Localization Standards Reader

By Dr. David Filip, davidf@davidf.org, david.filip@adaptcentre.ie

Release Notes 4.0.1 October 31, 2018

This is a copy released for teaching purposes only. The content of this reader was last updated in October 2018. New major version is planned approximately for autumn 2020.

Release Notes 4.0. October 05, 2018

This is a copy released for teaching purposes only. The content of this reader was last updated in October 2018. New major version is planned approximately for autumn 2020.

Content of this version is identical to

[Harvard citation]

Filip, D., 2019. Localization Standards Reader. *Multilingual*, 30 (1), pp. TBD.

[IEEE citation]
D. Filip, “Localization Standards Reader,” *Multilingual*, vol. 30, no. 1, pp. TBD, 2019.

This presentation structure is better suitable for teaching purposes.

About the Author, Disclaimer, and Focus

Author Background

David Filip is the Chair of OASIS XLIFF OMOS TC | Secretary and Lead Editor of OASIS XLIFF TC | former Co-Chair and Editor for W3C ITS 2.0 | GALA TAPICC Steering Committee Member | XLIFF TC Liaison at Unicode ULI TC | Convenor of JTC 1/SC 42/SG 2 Trustworthiness of AI | SC 42 Representative on ISO/IEC JAG Group on Trustworthiness | National mirror chair, NSAI TC 02/SC 18 AI | Head of the Irish National Delegation, ISO/IEC JTC 1/SC 42 AI | NSAI expert to ISO/IEC JTC 1/SC 38 Cloud Computing, ISO/IEC JTC 1 JETI, ISO TC 37/SC 3 Terminology management, SC 4 Language resources, SC 5 Language technology.

David works as a Moravia Research Fellow at the ADAPT Research Centre – Trinity College Dublin, Ireland.

Disclaimer

This Reader represents David's own expert opinion on the matter of standards in Localization. Nothing in this Reader, except when explicitly stated, can be construed as an opinion or recommendation of any of the standardization bodies, associations, consortia etc. that David works with. However, all info provided here is well backed with publicly accessible facts and data.

Focus of this Reader

This Reader does not purport to address everything around standards. To be informative and of practical use, there must be a focus. This focus is being provided via two limitations.

1) We are looking at standards that affect multilingual transformations of content
    a. We are looking at general content lifecycle standards only ad hoc and as far as they have bearing on multilingual transformations.
2) We are looking only at technical standards that are targeting actual technical interoperability, i.e. file or data formats and/or communication protocols.
    a. Abstract metadata, quality, or service level standards are discussed only marginally and only as long as they have bearing on real machine to machine interoperability in localization.
    b. Linguistic Standardization of particular Natural Languages (including English) is only mentioned as far as it touches on machine to machine interoperability in localization.

Terminology

AAMT

Asia-Pacific Association for Machine Translation. This is obviously not a traditional standardization body, yet their Standardization working group is the owner and maintainer of the Universal Terminology Exchange specification, see UTX.

ASTM International

Founded 1898, one of the oldest SDOs in existence, formerly known as American Society for Testing and Materials. ASTM International recently decided not to expand the acronym. Apart from materials and testing, ASTM also engages in service provision standards and this is where the SDO becomes relevant for the localization and translation industry. It's an individual membership based SDO. Companies can join but don't get more than one seat for joining, it's more like sponsorship if they do. ASTM Committee F43 Language Services and Products became relevant for localization and translation industry interoperability standardization by taking ownership of MQM development.

BCP

Best Current Practice – a normative track of RFCs at IETF. BCP # are persistent names that always point to the currently valid RFC issued under the normative BCP track.

Bitext

Text in two natural languages organized as an ordered succession of aligned source and target segment pairs. Such a structure is key for localization transformations, thus bitext standardization is a key to localization interoperability. Translation Memory and Term Base are considered special cases of Bitext where the order of entries is arbitrary.

DERCOM

Verband **DE**utcher **R**edaktions- und **CO**ntent **M**anagement System Herrsteller e.V. which means Association of German Manufacturers of Authoring and Content Management Systems. DERCOM is not a standardization body but a German trade association, they nevertheless decided to provide a German-producers-centric standardized and as they say "non-proprietary" solution to CMS and TMS interoperability. See COTI, which is the only standard produced by this Association.

Essential Patent

[used essentially in a standard] – A patent that cannot be avoided making a standard's implementation.

