Localizing Canon's User Documentation in Europe

Tim O'Donoghue

Abstract

Canon is a company that produces a wide range of products in the areas of imaging, information and industry. Canon's innovative products are marketed the world over and user documentation is a key component in each Canon product.

Canon Europa NV is one of Canon's regional headquarters. Even though the name implies responsibility for Europe alone, Canon Europa is also responsible for Eastern Europe, Russia, Africa and the Middle East.

One of Canon Europa's responsibilities is the localization of user documentation. This is quite a responsibility when you consider the number of languages spoken in Canon Europa's territories (over 25 major languages) and the number of products which Canon produces.

Business machines - copiers, faxes, printers, filing systems, etc. - account for a large proportion of Canon's sales, and for these products it is often necessary to produce user documentation in a number of different formats, both print and electronic, depending upon the product in question and the users needs. Canon Research Centre Europe, one of Canon's global R&D centres, has been working with Canon Europa over the past 5 years to develop an efficient localization and production process, in which:

- SGML is used as a publishing base to enable efficient and flexible publication;
- machine assisted translation, in the form of a translation memory, and translators' tools are used to enable efficient localization.

Dr. Tim O'Donoghue

Dr. Tim O'Donoghue leads the Solutions Applications Services group at Canon Research Centre Europe. He is a consultant to Canon Europa in the area of document localization and production.

Canon Europa NV
Bovenkerkerweg 59-61
1185 XB Amstelveen
The Netherlands
Phone: +31-20-545-8545
Fax:
Web: http://www.europe.canon.com/

Canon Research Centre Europe Ltd.
20 Alan Turing Road
Guildford GU2 5YF
UK
Phone: +44-1483-448844
Fax: +44-1483-448845
Web:http://www.cre.canon.co.uk/

Localizing Canon's User Documentation in Europe

Canon is a company that produces a wide range of products in the areas of imaging, information and industry. Canon's innovative products are marketed the world over and user documentation is a key component in each Canon product.

Canon Europa NV - one of Canon's regional headquarters - is responsible for Europe (Western, Central and Eastern), Russia, Africa and the Middle East. One responsibility is product localisation, which includes the translation of user documentation. This is quite a task when you consider the number of languages spoken in Canon Europa's territories - over 25 major languages - and the number of products which Canon produces.

Business machines - copiers, faxes, printers, filing systems, etc. - account for a large proportion of Canon's sales, and for these products it is often necessary to produce user documentation in a number of different formats, both print and electronic, depending upon the product in question and the users needs.

To enable this to be done effectively, Canon Europa has:

- implemented a SGML[1]-based process for flexibility and manageability
- developed and deployed a translation memory system - Adroit - to assist the localisation process

An Overview of the Production Process

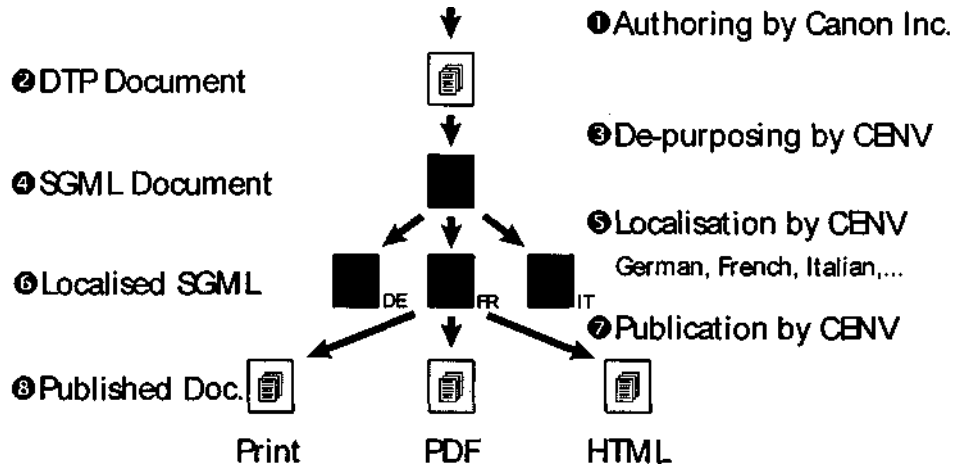

Figure 1: Overview of Production Process

Figure 1 gives an overview of the process by which Canon Europa (abbreviated CENV in the diagram) produces localised user documentation for business machines:

1) The original user documentation is authored by Canon Inc., Canon Europa's parent company based in Japan. The Japanese original is then translated into English by Canon Inc. and it is this English version which is delivered to Canon Europa.

