

Using Content Development Guidelines to Reduce the Cost of Localising Digital

Lorcan Ryan

Localisation Research Centre,
CSIS Department,
University of Limerick,
Limerick,
lorcan.ryan@ul.ie

Dimitra Anistasiou

Localisation Research Centre,
CSIS Department,
University of Limerick,
Limerick,
lorcan.ryan@ul.ie

Yvonne Cleary

Department of Languages and Cultural Studies,
University of Limerick,
Limerick,
yvonne.cleary@ul.ie

Abstract - *This article examines how content development guidelines can reduce the cost of localising digital content. The growth of digital content is examined initially, and a basic taxonomy of enterprise and personal content is described. The demand for localised content is explained next, from international audiences, who demand content customised for their own particular locales. The costs, as well as cost-reducing strategies, involved in localisation process are also described. The cost-reducing strategy of internationalisation is focused on in particular, with the three core processes of authoring, enabling and testing explained in detail. The article then describes how a Web 2.0 system called the Localisation Knowledge Repository (LKR) will be used to integrate content development guidelines into the internationalisation process. Finally, the benefits of the LKR are explained, including how the system facilitates the production of content for global audiences that is cheaper to translatable and requires less localisation testing.*

Keywords: Authoring, content development, digital content, internationalisation, localisation, pre-translation testing, technical communication, quality assurance, Web 2.0

1 Traditional and Digital Content

Content is a term that may be used to describe anything from a book, painting or video to an email, webpage or video game. Since the earliest cave paintings of the Stone Age, man has recorded ideas, information and opinions in a variety of physical written and illustrative media. Words, symbols and pictures have traditionally been captured on stone tablets, clay tokens, papyrus, vellum, canvas and paper over the years, and distributed as scrolls, magazines, books and paintings. Advances in technology, such as the invention of the earliest camera in the 1660s and first modern analog computer in 1930 (Encyclopaedia Britannica 2009), made it possible to store and communicate written and illustrative content in new formats such as photographic film, microfilm and magnetic tape. Traditional content is the term we use to refer to words, graphics, music or video published in physical, non-digital formats such as canvas, paper or microfilm.

Table 1: Traditional Content

Most content was published exclusively to these formats until the advent of the computer age, or information era, beginning in the 1980's.

1.1 What is Digital Content?

During 20th century, the invention of the computer led to a new method of capturing data in computer files. We refer to information stored in this way as digital content. Digital content is content is stored on hard drives or external storage media, published as a computer file or online, and accessed via a hardware device such as a computer, games console or mobile phone. It is viewed on display media such as computer monitors, televisions, mobile phone screens and eReaders, and shared via communication technologies such as the World Wide Web, email and Short Message Service (SMS).

A significant amount of content is initially developed using software development, desktop publishing, help authoring, web design or graphic design software, and published as computer files. Any content published as a computer file is regarded as digital content, regardless of whether it is printed at a later stage or not. A written document such as a software user guide, for example, may be published initially as a .DOC, .PDF or .HTML file, and uploaded to a support website for customers to access. However, the same user guide may also be printed for packaging with the software product CD-ROM. Indeed, people not familiar with digital content may prefer to read certain materials print format, such as books, photographs and timetables.

Table 2: Examples of Digital Content

1.2 Advantages of Digital Content

Traditional content is often converted to digital files with scanning equipment or conversion software; photographs, for example, may be scanned and saved as digital image files. Legacy content stored in physical media is generally digitised to avail of some of the following advantages associated with digital content (TidWiT 2009):

- i. Storage: Digital content uses binary digits as the basic unit of information storage, rather than physical material such as paper or canvas. This allows digital content to be stored in massive quantities, with a palm-sized hard disk able to store tens of millions of pages of digital content. Digital content is also durable, and does not usually degrade as quickly as traditional content (although some digital storage media may decompose over time).
- ii. Classification: It can be easier to access and retrieve specific information in digital content than its traditional counterpart. Search facilities may be used to locate data in digital content; with hyperlinks also helping users to navigate to relevant chunks of information. Digital content is easy to restructure if necessary, and metadata may also be added to make it easier to find and retrieve information. Metadata also reduces the storage requirement for digital content, as the same file does not have to be stored in multiple locations just because it refers to several different topics.
- iii. Accessibility: The connectivity afforded by the internet makes digital information much easier for publishers to distribute and for users to access. Content developers, for example, may now publish content online or email it directly to target audiences, rather than having to produce printed material. Digital content may also be made more accessible for disabled individuals by utilising assistive technologies such as screen reader software, Braille terminals, screen magnification software and speech processing applications.
- iv. Publishing & Reproduction Costs: The cost of publishing and reproducing digital content is much lower for digital content than its traditional content. Companies may create an electronic brochure and distribute it to thousands of customers simultaneously via email at relatively little time and cost. The same brochures would previously need to be printed, packed into envelopes and posted to each individual customer.

1.3 Factors Influencing the Growth of Digital Content

The volume of digital content produced today is significantly higher than ever before due to several factors, including the emergence of new electronic media, digitising of legacy content, user-generated content and corporate strategy. Each of these factors will be described in the following subsections.

1.3.1 Emergence of New Electronic Media

After the initial rise in popularity of the personal computer in the early 1990s, electronic media have evolved into sophisticated handheld devices such as internet-enabled mobile phones, PDAs, eReaders and portable DVD players. A significant portion of content is now published in both its traditional format and a multitude of additional digital

formats for new electronic media. Books, for example, are generally published as both hardback and paperback printed copies; but may also be published as e-books, audio books on CD or MP3 and so on.

1.3.2 Digitising of Legacy Content

A second factor for the huge increase in content production is the digitising of previously created content. Legacy content is usually converted to digital files with scanning equipment or conversion software; photographs, for example, may be scanned and saved as digital image files.