ETSI

European Telecommunications Standards Institute, "the mother of GSM" (Global System for Mobile communication, originally Groupe Spécial Mobile). One of the world's most influential producers of World Class (yet proprietary) Telecommunications standards. ETSI has a huge base of members in the Telecommunication Industry. Although ETSI has been successfully penetrating IT and IT Security standardization areas, its membership fees (and voting weight) calculation for Enterprises are still based on their annual telecommunications turnover.

ETSI ISG LIS

An Industry Specification Group that was formed in spring 2011 within ETSI to take over LISA OSCAR standards portfolio including related LISA intellectual property. The ISG first sat in August 2011, and had several closed door working meetings since. ISG LIS gave an update to the Localization Community in a public meeting at Localization World Seattle 2012. ETSI ISG LIS republished two of LISA OSCAR standards during its lifetime (See TMX and GMX-V). ETSI ISG LIS was closed down by ETSI on 21st August 2015 due to non-availability of a possible chair from an ETSI member organization.

FRAND

See RAND

GALA

Globalization and Localization Association is not a traditional standardization organization, however as the trade organization bringing together language service providers of all sizes and shapes, it is in a unique position to perform requirements gathering for important interoperability standardization in the industry. The latest GALA initiated standardization effort has been code named TAPICC and is a so far fairly successful attempt to resurrect previous OASIS, TAUS and LT-Innovate attempts at a Translation API (See TAPICC).

HTML

Hyper Text Markup Language is the key content format specification of the Open Web Platform owned by W3C. We are not specifically addressing HTML in this reader, as content formats are out of scope of this reader. For HTML 5's internationalization and localization aspects see ITS.

In 2014, W3C finally managed to publish HTML 5 as a Recommendation. This happened on 28th October 2014 and was a big milestone as it took a tad too long (since 2007) to bring the first HTML 5 version to full approval.

HTML 5.2 was approved as W3C Recommendation on 14th December 2017 https://www.w3.org/TR/2017/REC-html52-20171214/. HTML 5.2 superseded the HTML 5.1 Recommendation from 1st November 2016 https://www.w3.org/TR/2016/REC-html51-20161101/. The latest (always current) HTML 5.2 version (including possibly normative errata published later on) is at this persistent location https://www.w3.org/TR/html52/. The latest (always current) W3C published HTML version is at https://www.w3.org/TR/html/.

The Work in Progress version is 5.3 that is being worked on between W3C and the WHATWG. While the W3C published series of HTML 5, HTML 5.1, 5.2, 5.3 etc. is the official conservative succession of discrete standards, the cutting edge web pages and browsers are working with the HTML "living standard" continuously developed and published by the WHATWG https://html.spec.whatwg.org/multipage/. The latest revision of the HTML Living Standard was from 19th September 2018 at the time of writing this. The developer versions of the Living Standard are updated practically daily with the process known as "nightly builds".

W3C 5.3 is at a Working Draft stage in W3C, meaning the second lowest level of stability, a draft that invites a formal public review. Working Drafts are republished at least every six months until they are deemed stable enough to be frozen as a Candidate Recommendation. Thus HTML 5.3 has never been materially frozen for stability needed to progress towards higher levels of standardization. Same as HTML 5, HTML 5.1, and HTML 5.2, HTML 5.3 at W3C takes over proven and stable solutions from the WHATWG HTML Living Standard with the ultimate goal to standardize a stable version albeit a snapshot that will age rapidly compared to the Living Standard. The main aim of 5.1, 5.2, and 5.3 releases is to formalize the modularization of the HTML technology framework and to provide a fix reference for more conservative implementers of HTML such as governments and international agencies.

ICU TC

On 18th May 2016, the ICU Project Management Committee was transformed into the 4th Technical Committee of the Unicode Consortium, the ICU TC. This effectively changed an independent Open Source reference implementation project under the stewardship of IBM into a TC under direct control of the Unicode Consortium that also hosts the CLDR TC that produces the CLDR that the ICU is supposed to be a reference implementation of. This is arguably a conflict of interest situation as interoperability standardization should be implementation agnostic and reference implementations should be provided by membership or third party developers independently of the standard’s development and not directly controlled by the same SDO.

IETF

Internet Engineering Task Force is a non-membership standardization body. Contributors are strictly individuals that commit implicitly themselves (and their employers) simply by contributing without signing any formal contract. IETF creates Internet related Technical standards, protocols, processes and non-normative informative content as so called RFCs. IETF is institutionally and (partially) financially backed by the Internet Society.

IP

Intellectual Property.

IPR

Intellectual Property Rights.

