Linport as a Standard for Interoperability Between Translation Systems

|  |  |
| --- | --- |
| Alan K. MelbyBrigham Young UniversityDepartment of Linguistics & English Languageakmtrg@byu.edu  | Tyler A. SnowBrigham Young UniversityTranslation Research Grouptylerasnow@byu.edu  |

Abstract: There are many evolving technology standards in the translation/localization industry, such as XLIFF and the Internationalization Tag Set (ITS). The Linport project aims to define a standard container format for translation projects and tasks within projects. Linport, XLIFF, and ITS are complementary, with each focusing on different aspects of the international content production process. This paper presents the need for a container standard in addition to XLIFF and ITS and describes the status and likely future of Linport.

Container Formats 2

Shipping Containers 2

Translation Today Is Like Pre-Container Shipping 3

Linport: A Standardized Container for the Translation Industry 4

Elements of Linport 4

Structured Translation Specifications (STS) 5

History of Linport 6

Relationship to Other Formats 7

XLIFF 7

Internationalization Tag Set 2.0 10

Additional Linport advantages 11

Integration with Quality Assessment practices 11

Ease of Implementation 12

Linport is free. 12

Current and future development 12

How to get involved 13

# Container Formats

Despite the revolutionary changes in communication methods and efficiency of the late twentieth and early twenty-first centuries, being able to send a message to someone on the other side of the world still does not guarantee that *communication* can be achieved; language still serves as a barrier, which is why the translation industry exists. Although Internet communications have enabled greater speed and efficiency for translation, there remains a need for improvements in technological communication within this industry that can help eliminate many of the current manual bottlenecks that add time and expense to the translation process. These bottlenecks include manual processes such as opening files and saving the contents into supported formats, sending and receiving files and checking their integrity, negotiating unclear details about the translation process, and finding and ensuring resource availability. A substantial portion of translation costs is spend on management and manual tasks other than translation that are nonetheless currently needed to support translation. This situation in which many processes are essentially unchanged since the early 1990s suggests that new standards should be developed in order to promote interoperability and reduce the need for these low-value, repetitive tasks.

## Shipping Containers

A classic example of the problems caused by lack of interoperability, and the benefits that come from standardization, comes from the shipping industry. For thousands of years, countries and companies shipped products internationally. In doing so they have attempted to maximize trade profits by cutting costs and travel time. Efforts to streamline these processes were hindered by the wide variety of containers (boxes, bags, crates, barrels, etc.) in use. Due the wide variety of shapes and sizes involved, these often did not fit efficiently onto the trains, boats, and other vehicles that carried them to their final destinations. Because containers tended to be small enough for manual loading and offloading, loading operations were labor-intensive and error-prone, with high rates of breakage, misdirection, and other loss. Consequently, trucking, and train companies found it difficult to plan for the movement of incoming goods and minimize wastage. Goods would have to be unpacked from the holds of ships and then repacked into the cargo space on trucks or trains, wasting valuable time and money. All of these operations increased time-to-market and raised labor costs.

In response to the need to move large amounts of materiel in the Second World War, the United States introduced the first standardized shipping containers. Made from corrugated steel to standard dimensions with corner blocks that allowed them to be locked together, these containers could be loaded at their point of origin, manipulated with mechanized loaders, and shipped to their definition without the need to manually move their contents between containers. Ultimately, the international community agreed upon a standard for “intermodal shipping containers” with a size and construction based on these early U.S. containers. These standardized shipping containers could be easily moved from ship to train or semi-trailer truck without removing the contents. Since the dimensions and locking corner blocks were precisely defined and shippers knew precisely what to expect in advance, various modes of transportation were able to handle a single container. These improved measures of interoperability have enhanced the shipping industry’s productivity on a global scale, and international shipping is possible today at volumes orders of magnitude higher than was possible before the advent of intermodal shipping containers.