1.3.3 User-generated Content

The widespread adoption of the internet and the growth of Web 2.0 applications have created a new trend of users, rather than enterprises, publishing digital content. While digital content has traditionally been developed by professionals such as software developers and technical writers; millions of internet users are now creating and distributing digital information on a daily basis via instant messages, emails, forums and blogs. This combination of professional and social community publishers means that the volume of digital content is continuing to increase at a considerable rate.

1.3.4 Corporate Strategy

Enterprises often prefer developing and distributing digital rather than traditional content. Printing costs, for example, are reduced or eliminated by publishing technical support documentation online, rather than as paper-based product guides or user manuals. Distribution and packaging costs are also decreased by delivering software and other content to customers electronically, rather than sending them physical products. Digital content may also be distributed to users in a fraction of the time it takes to distribute traditional content.

1.4 The Growth Rate of Digital Content

If the world's rapidly expanding digital content was printed and bound into books it would form a stack that would stretch from Earth to Pluto 10 times (The Guardian 2009). As more people join the digital universe – through online access and internet-enabled mobile phones – the world's digital output is increasing at such a rate that those stacks of books are rising quicker than NASA's fastest space rocket (The Guardian 2009). The widespread adoption of home computers and usage of the World Wide Web in the 1990s accounted for the initial growth in production of digital content. This led to a boom in the computer hardware and software industry. According to the industry analyst INPUT, the total expenditure on software products in the US rose from \$250 million in 1970, to \$58 billion in 1995, to over \$100 billion in 2000. After recovering from the “dotcom bust” of 2001, the domestic demand for packaged software in the US exceeded \$126 billion in 2007 (International Trade Administration 2009).

Table3: Total Expenditure on Software in the US

Distribution of digital music also experienced rapid growth from the late 1990's on, with music downloads beginning to replace that of older media formats such as vinyl and cassette. The ratio of digital to analog sales in 2004 was roughly 1:99, but by 2007 it was roughly 1:9 (Recording Industry Association of America 2009). In 2008, physical album sales fell 20 percent to 362.6 million from 450.5 million, while digital album sales rose 32 percent to a record 65.8 million units (Tahoe Daily Tribune 2009). Another indicator of the growth in digital music is the number of digital music file formats, which increased from less than 10 in 2003 to over 100 in 2007.

The amount of websites and the volume of online documentation has increased steadily in the last 15 years. There were just 18,000 websites when internet monitoring company Netcraft began keeping track in August of 1995. By June 2009 however, there were over 70 million active websites (Netcraft 2009) as shown in Figure 1.

Figure1: Total Sites Across All Domains August 1995–June 2009(<http://news.netcraft.com>)

As of May 2009, the world's digital content is estimated to be 487 billion gigabytes. The digital universe is expected to double in size over the next 18 months, according to the latest research from technology consultancy IDC and sponsored by IT firm EMC, fuelled by a rise in the number of mobile phones (The Guardian 2009).

2 Enterprise & Personal Digital Content

Both digital and traditional content may be sub-divided into further categories, the following section examines how digital content may be classified as either enterprise or personal according to who has published it.

2.1 Enterprise Content

Enterprise content is usually developed by professionals such as software developers, help authors, web designers, technical writers, graphic designers and technical engineers for commercial purposes. Software products, including desktop applications, firmware, video games, operating system software, business packages and software development kits are a significant component of enterprise content. The top five software vendors in 2008 were Microsoft, IBM, Oracle, HP, SAP and Symantec (Software Top 100 2009).

The quality of enterprise content is usually very high, as poorly written content could create negative perceptions in customers. Enterprise content, therefore, is usually:

- Published by organisations such as Sony Music Entertainment, EMI, Universal Music Group and Warner Music Group (digital music), Fox Entertainment, Paramount Motion Pictures Group, Sony Pictures Entertainment and Time Warner (digital video), Google, MSN and Reuters (online documentation)
- High volumes published by relatively small number of content development professionals
- For commercial purposes
- Developed by professional authors
- Predictable content
- Static
- Published with professional tools such as software development kits, help authoring tools, web design packages and word processing solutions

2.2 Personal Content

Approximately 70% of the information in the digital universe is created by individuals rather than companies, and includes emails, photos, online banking transactions or postings on social networking sites (The Guardian 2009). This type of digital content, developed for social, non-commercial reasons, is known as personal content. Web 2.0 applications such as Google Wave, Blogger.com, YouTube, Facebook, World of Warcraft and Twitter (GoToWeb2.0 Web Applications Index 2009) enable users to instantly create, publish and share digital information via email, instant messaging, forums, blogs, social networking and massively multiplayer online games (MMOGs). Examples of users generating digital content include uploading videos to YouTube, publishing guitar tab on TabCrawler, sharing photographs on Picasa, sending emails via Gmail, posting blogs on Blogger.com, stating opinions on a personal website and creating “mod” games.

Table 4: Taxonomy of Digital Content

Personal content requires less quality control than enterprise content, as it is not published for commercial reasons and therefore doesn't normally undergo the rigorous testing and QA associated with enterprise content. Publishers of personal content may also use their own terminology and abbreviations. Online gamers, for example, use terms such as noob (a “newbie” or inexperienced gamer), leet (a sarcastic term referring to an “elite” player) and frag (to kill a computer game character), while SMS and instant messaging (IM) users regularly use abbreviations such as OMG (oh my god), GR8 (great) and LOL (laugh out loud).