IPR Mode

[of an SDO or TC] – Intellectual Property Right policies under which an SDO or TC has been chartered. As a general rule a Technical Committee’s Identity cannot possibly survive an IPR mode change. i.e. usually Technical Committees that want to change their IPR mode must re-charter, which requires dissolution of the original TC. This kind of obstruction is useful as a guarantee that standards related Intellectual Property Rights, such as the right to use Essential Patents, will be granted to standard implementers under legally stable conditions.

LDML

See UTS #35.

LISA

Localization Industry Standards Association, declared insolvent on February 28, 2011. See ETSI ISG LIS.

Non-Assertion

Non-Assertion is the most progressive of the RF IPR Modes. Under Non-Assertion, all of the parties participating in standardization effectively grant a Non-Assertion Covenant with respect to IP used essentially when implementing the standards developed within the scope of a TC thus chartered. This is the best IPR Mode with respect to Open Source development because any potentially needed license grant or negotiation is replaced by the Non-Assertion Covenant made upfront. OASIS TCs may be chartered under the Non-Assertion IPR Mode since 2008.

OASIS

Organization for Advancement of Structured Information Standards (formerly called SGML Open). One of the most important IT standardization consortia, based in Massachusetts USA; its sponsors include IBM, Cryptsoft, Intel, Microsoft, Oracle, SAP, and many more. Toolmakers and service providers are also well represented. OASIS concentrate on development of mostly XML based specifications that target interoperability and automation in specific areas, such as authoring, business transactions, web services, and localization. Recently the OASIS focus is on interoperability based on data models that can be expressed with different bindings or serializations. Work items also include high level reference architecture and orchestration standards. Other than XML approaches include abstract models, JSON based browser bindings, YAML etc.

Okapi Framework

Set of localization engineering transformation tools, an Open Source Reference Implementation of the most influential Localization Open Standards, such as XLIFF 1.2, XLIFF 2.0, XLIFF 2.1, TMX, TBX, SRX, ITS 1.0, ITS 2.0 and by virtue of the previous also of the OAXAL reference architecture model. Okapi team is largely driven by Argos Multilingual (Argos acquired ENLASOS in October 2017) Yves Savourel, the father of OpenTag and hence also XLIFF. In 2013, Vistatec, an Irish LSP, made significant contributions to the OKAPI framework, chiefly by releasing Ocelot (a translation and review editor) as part of the framework.

OLIF Consortium

Ad hoc industry consortium driven by SAP set up in 2000 to develop the OLIF standard. See OLIF standard.

OSCAR

LISA's Technical Committee (Special Interest Group) for actual standardization work. Explanation of the acronym is somewhat strained, meaning Open Standards for Container/Content Allowing Reuse. Dissolved along with LISA. See ETSI ISG LIS.

RAND

(also FRAND) – (Fair) Reasonable and Non-Discriminatory. IPR mode that allows owners charging for use of Essential Patents provided that the charge is reasonable and non-discriminatory. This is in contrast to the case where no RAND obligation exists and the patent holder has an effective monopoly over the technology and can theoretically set any conditions including refusing to grant a license at all. Unfortunately, both “reasonable” and “non-discriminatory” are relatively vague, so that their practical extent often needs to be ascertained by legal courts. OASIS TCs can be chartered as RAND, IETF, ISO, Unicode and others use RAND as their sole IPR mode. FRAND is the sole IPR mode of ETSI.

2013 saw some groundbreaking legal development with regard to standards offered under RAND IPR policies. Judge Robart (with a Washington District Court) ruled in a RAND case between Microsoft and Motorola on 25 April, 2013. 9th District Court upheld Judge Robart’s decision on July 30, 2015. Judge Robart with his **Findings of Fact** (FOF) in Microsoft vs Motorola was the first judge ever to provide a **rate setting opinion**. And not only that, the methodology of setting the rates is so well explained and logically constructed that it is being used as precedent not only by other US courts (Apple vs. Motorola) but also outside of US and by policy makers. Many analysts have offered their opinions; while Judge Robart’s ruling does not use a revolutionary methodology it is undertaken rigorously and consistently and the usual **Georgia-Pacific analysis** that uses a hypothetical rates negotiation as its basis has been **modified** to take into account specifics of standards setting under RAND terms. Most importantly, Judge Robart stated that the hypothetical rates negotiation takes place in the context of a RAND obligation and in the context of **royalty stacking** due to many patented technologies used essentially in a standard available under RAND conditions. The goal of the Standards Setting Organizations operating under RAND is to expand market for all its members through better interoperability, which is in public interest, and **withholding patents** – that by the way gained traction due to being included in a standard – is a **breach of contract**. As a consequence, because the FOF did set specific bands and rates, Judge Robart has been able to rule (in the Conclusions of Law following the FOF) that Motorola’s initial licensing offer had not been made in good faith and therefore did constitute a **breach of contract**, in particular the **obligation to license the essential patents under RAND conditions to all implementers**.

