential sources websites represents a complete graph or website graph as shown in Fig. 4.3(b). A total of 480 webpages are qualitatively compared for their propagations in (1:7) country-specific websites as shown in Table 4.3. The

Table 4.3: Statistics on Comparison in Website Graph.

| Each Category | | All Category | |
|---|---|---|---|
| Each 1 : 7 | Complete | Each 1: 7 | Complete |
| 20 | 20 * 8 = 160 | 3 * 1 * 20 = 60 | 3 * 8 * 20 = 480 |




(a) Propagation from country-specific website for India

(b) Propagation in Website Graph

(c) Propagation in country-specific website Pair

Figure 4.3: Comparison in Website Graph and Website Pair.

Table 4.4: Statistics on Comparison in Website Pair.

| **Each Category** | | **All Category** | |
|---|---|---|---|
| Each Pair | All Pair | Each Pair | All Pair |
| 20 | 28 * 20 = 560 | 3 * 1 * 20 = 60 | 3 * 28 * 20 = 1680 |

purpose of studying propagations in (1:7) country-specific websites is to examine the presence of scales in sharing content for specific content categories.

**(b) Website Pair**

Propagations in (1:1) country-specific website pair is also examined by comparing the content in the webpages of the corresponding websites. Fig. 4.3(c) shows the propagation in website pair representing India and United States. As shown in Table 4.4 for each website pair there are 60 comparisons of webpages making a total of 1680 comparison for all 28 website pairs. The purpose of studying propagations in (1:1) country-specific website pair is to examine coupling between websites in sharing content for specific content categories.

## 4.5 Analysis on Propagation in Content Categories

From examining propagation interesting results for scales and coupling in sharing for specific content categories are compiled.

### 4.5.1 Propagation in Website Graph

Table 4.5 and 4.6 present the qualitative results of comparing webpages for propagations in (1:7) country-specific websites over a website graph. The suitability of content globally, regionally and locally is represented with three cases: (a) propagation to all country-specific websites (b) propagation to some country-specific website and (c) no propagation. Scales in sharing are determined from such cases for specific content categories. Differences in scales while sharing content for specific content categories are revealed from comparing their propagations.

As illustrated in Table 4.6 and Fig. 4.4 out of 160 comparisons of webpages in "Corporate Information", 50% of cases are identified in which propagation occurs among all country-specific websites while 32% of cases in which propagation occurs in some websites and 18% of cases in which no propagation occurs among the websites. With more than 80% cases in which propagation occurs from at least a single country-specific website, the

Table 4.5: Propagation of Content Categories in Website Graph

| Propagation in 1: 7 Website* | Corporate Information | | | Product Information | | | Customer Support Information | | |
|---|---|---|---|---|---|---|---|---|---|
| | All | Some | None | All | Some | None | All | Some | None |
| IN | 10 | 7 | 3 | 3 | 7 | 10 | 4 | 3 | 13 |
| AU | 10 | 7 | 3 | 3 | 6 | 11 | 4 | 3 | 13 |
| UK | 10 | 9 | 1 | 3 | 10 | 7 | 4 | 4 | 12 |
| IE | 10 | 8 | 2 | 3 | 9 | 8 | 4 | 4 | 12 |
| US | 10 | 3 | 7 | 3 | 4 | 13 | 4 | 1 | 15 |
| CA | 10 | 6 | 4 | 3 | 5 | 12 | 4 | 1 | 15 |
| ME | 10 | 5 | 5 | 3 | 8 | 9 | 4 | 3 | 13 |
| ZA | 10 | 7 | 3 | 3 | 8 | 9 | 4 | 4 | 12 |
| **Total** | **80** | **52** | **28** | **24** | **57** | **79** | **32** | **23** | **105** |

* combination of all 1:7 propagation equals a website graph

suitability of content in "Corporate Information" is not limited within a single country. The result strongly suggest for the suitability of content related to "Corporate Information" globally in all country-specific websites.

Comparing webpages in "Product Information" revealed suitability mostly either for some countries or limited to a specific country. Only 15% cases in which propagations occur in all websites are identified which strongly suggest that content in "Product Information" is not globally suit-

Table 4.6: Summary: Propagation Cases for Content Categories

| Content Category | Propagation in Country-Specific Websites | | | | | | Comparison of Webpage |
|---|---|---|---|---|---|---|---|
| | All | % | Some | % | None | % | |
| a. Corporate Information | 80 | 50 | 52 | 32 | 28 | 18 | 160 |
| b. Product Information | 24 | 15 | 57 | 36 | 79 | 49 | 160 |
| c. Customer Support Information | 32 | 20 | 23 | 14 | 105 | 66 | 160 |
| **Total** | **136** | **28** | **132** | **28** | **212** | **44** | **480** |

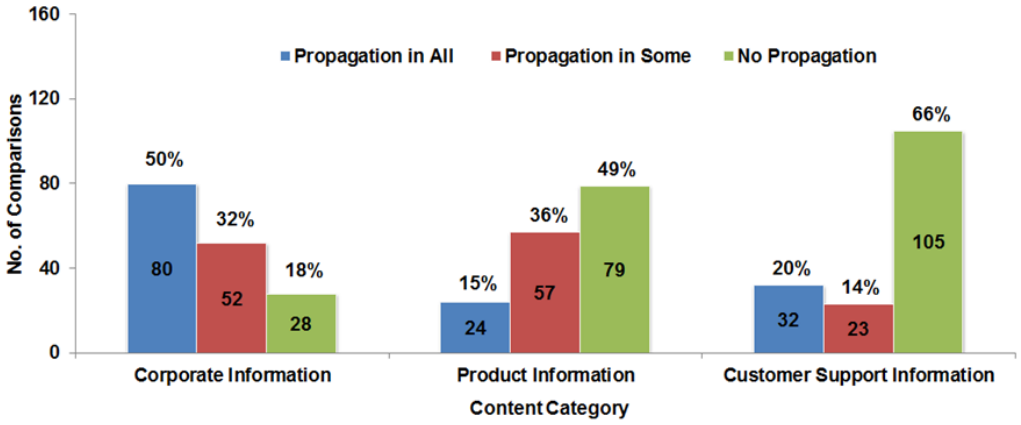



Figure 4.4: Comparison of Propagation in Website Graph.

able. However, 36% cases of propagation to some websites and 49% cases of no propagation are comparable to infer the suitability of "Product Information" both regionally and locally among countries. Contrary to this, comparing webpages in "Customer Support Information" strongly suggested the suitability of content locally within a country with 66% cases of no propagation among countries.

### 4.5.2 Propagation in Website Pair

Table 4.7, 4.8, 4.9 and 4.10 qualitative presents the result of comparing webpages for propagations in (1:1) country-specific website pair in content categories "Corporate Information", "Product Information" and "Customer Support Information" respectively. The coupling in sharing are represented from the occurrences of (a) propagation in website pair and (b) no propagation. The higher the occurrences of propagation for specific content categories indicates high coupling between websites while sharing content for specific categories. Comparing the webpages, the differences in coupling in country-specific website pair with respect to categories are revealed.

As illustrated in Table 4.10 and Fig. 4.5 out of 560 comparisons of webpages in "Corporate Information" among 28 country-specific website

Table 4.7: Propagation for Corporate Information in Website Pair.

| | IN | AU | UK | IE | US | CA | ME | ZA | Propagation | | | | Comparison |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Yes | % | No | % | of Webpage |
| IN | | **15** | **15** | **14** | **12** | **13** | **14** | **15** | 98 | 70 | 42 | 30 | 140 |
| AU | 15 | | **15** | **15** | **13** | **14** | **14** | **16** | 87 | 73 | 33 | 27 | 120 |
| UK | 15 | 15 | | **17** | **12** | **15** | **15** | **16** | 75 | 75 | 25 | 25 | 100 |
| IE | 14 | 15 | 17 | | **12** | **15** | **15** | **16** | 58 | 73 | 22 | 27 | 80 |
| US | 12 | 13 | 12 | 12 | | **11** | **11** | **13** | 35 | 58 | 25 | 42 | 60 |
| CA | 13 | 14 | 15 | 15 | 11 | | **14** | **15** | 29 | 73 | 11 | 27 | 40 |
| ME | 14 | 14 | 15 | 15 | 11 | 14 | | **15** | 15 | 75 | 5 | 25 | 20 |
| ZA | 15 | 16 | 16 | 16 | 13 | 15 | 15 | | - | - | - | - | - |
| | | | | | | | | | **397** | **71** | **163** | **29** | **560** |

Table 4.8: Propagation for Product Information in Website Pair.

| | IN | AU | UK | IE | US | CA | ME | ZA | Propagation | | | | Comparison |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Yes | % | No | % | of Webpage |
| IN | | **8** | **9** | **8** | **5** | **6** | **8** | **8** | 52 | 37 | 88 | 63 | 140 |
| AU | 8 | | **9** | **8** | **4** | **6** | **7** | **8** | 42 | 35 | 78 | 65 | 120 |
| UK | 9 | 9 | | **11** | **5** | **6** | **8** | **8** | 38 | 38 | 62 | 62 | 100 |
| IE | 8 | 8 | 11 | | **6** | **6** | **7** | **8** | 27 | 34 | 53 | 66 | 80 |
| US | 5 | 4 | 5 | 6 | | **5** | **5** | **5** | 15 | 25 | 45 | 75 | 60 |
| CA | 6 | 6 | 6 | 6 | 5 | | **8** | **8** | 16 | 40 | 24 | 60 | 40 |
| ME | 8 | 7 | 8 | 7 | 5 | 8 | | **10** | 10 | 50 | 10 | 50 | 20 |
| ZA | 8 | 8 | 8 | 8 | 5 | 8 | 10 | | - | - | - | - | - |
| | | | | | | | | | **200** | **36** | **360** | **64** | **560** |

pairs, 71% of cases with propagation occurring in website pairs are identified which suggest high coupling while sharing content related to corporate information.

Contrary to this, 75% of cases with no propagations in website pair are identified for Customer Support Information which suggest low coupling while sharing content related to supporting customer. Similarly for Product Information though the occurrences of no propagation are higher 64%,

Table 4.9: Propagation for Customer Support Information in Website Pair.

| | IN | AU | UK | IE | US | CA | ME | ZA | Propagation | | | | Comparison of Webpage |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Yes | % | No | % | |
| IN | | **5** | **5** | **5** | **4** | **5** | **5** | **5** | 34 | 24 | 106 | 76 | 140 |
| AU | 5 | | **6** | **5** | **4** | **4** | **5** | **6** | 30 | 25 | 90 | 75 | 120 |
| UK | 5 | 6 | | **7** | **4** | **5** | **6** | **7** | 29 | 29 | 71 | 71 | 100 |
| IE | 5 | 5 | 7 | | 4 | 5 | **5** | **6** | 20 | 25 | 60 | 75 | 80 |
| US | 4 | 4 | 4 | 4 | | **4** | **4** | **4** | 12 | 20 | 48 | 80 | 60 |
| CA | 5 | 4 | 5 | 5 | 4 | | **5** | **5** | 10 | 25 | 30 | 75 | 40 |
| ME | 5 | 5 | 6 | 5 | 4 | 5 | | **7** | 7 | 35 | 13 | 65 | 20 |
| ZA | 5 | 6 | 7 | 6 | 4 | 5 | 7 | | - | - | - | - | - |
| | | | | | | | | | **142** | **25** | **418** | **75** | **560** |

Table 4.10: Summary: Propagation for Content Categories.

| Content Category | Propagation in Website Pair | | No Propagation in Website Pair | | Comparison of Webpage |
|---|---|---|---|---|---|
| | *N* | % | *N* | % | |
| a. Corporate Information | 397 | 71 | 163 | 29 | 560 |
| b. Product Information | 200 | 36 | 360 | 64 | 560 |
| c. Customer Support Information | 142 | 25 | 418 | 75 | 560 |
| **Total** | **739** | **44** | **941** | **56** | **1680** |

the differences with occurrences of propagation are comparable (only 28% while in other categories the difference are >50 %). The coupling in a website pair while sharing content for "Product Information" tends to be neutral."Corporate Information" among 28 country-specific website pairs, 71% of cases with propagation occurring in website pairs are identified which suggest high coupling while sharing content related to corporate information. Contrary to this, 75% of cases with no propagations in website pair are identified for "Customer Support Information" which suggest low coupling while sharing content related to supporting customer. Similarly for "Product Information" though the occurrences of no propagation are higher

64%, the differences with occurrences of propagation are comparable (only 28 % while in other categories the difference are >50%). The coupling in a website pair while sharing content for "Product Information" tends to be neutral.