Personal content, therefore, is usually:

- Developed and published by individuals
- Small in volume, published by a huge number of individual users
- For social or personal purposes
- Developed by social users
- Non-predictable, unique content
- Dynamic
- Usually published online

3 Localisation

Localisation is the process of adapting products, services and associated documentation so that they are understandable, acceptable and functional in target locales. Content is made understandable in locales by accurately translating it, acceptable by taking cultural differences into account, and functional by post-translation testing and editing. Inaccurate translations may cause confusion, lack of cultural sensitivity may cause offence, and content malfunctions may cause user frustration. Localisation should, therefore, encompass cultural awareness and technical expertise as well as the core activity of translation.

3.1 Why Localise?

Translation of traditional content has been happening ever since the Hebrew Bible was translated into a Koine Greek version called the Septuagint between the 3rd and 1st centuries BC (Jobes and Silva 2001). Large scale translation of digital content, however, only began in the 1980s. Multinational corporations sought to increase international sales revenue by exporting translated software and documentation, but quickly realised that cultural and technical, as well as linguistic, issues required special attention. Therefore, to remain competitive in the global economy, organisations modified their exported products and services to give them the look and feel of locally-made products. This strategy is called localisation, and surveys show that nine out of ten businesses prefer to purchase products that have been adapted to their own language and market needs (Common Sense Advisory 2006).

Although digital content was initially translated by enterprises for business reasons, some organisations translate it for informative rather than commercial reasons. Parliaments, governments, councils and local authorities for example, translate digital documentation related to taxation and voting. The European Union (EU) produces legislation and documents of major public importance or interest in all 23 official EU languages (Europa Languages and Europe 2009). NGOs (not-for-profit organisations) including charities, foundations, social enterprises and humanitarian movements may also translate content to generate public awareness or collect donations for charitable causes. The World Health Organization (WHO) for example, publishes multilingual digital content with the aim of educating people about disease and promoting the general health of the world's population. Translators Without Borders also provide free translations to humanitarian organisations.

Internet users may also be motivated to volunteer translations for social reasons, such as prestige or the desire to share digital content with other users who speak their language. One Spanish user of Facebook for example, was responsible for translating almost 3 percent of the entire site for no other purpose than making it more accessible to other Spanish users (Facebook 2009). Facebook is also being translated into over 60 less widely spoken languages by its users, including Esperanto, Welsh and Afrikaans (Coyle 2009).

3.2 The Challenges of Localisation

There are several challenges involved in localising digital content; three of the most important can be classified under the headings of volume, cost, time and quality. The sheer volume of digital content, both enterprise and personal, coupled with the fact that 304 world languages have a million or more speakers (Ethnologue Languages of the World 2009), makes localising all of it a daunting prospect (Freij 2009). To furnish an example, the help system alone for Microsoft Office 2003 consists of over 700,000 words.

Commercial localisation is usually an expensive process involving investment in professionals and processes. Technologies such as machine translation systems also aim to reduce the cost of localisation projects in the long run, but may require a significant initial investment to install and maintain. Localising personal content may be less expensive if crowdsourcing is used; i.e. the content translated by users for no fee.

Multinational corporations often follow a strategy called "SimShip," or simultaneous launch and shipping of multilingual versions of their products in numerous locales. This puts enormous pressure on them to have localised versions of products ready in time for the source language market release. Some enterprise content may also require real-time translations such as live technical support or television subtitling. Users may request instant localisation of online communications with users in different locales such as emails, instant messages, forums, blogs and online gaming messages.

Enterprise content is usually developed for commercial reasons and either sold directly to customers (software products, digital music and so on) or produced to support the product offering (user assistance, marketing communications, legal documentation and so). High quality enterprise content, in both its original version and localised varieties, is essential therefore to maintain the image of the organisation (and product) with customers. Personal content is generally developed by users for social reasons, so the translation quality level is not as important.

3.3 Factors Affecting the Cost of Localisation

The cost of localising digital content is influenced by a number of factors, including the project scope, type of content and specific files involved. The scope of a localisation project is usually the most significant indicator of its cost. Scope can be measured in a number of ways including word counts, string counts, number of files and number of languages.

Enterprise content is typically subject to high quality translation, engineering and testing by localisation professionals. These types of projects usually involve significant levels of investment in the appropriate processes, professionals and technologies. Translating a simple message on a social networking website on the other hand, where quality is not an important consideration, is generally an inexpensive activity.

Some file types are more difficult to localise than others; flash files, for example, are time-consuming to prepare for translation. Image files with extensions such as JPEG and GIF may need to be sent to a graphic designer for localisation, as translator may not have the technical expertise to edit such files. These types of files add cost to localisation projects as they may be time-consuming or require additional specialised professionals to translate

3.4 The Cost of Localisation

While the main expense involved in localising user-generated content is usually confined to the cost of translating it, large-scale enterprise content localisation projects may incur several different types of costs related to localisation processes, professionals and technologies. Planning, coordinating and implementing processes such as translation, quality assurance and project management may require significant investment of resources.

Localisation projects may require the skills of a diverse range of professionals from translators, technical writers, software developers and help authors to localisation engineers, proofreaders, testers and project managers. Hiring and coordinating these professionals represents another significant cost in commercial localisation projects.

Another cost involved in localising enterprise content is investment in localisation technologies. Tools such as content management systems, desktop publishing applications, computer-aided translation (CAT) tools, software localisation suites, quality assurance (QA) software and project management programs are all used by organisations to complete localisation projects more productively and effectively.

3.5 Reducing the Cost of Localisation

This article focuses on reducing the costs associated enterprise content localisation projects. These types of localisation projects usually involve digital content publishers sending source language content to a language service provider (or freelance translator) for translation. After proofreading and testing, the localised versions of the digital content are published and distributed to international audiences. Companies attempt to reduce the costs associated with several aspects of enterprise content localisation projects, including the cost of localisation processes, professionals and technologies.