## Translation Today Is Like Pre-Container Shipping

The need for standardization applies not only to the shipping industry but to many other fields as well, including translation technologies. Translation technology companies constant vie for market share for their translation software. Companies like SDL, LingoTek, MultiLing, Kilgray, SYSTRAN, and XTM-Intl, among many others, all have an interest in enabling potential customers to use their tools. All of these translation software systems, however, have different interfaces that handle translation projects in different ways, often creating interoperability difficulties: a translation project created in one tool generally cannot be opened in another tool—at least not without significant processing and, often, manual intervention—and all participants in the supply chain are limited to a particular tool set. The result of this situation is that otherwise competent service providers may be excluded from participation if they do not happen to own or use a particular tool. (This situation is similar to that of a hypothetical world in which every shipping provider uses its own type of shipping container and a customer cannot get a container shipped the last ten kilometers from the rail depot because no trucking company happens to support that type of container, even as hundreds of trucks configured for other containers sit idle. As a result the company is forced to manually unload the contents of the container and pack them into another sort of container.)

For example, if one company starts a translation project using SDL tools and then subcontracts out of house to a freelance translator who uses MemoQ (Kilgray), the same problem ensues as found in the shipping container example. The various project components, such as source text, specifications, and file types, all have to be “unpacked” from the SDL format and “re-packed” or reformatted in the MemoQ’s format for the freelance translator to do his or her allotted portion of the project. Finally, the completed translation then has to returned in SDL format. Similar to the shipping industry’s need for standardized containers, the translation industry needs a standard “container” of its own, to allow for interoperability between the numerous translation software tools now available to professional translators and companies worldwide. Fortunately, a standardized translation project container will soon be available.

# Linport: A Standardized Container for the Translation Industry

Linport (**L**anguage **In**teroperability **Port**folio), an open, comprehensive and interoperable container solution for all translation processes and projects, is currently under development. A Linport container includes the details of an entire translation project and carries each of the individual components that comprise the various tasks relevant to a translation project. These tasks could include the initial translation of a text from one language to another, a translation revision, a review, or a proofreading task. Because Linport is an open format, each of these tasks would be accessible to any participant in the project who supports the Linport format (versus supporting a variety of proprietary formats). Upon task completion, they would be able to “pack” their goods back into the Linport container for further processing. Therefore, in keeping with the shipping example, a Linport portfolio represents an overall project view much like a shipping container: each process in the production chain knows exactly what to expect and can optimize processes around these expectations. At the same time, Linport containers can define particular translation tasks within the project. These tasks are comparable to smaller standardized packages that are shipped inside a larger container. One such example of a task-specific format is the TIPP format, which will be explained later.

## Elements of Linport

The current implementation of Linport is represented by a directory structure format containing two sub-folders;

1. *portinfo*. Similar to an HTML header element, the *portinfo* folder contains information about the portfolio as a whole, such as specifications and support files that apply to all relative subtasks, and universal identifiers to facilitate breakdown and reintegration of the portfolio.
2. *payload*. This folder contains any translated or untranslated documents, as well as the supporting resources needed for translation or any other required task, such as revision or review. Examples of these resources could include translation memory files, textual references, terminology files, and style guides.

Linport portfolios can contain almost any file format, as long as it fits into the predefined directory structure. To facilitate this methodology, the payload folder is divided into language folders, such as the “en” folder for English or “es” folder for Spanish. These folders then contain individual document folders, which house exactly one document in a “doc” folder and its supporting files (such as glossaries and translation memories) along with the document’s specifications, in an “assist” (support) folder, whose name is under discussion. In this manner, translation tools know how all files correlate with each other and can handle them appropriately. The directory structure outline can be found online at <http://dragoman.org/linport/ldm.txt> .