2) The delivered material is usually a DTP document, typically PageMaker or Quark XPress, depending upon the product group.

3) The first step of Canon Europa's process is to *de-purpose* this DTP document. De-purposing is the process by which the DTP document is transformed into a *publication-independent* form.

4) The publication-independent form chosen by Canon Europa is SGML. The specific SGML application used is *TranScribe-III,* a DTD developed specifically for Canon's business machine user documentation.

5) The English SGML document is localised into the necessary language versions for the areas in which the product is to be marketed. This typically ranges from 10 to 20 different language versions.

6) The result of localisation is a set of SGML documents, one for each of the target languages.

7) For each of the SGML documents, a *publication process* is applied to convert the SGML document into its final published form.

8) Print, PDF (Adobe's Portable Document Format) and HTML (HyperText Markup Language) are the three publication formats

currently produced by this process.

Why SGML?

Upon reception of the DTP document from Canon Inc., Canon Europa's first step is to de-purpose this material to ensure a *solid foundation* for the subsequent stages of production. De-purposing can be a very expensive task, but Canon Europa's experience has shown that effort invested early in the production process is recovered with "interest added" in the later localisation and publishing stages. This solid foundation simplifies the management of the localisation process and allows increased automation of the publishing processes.

One example of where SGML has been used to simplify the localisation process is in managing the consistency between hardware localisation and the translation of user documentation. User documentation cannot be translated in isolation, that is, without reference to the complete product kit, of which it is a part.

The complete localisation of a product requires the localisation of its various components, including:

- The product hardware. E.g. the fax machine, including the various user interface components (buttons, display messages on LCD panels, etc.)
- The product software. E.g. the printer driver for a Bubble Jet printer.
- The user documentation, which documents both the hardware and the software.

Importantly, the localisation of these various components must be *consistent.*

In the SGML document, (hardware) display messages are identified by the appropriate SGML tags. These tagged messages are then used to query a separate database in which the previously translated display messages are stored (as used to localise the hardware), and the correct translated message is interpolated into the translated user document. Using the same display message database for both the production of localised hardware and documentation ensures consistency between these components in the overall localised product.

A second example of where SGML has simplified the localisation process is in the handling of non-ASCII character data. Canon's TranScribe-III SGML application uses the SPREAD [2] Entities - which themselves are derived from Unicode[3] character set - to represent all non-ASCII characters in the SGML documents. And Unicode mapping tables – available from the Unicode Consortium [4] - are used to enable import and

export to data which is encoded with respect to a given character set or code page.

Adroit

Canon uses a machine-assisted translation system as part of its localisation process for user documentation. This system, named *Adroit,* was developed by Canon Research Centre Europe for Canon Europa. Adroit is based on the *translation memory* concept whereby a database of translations is used to assist in the translation of new documentation. New documents are compared against older documents, stored in the database, and only those areas identified as "new" need be manually translated. The areas identified as "old" can be translated using older previously-translated documents.

The Adroit database is built from previously translated user documentation, extracting translations from pairs of documents which are mutual translations of one another. The extraction of translations relies upon the structure of the documents being invariant between the pairs of documents. This necessary structure is provided by the logical structure imposed by the TranScribe-III SGML application and protected during translation (by only sending textual data to translators, not SGML structural data).

Development of Adroit started at Canon Research in 1991 and 1993 saw the first use of Adroit by Canon Europa. The initial database was created using archived typesetting data and is updated as and when new documents are fully localised. At time of writing, the database contains translations extracted from over 225,000 pages of user documentation added over the past five years. And, of course, this is growing every month.