## 4.6 Preferences with Content Categories

Referring to Fig. 4.4 and Fig. 4.5 from examining propagations in (1:7) websites among all country-specific websites, some websites or within a website; we compile the scales in sharing for specific content categories.

### 4.6.1 Scales in Content Categories

- Propagation of "Corporate Information" at a global scale suggests on knowledge sharing related to corporate information occurring among global communities. In such cases, the contribution of knowledge is permitted to occur from the participation of local communities while dissemination of up-to-date knowledge is to occur globally for sharing knowledge consistently among global communities.

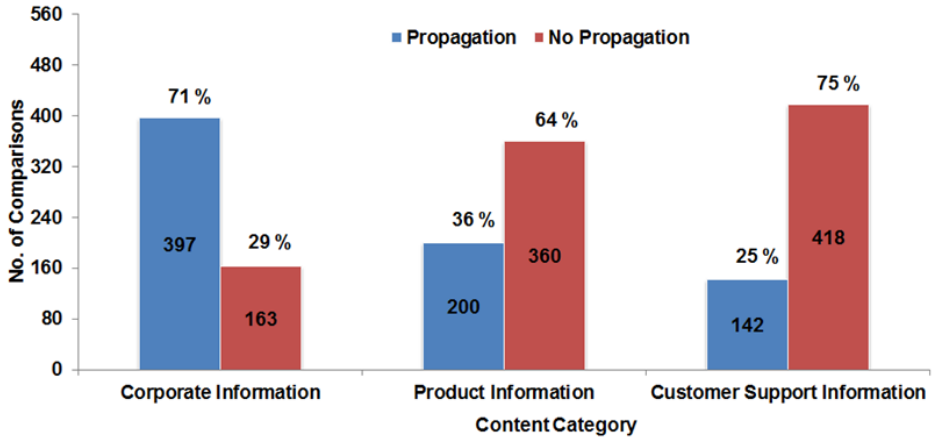



Figure 4.5: Comparison of Propagation in Website Pair.

- Propagation of "Product Information" both at regional and local scale suggest on dissemination of up-to-date knowledge either restricted among several countries within and across regions or limited to specific country when describing product specifications, usage and so on.
- Propagation of "Customer Support Information" at a local scale suggests on dissemination of knowledge restricted to local communities. As it is logical that suitability of content is limited to specific locale where it is produced, local scale also suggests for synchronization of content updates to occur within a country. For example, content synchronized in official languages (English and French) within a country for Canada.

Scales in sharing globally, regionally and locally are also useful to enable content consistencies in knowledge sharing. The restrictions in publication of content and their description in specific languages associated with scales also shows preferences that differ for specific content categories.

### 4.6.2 Coupling in Content Categories

Priorities for content consistency with respect to content categories while sharing are inferred from the coupling between websites.

- High coupling in website pair for sharing content related to "Corporate Information" suggests on contribution for knowledge occur frequently and consistency has to be strictly enforced while sharing such content.
- Low coupling in website pair for sharing content related to "Customer Support Information" suggests the contribution for knowledge occurs less frequently and the policy for consistency is not strictly enforced while sharing such content.
- Similarly, neutral coupling in website pair for sharing content related to Product Information suggests that the policy for consistency to be moderately enforced while sharing such content.

Coupling in sharing are useful in setting priorities for content consistency in the presence of several content categories. For example, content consistency for corporate related information have higher priorities in comparison to product related or customer support related information while sharing. From propagations in (1:7) websites and (1:1) website pairs, the results in Table 4.11 compiles the traits for specific categories: scales and coupling which are indication on preferences for sharing that vary for specific content categories and the need for content consistencies constraints.

## 4.7 Guidelines on Content Consistency

The findings on scales and coupling in sharing are useful in specifying guidelines for web manager as they share the content with communities. Following guidelines are compiled from this analysis.

- Global consistency is needed when sharing corporate related information, Local consistency is needed when sharing customer support related information, and Regional consistency is needed when sharing product related information.
- Corporate related information is more vulnerable to inconsistency as it is shared globally and hence need higher priority for consistency compared to other content categories.
- Product related information is more vulnerable to become inconsis-

Table 4.11: Summary: Preferences in Sharing for Content Categories.

| Content Category | Scales in Sharing | Coupling in Sharing |
|---|---|---|
| a. Corporate Information | Global | High |
| b. Product Information | Local and Regional | Neutral |
| c. Customer Support Information | Local | Low |

tent for specific region so region specific policy is needed. In doing so consistency is to be achieved only in languages that are offered within specific regions.

- Customer support information is more vulnerable to become inconsistent in local languages offered within a community so consistency is needed to restrict in limited languages used within a locality.

- High coupling occurs from frequent interaction for sharing specific content which increases the chances of inconsistency. Since corporate related information is shared more frequently and the websites interact more when sharing such information, the priority for its consistency is increased.

- Coupling between websites decreases as scales in sharing is restricted from global to local communities. This means local consistency becomes a priority as the websites become independent. Since coupling is found to decrease when sharing customer support related information, consistency in local languages is suited for such content.

The guidelines provide a general idea on the consistency needed at global, regional and local scales when sharing specific content categories and also on the restriction to specific languages that are used by the communities. Considering such guidelines web manger is better prepared to execute consistency policy and promote consistency in the content shared with their customer via country-specific websites.

## 4.8 Hypothesis Verification

As illustrated in Table 4.5, Table 4.10, Fig. 4.4 and Fig. 4.5 the differences in propagation both in terms of scale and number of occurrences among country-specific websites is identified for the content categories. Suitability for sharing content related to “Corporate Information” globally in all

countries, sharing content related to "Customer Support Information" locally within specific country and sharing content related to "Product Information" both regionally and locally suggest on scales in sharing that restrict the publication of content and their description in specific languages.

Similarly, the occurrences of propagations in websites are also found to vary with respect to content categories. High coupling between websites while sharing content for "Corporate Information" indicate that corporate information are shared more frequently and so consistency has to be strictly enforced while sharing such content.

From identifying traits such as scales and coupling in sharing that vary with content categories, we verified that propagation among county-specific websites is constrained for specific categories in knowledge sharing. Hence the preferences for sharing specific content categories vary as we find that the communities prefer to share corporate related information globally while customers support related information locally and so on.

## 4.9 Summary

The focus in the chapter is to support content consistency in knowledge sharing where customization is essential and exact correspondences in shared content is not a compulsion. In such cases, the preferences shown by communities in sharing is a deciding factor to incorporate the constraints in content consistency. In the absence of such preferences, propagating content updates becomes difficult as there is no prior information on restriction in propagation resulting in content inconsistencies such as global or local inconsistency in the shared content among communities which are also shown in the example. Towards determining preferences in sharing among communities, this chapter contributed from analysis based on propagation to qualitatively compared webpages from specific content categories and examined their propagation in website graph and website pairs.

From examining propagation in (1:7) country-specific websites, we revealed scales in sharing that varied with specific content categories. Re-

sult suggested that “Corporate Information” tend to be shared globally and “Customer Support Information” tend to be shared locally while “Product Information” tends to be locally and regionally suitable for sharing. Examining propagation in (1:1) country-specific website pairs revealed coupling in sharing due to differences in the occurrences of propagation for specific content categories. Result showed tendency for high coupling in websites while sharing content for “Corporate Information” and suggested content consistency to be strictly enforced for such content. From revealing traits such as scale and coupling in websites, this chapter expanded our understating of preferences in sharing that differ for specific content categories. The preferences are useful as guidelines for web manager to promote consistency in cross-site content and important in multi-language knowledge sharing system catered to customization in knowledge sharing.

# Chapter 5

# Determining Preferences in Sharing with Geographic Regions

Communities that share content comprising several categories are found to show specific preferences on scales and coupling in the previous chapter. Though country-specific websites offer several such content categories; those websites also represent geographic regions such as Europe, Asia Pacific, North America and so on. On this consideration, the presence of inconsistent content in websites has the potential to cause regional discrepancies, for example inconsistent content in product usage or specification for customer in Asia Pacific and North America. To avoid such discrepancies, it is necessary to have an understanding of the underlying preferences among communities in sharing within or beyond a specific region. This chapter undertakes analytical studies to determine the preferences in sharing by examining the propagation of content among country-specific websites within a geographic region and among geographic regions. Traits such as coupling and scales in sharing that vary for specific geographic region and specific content categories are revealed which showed specific preferences among communities and raised the prospect for avoiding regional discrepancies.

## 5.1 Background

Though websites offer a direct channel for global brands to communicate their businesses in the international market, it is a delicate medium when it comes to enhancing relation with customer [Argenti and Druckenmiller, 2004]. Inconsistency from the unavailability of up-to-date content or the presence of conflicting content in country-specific websites have the potential to risk brand image. More severe are inter-regional and intra-regional discrepancy caused from conflicting content shared with customer in same or different geographic region, for example mismatch in information related to product usage, specification for customer in Asia Pacific and North America. This chapter foresees the occurrence of regional discrepancies in information shard via country-specific websites.

**Definition.** Regional discrepancy in cross-site content is said to occur when content published in website $w_1$ representing geographic region $r_1$ and website $w_2$ representing geographic regions $r_2$ share inconsistent content inside region $r_1$ or between regions $r_1$ and $r_2$; i.e. $Semantic(Content, w_1) \neq Semantic(Content, w_2)$ due to (a) missing information in one of websites either $w_1$ or $w_2$ (b) updates not propagated from website $w_1$ to $w_2$ or vice versa and (c) conflicting information published between websites $w_1$ and $w_2$.

Previous researches mostly studied content and design features in corporate websites among cultural groups, industry or product types with mixed results for standardization and customization [Halliburton and Ziegfeld, 2009, Shin and Huh, 2009, Fletcher, 2006]. Geographic perspective is relatively less explored. However the persistence of cultural differences in geographic region is also noted in past studies which showed varying preferences in the use of messaging service among communities from North America and Asia [Kayan et al., 2006]. The perception on website effectiveness that varied for customer from North America and European region; the differences between how customers in North America perceive marketing stimuli as compared to customers in other parts of the

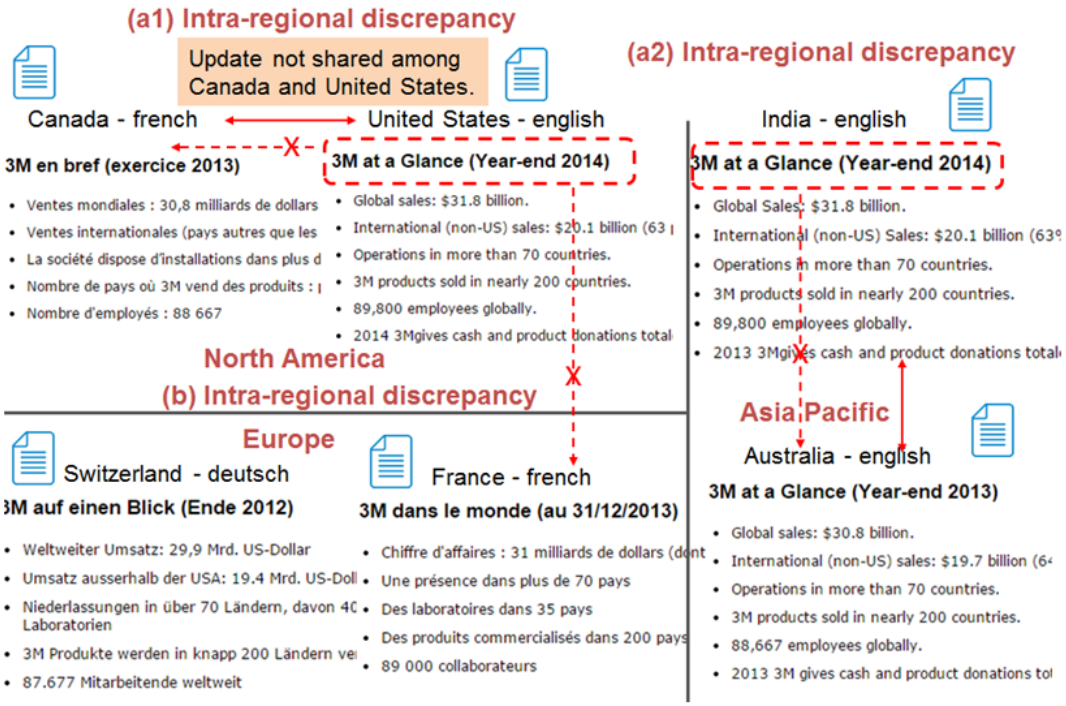



Figure 5.1: Intra- and Inter-regional Discrepancy in Cross-Site Content.

world in [Chakraborty et al., 2005, Lynch and Beck, 2001] are motivating to consider geographic aspects in sharing.