REST, RESTful

REST means Representational State Transfer, this is a widely known web services architectural best practice that has however never been officially standardized. Unlike SOAP, which is a specific standardized protocol, there is no single REST standard or even a single best practice, REST is rather an architectural style for interoperability via APIs (Application Programming Interface) on the Internet that can have varied instantiations and implementations. REST makes use of other standards, HTTP (Hyper Text Transfer Protocol and URI (Uniform Resource Identifier) are at the core, the exchanged messages are usually in the form of HTML, XML, or JSON.

RF

Royalty Free. IPR mode that mandates and guarantees Royalty Free use of Essential Patents in order to implement a standard. OASIS TCs can be chartered as RF (as is XLIFF TC) and RF is the sole IPR mode of W3C. See also Non-Assertion, which is a special case of RF.

RFC

Request for Comments – IETF Recommendation of a technical or process solution that can be normative or informative. RFC # are not persistent identifiers. RFCs may be updated or obsoleted by other RFCs with higher sequential number.

SC

Subcommittee. Capacity of subcommittees to develop standards or not fully depend on the mandate that the parent TC give them. The extent of the mandate is restricted by the parent TCs scope and may be also restricted by the policies of the governing SDO. We don't use SC in its other common meaning Steering Committee.

SDO

Standards Developing Organization, is considered synonymous with SSO.

SSO

Standards Setting Organization, is considered synonymous with SDO.

TC

Technical Committee – Standardization bodies or consortia usually own, create, maintain, and update technical standards through purpose specific Technical Committees. Many of them indeed do call them Technical Committees or Subcommittees in their organizational structures (e.g. OASIS, Unicode, ISO), other don't (e.g. W3C). Here we use Technical Committee in the generic sense, as any formal group that owns (creates/maintains/updates) a technical standard, even though it might be called something else in its home organization (Industry Specification Group, Working Group, Special Interest Group etc.)

UAX

Unicode Standard Annex, see Unicode Consortium. UAX # are persistent names that always point to the actual Revision #.

Unicode Consortium

Unicode Consortium. Home of the Unicode Standard and CLDR. Unicode also publishes specialized Technical Reports, UTR#. Some of these reports are Unicode Annexes, UAX#, i.e. essential parts of the current Unicode Standard release, however published as separate documents. Unicode also produces Unicode Technical Standards (UTS#). These are independent of the Unicode Standard and currently specify Unicode Standard or CLDR related algorithms, formats, mark up languages etc. Unicode Consortium currently has four (3) TCs: UTC, CLDR TC, and ICU TC. ULI TC has been transformed into a CLDR TC Subcommittee, so that it is now CLDR TC/ULI SC.

ULI

Unicode Localization Interoperability TC. The 3rd Unicode Consortium Technical Committee, formed in April 2011. However, ULI TC has been transformed in early 2018 into a CLDR TC Subcommittee, so that it is now CLDR TC/ULI SC.

ULI has not chartered creating its own standards, instead it is looking into localization interoperability related standards behaviors and profiling. ULI has been engaging in revisions of UAX #29 behavior and related SRX and XLIFF developments (see the respective standards). ULI driven segmentation exceptions are being released as part of CLDR as of Release 24. ULI charter is rather broad and the committee now considers relationships of the segmentation behavior with XLIFF and SRX. Recently ULI exploited linked open data and NLP techniques to mine data for its segmentation exceptions algorithms. After the ETSI ISG LIS disbanded in September 2015, ULI TC agreed to host SRX as publicly available specification.

UTR

Unicode Technical Report, see Unicode Consortium. UTR # are persistent names that always point to the actual Revision #.

UTS

Unicode Technical Standard, see Unicode Consortium. UTS # are persistent names that always point to the actual Revision #.

W3C

Worldwide Web Consortium. W3C owns many key standards including XML, HTML (see HTML) and CSS. Most relevant for our industry is the Internationalization Activity, which – apart from its core WG working as an Internationalization gatekeeper for other W3C groups – hosts ITS IG. ITS IG oversees usage and development needs for ITS 2.0 that had been developed and published by the MultilingualWeb-LT WG. Note that W3C has loosened the activities structure in 2012 and groups can be now part of more than one Activity. The Internationalization activity has received a boost in 2018 and works on several specifications of layout requirements for specific scrips, such as Arabic, Korean. This development has been encouraged by the success of Japanese and Chinese layout requirements. This development is especially prominent after IDPF merged with W3C in early 2017 and W3C became the home of the EPUB format.