Each new English user document is processed by Adroit by comparing its text content sentence by sentence against the English text stored in the database. If this comparison yields a match — either exact or approximate - then the matched text is replaced with a translation obtained from the database. The result of Adroit processing is a partially-translated document, containing both (untranslated) English text and target language text (obtained from the database). On average, over the course of a year, we experience translation rates of approximately 40%, although this does vary greatly depending upon the similarity for a given document.

When Adroit was initially developed for Canon Europa, there weren't many other translation memory systems available in the marketplace. Of course times have changed and now there are many products offering a similar functionality: IBM's Translation Manager II, Star's Transit (used by Canon, but not for its translation memory capabilities), Trados' Translators Workbench, etc. However, Adroit is still a competitive choice for Canon Europa since it particularly tailored to Canon's needs (in its handling of display message, for example) and it has a number of features

useful to Canon but not yet found in other translation memory products, including:

- A hierarchically structured translation database, rather than a single large database into which all translations are deposited. The database is structured to reflect the different types of documents from which the translations were obtained. For example:
  - **copier/black+white/NP6112**
  - **fax/multi-function/MPC30/hardware**
  - **fax/multi-function/MPC30/software**
- The ability to quickly compare a new document against the hierarchical database to determine which segments of the database should be used for memory translation.

Transit

The output of Adroit is a partially translated document, that is, a document which contains sentences both in the source language (English) and the target language (as obtained from the translation memory). This document is sent to a translator.

Adroit only sends sentential data **(#PCDATA** - in SGML parlance - and those minor elements which can appear in sentences, such as font changes etc.) and a majority of the SGML structure is retained by Adroit. When returned by the translator, the fully translated text is merged with the SGML structure of the original English document to yield a fully translated TranScribe-III document instance.

This process of separating SGML structure from text content in the translation phase is convenient both for Canon and its translators:

- It ensures that translators cannot disturb the SGML structure, which is essential for correct updating of the Adroit translation database.
- It allows translators to concentrate on translating the text without having to locate it in the often complex SGML structure.

The translator performs three types of tasks when working on a document sent from Adroit:

1) Translating, from scratch, any text in the source language.

2) Proof reading the exact translations supplied by Adroit.

3) Checking, and possibly modifying, the approximate translations supplied by Adroit.

It is possible, in the case of **2),** for the translator to modify an exact translation supplied by Adroit. However, this seldom occurs and when it does, it is usually justifiable.

Translators use the Star's *Transit* system which provides them with a suitable interface to perform these and other localisation related tasks. A screen shot of the Transit system is shown in Figure 2:

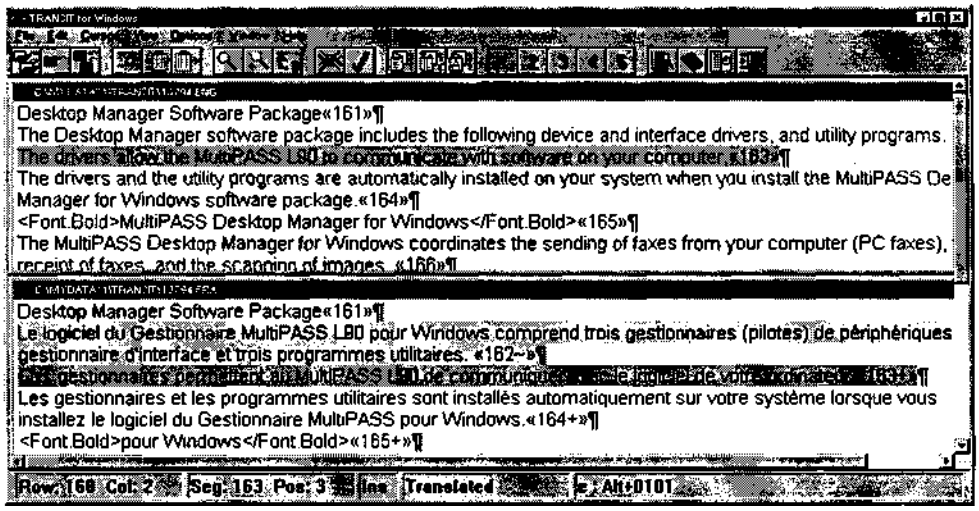

Figure 2: Screen shot of the Translator's Interface

- The original English is shown in the top pane with the partial translation produced by Adroit in the bottom.