3.5.1 Localisation Processes

Organisations attempt to decrease the cost of the localisation projects by adopting cost-reducing strategies for component tasks such as translation, quality assurance and project management. Enterprises may attempt to reduce the cost of translation by using content development guidelines, best practices and standards to produce high quality digital content, which is easier to translate for both human translators and machine translation systems. Companies may also use technology to reduce the cost of translation. Machine translation systems for example, are used to automatically translate digital content, while translation memory tools recycle previously translated language strings into new projects.

Organisations may adopt a strategy called to internationalisation to reduce the cost of localisation quality assurance (QA). Internationalisation is the process of generalising a product or document so that it can handle multiple languages and cultural conventions without the need for redesign. Internationalisation tactics such as writing for global audiences, enabling content for different locales and pre-translation testing ensure that digital content requires as little post-translation testing and localisation QA as possible. Enterprises may also use technologies such as localisation QA software, desktop publishing applications and software testing tools to automate QA procedures, and therefore reduce the costs associated with the process.

Companies may invest in project management software and workflow tools to reduce the cost of project planning, resource allocation and communication. Project managers may also experiment with new workflows to improve the efficiency and effectiveness of the overall localisation process.

3.5.2 Localisation Professionals

Companies may attempt to reduce the cost of hiring localisation professionals by using freelance translators or crowd-sourcing rather than procuring the more expensive services of language service providers (LSPs) or translation agencies. Localisation technologies such as machine translation systems may also be used to replace the human element and automate certain tasks.

3.5.3 Localisation Technologies

Enterprises attempt to reduce the cost of investing in localisation technologies in three important ways. Firstly, enterprises may choose to purchase inexpensive commercial technologies, even if the cheaper tools do not have as comprehensive a feature set as some of their more expensive alternatives. Localisation technology varies in price from CAT tools such as MemoQ costing less than a €1,000, to sophisticated content management systems which may cost several hundred thousands euros. A second method of reducing the cost of localisation technology is investing in open source rather than commercial tools. Open source software makes the source code available to the general public, and has relaxed or non-existent copyright restrictions. It is usually free to download and use, although costs may be incurred during installation, support and customisation. Finally, some companies simply develop their own proprietary localisation solutions, rather than use either commercial or open source tools. Although the research and development cost involved in this approach is initially quite steep, in the long run the enterprise does not have the expense of product upgrades, training or support contracts.

4 Internationalisation

This article focuses on reducing the cost of localisation processes by implementing internationalisation guidelines using Web 2.0 technology. Internationalisation is the process of generalising a product or document so that it can handle multiple languages and cultural conventions without the need for redesign. LISA (2009) suggests that it consists primarily of abstracting the functionality of a product away from any particular culture, language or market so that support for specific markets and languages can be integrated easily. If a product has not been internationalised in advance, it can take twice as long and cost twice as much to localise, and such added expense may make it uneconomical to localise at all (LISA 2009).

Internationalisation takes place at the level of program design and document, before the translation process (Figure 2). Enterprises should define which regions digital content will be distributed to before implementing the strategy; internationalising content for FIGS (French-Italian-German-Spanish) locales, for example, is significantly different from internationalising content for BRIC (Brazilian-Russian-Indian-Chinese) locales.

Figure 2: The Global Product Development Cycle (Localization Industry Standards Association 2009 <http://www.lisa.org/What-Is-Globalization.48.0.html>)

Internationalisation and localisation are necessary due to variances in different locales, and to ensure that products, services and documentation are understandable and acceptable in different regions regardless of language or culture.

4.1 Diversity in Different Locales

Diversity in different locales can make content (which has not been prepared for translation) extremely costly and time-consuming to localise. We will describe three main types of diversity that distinguish locales; linguistic, cultural and technical diversity.

4.1.1 Linguistic

Linguistic diversity refers to variations in the language and writing conventions used in different locales. Sometimes even a single language may contain multiple regional varieties which should be considered in the content development process. Spanish, for example, with over 400 million speakers worldwide, has regional varieties such as European Spanish, Castilian Spanish, Latin American Spanish, Standard Spanish, International Spanish and Neutral Spanish. Although all of these different forms are more or less understood in different locales, each variant has a unique vocabulary, pronoun usage and tense preference. “Costo,” for example, is the Latin America Spanish term for cost,

whereas in Spain, it refers to hashish in informal speaking (Tek Translation 2009). Spelling variants may also exist for different countries where the same language is spoken; the word “localisation” in English (Ireland) for example, is spelled “localization” in English (U.S.). Another issue authors should consider is the meaning that a particular term in the source language has in other locales. General Motors discovered this to their detriment when promoting their Chevrolet Nova automobile in Spanish-speaking markets – “no va” means “doesn’t go” in Spanish!

Writing conventions for currency, time, dates, weights and measurements may also differ depending on the locale. In France, for example, dates are written in the format day/month/year, but in Germany they are written day.month.year and in the Netherlands they are written day-month-year. In the United States however, dates are written month/day/year, and in Scandinavia (and some parts of Asia) are written year-month-day. The latter is defined by ISO 8601 as the international standard for writing dates (International Organization for Standardization 2009).

Considering issues such as weight measurement units during the content development process is also important. In 1983 an Air Canada Boeing 767 jet nicknamed the “Gimli Glider” completely ran out of fuel at 41,000 feet, halfway through its journey from Montreal to Edmonton. The pilot managed to crash-land the plane and nobody aboard was seriously injured, but an investigation was held immediately to deduce why this near-disaster occurred. The investigation found that the fuel requirement for the flight was set in metric units (20,000kg) while the local flight crew, who were used to calculating in imperial units, filled the plane with an incorrect fuel level (20,000lbs) which was insufficient for the flight (Aviation Safety Network 2009).