## Structured Translation Specifications (STS)

Structured Translation Specifications (STS) enhance Linport’s ability to store and transfer the information necessary for translation tasks. They allow companies or translation project managers to specify important metadata about the translation itself, such as the target audience and intended use of the translation. An STS file includes 21 important translation parameters that should be formulated either during or (preferably) before initiating a translation task or project. At the highest level, these specifications deal with three aspects of translation: the product (that is, deliverables such as a target text), the process used to obtain the product, and other aspects of the project, such as due dates and compensation. The 21 translation parameters whose values form specifications are provided in Table 1 and are also available at <http://ttt.org/specs>. The translation parameters that provide the framework for an STS are from an international guidance document for translation project (ISO/TS 11669).

|  |
| --- |
| A. Linguistic *product* parameters [1–13] Source-content information [1–5][1] textual characteristics a) source language b) text type c) audience d) purpose [2] specialized language a) subject field b) terminology [3] volume [4] complexity [5] origin Target content requirements [6–13][6] target language information a) target language b) target terminology [7] audience [8] purpose [9] content correspondence [10] register [11] file format [12] style a) style guide b) style relevance [13] layoutNote: process and project parameters can be separated from the product parameters.B. Production *process* tasks [14–15][14] typical production tasks a) preparation b) initial translation c) in-process quality assurance [15] additional tasksC. *Project* Environment [16–18][16] technology [17] reference materials [18] workplace requirementsD. *Project* Stakeholder Relationships [19–21][19] permissions a) copyright b) recognition c) restrictions [20] submissions a) qualifications b) deliverables c) delivery deadline [21] expectations a) compensation b) communication |

Table 1. Translation Parameters for Product, Process, and Project

## History of Linport

Linport is an example of a conglomeration of companies and organizations working together. The project comes from three main project streams. In March 2011 many of the organizations that participated in the now-defunct Localization Industry Standards Association (LISA) agreed that a container type format was needed in the translation/localization industry. A small team was designated to develop a proposal for translation containers. The first presentation of their work was given a month later in Torino, Italy at the JIAMCATT translation technology conference (JIAMCATT 2011). After the presentation, a representative of the European Commission’s Directorate General for Translation (DGT) indicated that their organization already had worked on a similar project, known as the Multilingual Electronic Dossier (MED) project. MED aimed to represent an entire translation project in a translation “dossier.” After a series of discussions, the Container project and the MED project were merged to form the Linport project in July 2011, hosted by LTAC Global, a non-profit organization. It was decided that the Linport container would be called a *portfolio* and would contain all data pertaining to a translation project, be it an authoring, translation, or publication project.

In 2010, unbeknownst to the Linport project, an initiative called Interoperability Now! (or IN! for short) began work on another similar project. The participants in the Linport project and those involved in the IN! project became aware of the overlap in the fall of 2011 and then held a series of discussions. By mid 2012, IN! agreed to integrate its own “container” format into the Linport project.

IN!’s primary contribution to the Linport project is the Translation Interoperability Protocol Package (TIPP) and XLIFF:doc, which is a subset of XLIFF 1.2. TIPP represents a single translation task to be performed using exactly two languages in a translation workflow. A Linport portfolio, by contrast, addresses the whole translation project, potentially involving many languages and tasks. By design, a Linport portfolio can be broken down into multiple TIPP task packages, which could be accessed, translated, and reintegrated back into the portfolio for transportation to another translation tool. (History by Melby et al. 2012, *Multilingual Magazine*).

TIPP was designed with XLIFF in mind. Information about the TIPP format, including the parser tool, can be found online at:

<http://code.google.com/p/interoperability-now/>.

# Relationship to Other Formats

Linport portfolios are designed to efficiently move translation data between translation environment tools. The need that Linport addresses is not new, and other specifications address part of this need. However, Linport operates at a different level than these other specifications. Two specifications are particularly worth note: OASIS’ XML Localization Interchange File Format (XLIFF) and the W3C’s Internationalization Tag Set 2.0 (ITS 2.0) specification.

## XLIFF

XLIFF has emerged as one of the most important localization-related standards. XLIFF provides a standardized file format for storing, translating, and reexporting content taken from other native-format files. For example, if a company has a proprietary format for embedded UI strings, it can devise an XLIFF export routine for that format and generate an XLIFF file. With that XLIFF file, it can then work with any of the many translation service providers that support XLIFF, even if those providers know nothing about the internal format.