- The background colour is used to indicate the different types of text: grey for the current sentence, white for a sentence exactly translated by Adroit, blue for an approximate translation and yellow for untranslated text.

- Elements, and other protected areas are coloured blue. These cannot be edited by the translator.

Transit does contain its own translation memory facility, but this is not directly used since this function is provided by Adroit. Within Canon's localisation workflow, Transit's role is primarily as an editing tool, allowing translators to work with the output of Adroit. However, Transit's translation memory facility is used by the translators who build their own memories. These local memories allow the translator to handle repetition in the current translation job as well as providing a memory for whatever other translation work they undertake.

Future versions of Adroit will support other translators environments – Trados' Translators Workbench for example - so that translators can use their chosen tool rather than one specified by Canon.

At what cost SGML?

While there are clear benefits - both in localisation and publication - from using SGML, there is also a cost: *conversion.* The current de-purposing processes are highly dependent upon the DTP material received from Canon Inc. and often a great deal of manual effort is required in this phase. However, this effort can be reduced: If the original DTP material is authored in a structured manner, e.g. by using a standard set of styles, then it can be - partially at least - mapped via a *filter* onto an equivalent SGML structure.

Canon Europa is aiming to increase the amount of automation in this de-purposing phase by influencing upstream document production at Canon Inc. Our short term aim is for Canon Inc. to implement a set of DTP authoring guidelines that will assist the automatic conversion from DTP formats to SGML or other document formats. A longer term goal is of course to push SGML further upstream, eventually to the original Japanese authoring environment, but there are many issues which need to be addressed before this can happen.

The **Woes of System Integration**

Within the workflow of Canon's document localisation and production process there are a number of discrete components which must be glued together so that they can work in harmony:

- Adroit, the translation memory
- Transit, the translators working environment
- The authoring environment, SGML or otherwise, used to create the documentation
- The typesetting systems which are used to publish the documentation
-

This integration is typically enabled by creating various conversion programs - *filters* - which covert one component's data format into a format suitable for another component. Within Canon's workflow there are many such filters, a prime example being the two filters that enable Adroit to 'export data to' and 'import data from' Transit. If Adroit were to support a second system for translators - Translators Workbench for example - then this would require the creation of a further two filters to handle Trados' data format.

One of they problems facing system designers and integrators is that the many components which constitute a complete localisation system are not

immediately compatible; they have to be glued together in some way. *No two programs speak the same language;* each has its own API or data format.

The ideal situation for integrators is where the various components - authoring tools, translation memories, translation environments,... - can be treated as modular objects, each sharing a standard API and a standard format for data interchange. How long until this pipe dream is reality? Who knows? The development of localisation-aware authoring environments will be crucial. There are already quite promising industry initiatives such as OpenTag[5], in which companies from different areas of the localisation and documentation industry are developing an XML[6]-based standard that will support open data encoding methods during the localisation process. If such initiatives are successful then designing filters and 'system integration glue' will be a thing of the past, and designing localisation workflows will be more like playing with Lego. This will result in systems which are easier to implement and more reliable in operation.

References

1) The Standard Generalised Markup Language. ISO 8879:1986.

2) See **http://www.allette.corn.au/sgml/ercs/allent.htm** for details of the SPREAD (Standardisation Project Regarding East Asian Documents) Public Entity Set.

3) The Unicode character encoding standard, an international character code for information processing. ISO/IEC 10646-1:1993.

4) See **http://www.unicode.org/** for details of the Unicode Consortium.

5) See **http://www.opentag.org/** for details of the OpenTag Initiative.

6) Extensible Markup Language is a data format for structured document interchange on the Web. See **http://www.w3.org/XML/** for details.