## 5.2 Regional Discrepancy in Cross-Site Content

Fig. 5.1 illustrates discrepancies in information shared with customer residing in several geographic regions that possibly occurs while managing websites in global brands. For illustration purpose, we examine the content "3M at glance" from country-specific websites managed in global brand '3M' for United States, Canada, France, Switzerland, India and Australia representing geographic regions: North America, Europe and Asia Pacific. Following problems are compiled.

### 5.2.1 Intra-regional Discrepancy

As illustrated only country-specific websites for US and India offer latest information for the year 2014 while the websites for remaining countries offer information for the previous year. Intra-regional discrepancies are highlighted from the lack of sharing latest information among countries inside same geographic region such as Asia Pacific (India and Australia) and North America (United States and Canada).

### 5.2.2 Inter-regional Discrepancy

Though latest information is partly accessible to customer in countries within Asia Pacific and North America, the latest information is not available at all to customer accessing websites in Europe highlighting the occurrence of inter-regional discrepancies. In addition to absence of latest information in the websites within Europe; conflicts in information shared for the previous year 2013 among country-specific websites for France, Australia and Canada is also observed. The statistics in 'global sales' and 'number of employees' offered in websites for France conflicts with content offered both in Canada and Australia. Inter-regional discrepancies from the presence of conflicting content in countries among region such as Asia Pacific, North America and Europe leads to customer accessing contradictory information shared among these region. The occurrences of regional discrepancies in shared information due to content updates not propagated or content conflict illustrated from this example motivates for examine propagation for restriction in sharing for specific geographic regions as existent in the country-specific websites which are useful in the design of multi-language knowledge sharing system.

## 5.3 Hypothesis

The strategy for promoting business globally is a daunting task for managers who often rely on tools to make critical business decisions such as

CAGE tool with cultural, administrative, geographic and economic measures [Ghemawat, 2001, Nachum and Zaheer, 2005]. For example opportunities for global business investment in a host/target country increases from the closeness in terms of shared language, past colony-colonizer links, common border and so on. Notably the influence of cultural factors is also seen in the web presence in deciding what constitutes website localization to the severity of localization effort. For instance, the increasing cultural distances and closeness in the physical distances between home and target market is found to impact the multinational's decision in launching local sites for a specific market [Vrontis et al., 2012]. Though web seems culturally and physically neutral medium, the cultural theorists also have diverging opinion from standardized "one size fits all" to depiction of cultural relevance with the local market. Such opposing view also influences the managerial duties for managing design and content in websites and probably on sharing with the option to either decentralize localization responsibilities to country offices or centralize in home country [LionBridge, 2009].

Though previous researches have explored preferences for standardization or localization in the design and content among cultural groups, industry, product types and so on; the geographic factor is relatively less explored. The depiction of preferences in the use of instant messaging that varies among North America and Asia [Kayan et al., 2006]; differences in the perception of customer to marketing stimuli and website effectiveness from specific region such as North America and Europe are supportive of geographic consideration in content and design features [Chakraborty et al., 2005, Lynch and Beck, 2001]. Catered to specific region, the location specificity in knowledge sharing among certain geographic region also put forth that relevance of knowledge is confined within specific region and transferring the same knowledge to other region is a futile practice. Grounding on this notion, content shared among websites is presumed to depict either standardization with same content offered or localization with localized content not shared among websites. As cultural differences in geographic regions and location specificity in knowledge sharing are depicted in previous researches, the suitability of content

for specific region and their restriction in sharing among country-specific websites is also presumed to differ with geographic regions. To shed light on restriction on sharing with geographic regions, we set the following hypothesis.

**Hypothesis 1.** Propagation among country-specific websites is constrained by geographic regions: Asia Pacific, Europe, North America and so on

As several categories such as corporate related or product related information and so on are published in websites, the restriction in sharing content for specific category is also presumed to vary with geographic regions. We set the following hypothesis.

**Hypothesis 2.** Propagation of content categories such as corporate related or product related information among country-specific websites is constrained by geographic regions.

The goal from the stated hypotheses is to uncover traits from examining propagation of content among websites targeted for specific region. The contribution will be towards determining the preferences in sharing among communities that represent specific geographic regions which shed light on customization in knowledge sharing.

## 5.4 Outline on Methodology

Fig. 5.2 gives an outline on methodology for examining sharing among countries in specific geographic regions from sampling webpages to comparing webpages in their country-specific websites.

### 5.4.1 Websites and Geographic Regions

We selected websites from 10 global brands (Nivea, 3M, Starbucks, Acer, Samsung, KPMG, HP, Nestle, Avon, and John Deere) that are listed in the web globalization report card [Yunker, 2014]. From each global brand we

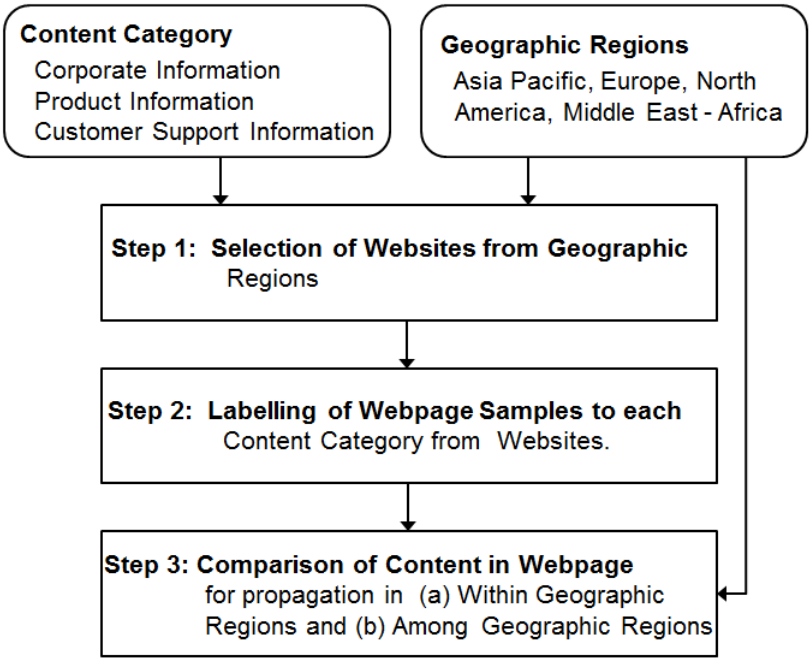



Figure 5.2: Outline on Examining in Geographic Regions.

Table 5.1: Statistics on Websites and Geographic Regions.

| Brand | Geographic Region | Website in each brand | Content Category | Total | |
|---|---|---|---|---|---|
| | | | | Website | Webpage |
| 10 | 4 | 4 * 2 = 8 | 3 | 10 * 8 = 80 | 10 * 48 = 480 |

sampled 8 country-specific websites from geographic regions: Asia Pacific, North America, Europe and Middle East-Africa that represent countries India, Australia, United Kingdom, Ireland, United States, Canada, Middle East and South Africa respectively. We collected a total of 80 country-specific websites are the source for sampling webpages (Table 5.1). Four regions particularly Asia Pacific, North America, Europe and Middle East-Africa are chosen as global brands are found to categorize their country-

specific websites into these regions. Also, cultural differences among chosen regions are identified in previous researches[Kayan et al., 2006, Chakraborty et al., 2005]. We also use content categories comprising of "Corporate Information" in sampling webpages that provide background information of a company such as mission statements, history and its people; "Product Information" in sampling webpages on description, usage, and specification of product and "Customer Support Information" in sampling webpages on ways to contact company or find answer to queries.

### 5.4.2 Comparison Within and Among Geographic Regions

From country-specific websites in each geographic region, webpages that offer content in English language is manually analyzed to label them to specific content categories. We collected 48 webpage samples from website in each global brand making a total of 480 webpage samples Table 5.1. We qualitatively compared webpages to determine whether propagation occur or do not occur among the websites within and beyond geographic regions. Propagation is said to occur among the websites upon the presence of exactly same paragraph or comparable paragraph in their webpages. Similarly, no propagation among websites is assigned when content is not same in the webpages of the corresponding website or webpage do not exist.

Table 5.2: Statistics on Comparison within Geographic Regions.

| **Each Region** | | **All Regions** | |
|---|---|---|---|
| **Each Category** | **All Category** | **Each Category** | **All Category** |
| 20 | 3 * 20 = 60 | 4 * 20 = 80 | 3 * 80 = 240 |

Table 5.3: Statistics on Comparison among Geographic Regions.

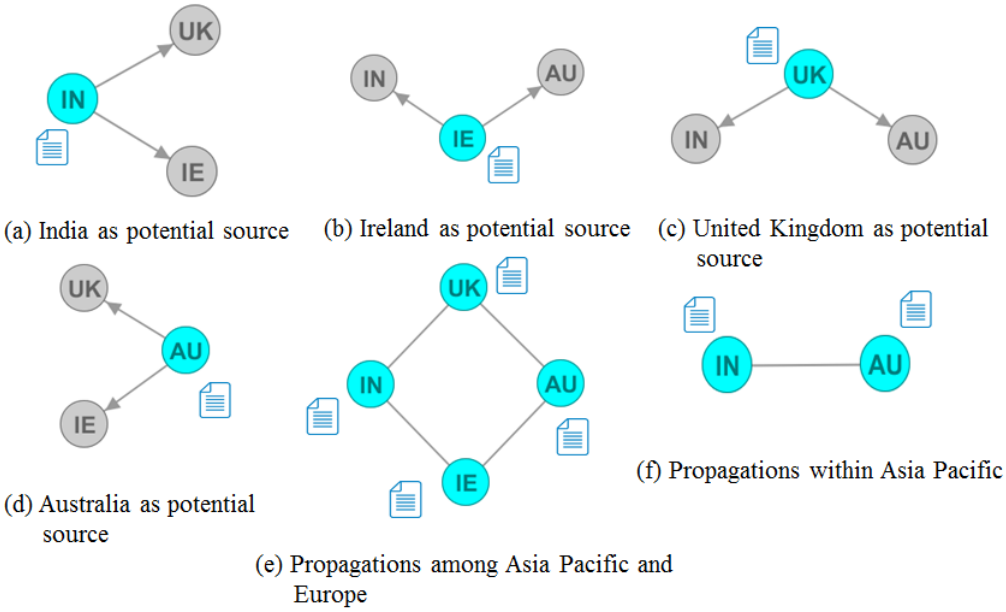
| Each Region | | All Regions | |
|---|---|---|---|
| **Each Category** | **All Category** | **Each Category** | **All Category** |
| 80 | 3 * 80 = 240 | 6 * 80 = 480 | 3 * 480 = 1440 |

**(a) Within Geographic Regions**

Fig. 5.3(f) illustrates an example of propagation occurring within Asia Pacific with propagation of content in websites between India and Australia. As the information on source website where the content first originates in its webpage is not publicly accessible; each country-specific website is considered as a potential source for publishing content in its webpage and sharing with the remaining websites (cyan colored node represents the source website). A total of 240 comparisons of webpages are performed to check for propagation occurring within all four geographic regions as shown in Table 5.2.

**(b) Among Geographic Regions**

From four geographic regions there are 6 possible inter-regional comparisons for propagation among region such as Asia Pacific - Europe, Asia Pacific - North America and so on. Fig. 5.3(a) depicts India as the potential source website with content propagated to countries in Europe (United Kingdom and Ireland). From choosing each country as a potential source website and comparing webpages the graph in Fig. 5.3(e) represents propagation occurring among countries in Asia Pacific and Europe. For each such inter-regional pair, 240 comparisons of webpages are performed. Such comparisons are repeated to check for propagation in remaining inter-regional pairs. A total of 1440 comparisons of webpages are performed to check for propagation among geographic regions as shown in Table 5.3.




(a) India as potential source (b) Ireland as potential source (c) United Kingdom as potential source

(d) Australia as potential source (f) Propagations within Asia Pacific

(e) Propagations among Asia Pacific and Europe

Figure 5.3: Propagation in Geographic Regions.