In a similar incident in 1998, the Mars Climate Orbiter was lost due to a navigation error when a subcontractor used imperial units (pound-seconds) instead of the metric units (newton-seconds) as specified by NASA (National Aeronautics and Space Administration 2009). Following this incident, NASA reverted back to using imperial units as their only system of measurement and continues to do so.

As well as writing conventions, other linguistic issues to consider are:

- Writing direction (Persian, Hebrew and Arabic are bi-directional languages running from right to left)
- Spacing rules (spaces are not used to separate words in Tibetan)
- Sorting order (if the content contains alphabetically-sorted lists, these will have to be reordered for each language)

4.2.2 Cultural

The culture of a particular locale consists of the shared attitudes, values, goals, and practices that characterise the area. No two locales share identical cultures, and it is important to capture this diversity during the content development process to avoid unintentionally irritating, offending or frustrating users in different locales with inappropriate content. Important cultural aspects that authors should consider are religious beliefs, political attitudes, colour associations, national holidays, sacred symbols, role of the family and so on.

The most famous example of publishing culturally inappropriate content in recent times occurred in 2005 when Danish newspaper Jyllands-Posten included 12 cartoons depicting the Islamic prophet Muhammed. This led to protests from many Muslims who felt the cartoons were Islamophobic, racist and blasphemous to people of the Muslim faith. Although the publishers of the cartoons maintained they were intended to be humorous and did not discriminate against Muslims, anger remained over the offensive images. Consumer boycotts of Danish products were organised, several Danish embassies were attacked and death threats were issued to the illustrators of the cartoons. Although the Danish Prime Minister apologised for any offense caused by the cartoons, tension remained between Denmark and some sections of the Islamic world.

As well as having knowledge of political and religious beliefs, it is also useful for content authors to understand the subtle nuances of target locales such as colour associations. Purple, for example, is associated with royalty and wisdom in western cultures, but represents mourning in some Asian countries. By incorporating cultural awareness into the authoring process, the risk of publishing content that is offensive to users in different locales is minimised.

4.3.3 Technical

In addition to linguistic and cultural diversity, content authors should also be aware of technical variances in different locales. Character encoding is one of the most important technical issues to consider when developing digital content for global audiences. A character encoding system consists of a code that pairs a sequence of characters from a given character set with a sequence of natural numbers. Authors should ensure that the system being used supports all characters in the language in which the content is being developed.

Another important issue to consider is whether keyboard shortcuts and hotkey combinations will work correctly in different locales. One must also be aware of the technical infrastructure, mobile devices and so on used in different locales. Online content, for example, should function correctly in different locales regardless of whether it is run on a Windows or Mac operating system or accessed via Internet Explorer or Firefox. Software applications may have additional technical concerns such as string concatenation or hard-coded strings.

The Localisation Industry Standards Association (LISA) estimates that content developed without consideration for the linguistic, cultural and technical variances in different locales takes twice as long to develop and costs twice as much to localise (LISA 2003). This paper proposes incorporating guidelines into internationalisation strategies to ensure the development of high quality digital content that takes into account the variances between different locales. Before examining how these may be incorporated to reduce the cost of localisation, it is necessary to examine each of the activities involved in the internationalisation process.

4.2 Author-Enable-Test Strategy

The internationalisation of digital content consists of several key tasks, the most significant of which are authoring, enabling and testing. The implementation of these tasks during the digital content development process is referred to as an AET Strategy in this article. The terms AET Strategy and internationalisation therefore, are used almost synonymously in this article; the main difference being that the latter refers to a broad strategy of preparing any product, service or content for translation, while the former refers to three specific processes implemented to make digital content as translatable as possible (Table 5).

Table 5: Components of an AET Strategy

4.2.1 Authoring

Authoring is the process of writing or constructing an electronic document or system. Authoring of source language content may also be considered to be the first step in the localisation process, with its quality level having a significant impact on the cost of translation and localisation QA. The main objective of the authoring process is to develop source language digital content that is usable and translatable. Usable content increases satisfaction among local users, while translatable content decreases the cost of localising the content for international users.

Content authoring research is based on the academic field of technical communication, with a focus on linguistics, content development guidelines, cultural research and technical writing standards. Specific professionals are dedicated to authoring different types of enterprise content; software applications, for example, are generally authored by software developers while digital documentation is usually generated by help authors, web designers and technical writers. These professionals typically use authoring tools such as Microsoft Word, Adobe Framemaker, Adobe Dreamweaver and Madcap Flare to develop digital content. These tools allow authors to check the linguistic quality of the content they are creating, although issues specific to localisation (such as character encoding) may not always be included in these validation checks.

Content development guidelines published by researchers, organisations or professional authors assist the authoring process. These guidelines usually develop from academic research, industry best practices and proposed standards. Authors also use controlled natural languages (CNLs) to prepare content for localisation by deliberately restricting the grammar and vocabulary to reduce or eliminate ambiguity and complexity. CNLs, such as Simplified English and E-Prime, are adopted by some organisations to increase the readability and translatability of digital content. Simplified English offers a carefully limited and standardized subset of English designed reduce ambiguity, make human translation easier and facilitate machine translation. E-Prime attempts to generate similar benefits by eliminating all forms of the verb to be: "be", "is", "am", "are", "was", "were", "been" and "being" (and their contractions). Authors may also use corporate dictionaries, glossaries and terminology databases to improve the consistency of content developed. Some organisations issue style sheets to authors to ensure the content they develop is of a high quality.