XLIFF thus represents an important level of abstraction that helps move away from the need to support many proprietary file types in the translation process: if there is an XLIFF import/export routine for a format, then any XLIFF-compliant tool will, in principle, work with it. However, XLIFF does not address the broad issue of how to package multiple files and convey project-related metadata, nor can XLIFF include all of the related resources that might be required for translation: translation memories, terminology resources, reference materials, and project instructions. These sorts of materials are often the ones most likely to be lost or require manual handling in the translation process, so XLIFF cannot directly address the costs they impose on the process.

Linport operates as a higher level. It does not, itself, impose restrictions on file formats or content and does not directly ensure that content can be accessed (although specific Linport profiles, see below, may have additional restrictions). Rather Linport functions as a structured container for content with metadata. A novice user of Linport could use it to contain a large number of native Photoshop files with the expectation that one of the tasks will be to extract text from them and send it on. The Linport committee, however, strongly encourages Linport users to work with XLIFF as the “payload” format to help ensure maximum interoperability.

XLIFF users will find that a Linport portfolio can incorporate XLIFF files at its very heart. A Linport portfolio can contain any number of translatable and/or already translated documents. Although any bi-text, monolingual, or multilingual document can be contained within a Linport portfolio, it is anticipated that XLIFF will be the most common format used. In addition, Linport “profiles” can impose restrictions on possible content formats to improve interoperability, and it is anticipated that “strict” Linport profiles will mandate the use of XLIFF. The TIPP profile, for example, requires XLIFF:doc.

An XLIFF document will often be accompanied by non-XLIFF supporting files, such as terminology files (e.g., TBX files), translation memory files (e.g., TMX files), or translation metadata (e.g., an STS file). All of these files can be optionally grouped together within a TIPP package or directly into the payload folder of a Linport portfolio. In this way, several XLIFF documents with their support files can easily be packaged together into one Linport portfolio, as shown in the diagram below.

Figure 1. Linport portfolio organization

Linport does not replace XLIFF. The two formats work together to promote and enhance organization and interoperability in the translation workflow. The relationship can be summarized in the following points.

* *Linport adds a new level of abstraction to the XLIFF format.* XLIFF is designed to facilitate and simplify the translation of a source text to into a target language. Linport, by contrast, is designed to organize the entire translation process (including multiple files and greater numbers of languages) and facilitate the transmission of all resources before, during, and after the actual translation. Entire translation projects can be efficiently organized, broken down into translation tasks (which may involve XLIFF), and finally reintegrated back into a project portfolio for revision, review or final publication.
* *Linport provides project level interoperability for both XLIFF and non-XLIFF translation projects.* Whether service providers work with a pure XLIFF-style project in house or with multiple freelance translators, Linport provides a standardized way to move project data between all stages of the translation workflow.
* *Linport can handle multiple file types.* Linport allows XLIFF to easily work with non-XLIFF file types. Translation file formats for source texts may range from XLIFF to DOCX to desktop publishing formats. Glossary formats are equally diverse. Linport’s standardized file structure will help machines and humans quickly associate all of the various parts of a translation project. Future Linport-aware XLIFF tools will be able to convert translation data to and from XLIFF format with ease. International standards for glossaries and terminology databases will be taken into account.
* *Linport incorporates ISO/TS 11669.* As discussed above, the Structured Translation Specifications (STS) implement the 21 parameters defined in ISO/TS 11669 and use of STS is required for Linport. XLIFF processes benefit from the information contained in an STS, but XLIFF files themselves cannot contain STS files, necessitating a separate communication process for specifications. By consolidating STS (and other metadata) into a single container with an XLIFF payload, Linport helps ensure the integrity of the entire process and that all parties receive the relevant information needed to promote quality.

Taken together these points demonstrate that an ideal ecosystem would support both XLIFF and Linport, with both formats handling their respective strengths in tandem to help deliver process resources and automation (Linport) and low-level content interoperability (XLIFF).

## Internationalization Tag Set 2.0

The Internationalization Tag Set 2.0 (ITS 2.0) is the 2013 update of ITS 1.0. ITS provides a robust and flexible set of metadata categories that support the internationalization and localization of XML and HTML5 content.