## 5.5 Analysis on Propagation in Geographic Regions

Results from comparing webpages for propagation within and among geographic regions are compiled in this section. Further result from examining specific content categories and their propagation for specific geographic region is also presented.

### 5.5.1 Propagation within Geographic Regions

The qualitative result of comparing webpages for propagation among country-specific websites inside each geographic region is compiled in Table 5.4. The occurrence of propagation and no propagation among websites is used as a measure for coupling in sharing content among countries within region. The higher the occurrences of propagation within a region indicates high coupling in country-specific websites within that region. Comparing

Table 5.4: Summary: Propagation within Geographical Regions.

| Within Geographic Regions | Occurrences of Propagation | | | |
|---|---|---|---|---|
| | Yes | % | No | % |
| Asia Pacific | 29 | 48 | 31 | 52 |
| Europe | 35 | 58 | 25 | 42 |
| North America | 20 | 33 | 40 | 67 |
| Middle East - Africa | 32 | 53 | 28 | 47 |

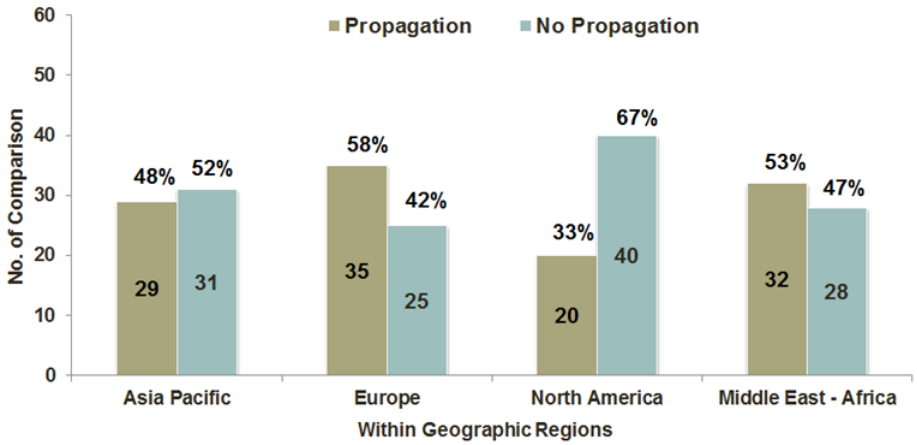



Figure 5.4: Comparison of Propagation within Geographic Regions.

the webpages from country-specific websites inside each geographic region, differences in coupling in websites are revealed.

As shown in Fig. 5.4 the number of occurrences of propagation and no propagation among website are comparable within Asia Pacific and Middle East-Africa. However for websites within North America, the occurrence of propagation tends to be less 33% while majority of cases 67% show no propagation occurring among websites. In contrast the number of occurrences of propagation tends to be higher 58% among websites in Europe with some cases 42% of no propagation. The differences in the occurrences of propagation and no propagation suggest on coupling in websites that vary with each region.

Table 5.5: Summary: Propagation among Geographic Regions.

| | Asia Pacific | | | | Europe | | | | North America | | | | Middle East - Africa | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Yes | % | No | % | Yes | % | No | % | Yes | % | No | % | Yes | % | No | % |
| **Asia Pacific** | | | | | **107** | **45** | **133** | **55** | **78** | **33** | **162** | **68** | **105** | **44** | **135** | **56** |
| **Europe** | 107 | 45 | 133 | 55 | | | | | **83** | **35** | **157** | **65** | **110** | **46** | **130** | **54** |
| **North America** | 78 | 33 | 162 | 68 | 83 | 35 | 157 | 65 | | | | | **87** | **36** | **153** | **64** |
| **Middle East - Africa** | 105 | 44 | 135 | 56 | 110 | 46 | 130 | 54 | 87 | 36 | 153 | 64 | | | | |

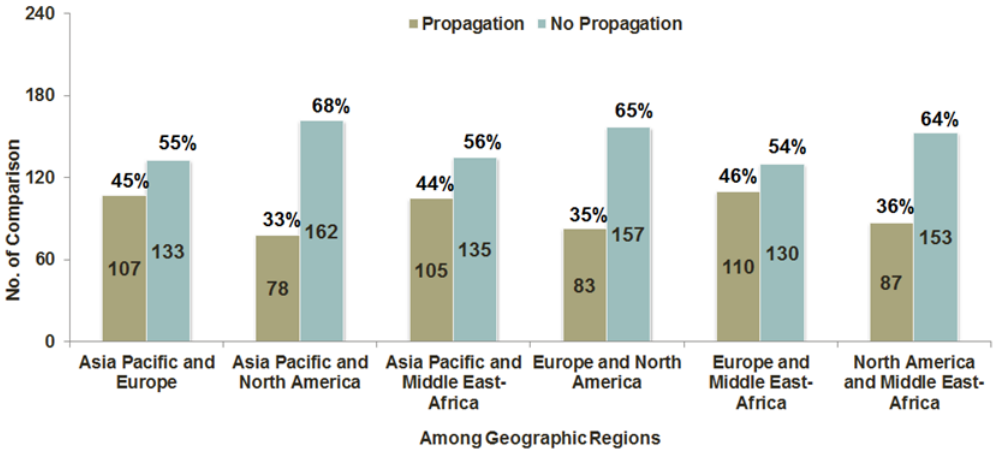



Figure 5.5: Comparison of Propagation among Geographic Regions.

### 5.5.2 Propagation among Geographic Regions

Table 5.5 compiles the qualitative result of comparing webpages in websites for propagation among geographic regions. As illustrated in Fig. 5.5 the number of occurrences of propagation and no propagation is comparable among Asia Pacific, Europe and Middle East-Africa. However, there tend

to be noticeable differences in the number of occurrences of propagation and no propagation while sharing content with countries in North America. Less than 40% cases of propagation occurring from countries in Asia Pacific, Europe and Middle East-Africa with countries in North America and more than 60% cases of no propagation in websites for sharing content from Asia Pacific, Europe and Middle East- Africa with customer in North America. Similar to coupling inside geographic region, differences in coupling in websites are also identified that vary for sharing content with other region.

### 5.5.3 Propagation of Content Category in Geographic Regions

The qualitative results of comparing webpages for specific content categories "Corporate Information", "Product Information" and "Customer Support Information" in websites among several geographic regions is shown in Table 5.6. The occurrences of propagation tend to be higher among regions Asia Pacific, Europe and Middle East-Africa for sharing 'corporate related information' in comparison to North America as shown in Fig. 5.6. Higher occurrences of no propagation among region while sharing "Product Information" and "Customer Support Information" are also seen in Fig. 5.7 and Fig. 5.8. More than 70% no propagation cases are found for product related information with countries in North America while no propagation cases are higher (more than 70%) among all four geographic regions.

## 5.6 Preferences with Geographic Regions

Referring to Fig. 5.4, Fig. 5.5, Fig. 5.6 , Fig. 5.7 and Fig. 5.8 insights into coupling in websites as well as scales in sharing specific content categories within and beyond geographic regions are compiled that give rise to preferences in sharing among communities.

Table 5.6: Summary: Content Category among Geographic Regions.

| Among Geographic Regions | Corporate Information | | | | Product Information | | | | Customer Support Information | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Yes | % | No | % | Yes | % | No | % | Yes | % | No | % |
| Asia Pacific and Europe | 58 | 73 | 22 | 28 | 31 | 39 | 49 | 61 | 18 | 23 | 62 | 78 |
| Asia Pacific and North America | 46 | 58 | 34 | 43 | 16 | 20 | 64 | 80 | 16 | 20 | 64 | 80 |
| Asia Pacific and Middle East -Africa | 58 | 73 | 22 | 28 | 29 | 36 | 51 | 64 | 18 | 23 | 62 | 78 |
| Europe and North America | 49 | 61 | 31 | 39 | 17 | 21 | 63 | 79 | 17 | 21 | 63 | 79 |
| Europe and Middle East - Africa | 61 | 76 | 19 | 24 | 27 | 34 | 53 | 66 | 22 | 28 | 58 | 73 |
| North America and Middle East - Africa | 46 | 58 | 34 | 43 | 23 | 29 | 57 | 71 | 18 | 23 | 62 | 78 |

### 5.6.1 Coupling within Geographic Regions

- Referring to Fig. 5.4 , with majority cases of propagation almost 60% occurring among country-specific websites in Europe, high coupling in websites among countries in Europe is suggested. Similarly, tendency for no propagation, more than 60%, occurring among country-specific websites in North America suggest low coupling in sharing content among countries in North America.

- Results also revealed that global brands tend to show high coupling in their country-specific websites when sharing content targeted for customers from European region while preferred less coupling in websites targeted for customers in North America.

As previous researches accounted for cultural differences in geographic regions, the findings on differences in coupling further supported existing differences in sharing content for specific region.

- Another interesting finding from differences in coupling is that country-specific websites in North America tend to be more autonomous in comparison to websites in Europe.

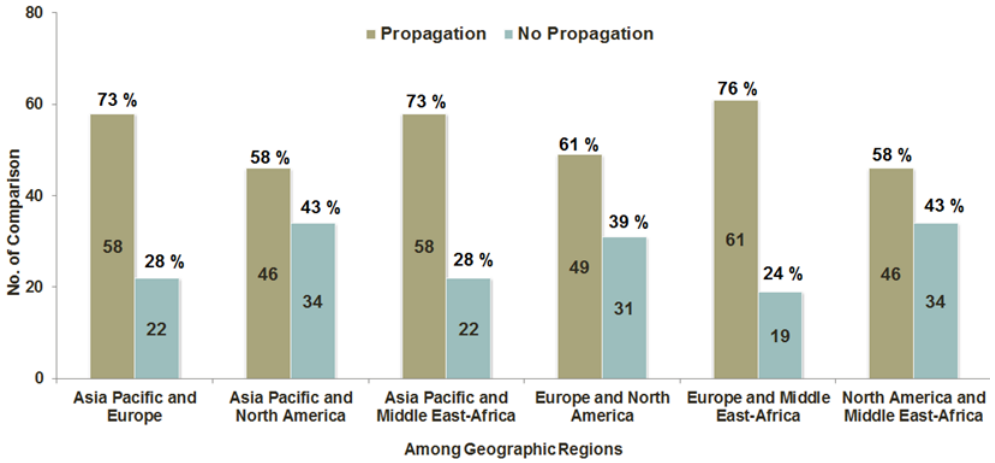



Figure 5.6: Propagation for Corporate Information.

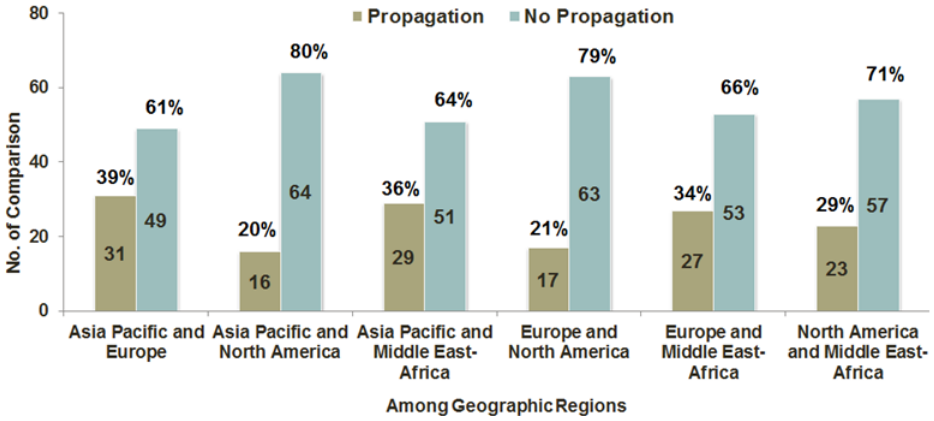



Figure 5.7: Propagation for Product Information.

English being a dominant official language in North America while several official languages is used in Europe, the low coupling in websites within North America seems reasonable as the customer will probably access English content if not available in their mother language. On the other hand, higher coupling in websites within Europe also seems reasonable as content is available in multiple official languages in several countries, more interaction is needed to assure that content is shared with customer from most of

the countries in Europe.