WG

Working Group, W3C and IETF term for Technical Committee that is mandated to produce W3C recommendations (i.e. standards in W3C speak) or RFCs (that can be normative or informational in IETF). Working Groups also exist in ISO TCs and SCs. ISO Working Groups work in the so called "expert mode" and draft standards in early stages before these are progressed or rejected in the "national body mode" by the parent TC or SC.

Overview of Relevant Technical Standards in Alphabetical Order

BCP 14
RFC 2119 - Key words for use in RFCs to Indicate Requirement Levels
RFC 8174 - Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words

Standard Short Name

BCP 14

Owner

IETF

IPR Mode

RAND

Current version

For a very long time (20 years), BCP 14 had just the RFC 2119 as its content, so that many standards makers treated “BCP 14” and “RFC 2119” as synonyms. With the addition of RFC 8174 (RFC 2119 Clarification) it became important to distinguish if a standardization work product references BCP 14 as a whole or just the RFC 2119. RFC 8174 makes it clear beyond any reasonable doubt that the IETF notion of normative keywords is typography driven. “MUST” (or any other of the defined keywords) has its normative meaning only as far as it is printed in UPPERCASE, as opposed to the ISO notion of normative keywords that is concept driven and in fact forbids typographical distinction of normative keywords.

RFC 2119, released for unlimited distribution in March 1997.

RFC 8174 updates RFC 2119 and was released for unlimited distribution in May 2017.

Current WIP

In fact, it is undesirable to make changes to this BCP because so many normative texts across IETF, W3C, OASIS etc. depend on the meaning of the normative keywords as set out here. It is questionable whether adding RFC 8174 was a good thing, as it deepened the rift between the IETF and ISO notions of normative keywords. It is too early to tell what the impact of the clarification was. Before, if UPPERCASE was omitted on a keyword in a spec you could fix it in later publishing stages as it was considered editorial. In theory, it sounds clear to tell that only UPPERCASESD keywords are keywords but in practice, it will make corrections difficult as changing from plain to UPPERCASE has been made to a material change by this supposed clarification. This clarification has some potential to cause interpretation damage. Until UPPERCASING was made mandatory, editors were dutifully aiming to avoid the keywords and use non-normative semantic variants. Now, many will stop doing that citing RFC 8174, causing lot of confusion in ordinary readers. It may well happen that this clarification will have a knock on effect and might cause more changes to this BCP in the feature.

Short Description

BCP 14 defines the standardization specific meaning of normative keywords such as MUST, MUST NOT, OPTIONAL, REQUIRED, RECOMMENDED etc. RFC 2119 is the most common normative reference in other specifications throughout Information Technology standardization bodies. It is worth noting that ISO uses a slightly different set of normative keywords without any typographic distinction.

Localization related standards (same as other industry standards) such as ITS and XLIFF are using the BCP 14 key words to make their normative statements that create the basis of conformance statements, testing, and verification.

BCP 47
Tags for Identifying Languages and Matching of Language Tags

Standard Short Name

BCP 47

Owner

IETF

IPR Mode

RAND

Current version

The current version consists of:

RFC 5646, Tags for Identifying Languages, released for unlimited distribution in Sep 2009.
RFC 4647, Matching of Language Tags, released for unlimited distribution in Sep 2006.

BCP 47 is a persistent name that always points to the latest released RFCs in the track, no matter what the current RFC numbers.

Current WIP

N/A

BCP 47 itself is stable, which is important for backwards and forwards compatibility. New tags are being continuously registered via registration authorities specified in the standard. Most current developments are connected to Unicode Extensions for BCP 47; see BCP 47 Extension U and BCP 47 Extension T.

Short Description

BCP 47 is a normative IETF track that compiles recommendations how to create a unique language tag from codes defined in several other normative sources (including ISO two and three letter codes, script, country and region codes). It is frequently normatively referenced from OASIS, Unicode, W3C etc. standards.

BCP 47 Extension T
Transformed Content

Standard Short Name

BCP 47 Extension T

Owner/ Maintainer

IETF/ Unicode Consortium

IPR Mode

RAND

Current version

Informational RFC 6497, published in Feb 2012.