Perhaps the simplest example of an ITS data category is the “translate” category, which can take two values—yes and no—and provides instruction to translation processes whether to translate content or not. For example, a website might contain a regulatory notice that should not be translated into another language. In such a case, the HTML code could add “translate=no” to the <p> element containing the notice and automatic translation process would know not to translate it. (In fact, translate has been adopted as a core feature in HTML5 and all of the major online translation engines now support it.)

Other ITS data categories include the following:

* **Localization notes**, which provide guidance for how content should be treated in localization processes.
* The ability to mark and identify **terminology**.
* The ability to identify **text directionality** in formats that do not natively support it.
* Information about the **language** of text.
* A mechanism to specify whether XML **elements within text** should interrupt text flow or be seen as part of the surrounding text.
* Information on the **domain** (subject field) of text
* The results of automated **text analysis** processes (such as named entity extraction).
* The ability to tag text (**locale filter**) for translation into specific locales (or to exclude it from translation for specific locales)
* A mechanism to identify the **provenance** (source) of text, e.g., to state that particular content was translated with a particular system.
* A mechanism to point to **external resources** that contain the translatable text associated with a particular piece of content, such as a graphic, that cannot be translated directly.
* The ability to indicate that particular elements in a file format should contain the translation of particular other elements (**target pointer**).
* A mechanism to add **ID values** to content that lacks them to support translation requirements.
* Indications as whether or not to **preserve spaces** in content.
* A set of metadata on **localization quality issues** and **localization quality ratings**.
* Information on the internal confidence of machine translation systems (**MT confidence**) that can be used to flag content for further attention
* A mechanism to specify the **allowed characters** in a given piece of content (e.g., to state that a translation should not allow certain forbidden characters)
* A way to indicate the maximum **storage size** for content, such as database entries, that must conform to size limitations.

This extensive list adds important standard capabilities to XML and HTML5 that are needed for localization. However, even though some of the functionality listed above overlaps Linport functionality, it functions at a much lower level. For example, Linport might specify that the **domain** of an entire project is “legal,” but ITS 2.0 can specify that a short section within a document in fact belongs to another domain (e.g., there is a lengthy quotation from a company’s earnings report [domain = financial] embedded in a legal brief).

As was the case with XLIFF, ITS 2.0 is entirely complementary with Linport. While Linport handles the high-level project details, ITS provides tagging at the block or span level within documents to support specific requirements. From an interoperability standpoint, the ideal situation would be if Linport served as the overall container and project format, XLIFF was used as the payload (content) format, and ITS markup served to add internationalization and localization capabilities within the content.

# Additional Linport advantages

In addition to the advantages and benefits described above, Linport has a number of other advantages, discussed in this section.

## Integration with Quality Assessment practices

Linport allows for translation quality assessment metrics such as QTLaunchPad’s Multidimensional Quality Metrics (MQM). Linport can easily incorporate any quality assessment metric, thus allowing enhanced interaction between translators and reviewers in order to produce higher quality translations. MQM (see http://www.qt21.eu/launchpad/) is an emerging translation quality assessment metric format that is customizable for different projects or documents. QTLP or any other metric format can be contained in a Linport portfolio.

There is increasing awareness in the translation community that translation quality is not a one-size-fits-all proposition and that quality measurements must be adapted to meet particular needs. The 21 parameters defined in the STS provide a way to define quality expectations up front, even before translation begins. MQM uses a subset of the 21 parameters to help users to define relevant quality metrics that assess the issues that matter for a particular task (see <http://bit.ly/dim2issue>). Because Linport includes these parameters at a basic level, it helps not only to set expectations, but to provide relevant ways of ensuring that expectations are met.

## Ease of Implementation

*Linport is a solution that is easy to implement.* Online tools to create and validate a Linport portfolio are currently under development, but all that is really needed to build a Linport portfolio is an operating system with a directory structure. A Linport-style portfolio can be hand-generated in five minutes or less. As long as the Linport guidelines are understood, any reasonably tech-savvy individual can create a viable Linport portfolio.

In contrast to typical complex specifications, Linport portfolios can be unpacked and understood with a minimum of explanation and effort. This simplicity makes it easy for translation tools to add Linport support, helping lower the burden required to implement the specification and reap its benefits. While complex process-driven implementations are possible, basic implementation is straightforward. In addition, the Linport committee is creating a set of free tools for working with Linport portfolios.