It is also essential for authors to consider cultural nuances when developing digital content for international audiences. These cultural aspects may be collectively examined by constructing a PESTEL analysis for each locale which includes:

- Political Considerations (government type, political history)
- Economic Considerations (purchasing power, standard of living)
- Socio-cultural Considerations (language, religion, attitudes, customs, colours, myths, symbols, fashion, education, role of the family)
- Technological Considerations (technical expertise, computer hardware)

- Environmental Considerations (natural resources, attitude to the environment)
- Legal Considerations (legal implications such as financial regulations & FDA requirements)

Poorly authored digital content, therefore, may contain incorrect grammar and punctuation, unclear language or inconsistent terminology. Time, dates, currency and measurement units may be used inconsistently throughout the content. References to religion, politics and symbols also have the potential to be offensive to users in different locales. Problems like these not only reduce the quality of the source language content for local users, but also make it more difficult to translate for international users.

4.2.2 Enabling

Enabling is the process of preparing digital content at a technical level so that it can handle multiple languages. Translating digital content that is not properly enabled for localisation may result in errors such as poor layout, clipped text and overlapping controls, characters not displaying correctly or international keyboards not working correctly with the content. Enabling aims to prepare content for localisation such that the process of locating and rectifying post-translation errors is less costly and time-consuming. Some of the main tasks involved in enabling digital content for localisation are:

- Using the Unicode character encoding standard to ensure that all international character sets are supported in the content, including bi-directional scripts such as Arabic and Hebrew
- Designing software user interfaces and document layouts to accommodate for the expansion of translated text
- Setting maximum limits for string lengths to avoid layout problems

Software engineers, web developers or dedicated internationalisation professionals are usually responsible for enabling digital content for localisation at a technical level. These professionals use content development tools (software development kits, help authoring software, web design packages, desktop publishing applications) or dedicated internationalisation tools (such as Globalyzer Diagnostics) to assist them in completing enabling tasks.

4.2.3 Testing

Testing, as a part of an AET internationalisation strategy, refers to checking the linguistic, cosmetic and functionality quality of digital content, prior translating it into different languages. Linguistic quality refers to how readable and translatable the content is, as well as how culturally appropriate it is for different locales. Linguistic quality is usually checked by a technical writer or editor. The main tasks involved in checking the linguistic quality of the source language content are:

- Proof-reading to ensure clarity of expression, grammar and punctuation
- Consistency checking to ensure adherence to corporate glossaries or termbases
- Checking language formatting such as time, date, weight and measurement formatting
- Verifying word and sentence lengths are appropriate for human or machine translation

Enterprises using CNLs, style sheets, corporate terminology databases or authoring tools usually spend far less time testing linguistic quality before translation than organisations that have no structured global content development process. Digital content that has not been tested for linguistic quality prior to translation, is more time-consuming for human translators to translate, and more difficult for machine translation systems to process.

Cosmetic quality refers to how visually consistent and aesthetically pleasing digital content is; this is usually checked by a technical writer, help author, software engineer or desktop publishing specialist. The main tasks involved in testing the cosmetic quality of digital content, before translation, are:

- Visual inspection of software user interfaces and document layouts to ensure there is room for text to expand after translation
- Pseudo-translation to preview the impact of translation is likely to have on the source language digital content
- Verifying that the character encoding is appropriate for display in different locales

Testing the functionality of digital content is essential to ensure a high standard of quality, regardless of whether the content will be localised or not. Digital content published with functionality errors will not only have to have these rectified in the source language version, but also in each localised version. Test engineers are usually responsible for testing the functionality of source language digital content, completing tasks such as:

- Test compiling, virus scanning and checking functionality of software applications
- Checking the operability hyperlinks and search boxes of online documentation

5 The Localisation Knowledge Repository (LKR)

A web application called the Localisation Knowledge Repository (LKR) is proposed as a method of incorporating content development guidelines into the AET process, with the aim of making digital content more translatable and less expensive to localise. The LKR (currently in development) consists of three distinct sections, the Digital Library, Test Area and Virtual Community. Each area of the LKR is based on user requirements generated from active PhD research projects running in the Localisation Research Centre (LRC) located in the University of Limerick.

5.1 LKR Digital Library

The LKR digital library is an online repository of content development guidelines, cultural guidelines and relevant industry standards. The Digital Library is initially populated with content development data generated through primary and secondary research. The LKR also incorporates a feedback loop enabling users to upload content development and cultural guidelines. Any uploaded guidelines are reviewed by a moderator before being published to the Digital Library in order to maintain the quality of the LKR system. Another means of ensuring the usefulness and relevance of the guidelines is by providing users with a rating system where they may publish a comment about each guideline and rate it on a five-point scale.

5.1.1 Content Development Guidelines

Content development guidelines are instructions, principles and best practices for content developers writing digital content for international audiences. These guidelines are compiled from primary research, existing literature and industry best practices. Most of the guidelines are sourced from the academic field of technical writing, but other relevant areas include internationalisation, web design, help authoring, software development, document design, linguistics, controlled natural languages (CNLs) and terminology management. Additional secondary research is generated from company reports and case studies. Primary research is generated from interviews, focus groups, surveys, and usability testing with technical writers, help authors and document developers.

Content guidelines stored in the LKR Digital Library are classified into five categories for easy access:

- Content (language style, voice and tone, punctuation and grammar, sentence length, terminology, graphics)
- Presentation (font type and size, use of colour, blank space, page layout)
- Navigation (table of content, navigation maps, reading sequences, search boxes, indexes)
- Accessibility (access medium, features for disabled users)
- Other Issues (functionality on different hardware, operating systems and web browsers)

5.1.2 Cultural Guidelines

The second section of the Digital Library enables users to define a particular locale and access a selection of significant cultural considerations associated with it. Only cultural research significant to digital content development and localisation is stored in the Digital Library, as creating a repository detailing cultural aspects of every world locale would make the system unwieldy and unmanageable. Relevant guidelines, therefore, are extracted from cultural research conducted on each locale, and published in the cultural guidelines section of the LKR Digital Library. The cultural guidelines are classified according to a PESTEL analysis (see 4.2.1).