### 5.6.2 Coupling among Geographic Regions

- Less than 40% cases of propagation occurring from countries in Asia Pacific, Europe and Middle East-Africa with countries in North America and more than 60% cases of no propagation suggest low coupling in websites for sharing content from Asia Pacific, Europe and Middle East- Africa with customer in North America. (Fig. 5.5)
- Differences in coupling also suggest that global brand tends to show preference for sharing content mostly among markets in Asia Pacific, Europe and Middle-East Africa while prefer to offer specialized content for markets in North America not shared with other region.
- The low coupling with websites in North America also suggest that country-specific websites within North America tend to have less interaction with websites from other region.

Though, English is globally used, languages other than English is used in region outside North America which could possibly reduce interactions

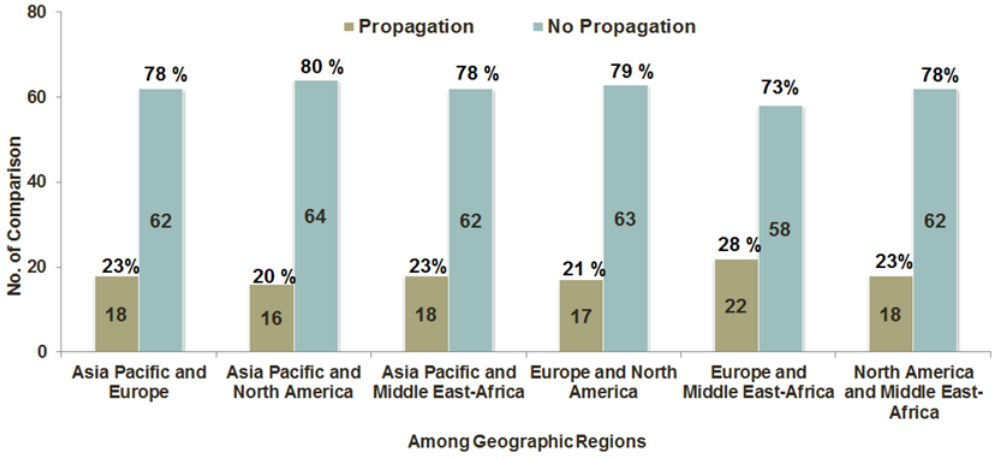



Figure 5.8: Propagation for Customer Support Information.

among websites from North America and remaining regions.

### 5.6.3 Scales and Coupling in Content Categories

- Occurrences of higher propagation in websites among regions while sharing content for "Corporate Information" outside its specific region suggests high coupling in websites sharing 'corporate related information' from one region to another region. In fact propagation is higher (more than 70%) among Asia Pacific, Europe and Middle East-Africa (Fig. 5.6).

- Higher occurrences of propagation among region also suggest "Corporate Information" to be globally suitable for customer in all regions.

- Higher occurrences of no propagation among region for "Product Information" and "Customer Support Information" suggest that such content tend to be region specific and either locally or regionally suitable (Fig. 5.7 and Fig. 5.8).

- Low coupling in country-specific websites is also suggested for the higher occurrences of no propagation in sharing content describing the product or contact with its customer.

- Noticeable differences with less than 30% occurrences of propagation and more than 80% no propagation while sharing "Product Information" with countries in North America also suggested that websites in North America are more likely to prefer region-specific content when describing the specification or usage of the product (Fig. 5.7).

- The differences in the occurrences of propagation and no propagation for "Customer Support Information" seem to be consistent among all regions suggesting websites in all regions are more likely to prefer offering region-specific content when describing customer support (Fig. 5.8).

## 5.7 Guidelines on Content Consistency

From identifying traits as scales and coupling in websites, the analysis depicted preferences in sharing among communities that vary with specific geographic region and specific content categories shared in those regions which are important as guidelines for web manager. Following guidelines are compiled from this analysis.

- Websites inside European region are found to prefer more interaction while sharing content so intra-regional consistency is needed when sharing content in European market.
- Websites inside North America tend are found to be autonomous and prefer to have less interaction with websites from other regions so local consistency is needed when sharing content in North American region.
- Websites among Asia Pacific, European and Middle East-Africa are found to prefer more interaction in their shared content so inter-regional consistency is needed to share consistent content among those regions.
- Corporate related information is found to be share more frequently among Asia Pacific, European and Middle East-Africa so inter-regional consistency is needed when sharing such content.
- Product related information is found to be specialized for North America but are shared among other regions so local consistency is suited inside North America and inter-regional consistency is suited among remaining regions.
- Customer support information is found to be specialized in each region so intra-regional consistency is suited when sharing such content.

These guidelines provide a general idea on the consistency needed when sharing content within and beyond specific geographic regions and also on

the restriction to specific languages that are used by the communities. Considering such guidelines web manger is better prepared to customize consistency policy for specific geographic region and avoid regional discrepancies in the content shared with their customer via country-specific websites.

## 5.8 Hypothesis Verification

The occurrences of propagation and no propagation among websites are found to vary within and beyond the geographic regions. High coupling in websites for sharing content among countries in Europe due to higher occurrences of propagation while low coupling between websites among countries in North America are found. Propagation among geographic regions also suggested differences in coupling in websites while sharing content with other region. Results are convincing in illustrating propagation varying among websites within and beyond geographic regions depicting differences in coupling and supporting hypothesis H1 for constraints due to geographic region.

Similarly comparing webpages for specific content categories also revealed propagations that vary in coupling and scales. Higher occurrences of propagation while sharing content for "Corporate Information" and higher occurrences of no propagation among country-specific websites while sharing content for "Product Information" and "Customer Support Information" suggested different levels of scale in sharing such content. Results are convincing in illustrating propagation that vary for specific content categories among geographic regions supporting hypothesis H2.

## 5.9 Summary

The focus in this chapter is to expand our understanding on managerial preferences when it comes to sharing catered to specific geographic regions as depicted in the country-specific websites of global brands. Ignoring such preferences while managing websites can result in inconsis-

tent content shared among region causing regional discrepancies, for example inter-regional discrepancies in information shared with customer from North America and Europe. To deal with regional discrepancies, this chapter contributes from the analytical studies by qualitatively comparing content in webpages and examining their propagation among websites within and beyond geographic regions. From examining propagation within geographic regions high coupling in websites among countries in Europe while low coupling in websites in North America are revealed which suggested that websites inside North America tend to be autonomous and participate less in sharing. Higher propagations inside European region also revealed global brands preferences for sharing most of its content among countries in Europe compared to other region. This raised an important concern that among all regions customer inside European region is more vulnerable to intra-regional discrepancies from inconsistency in shared content.

Similarly, examining propagation among geographic regions also revealed differences in coupling with low coupling in websites while sharing content with countries in North America. This further supported the autonomous nature of websites in North America showing preferences for less interaction with websites from other region. However, websites from Asia Pacific, Europe and Middle-East Africa are found to participate mostly in sharing among themselves; hence more vulnerable to inter-regional discrepancies from sharing inconsistent content with customer in these region. The inspection of specific content categories also revealed tendencies for sharing corporate related information globally in all regions while information related to product and customer support within specific region. It is found that websites in North America have higher preferences for specialized product related information not shared with other region; while customer support related information are specialized inside all regions and not shared.

The revelation of preferences in sharing content within and beyond geographic regions offered guidelines on consistency policy customized for specific region. For example, rigid policy for content consistency is suited for sharing content with customer in European market; whereas policy can be lenient while sharing content with customer in North America. From

revealing traits such as coupling and scales the preferences among communities in sharing with respect to specific geographic regions and specific content categories are determined in this chapter which is useful in designing multi-language knowledge sharing system that offers customization in knowledge sharing.

# Chapter 6

# Supporting Consistency in Customized Knowledge Sharing

Grounding on the analytical results of previous chapters which depicted restrictions in propagation for specific content categories and geographic regions this chapter presents a technique to support content consistency allowing community preferences in knowledge sharing. The technique is based on the concept of propagating content updates restricted to specific languages or communities employed as pattern of sharing in the delivery of knowledge. Allowing community to specify preferences with pattern of sharing the technique enables content updates to propagate where necessary which means exact correspondence in shared content is made only where necessary. Though techniques proposed in chapter 1 focused only on multilingual content, the technique in this chapter is applicable for both multilingual and monolingual cases. The advantage of proposed technique is its simplicity with support for either automation or employed manually for consistent knowledge sharing.

## 6.1 Background

Community preferences in sharing is found to persist in knowledge sharing from the presence of traits such as scales and coupling that vary with

the categories of content shared and to specific geographic regions. Former analyses clearly showed that in the absence on information related to scale and coupling, it is not known how and where to propagate content updates which result in global and local inconsistency as well as regional discrepancy in content shared in websites i.e. cross-site content. However since traits such as scales provide information on the spread of content and the need for content consistency globally, regionally or locally, the propagation of content updates can be restricted based on scales.

Similarly, coupling provides information on priorities on content consistency especially where there are several categories to share with various geographic regions. Such traits especially scales provide information on the restriction of content updates to specific communities associated with specific content categories and geographic regions. In the light of community preferences, the concept of propagating content updates restricted to specific communities depending on the content categories and geographic regions is used as a technique to promote consistency.

In the propagation of content among country-specific websites the sharing is obvious with either one of three cases: (a) propagation to entire country-specific websites (b) propagation to some country-specific websites and (c) no propagation. Information on the suitability of content either to entire countries or some counties or even to a single country is gathered from such cases. In globalization studies such cases confirm the suitability at global, regional or local scales. Interesting results on the suitability of content when shared in geographic regions and the confines on the propagation of content updates can be based on these cases.

## 6.2 Pattern of Sharing

Three cases (a) propagation to entire country-specific websites (b) propagation to some country-specific websites and (c) no propagation are generalized as the pattern of sharing.

### 6.2.1 Internationalization

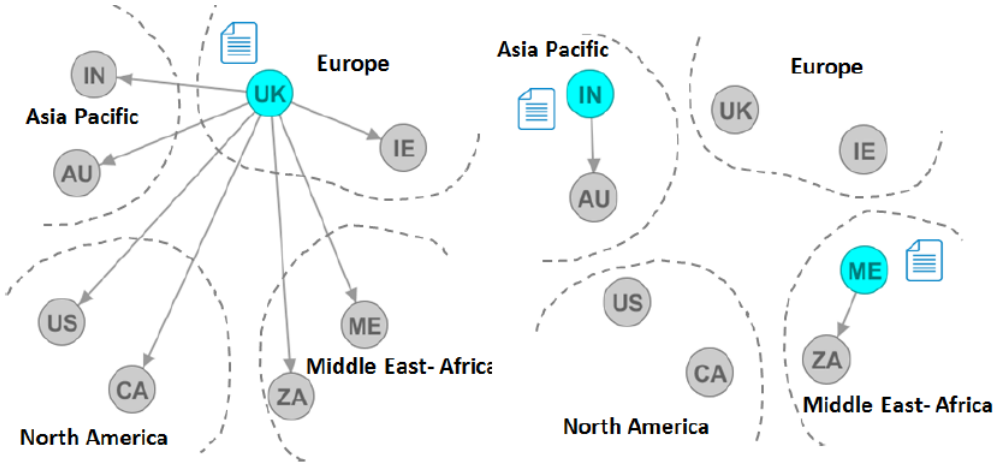
Previous studies [Esselink, 2000] have presented views on 'Internationalization' as the processes of generalizing a product for handling multiple languages and cultural conventions. With respect to content shared among country-specific websites, this represent content suitable at a global scale or content that can be produced in multiple languages for global communities. This also represents the propagation of content among entire countries from the publication of content globally and description in all available language offered in global brand. Fig. 6.1(a) in which content "About Us" managed in global brand "John Deree" is propagated to countries in all region resembles Internalization in sharing. For up-to-date knowledge sharing, integrating Internationalization pattern in delivery of knowledge while sharing content for corporate related information enables propagation of content updates globally among countries and in several languages.

### 6.2.2 Regionalization

In context to globalization, the view on regionalization represents a world that becomes less interconnected with a stronger regional focus and is interesting for researches on market segmentation [Rugman and Verbeke, 2004]. Inter-regional and intra-regional suitability in content shared as identified for produced related information"Gear S" generalizes as Regionalization pattern in the delivery of knowledge regionally among countries Fig. 6.1(b).