Current WIP

Extension T is regularly maintained as part of the CLDR release cycle, see CLDR

Short Description

Extension T is possible via the extensibility mechanism defined in BCP 47 (RFC 5646) itself. Extension T has been specified in an Informational RFC 6497, so that it has not the Internet Standard Status from the point of view of IETF. It has however normative status within Unicode Consortium as it is being maintained as part of CLDR, which is their major normative deliverable. This extension allows for additional tags specifying, from which other language, locale or script the content at hand had been transformed. Extension T is not recommended for usage in structured environments such as XML, where this type of metadata can be specified using mark up solutions rather than a single text field. Note that extension T is appending the information about the originating language/locale with a leading "–t", which obviously means that the BCP tag starts with the target locale and the source locale is being appended. This makes sense given the structure of BCP 47 tags, but may be perceived as contrary to the customary listing order of source and target languages, so e.g. "EN-t-IT" actually means that the tagged content is English but was transformed from Italian, not the other way round.

BCP 47 Extension U
Unicode Locale Extension ('u') for BCP 47

Standard Short Name

BCP 47 Extension U

Owner/ Maintainer

IETF/ Unicode Consortium

IPR Mode

RAND

Current version

Informational RFC 6067, published in Dec 2010, maintained by Unicode Consortium as part of CLDR, see CLDR.

Current WIP

Extension U is regularly maintained as part of the CLDR release cycle, see CLDR. ULI TC delivers the input exception data for sentence breaking mechanisms per different locales for the periodic CLDR releases.

As result, sentence breaking behaviors driven by different exception data can be specified through assigned keys under the extension ‘u’ mechanism.

Notably the canonicalization algorithm included in RFC 6067 is slightly out of date. The current provision is specified in UTS #35. The canonicalization algorithm should be therefore updated in RFC 6067.

Short Description

Extension U is possible via the extensibility mechanism defined in BCP 47 (RFC 5646) itself. Extension U has been specified in an Informational RFC 6067, so that it has not the Internet Standard Status from the point of view of IETF. It has however normative status within Unicode Consortium as it is being maintained as part of CLDR, which is their major normative deliverable. Additional tags for fine grained locale identification are appended to the BCP 47 tag with leading “-u”.

CLDR

Standard Short Name

CLDR

Owner

Unicode Consortium

IPR Mode

RAND

Current version

v34 (released on October 15, 2018)

v34.1 (expected in June 2019).

Current Work In Progress

Submission period for version 35 will start in April, 2019 and the release is currently planned for October 2019. CLDR is being released on a regular semi-annual schedule, whereas the cycle starting in Q4 of each year is focused on tooling and bug fixing and usually skips the public data submission phase.

Short Description

Unicode Common Locale Data Repository, http://cldr.unicode.org/. Standard repository of internationalization building blocks, such as date, time and currency formats, sorting (collation) rules, etc.

CLDR is not a standard in a classical sense. It is – as the name suggests – a repository, that is being constantly updated and released on a rolling basis following its Data Release Process (http://cldr.unicode.org/index/process#TOC-Data-Release-Process)

Unicode extensions U and T are maintained as part of the CLDR regular release cycle. CLDR also contains segmentation exception data provided by the ULI SC as of release 24 (ULI TC then).

COTI
COmmon Translation Interface

Standard Short Name

COTI

Owner

DERCOM

IPR Mode

Unclear. The specification, documentation (chm), validation artifacts (wsdl, xsd) are available for free, but no restrictions seems to have been specified with regards to licensing of IPR essential to implementing the specification. The specification says to lookup “the intellectual property rights section of the technical committee web” at http://www.dercom.de/ but there seems to be no such a section. This unfortunately means that not even (F)RAND restrictions apply for licensing offers made for essential IPR by their respective owners. This is at least until DERCOM actually places a publicly visible IPR information.

Current version

COTI Specification Version 1.1.1, from July 07, 2017, which is a minor update from COTI Specification Version 1.1, from April 29, 2016. The 1.0 Version was first approved on May 25, 2014 and was subject to only minor editorial fixes by September 2015. Version 1.1 brought some additional material such as the XML Schema, but also changes in the WSDL artifacts.

Current WIP

There doesn’t seem to be a specific plan to further develop the specification. Rather it seems that DERCOM makes changes and fixes as its membership gradually progresses in fulfilling its obligation to implement COTI level 1 through COTI level 3 compliance. The wsdl changes in the version 1.1 were most probably driven by implementation experience of the DERCOM membership.