## Linport is free.

Linport is non-profit and non-proprietary. Any translation company or individual can use Linport royalty free. The Linport project is currently managed by LTAC Global, a non-profit organization, and the Linport “blueprint” for a standard will be turned over to an existing industry standards body such as OASIS or ETSI when it is reasonably complete. (The Linport committee has chosen initially to operate outside of a standards organization in order to reduce any administrative overhead and costs. When the specification is complete and fully functional, it will be turned over to a standards body where it is anticipated additional changes will be made.)

# Current and future development

Active development of Linport is ongoing and much future work remains to be done. The portfolio data model needs to be refined and formalized, with a schema and validation system, then submitted to a standards body. Various software projects are being developed such as an online portfolio builder, a “splitter” tool that breaks a portfolio into TIPPs, and a merger tool that integrates TIPP responses back into a portfolio.

## How to get involved

There are many ways to get involved in the ongoing development of Linport. You can join the Linport community and participate in monthly Linport conference calls by going to <http://www.linport.org> or joining the GALA Linport community group at <http://www.gala-global.org> (Search for Linport in their search box.) Or contact the authors, Tyler Snow (tylerasnow@byu.edu) and Alan Melby (akmtrg@byu.edu) to contribute real project data that can be used to test the Linport portfolio model. All contributed data must be non-confidential. You can also contribute by testing apps developed for Linport, developing your own Linport applications, or introducing Linport into your company’s translation workflow as an early adopter. The Linport community is open and thankful for any support you and your company are willing to provide.

**References**

International Annual Meeting on Computer-Assisted Translation and Terminology (JIAMCATT). 2012. Retrieved from <http://jiamcatt.org/>.

Linport: The Language Interoperability Portfolio Project. 2012. Retrieved from <http://linport.org>.

Melby, Alan. 2011. “The Seoul of Standards and You.” *The ATA Chronicle.* Oct: 12-16.

Melby, Alan. Chandler, Brian. Lommel, Arle. 2012. “Linport addresses translation package compatibility.” *Multilingual Magazine*. July/Aug: 45-47.

Melby, A., Lommel, A., Rasmussen, N., & Housley, J. 2012. “The Language Interoperability Portfolio (Linport) Project: Towards an Open, Nonproprietary Format for Packaging Translation Materials.” (*Unpublished article submitted to JIAL).*

Structured Specifications and Translation Parameters. 2012. Retrieved from www.ttt.org/specs.

**Glossary of Acronyms Used**

**DGT** – **D**irectorate **G**eneral for **T**ranslation

›A part of the EC; JIAMCATT Partner

›ec.europa.eu/dgs/translation

**EC** – **E**uropean **C**ommission

**ETSI** – European Telecommunications Standards Institute

›www.etsi.org

**IN!** – **I**nteroperability **N**ow**!**

›A group working to improve the interoperability of tools and technology within the localization industry

›code.google.com/p/interoperability-now

**ISO** – **I**nternational **S**tandards **O**rganization

›JIAMCATT Partner

›www.iso.org

**Linport** – The **L**anguage **In**ter-operability **Port**folio Project

›www.linport.org

**LISA** - **L**ocalization **I**ndustry **S**tandards **A**ssociation

›*Ceased to exist March 2011*

**OASIS** - **O**rganization for the **A**dvancement of **S**tructured **I**nformation **S**tandards

›www.oasis-open.org

**QTLP** or **QTLaunchPad** - Quality Translation Launch Pad.

EC-CORDIS-PF7-LT project 296347 cordis.europa.eu/projects/rcn/103949\_en.html (2012-07-01 to 2014-06-30)

›www.qt21.uk

**TAUS** – Translation Automation

›www.translationautomation.com

**TIPP** – Translation Interoperability Protocol Package – an **IN!** project

**XLIFF** – **X**ML **L**ocalisation **I**nterchange **F**ile **F**ormat

›XLIFF 1.2

›XLIFF 2.0

›XLIFF 1.2: docs.