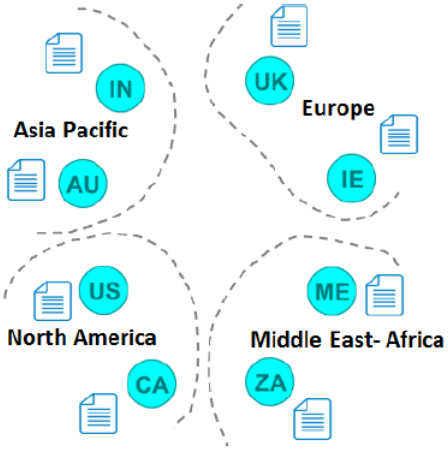
### 6.2.3 Localization

The view on localization is towards the process for making a product linguistically and culturally appropriate to the target locale (country/ region and language) where it will be used and sold [Esselink, 2000]. For content shared among country-specific website, localization represents the suitability of content at local scale without any necessities for the propagation of content updates among country-specific websites. Fig. 6.1(c) represents content "Contact Us" managed in Starbuck which is locally managed in




(a) Content "About Us" in John Deree propagated to all websites

(b) Content "Gear S" in Samsung propagated to some websites




(c) Content "Contact Us" in Starbucks propagated to none

Figure 6.1: Propagation among Country-Specific Websites in Global Brand.

each country. However integrating Localization in the delivery of knowledge enables consistent knowledge sharing from the propagation of content updates in multiple official languages within a country. Content consisten-

cies for customer support related information are promoted due to their local scales from integrating Localization pattern in the delivery of knowledge.

The combination of pattern of sharing in publishing webpages comprising component content that is to be shared globally, locally and regionally is also possible. Glocalization in the delivery of knowledge with content communicating both globally and locally [Maynard and Tian, 2004, Svensson, 2001] enabling knowledge sharing for both global and local communities is achieved from integrating Internationalization and Localization pattern. The propagation of content updates from the combination of global and local scales provides possibilities for knowledge sharing in which globally relevant knowledge are reused among countries while locally relevant knowledge from a specific country are referenced to generate locale-specific content for another country. The global and local dimension of knowledge development and sharing in [Adenfelt and Lagerström, 2006, Almeida and Phene, 2004] is also illustrated from such pattern.

Integrating pattern of sharing (a) Internationalization (b) Regionalization (c) Localization and their combinations in the delivery of knowledge; consistency in knowledge sharing is supported from the propagation of content updates restricted to global, regional or local communities and in specific languages. Such pattern of sharing can be applied to content either (i) automatically by employing text mining approaches to identify specific categories or (ii) manually by generating policy while sharing among countries.

## 6.3 Formalizing Rules

To describe the rules associated with pattern of sharing, we will use the following notations. Collection of website W is published in a global organization where each country-specific website $W_j \in W$ is targeted for specific country $j$ . Collection of language L is used in the organization with an official language $L_j \in L$ for a specific country. $W_j^i$represents a country-specific websites for country $j$ offering content in language $i$. R represents a geo-

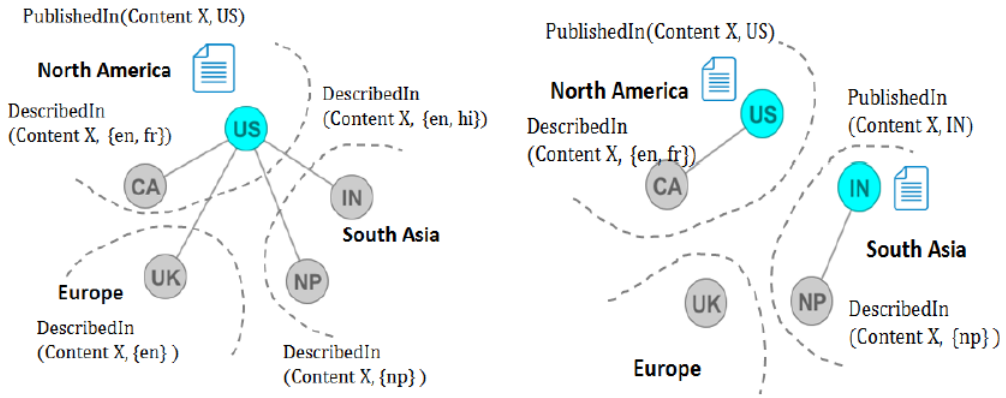
Table 6.1: Formalized Rules in Pattern of Sharing

| Pattern | Rule in Collaboration |
|---|---|
| **Internationalization** | **Rule 1:**<br>$\forall\, x : isContent(x) \wedge isSharedApplying(x, Internationalization) \Longrightarrow isPublishedIn(x, \mathrm{W}) \wedge isDescribedIn(x, \mathrm{L})$<br>**Description:** Content shared applying Internationalization pattern is published in entire websites and is described in entire languages offered among the websites. |
| **Localization** | **Rule 2:**<br>$\forall\, x, \exists\, j : isContent(x) \wedge isSharedApplying(x, Localization) \Longrightarrow isPublishedIn(x, \mathrm{W}_j) \wedge isDescribedIn(x, \mathrm{L}_j)$<br>**Description:** Content shared applying Localization pattern is published in the website of specific country $j$ and described only in the languages offered in that particular country. |
| **Regionalization** | **Rule 3:**<br>$\forall\, x, \exists j \in \mathrm{R} : isContent(x) \wedge isSharedApplying(x, Regionalization) \Longrightarrow isPublishedIn(x, \mathrm{W}_j) \wedge isDescribedIn(x, \mathrm{L}_j)$<br>**Description:** Content shared applying Regionalization pattern is publish at the specific website of country $j$ belonging to specific region R and described in the languages offered in countries from the region R. |

graphic regional to which a country $j$ belongs.

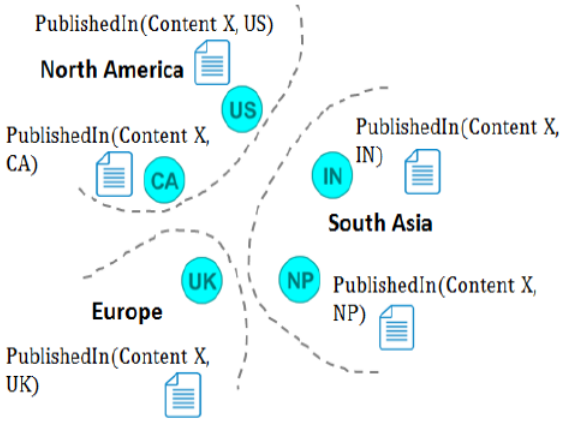
Following this formalization, a country-specific website for Canada is $W_{j=ca}$ and the official languages of Canada $L_{j=ca} = \{en, fr\}$ are English and French. The constructs used in formalizing rules are as following. Content shared among countries with a specific pattern is represented with $isSharedApplying(Content, Pattern)$. Content published in a website and described in specific language is represented with $isPublishedIn(Content, Website)$ and $isDescribedIn(Content, Language)$ respectively. Rules associated with the pattern of sharing in the delivery of knowledge are illustrated in Table 6.1. The publication of content at specific websites and their description in specific languages represent the scales that

restrict the propagation of content updates. The pattern of sharing applied in a collaborative setting is explained next.




(a) Internationalization Pattern

(b) Regionalization Pattern




(c) Localization Pattern

Figure 6.2: Pattern of Sharing applied in Delivery of Knowledge.

## 6.4 Applying Pattern of Sharing

For illustration purpose, web mangers managing websites for Canada (CA), United Kingdom (UK), United States (US), India (IN), Nepal (NP) with content offered in their official languages $L_{j=ca} = \{en, fr\}$, $L_{j=uk} = \{en\}$, $L_{j=us} = \{en\}$, $L_{j=in} = \{hi, en\}$ and $L_{j=np} = \{np\}$ are considered as participants in knowledge sharing (Fig. 6.2). In this example, English language is shared among UK, US and India; while both India and Canada also have multiple official languages. Consistent knowledge sharing among countries is required in both shared and unshared languages where needed from the delivery of knowledge.

Fig. 6.2(a) represents 'Internationalization' pattern applied for the delivery of knowledge from web manager for United States while sharing content with remaining countries. From Rule 1 in Table 6.1 , the content is published at United States and the description of content is shared in several languages (English, French, Hindi, Nepali) offered in websites for UK, CA, IN and NP. Updating content at any country-specific website for (United Kingdom, Canada, India, Nepal, United States) and in any languages will result in the delivery of updates at all websites from such pattern. This is depicted from the undirected propagation among the countries. Synchronization of content updates among countries in several official languages (both shared and unshared) is achieved from restricting the propagation globally.

The delivery of knowledge in Fig 6.2(b) represents 'Regionalization' pattern applied by web managers for United States and India to share content regionally with countries Canada and Nepal respectively. From Rule 2 in Table 12, the description of content is shared in several languages (English and French) offered among US and Canada and in languages (English, Hindi and Nepali) offered among Nepal and India. Content updates are shared regionally between countries such as within South Asia rather than with other regions. Inconsistencies in regionally shared content are avoided from the propagation of content updates limited to countries within regions and languages offered in those regions.

Fig. 6.2(c) represents 'Localization' pattern applied for the delivery of

knowledge locally. Such pattern is useful for restricting the propagation of content updates locally for countries offering multiple official languages. For example, content updates are synchronized locally in languages (English and Hindi) for India without sharing with remaining countries. With such pattern, the content is not shared with other countries even when a language is shared and updates are confined within a country and only in its languages. Inconsistencies in local knowledge sharing are avoided from applying Localization pattern in the delivery of knowledge. From the proposed pattern of sharing in the delivery of knowledge, we illustrated the techniques that can propagate content updates confined to specific communities globally, regionally and locally for content consistencies in knowledge sharing.

## 6.5 Summary

The focus in this chapter to propose technique that supports content consistency in customized knowledge sharing where exact correspondence in shared content is not always required. As community preferences in sharing that vary for specific content categories and specific geographic regions are depicted for customized knowledge sharing, the techniques allow community to specify their preferences as pattern of sharing. To support content consistency the technique is based on the concept of propagating content updates restricted to specific language or communities associated with pattern of sharing.

From generalizing three cases of propagation to entire country-specific websites, propagation to some country-specific websites and no propagation, pattern of sharing (a) Internationalization (b) Regionalization and (c) Localization is proposed with rules for restriction in the publication and description of content in specific languages. Several benefits is achieved from such pattern i) scaling the suitability of content for global, regional or local communities ii) propagation of content updates confined for consistent content sharing between the communities iii) synchronization of con-

tent updates globally in several languages offered on all country-specific websites or in the languages offered within a country-specific website. Integrating pattern of sharing to specific content categories either automatically or manually, inconsistency from missing content, content not updated and conflicting content among countries which are considered as problematic in knowledge sharing are avoided from propagation before even advancing for translation. The advantage of the technique is its simplicity and reliance on propagating content updates to support content consistency which makes it applicable for both monolingual and multilingual cases.

# Chapter 7

# Conclusion

Unprecedented growth in online collaboration is an opportunity for diverse communities to participate in knowledge sharing, example from resource rich to resource poor communities or vice versa. However inconsistency primarily from cases such as content omitted or content updates not shared and the presence of conflicting content is problematic for knowledge sharing among communities. Since it is impractical to propose consistency rules in multi-language knowledge sharing system in advance it is essential to focus on mentioned cases as potential cause for inconsistency in the shared content. Having said that though such cases may seem trivial at first eventually the complexity is increased as communities participate for knowledge sharing in their own languages and so inconsistent content is bound to occur in several languages. Further such cases also have the potential to cause inconsistencies at global and local scales leading to globally and locally shared inconsistent content among communities. Regional discrepancies due to inconsistent content shared with communities from several geographic regions are also equally anticipated in knowledge sharing.

In addition to the consequences from sharing inconsistent content in knowledge sharing the constraint in content consistency is another concern. The constraint due to diverging view on supporting content consistency is reliant on the knowledge sharing goals of the communities which is not uniform. Where the goal is to leverage knowledge equally among communities

a rigid consistency policy with exact correspondence in shared content is preferred while where the goal is to customize knowledge sharing there is a need to restrict sharing to specific language and specific communities with a non-rigid consistency policy which mean exact correspondence is preferred only where necessary.

## 7.1 Summary of Contributions

Grounding on the consequences from sharing inconsistent content and the constraints in content consistency arising from the disparate knowledge sharing goals of the communities this thesis makes following contributions in the design of multi-language knowledge sharing system.