Short Description

DERCOM claims that it had to step up to protect and promote the interests of its membership and – at the same time – to let the translation providers use the TMS of their choice. DERCOM argues that no such interface exists and it is true that a common interface is better than the present jungle of proprietary APIs on both ends that can only provide interoperability after some custom development on both ends for EACH new interface. Arguably this can be addressed by an Enterprise Service Bus or a messaging architecture, but not every CMS owner has the resources or technical muscle to do that.

COTI level 3 is SOAP based and provides a quite sophisticated state machine that provides synchronous support for translation orders, cancellations and even updates of existing orders, which is notoriously tricky. While DERCOM is right that no translation webservice interface has been standardized, it is surprising that COTI only standardized the business metadata interface and doesn't say anything about payload standardization, standardization over a canonical data model such as XLIFF lends itself, but COTI actually does not mandate any data model restrictions with regards to the payload. So actually this specification only solves the CMS part of the equation. The bundle is thrown over the wall and it's up to the LSP to figure out what – if any – is the payload structure, what is localizable and what not and so on.

While German CMS providers might be happy to implement always synchronous SOAP webservices, the general trend seems to be rather towards RESTful interfaces and microservices and it might hinder the adoption of COTI that the level 3 strictly enforces synchronous SOAP calls.

Level 1 is actually the specification of the payload ZIP package, its structure and manifest.. Sounds familiar? People just keep reinventing this very wheel, albeit each time with subtle differences that don't allow for automation among those "standard" ZIP packages. Electronic Dossier, Linport, TIPP, and now COTI level 1, all of these just defined a folder structure that they think a translation request and response should contain without talking to each other. Alan Melby and others invested some diplomatic effort into making the Electronic Dossier people talk to the Linport people and merge the Linport project with TIPP. But even before this could bear fruit, DERCOM COTI brought another "standard" folder structure.

COTI level 2 mandates automated exchange of the structured ZIP packages over hot (watched) folders. Whatever is placed in a hot folder is up for grabs for the receiving end. The TMS watches the CMS's hot folder for new projects and the CMS watches the TMS's hot folder for responses. Sounds like 90s?

So the interesting contribution to the topic of standardized Translation Webservices (which was by the way a name of an OASIS Technical Committee convened in late 2002 which attempted to standardize just this but was closed without fruition back in 2008) is in the level 3 compliance that requires SOAP and doesn't allow or specify any other admissible bindings, which would be certainly advisable.

Internationalization Tag Set

Standard Short Name

ITS

Owner/ Maintainer

W3C MultilingualWeb-LT Working Group/ ITS Interest Group

W3C ITS Interest Group (IG within Internationalization Activity) has become the informal maintainer of ITS 2.0 after the MultilingualWeb-LT Working Group mandate expired, however Interest Groups in W3C cannot create normative deliverables. Therefore, a new Working Group will have to be formed to commence work on the successor standard, as soon as the ITS IG identifies industry need for a successor

Version. However, the status of the ITS IG is unclear as its Charter expired on 31st December 2018 and it is not clear at the moment if the Charter can be extended. In case ITS IG lapses, stakehoders of ITS should initiate a W3C community group to look after the adoption and maintenance of ITS 2.0. The main purpose of such a group is to see if there's a need and momentum to create a new feature or maintenance release of the Internationalization Tag Set.

IPR Mode

RF

Current version

2.0, published as W3C Recommendation on 29th October 2013.

Current Work In Progress

New major version 2.0 has been published in W3C as full Recommendation on 29th October 2013, which means that work on another major Version is not imminent. This said, we need to note that ITS 2.0 covered a number of new areas with non-normative mappings. These include the RDF and the XLIFF mappings. See XLIFF for inclusion of the ITS functionality as an XLIFF 2.1 module. Current collaborations and liaisons include OASIS DITA, DockBook, XLIFF, and XLIFF OMOS TCs, as well as GALA TAPICC.

ITS IG is worked on further elaborating and maintaining other informative mappings, such as NIF or MQM. However, the MQM mapping aspect transitioned to the MQM W3C Community Group. ITS IG also maintained ITS extensions that support categories that could not be normatively specified for various reasons within the ITS 2.0 publishing schedule, for instance the Readiness data category.

Short Description

ITS 2.0 comprises 19 metadata categories compared to mere 7 in ITS 1.0. Also ITS 1.0 metadata categories were primarily designed for internationalization of XML content. Nevertheless, as abstract data categories, ITS can be implemented in non-XML environments and also can have a life cycle throughout the localization roundtrip of content. Importantly, ITS 2.0 normatively specifies usage of the old and new ITS data categories for XML and HTML 5 content, also new categories have been introduced that explicitly address the localization roundtrip, for instance localization quality assurance related data categories Localization Quality Rating and Localization Quality Issue. Importantly, ITS 2.0 is listed in the JTC 1 Big Data Standards Roadmap (ISO/IEC TR 20547-5) as a key automation enabler for Big Data and analytics dealing with human language.