5.1.3 Content Development Standards

A standard is an established norm or requirement outlining necessary criteria, methods, processes or practices. Standards may be developed by corporations, trade unions, industry associations or dedicated organisations. Several organisations develop and publish standards relevant to content development and localisation including the Localization Industry Standards Association (LISA), Organization for the Advancement of Structured Information Standards (OASIS), World Wide Web Consortium (W3C), International Organization for Standardization (ISO) and the Unicode Consortium. Relevant content development and localisation standards are included in the LKR Digital Library including:

- Localisation standards (XML Localization Interchange File Format (XLIFF), Segmentation Rules eXchange (SRX) and Translation Memory eXchange (TMX))
- Time and date formatting standards (ISO 8601)
- Usability standards (ISO 9241)
- Character encoding standards (Unicode)
- Web standards (ISO 8879:1986 SGML)

These standards help content developers publish consistent, high quality digital content.

5.2 LKR Test Area

The Test Area enables content developers to use the LKR system as a test bed to check the quality of digital content. The user accesses the LKR Test Area, opens a new project and specifies which files to check. The LKR then parses these files, showing the user a list of all the language strings in the file. After all the relevant files have been “checked-in” (i.e. a copies of the source files are uploaded to the LKR website), users may select a View Project Statistics option to display attributes about the project such as number of files, number of sentences, number of words, number of duplicate words, number of unedited strings, number of edited strings and number of signed off strings. A Generate Report option enables users to create project reports containing vital project statistics.

Figure3:LKRViewProjectStatisticsdialogbox

Users may also select a Check Content option to check the project files for a predefined list of quality issues, from poor spelling and grammar to repetition, inconsistency and broken tags. The results of the search are displayed in a list for the user, who has the ability to rectify any issues in an editing window.

Figure4:LKRCheckContentdialogbox

Once the relevant changes have been made, users can save the project and select the Export File option to generate an edited version of the original source file (in the same file format). Users can also choose the Pseudo-Localise option to export a file with a predefined level of text expansion to simulate the impact that translation might have on the source files.

5.3 LKR Virtual Community

The final component of the LKR is the Virtual Community. It consists of three sections:

- Forums Area (where LKR users can share ideas and opinions)
- Resources Area (where LKR users may upload and download resources such as style sheets, glossaries, termbases, corporate dictionaries and relevant multimedia files)
- Connect Area (where LKR users may contact other users by email or instant messaging)

5.4 Web 2.0 Features of the LKR

The LKR is an online repository that uses a selection of Web 2.0 features (O’Reilly Media 2009):

- User-generated content (users upload their own guidelines and resources to share with others)
- Crowdsourcing (users maintain the quantity of LKR data by uploading guidelines and resources, and maintain its quality by rating the guidelines and commenting on them (Trieloff 2007))
- Social networking (users express ideas and opinions via the Forums Section in the Virtual Community, and connect to other content developers via email and instant messaging)
- Web applications (the LKR Test Area is a functional system operating as a web, rather than a desktop, application)
- Customisation (when users create a LKR log-in profile, they define the type of content they develop, locales they work with and so, so that the data displayed to them by the LKR system is as useful and relevant as possible)

5.5 Using the LKR to Integrate Content Development Guidelines into the AET Process

Integrating content development guidelines into the AET process reduces the cost of localising digital content by making it more translatable. The LKR system enables content developers to integrate these guidelines into their workflows by providing a free, accessible database of the most up-to-date content development instructions, based on academic research and industry best practices. It also enables them to check-in the files they are working on, and conduct a customised quality check in the LKR Test Area. Users may customise the type what type of quality checks are run in the LKR Test Area; a software developer, for example, may check for overlapping controls a software user interface, while a technical writer may check for incorrect spelling or inconsistent terminology. The LKR, therefore, assists professionals in incorporating content development guidelines into the AET process in the following ways:

- Authoring: The LKR Digital Library gives content developers access to a vast repository of content development guidelines, cultural guidelines and relevant industry standards. They may also download glossaries, termbases and style sheets from the Resource Area, and contact other content developers for advice via the LKR Virtual Community.
- Enabling: The LKR Test Area supports the process of enabling digital content for localisation. LKR users, for example, may set maximum character limits on language strings to reduce the risk of their post-translation expansion

corrupting the layout of the content. The LKR Digital Library also contains guidelines on internationalisation and enabling content for international audiences.

- Testing: The LKR Test Area enables users to perform automatic linguistic testing (consistency checking, checking language formatting, verifying sentence lengths), cosmetic testing (pseudo-translation) and functionality testing (exporting edited versions of project files to check that they function correctly).

The LKR may be useful in the following scenarios:

- A web developer creating or updating web content
- A technical writer developing an online manual
- A help author designing a web help system
- A content specialist writing an e-learning course
- A marketing executive announcing a new product in a HTML email flyer

5.6 Benefits of the LKR

The LKR, as proposed, has the potential to deliver a number of benefits to its users:

- Reduce the cost of developing and localising digital content for global audiences
- Increase the quality of source language enterprise content for local user
- Improve the translatability of enterprise content (both for human translators and machine translation technologies)
- Provide a free system that marries industrial practice with academic research, and is maintained by its user base
- Preserve minority languages by making it easier to translate from and into more widely-spoken and commercially viable languages
- Bridge the digital divide by making it easier to develop digital content suitable for multiple locales (not just developed economies)
- Create an international network of content developers to share knowledge and opinions
- Inject cultural considerations into the localisation process (some view it as a purely technical activity)
- Provide an alternative to commercial desktop applications sold for profit (the LKR will be a free web-based application based on academic research, industry standards and enterprise best practice; incorporating Web 2.0 features such as customisation, social networking and crowdsourcing).