**(1) Support for content consistency in leveraging knowledge equally among communities.**

The difficulty in leveraging knowledge equally is elevated from the participation of communities for knowledge sharing in several languages. Due to which inconsistency even from trivial cases such as omitted content, updated content not shared and content conflict is problematic with the potential to occur in several languages and shared with several communities. Such inconsistencies are undesirable among communities that prefer to share knowledge equally in diverse languages. To deal with content inconsistency in multiple languages or multilingual content, the contribution is made with a process-based technique proposed to detect the presence of new content, updated facts or information and content conflict between languages. Based on the concept of synchronizing user editing activities, state transition model is proposed which is used to model multilingual content with states, action performed on them and the set of transition functions. Inconsistency detection rules are designed to represent state of the multilingual content leading to inconsistency in multilingual content. Experimental results from applying the proposed technique to the test set of revision his-

tories in multilingual Wikipedia articles showed satisfactory results with an average precision of 88% and a recall of 86% in detecting inconsistency. The proposed technique is useful in multi-language knowledge sharing system to support content consistency in leveraging knowledge equally.

**(2) Guidelines on content consistency in knowledge sharing among communities with content categories.**

To support customization in knowledge sharing the contribution is made towards understanding the preferences in sharing specific content categories such as 'corporate related information' or 'product related information' and so on among communities, in other words the need to restrict sharing content in specific languages or to specific communities. Given that several content categories are published in websites and shared among communities analytical study is undertaken to examine the influence of specific content categories on preferences in sharing. The approach based on propagation is proposed to qualitatively compare webpages and examine their propagations among country-specific websites first in a website graph (interconnecting the available websites) and second in website pairs. 480 webpages from 80 websites representing 10 global brands (Nivea, 3M, Starbucks, Acer, Samsung, KPMG, HP, Nestle, Avon, John Deree) are analyzed. A total of 480 comparisons of webpages in website graph and 1680 comparisons in website pair are performed. Traits such as coupling and scales in sharing are revealed that vary for specific content categories indicating the reason for preferences in sharing.

Examining propagation in website graph revealed that "Corporate Information" has tendency to be shared globally and "Customer Support Information" has tendency to be shared locally while "Product Information" tends to be locally and regionally suitable for sharing. Implication of such findings is the guidelines on the content consistency constraints needed for specific content, example: global consistency required for 'corporate related information' while local consistency required for customer support related information. Further examining propagation in website pair also revealed

coupling in websites with high coupling for 'corporate related information' and decrease in coupling as the scales is sharing reduces to local. Implication of such finding is the guidelines on setting priority while enforcing content consistency where high coupling means higher priority for content consistency which is needed for 'corporate related information'. The guidelines are useful in dealing with global and local inconsistency in shared content while knowledge sharing among communities.

**(3) Guidelines on content consistency in knowledge sharing among communities with geographic regions.**

Though several content categories are offered in country-specific websites those websites also represent several geographic regions such as Europe, Asia Pacific, North America and so on. Regional discrepancies are seen to occur from inconsistent content shared among communities from several geographic regions, for example inconsistent content in product usage or specification for customer in Asia Pacific and North America. To deal with regional discrepancies, the contribution is made towards understanding the underlying preferences among communities in sharing within or beyond geographic regions with analytical studies. Traits such as coupling and scales in sharing that vary for specific geographic region and specific content categories are revealed which showed specific preferences among communities. A total of 240 comparisons of webpages within region and 1440 comparisons among region are performed.

Examining propagation within geographic regions revealed high coupling in websites among countries in Europe and low coupling in websites inside North America which revealed that websites in Europe tend to more dependent and prefer to share most content in comparison to websites in North America. Websites inside North America tend to be autonomous and prefer to participate less in sharing. This raised an important concern as guidelines that among all regions customer inside European region is more vulnerable to intra-regional discrepancy from inconsistency in shared content and content consistency is required to be strictly enforced when sharing

content in European region. However examining propagation among geographic regions further supported the autonomous nature of websites in North America showing preferences for less interaction with websites from other region. Additionally websites from Asia Pacific, Europe and Middle-East Africa were found to prefer sharing most content with each other, a concern for inter-regional discrepancy. Further inspection of specific content categories also revealed tendencies for sharing 'corporate related information' globally in all regions while information related to 'product and customer support' within specific region. Interestingly it is found that websites in North America prefer specialized 'product related information' not shared with other region while 'customer support related information' are specialized inside all regions and not shared. Implication of such findings is the guidelines on content consistency customized for specific region.

**(4) Support for content consistency in customized knowledge sharing among communities.**

The analytical results have already depicted restriction in propagation while sharing specific content categories and to specific geographic regions due to community preferences. Inconsistency in content shared among websites or cross-site content is also found to occur globally or locally and even regional discrepancy is possible in cross-site content. To support content consistency allowing community preferences in customized knowledge sharing, the contribution is made with a technique that is based solely on the concept of propagating content updates restricted to specific languages or specific community. From generalizing cases of propagation, pattern of sharing (a) Internationalization (b) Regionalization and (c) Localization with rules for restricting the publication and description of content to specific languages or community is employed for the delivery of knowledge. Integrating pattern of sharing as community preferences content inconsistencies are dealt from scaling the suitability of content to global, regional or local communities, propagation of content updates confined to specific community and supporting content consistency globally in several languages or limited to

languages for specific community. The contribution from the proposed techniques is also its simplicity and the possibility to apply automatically from identifying specific content categories or manually as policies.

With the contribution from techniques and guidelines in dealing with inconsistency in multilingual content and cross-site content for the design of multi-language knowledge sharing system additional contributions are made by supporting resource deprived communities in knowledge sharing.

**(5) Support for content consistency without reliance on language processing.**

Inconsistencies in multilingual content are tackled in previous researches mostly with techniques from language generation to language processing which limits their suitability to resource poor languages. The problem surfacing limited support to resource poor languages is due to (a) dependence on language processing (b) necessity for massive linguistic corpuses in training systems which is unfortunately not available for resource deprived communities. The essence of multi-language knowledge sharing is not truly achieved when communities with limited language resources are not involved in knowledge sharing. To support content consistency in variety of languages, the techniques with no reliance on language processing is contributed from this thesis.

First the techniques to detect inconsistency in multilingual content while leveraging knowledge equally is based on user editing activities and secondly the techniques to support content consistency in cross-site content for customized knowledge sharing is based on restriction in propagating content updates. Both techniques do not depend on language processing making them applicable to variety of languages. The advantage is the support for knowledge sharing from resource rich to resource poor languages or vice versa. The contribution is also towards encouraging participation of resource deprived communities in knowledge sharing.

**(6) Simplicity in Technique for content consistency.**

To support the notion of "multi-language" by enabling content consistency among communities that participate in a common language and different languages, the proposed techniques catered to monolingual and multilingual cases. The techniques for content consistency are based on user editing activity and propagation of content updates which is not targeted to specific language and simple to integrate with translation.

From the techniques that are simple and applicable in variety of languages along with the guidelines for content consistency to deal with inconsistency in multilingual content, global and local inconsistency as well as regional discrepancy in cross-site content; this thesis contributed in the design of multi-language knowledge sharing system catered to knowledge sharing goals of communities both in leveraging knowledge equally and customization in knowledge sharing.

## 7.2 Future Directions

In the light of supporting multi-language knowledge sharing following future directions are suggested.

- *Extending analysis from increasing data sample to cover more content categories and geographic regions.*

  We aim to extend the analysis in Chapter 4 and 5 by increasing the data samples such as the number of global brands from various industrial sectors. We also plan to increase the number of country-specific websites and geographic regions. Particularly, the quantity of country-specific websites within each region can be increased to generalize the findings for all countries. Further content categories can also be increased to cover analysis on preferences in sharing for more information such as employment, financial, social responsibility and so on.

- *Examining propagation by replacing manual comparison of webpages with existing content analysis techniques.*

  We aim to apply content analysis techniques such as text mining or lexical mining for extracting geographic references [Quercini et al., 2010] to compare the presence of same content among websites and examine propagation in Chapter 4 and 5. We can apply such technique to large set of data samples and base our analysis on statistical results.

- *Extending analysis on propagation from single language to distinct language pair.*

  Country-specific websites that offer content in English language are used for examining propagation in Chapter 4 and 5. We aim to examine propagation in distinct language pair among websites so that we can determine preference in sharing specific content categories to specific region with respect to different languages.

# Bibliography

[Adar et al., 2009] Adar, E., Skinner, M., and Weld, D. S. (2009). Information arbitrage across multi-lingual wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 94–103. ACM.

[Adenfelt and Lagerström, 2006] Adenfelt, M. and Lagerström, K. (2006). Knowledge development and sharing in multinational corporations: The case of a centre of excellence and a transnational team. *International Business Review*, 15(4):381–400.

[Al Assimi and Boitet, 2001] Al Assimi, A.-B. and Boitet, C. (2001). Management of non-centralized evolution of parallel multilingual documents. In *Proc. Internationalization Track, 10th International World Wide Web Conference, Hong Kong*.

[Almeida and Phene, 2004] Almeida, P. and Phene, A. (2004). Subsidiaries and knowledge creation: The influence of the mnc and host country on innovation. *Strategic Management Journal*, 25(8-9):847–864.

[Ambos and Ambos, 2009] Ambos, T. C. and Ambos, B. (2009). The impact of distance on knowledge transfer effectiveness in multinational corporations. *Journal of International Management*, 15(1):1–14.

[Ardichvili et al., 2006] Ardichvili, A., Maurer, M., Li, W., Wentling, T., and Stuedemann, R. (2006). Cultural influences on knowledge sharing through online communities of practice. *Journal of knowledge management*, 10(1):94–107.

[Argenti and Druckenmiller, 2004] Argenti, P. A. and Druckenmiller, B. (2004). Reputation and the corporate brand. *Corporate Reputation Review*, 6(4):368–374.

[Bao et al., 2012] Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., and Gergle, D. (2012). Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084. ACM.

[Barnes and Vidgen, 2002] Barnes, S. J. and Vidgen, R. T. (2002). An integrative approach to the assessment of e-commerce quality. *J. Electron. Commerce Res.*, 3(3):114–127.

[Bey et al., 2006] Bey, Y., Boitet, C., and Kageura, K. (2006). The transbey prototype: an online collaborative wiki-based cat environment for volunteer translators. In *LREC-2006: Fifth International Conference on Language Resources and Evaluation. Third International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III)*, pages 49–54.

[Bouayad-Agha et al., 2002] Bouayad-Agha, N., Power, R., Scott, D., and Belz, A. (2002). Pills: Multilingual generation of medical information documents with overlapping content. In *LREC*.

[Brodie and Mylopoulos, 2012] Brodie, M. L. and Mylopoulos, J. (2012). *On knowledge base management systems: integrating artificial intelligence and database technologies*. Springer Science & Business Media.

[Bronner et al., 2012] Bronner, A., Negri, M., Mehdad, Y., Fahrni, A., and Monz, C. (2012). Cosyne: Synchronizing multilingual wiki content. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, page 33. ACM.

[Chakraborty et al., 2005] Chakraborty, G., Srivastava, P., and Warren, D. L. (2005). Understanding corporate b2b web sites' effectiveness from

north american and european perspective. *Industrial Marketing Management*, 34(5):420–429.

[Climent et al., 2003] Climent, S., Moré, J., Oliver, A., Salvatierra, M., Sànchez, I., Taulé, M., and Vallmanya, L. (2003). Bilingual newsgroups in catalonia: A challenge for machine translation. *Journal of Computer-Mediated Communication*, 9(1):0–0.

[Cormican and Dooley, 2007] Cormican, K. and Dooley, L. (2007). Knowledge sharing in a collaborative networked environment. *Journal of Information & Knowledge Management*, 6(02):105–114.

[Daryanto et al., 2013] Daryanto, A., Khan, H., Matlay, H., and Chakrabarti, R. (2013). Adoption of country-specific business websites: The case of uk small businesses entering the chinese market. *Journal of Small Business and Enterprise Development*, 20(3):650–660.

[Désilets et al., 2006] Désilets, A., Gonzalez, L., Paquet, S., and Stojanovic, M. (2006). Translation the wiki way.

[Easterbrrok and Nuseibeh, 1996] Easterbrrok, S. and Nuseibeh, B. (1996). Using viewpoints for inconsistency management. *Software Engineering Journal*, 11(1):31–43.

[Esselink, 2000] Esselink, B. (2000). *A practical guide to localization*, volume 4. John Benjamins Publishing.

[Faigley and Witte, 1981] Faigley, L. and Witte, S. (1981). Analyzing revision. *College composition and communication*, pages 400–414.

[Fletcher, 2006] Fletcher, R. (2006). The impact of culture on web site content, design, and structure: An international and a multicultural perspective. *Journal of communication management*, 10(3):259–273.

[Fong Boh et al., 2013] Fong Boh, W., Nguyen, T. T., and Xu, Y. (2013). Knowledge transfer across dissimilar cultures. *Journal of Knowledge Management*, 17(1):29–46.

[Ghemawat, 2001] Ghemawat, P. (2001). Distance still matters. *Harvard business review*, 79(8):137–147.

[Hajlaoui and Boitet, 2005] Hajlaoui, N. and Boitet, C. (2005). A" pivot" xml-based architecture for multilingual, multiversion documents: parallel monolingual documents aligned through a central correspondence descriptor and possible use of unl. *Research on Computing Science*, 12:309–326.

[Hall, 1997] Hall, S. (1997). The local and the global: Globalization and ethnicity. *Cultural politics*, 11:173–187.

[Halliburton and Ziegfeld, 2009] Halliburton, C. and Ziegfeld, A. (2009). How do major european companies communicate their corporate identity across countries?-an empirical investigation of corporate internet communications. *Journal of Marketing Management*, 25(9-10):909–925.

[Hartley and Paris, 1997] Hartley, A. and Paris, C. (1997). Multilingual document production from support for translating to support for authoring. *Machine Translation*, 12(1-2):109–129.

[He, 2001] He, S. (2001). Interplay of language and culture in global e-commerce: a comparison of five companies' multilingual websites. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 83–88. ACM.

[Hillier, 2003] Hillier, M. (2003). The role of cultural context in multilingual website usability. *Electronic Commerce Research and Applications*, 2(1):2–14.

[Hofstede and Hofstede, 2001] Hofstede, G. H. and Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage.

[Hopcroft, 1979] Hopcroft, J. E. (1979). *Introduction to automata theory, languages, and computation*. Pearson Education India.

[Huberdeau et al., 2008] Huberdeau, L.-P., Paquet, S., and Désilets, A. (2008). The cross-lingual wiki engine: enabling collaboration across language barriers. In *Proceedings of the 4th International Symposium on Wikis*, page 13. ACM.

[Huizingh, 2000] Huizingh, E. K. (2000). The content and design of web sites: an empirical study. *Information & Management*, 37(3):123–134.

[Inaba et al., 2007] Inaba, R., Murakami, Y., Nadamoto, A., and Ishida, T. (2007). Multilingual communication support using the language grid. In *Intercultural Collaboration*, pages 118–132. Springer.

[Ishida, 2011] Ishida, T. (2011). *The language grid: Service-oriented collective intelligence for language resource interoperability*. Springer Science & Business Media.

[Jones, 2008] Jones, J. (2008). Patterns of revision in online writing a study of wikipedia's featured articles. *Written Communication*, 25(2):262–289.

[Kale, 1991] Kale, S. H. (1991). Culture-specific marketing communications: An analytical approach. *International Marketing Review*, 8(2).

[Kayan et al., 2006] Kayan, S., Fussell, S. R., and Setlock, L. D. (2006). Cultural differences in the use of instant messaging in asia and north america. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 525–528. ACM.

[Kumaran et al., 2008] Kumaran, A., Saravanan, K., and Maurice, S. (2008). wikibabel: community creation of multilingual data. In *Proceedings of the 4th International Symposium on Wikis*, page 14. ACM.

[Kussmaul and Jack, 2009] Kussmaul, C. and Jack, R. (2009). Wikis for knowledge management: Business cases, best practices, promises, & pitfalls. In *Web 2.0*, pages 1–19. Springer.

[LionBridge, 2009] LionBridge (2009). Buidling a global web strategy best practices for developing your international online brand.

[Lynch and Beck, 2001] Lynch, P. D. and Beck, J. C. (2001). Profiles of internet buyers in 20 countries: Evidence for region-specific strategies. *Journal of International Business Studies*, 32(4):725–748.

[Main, 2001] Main, L. (2001). The global information infrastructure: Empowerment or imperialism? *Third World Quarterly*, 22(1):83–97.

[Maynard and Tian, 2004] Maynard, M. and Tian, Y. (2004). Between global and glocal: Content analysis of the chinese web sites of the 100 top global brands. *Public Relations Review*, 30(3):285–291.

[Nachum and Zaheer, 2005] Nachum, L. and Zaheer, S. (2005). The persistence of distance? the impact of technology on mne motivations for foreign investment. *Strategic Management Journal*, 26(8):747–767.

[Nguyen, 2007] Nguyen, N. T. (2007). *Advanced methods for inconsistent knowledge management*. Springer Science & Business Media.

[Nuseibeh et al., 2001] Nuseibeh, B., Easterbrook, S., and Russo, A. (2001). Making inconsistency respectable in software development. *Journal of Systems and Software*, 58(2):171–180.

[Okazaki and Alonso Rivas, 2002] Okazaki, S. and Alonso Rivas, J. (2002). A content analysis of multinationals' web communication strategies: Cross-cultural research framework and pre-testing. *Internet Research*, 12(5):380–390.

[OLeary, 2009a] OLeary, D. (2009a). Multilingual knowledge management. In Bramer, M., editor, *Artificial Intelligence An International Perspective*, volume 5640 of *Lecture Notes in Computer Science*, pages 133–156. Springer Berlin Heidelberg.

[O'Leary, 2008] O'Leary, D. E. (2008). A multilingual knowledge management system: A case study of fao and waicent. *Decision Support Systems*, 45(3):641–661.

[OLeary, 2009b] OLeary, D. E. (2009b). Wikis:from each according to his knowledge. *Online Communication and Collaboration: A Reader*, page 89.

[Olson et al., 2005] Olson, J. S., Grudin, J., and Horvitz, E. (2005). A study of preferences for sharing and privacy. In *CHI'05 extended abstracts on Human factors in computing systems*, pages 1985–1988. ACM.

[Palmer, 2002] Palmer, J. W. (2002). Web site usability, design, and performance metrics. *Information systems research*, 13(2):151–167.

[Paris and Vander Linden, 1996] Paris, C. and Vander Linden, K. (1996). An interactive support tool for writing multilingual manuals. *Computer*, 29(7):49–56.

[Power et al., 2003] Power, R., Scott, D., and Hartley, A. (2003). Multilingual generation of controlled languages.

[Quercini et al., 2010] Quercini, G., Samet, H., Sankaranarayanan, J., and Lieberman, M. D. (2010). Determining the spatial reader scopes of news sources using local lexicons. In *proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 43–52. ACM.

[Robbins and Stylianou, 2003] Robbins, S. S. and Stylianou, A. C. (2003). Global corporate web sites: an empirical investigation of content and design. *Information & Management*, 40(3):205–212.

[Rugman and Verbeke, 2004] Rugman, A. M. and Verbeke, A. (2004). A perspective on regional and global strategies of multinational enterprises. *Journal of International Business Studies*, 35(1):3–18.

[Scott and Evans, 1998] Scott, D. and Evans, R. (1998). Multilingual document management without translation using natural language generation in the multilingual information society. In *Speech and Language*

*Engineering-State of the Art (Ref. No. 1998/499), IEE Colloquium on*, pages 9–1. IET.

[Shin and Huh, 2009] Shin, W. and Huh, J. (2009). Multinational corporate website strategies and influencing factors: A comparison of us and korean corporate websites. *Journal of Marketing Communications*, 15(5):287–310.

[Singh, 2011] Singh, N. (2011). *Localization strategies for global e-business*. Cambridge University Press.

[Singh et al., 2005] Singh, N., Kumar, V., and Baack, D. (2005). Adaptation of cultural content: evidence from b2c e-commerce firms. *European Journal of Marketing*, 39(1/2):71–86.

[Singh and Pereira, 2005] Singh, N. and Pereira, A. (2005). *The culturally customized web site*. Routledge.

[Singh et al., 2009] Singh, N., Toy, D. R., and Wright, L. K. (2009). A diagnostic framework for measuring web-site localization. *Thunderbird International Business Review*, 51(3):281–295.

[Sommers, 1980] Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College composition and communication*, pages 378–388.

[Sousa et al., 2010] Sousa, F., Aparicio, M., and Costa, C. J. (2010). Organizational wiki as a knowledge management tool. In *Proceedings of the 28th ACM International Conference on Design of Communication*, pages 33–39. ACM.

[Sumathi and Esakkirajan, 2007] Sumathi, S. and Esakkirajan, S. (2007). *Fundamentals of relational database management systems*, volume 47. Springer.

[Sun, 2001] Sun, H. (2001). Building a culturally-competent corporate web site: an exploratory study of cultural markers in multilingual web design. In *Proceedings of the 19th annual international conference on Computer documentation*, pages 95–102. ACM.

[Svensson, 2001] Svensson, G. (2001). glocalization of business activities: a glocal strategy approach. *Management decision*, 39(1):6–18.

[Tanaka et al., 2011] Tanaka, R., Murakami, Y., and Ishida, T. (2011). Cascading translation services. In *The Language Grid*, pages 103–117. Springer.

[Tixier, 2005] Tixier, M. (2005). Globalization and localization of contents: Evolution of major internet sites across sectors of industry. *Thunderbird International Business Review*, 47(1):15–48.

[Tonella et al., 2002] Tonella, P., Ricca, F., Pianta, E., and Girardi, C. (2002). Restructuring multilingual web sites. In *Software Maintenance, 2002. Proceedings. International Conference on*, pages 290–299. IEEE.

[Tonella et al., 2006] Tonella, P., Ricca, F., Pianta, E., and Girardi, C. (2006). Automatic support for the alignment of multilingual web sites. *Journal of Software Maintenance and Evolution: Research and Practice*, 18(3):153–179.

[Traicu and Prostean, 2012] Traicu, M. H. and Prostean, G. (2012). A model of translation management systems for multilingual documents. In *Applied Computational Intelligence and Informatics (SACI), 2012 7th IEEE International Symposium on*, pages 115–118. IEEE.

[Vrontis et al., 2012] Vrontis, D., Shoham, A., and Shneor, R. (2012). Influences of culture, geography and infrastructure on website localization decisions. *Cross Cultural Management: An International Journal*, 19(3):352–374.

[Wagner, 2004] Wagner, C. (2004). Wiki: A technology for conversational knowledge management and group collaboration. *The Communications of the Association for Information Systems*, 13(1):58.

[Wiese et al., 2011] Wiese, J., Kelley, P. G., Cranor, L. F., Dabbish, L., Hong, J. I., and Zimmerman, J. (2011). Are you close with me? are you nearby?: investigating social groups, closeness, and willingness to share. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 197–206. ACM.

[Yamashita and Ishida, 2006] Yamashita, N. and Ishida, T. (2006). Effects of machine translation on collaborative work. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 515–524. ACM.

[Yunker, 2002] Yunker, J. (2002). *Beyond borders: Web globalization strategies*. New Riders.

[Yunker, 2014] Yunker, J. (2014). The 2014 web globalization report card.

# Publications

## Major Publications

### Journals

1. **Amit Pariyar**, Yohei Murakami, Donghui Lin, and Toru Ishida, "Inconsistency Detection in Multilingual Knowledge Sharing," *Journal of Information and Knowledge Management*. Vol. 13-3, December 2014.

### International Conferences

1. **Amit Pariyar**, Donghui Lin and Toru Ishida, " Tracking Inconsistencies in Parallel Multilingual Documents," *In Proceedings of 2013 International Conference on Culture and Computing (Culture and Computing 2013)*, pp.15-20. IEEE, September 2013.

2. **Amit Pariyar**, Yohei Murakami, Donghui Lin and Toru Ishida, "Content Sharing in Global Organization: A Cross-Country Perspective," *The Third ASE International Conference on Social Informatics (SocialInformatics 2014)*, Cambridge, USA. ASE, December 2014.

3. **Amit Pariyar**, Yohei Murakami, Donghui Lin and Toru Ishida, " Content Sharing in Global Brand from Geographic Perspective," *International Conference on Culture and Computing (Culture and Computing 2015)*. IEEE, October 2015. (**Best Paper Award Nominated**)

## Workshop

1. **Amit Pariyar**, Donghui Lin, and Toru Ishida, "Consistency Management in Parallel Multilingual Documents," *2013 Information Processing Society of Japan*. IPSJ, March, 2013.