These potential benefits will hopefully address the linguistic, technological and connectivity issues encountered by content developers, enabling them to publish highly internationalised and usable digital content more productively and cost-effectively.

6 Conclusion

We started off by distinguishing between traditional and digital content in Section 1, and went on to classify digital content as either enterprise or personal in Section 2. In Section 3 we examined why enterprises localise digital content, and the costs involved in doing so. Internationalisation is examined in Section 4, with the three core processes of authoring, enabling and testing explained in detail. Section 5 describes how content development guidelines can be implemented into the digital content production process via a Web 2.0 system called the LKR. These guidelines ensure that content is generated for global rather than local audiences, and is less costly to translate and test due to improved translatability (for human translators and machine translation systems) and quality respectively.

The success of the LKR ultimately depends on how well it is embraced by the content developer community. It has the potential to pool the collective knowledge of content developers worldwide, and dramatically increase the quality and consistency of the enterprise content being published worldwide. A centralised, online repository of data from subject matter experts in different regions could have enormous benefits for multinational corporations developing digital content for international audiences, and could potentially have an impact on the service fee charged by language vendors in the future.

Despite this promising initial reaction to the concept of the LKR, there are several challenges to overcome. The first challenge is technical; parsers will have to be developed to support the myriad of file formats that content developers currently work with. The second challenge is the usability of the system; it will have to be user-friendly and accessible, allowing users to customise it so that they only access those guidelines which are relevant to their particular projects. Finally, the quality and relevance of the data in Digital Library depends very much on user contributions, interaction and feedback.

References

- Aviation Safety Network (2009) 'Accident description', Retrieved 19 June 2009, from <http://aviation-safety.net/database/record.php?id=19830723-0>
- Common Sense Advisory (2006) 'Report on global consumer online buying preferences, showing the impact of language, nationality, and brand recognition', Retrieved 19 June 2009, from http://www.common senseadvisory.com/news/pr_view.php?pre_id=39
- Coyle (2009) 'Facebook gets set for an Irish language lesson', Retrieved 19 June 2009, from Times Online <http://www.timesonline.co.uk/tol/news/world/ireland/article5489404.ece>
- Encyclopaedia Britannica (2009) 'History of computing,' Retrieved 19 June 2009, from <http://www.britannica.com/EBchecked/topic/130429/computer/216032/Invention-of-the-modern-computer>
- Ethnologue Languages of the World (2009) 'Ethnologue language name index', Retrieved 09 July 2009, from http://www.ethnologue.com/ethno_docs/distribution.asp?by=size
- Europa Languages and Europe (2009) 'Is every document generated by the EU translated into all the official languages?', Retrieved 19 June 2009, from <http://europa.eu/languages/en/document/59#5>
- Facebook (2009) 'Facebook releases site in Spanish; German and French to follow', Retrieved 19 June 2009, from <http://www.facebook.com/press/releases.php?p=16446>
- Freij, N. (2009) 'Web 2.0 and localization', Retrieved 19 June 2009, from <http://blog.globalvis.com/2008/04/web-20-and-localization.html>
- GoToWeb2.0 Web Applications Index (2009) 'Web 2.0 tools and applications', Retrieved 19 June 2009, from <http://www.go2web20.net>
- International Trade Administration (2009) 'Computer software industry 2008', Retrieved 19 June 2009, from http://www.ita.doc.gov/investamerica/computer_software.asp
- International Organization for Standardization (2009) 'Numeric representation of dates and time', Retrieved 19 June 2009, from http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm
- Jobes, K. and Silva, M 2001 Invitation to the Septuagint ISBN 1-84227-061-3, (Paternoster Press, 2001).
- LISA (2003) 'LISA industry primer 2003', Retrieved 19 June 2009, from <http://www.cit.gu.edu.au/~davidt/cit3611/LISAprimer.pdf>
- National Aeronautics and Space Administration (2009) 'Mars Climate Orbiter', Retrieved 19 June 2009, from <http://solarsystem.nasa.gov/missions/profile.cfm?MCode=MCO&Display=ReadMore>
- Netcraft (2009) 'June 2009 web server survey', Retrieved 19 June 2009, from <http://news.netcraft.com>
- O'Reilly Media (2009) 'What Is Web 2.0? Design patterns and business models for the next generation of software', Retrieved 19 June 2009, from <http://oreilly.com/web2/archive/what-is-web-20.html>
- Recording Industry Association of America (2009) 'Consumer purchasing trends', Retrieved 19 June 2009, from http://www.riaa.com/keystatistics.php?content_selector=consumertrends
- Software Top 100 (2009) 'The world's largest software companies', Retrieved 19 June 2009, from <http://www.softwaretop100.org/list.php?page=1>
- Tahoe Daily Tribune (2009) 'Another tough year for the music industry', Retrieved 19 June 2009, from <http://www.tahoe-dailytribune.com/article/20090108/ENTERTAINMENT/901079974/1005/NONE&parentprofile=1056&title=Another%20tough%20year%20for%20music%20industry>

Tek Translation (2009) 'Spanish language variations', Retrieved 19 June 2009, from http://www.tektrans.com/docs/Tek_Educational_Best_Practice_-_Spanish_Variations.pdf

The Guardian (2009) 'Internet data heads for 500bn gigabytes', Retrieved 19 June 2009, from <http://www.guardian.co.uk/business/2009/may/18/digital-content-expansion>

TidWiT Digital Content Marketplace (2009) 'What is digital content?', Retrieved 19 June 2009, from <http://www.tidwit.com/WhatIs.aspx>

Trieloff, L. (2007) 'Living in a multilingual world: Internationalization for Web 2.0 applications', Retrieved 19 June 2009, from <http://www.slideshare.net/lars3loff/living-in-a-multilingual-world-internationalization-for-web-20-applications>