Owners of ITS decorated content want their internationalization and localization related metadata to inform the roundtrip and make it to the target content in a meaningfully processed state that allows for drilling down into the process and for reconstructing the audit trail. Localization workflow managers should pay attention to information flows directed by the ITS data categories introduced by their customers up in the tool chain. There is also potential to introduce automated or semi-automated ITS decoration steps before extraction, or to introduce relevant XLIFF mappings of ITS data categories recorded on translations during the localization roundtrip that can be imported back into ITS in XML or HTML on merging back of localized content. Categories like Translate (that had become a native HTML 5 attribute), Elements Within Text, Locale Filter, Target Pointer, or External Resource should drive extraction and merging back of localizable content. Terminology and Text Analytics disambiguation markup should be passed onto human and machine translators. Proper interpretation of Directionality markup is a must for sound handling of bidirectional content using Arabic or Hebrew scripts. Self-reported Machine Translation Confidence should be passed onto the content recipients, possibly along with quality assurance related metadata (Localization Quality Issue and Localization Quality Rating).

The above considerations are especially valid when hooking up existing localization workflows upwards into the tool chain. Existing workflows should introduce mappings of ITS data categories used in source content, so that the metadata flow is not broken throughout the content life cycle, of which localization workflow is a critical value adding segment.

All current ITS 2.0 categories: Translate (flag indicating translatability or not), Localization Note (for alerts, hints, instructions), Terminology (to identify terms and non-terms and optionally provide definitions), Directionality (manages left to right/right to left display behaviors of content portions), Language Information (BCP47 language tags on relevant content portions), Elements within Text (shows which elements break flow or not to help encode segmentation), Domain (to identify content theme or topic for better choice of relevant services or creation of training corpora), Text Analysis (for inclusion of automatically provided disambiguation data and term candidates), Locale Filter (indicates which portions of text are relevant for which locales), Provenance (tracks agents – machine or human – involved in content transformations) , External Resource (global pointers to external localizable resources such as images or other binary data), Target Pointer (identifies transformation targets in multilingual documents), ID Value (rule to provide unique content identifiers to be maintained throughout transformations), Preserve Space (specifies whitespace handling) , Localization Quality Issue (provides a way to mark up and classify specific language quality assurance issues), Localization Rating (quality rating expressed a single 0-100 score), MT Confidence (provides self-reported MT quality score as a single 0-1 number), Allowed Characters (specifies restrictions on character data using simple regular expressions), Storage Size (a simple encoding based restriction mechanism to make sure that translations fit restricted database fields, forms and so on).

ICU - International Components for Unicode

Standard Short Name

ICU

Owner

Unicode Consortium

IPR Mode

RAND

Current version

ICU 63 for Unicode 11.0 and CLDR 34

ICU4J 63.1 released on October 10 2018. This URL http://icu-project.org/apiref/icu4j/index.html always points to the current official version.

ICU4C 63.1 released on October 10 2018. This URL http://icu-project.org/apiref/icu4c/ always points to the current official version.

Current Work In Progress

ICU (both the Java and C projects) is being developed continually to address bug fixes and dependencies. Major releases are driven by the releases of the Unicode Standard and changes in CLDR data. New major release should come after stabilization of CLDR 35 data in May 2019. Interestingly, in summer 2018, all ICU infra migrated from SVN to GitHub.

Short Description

ICU was until May 2016 an IBM driven open source project, in fact the most important reference implementation of both Unicode and CLDR. As ICU provides internationalization libraries, it actually consists of two subprojects ICU4C and ICU4J that provide libraries for C and C++ and Java respectively. Spectacularly, ICU is the common denominator of both Android and iPhone Operating Systems. ICU 58 was the first ICU version released under the ICU TC governance.

JLIFF (JSON Localization Interchange Fragment Format)

Standard Short Name

JLIFF

Owner

OASIS XLIFF OMOS TC

IPR Mode

Non-Assertion (RF)

Current version

Pre-release

Current Work In Progress

In summer 2018, JLIFF JSON 2.0 and 2.1 schemas were completed by the committee on the XLIFF OMOS TC GitHub repository. The TC started working on a prose specification to be reviewed hopefully in early 2019. JLIFF schema has been attracting implementers even before becoming stable. Notably, Vistatec released an open source implementation of core JLIFF in March 2018 https://github.com/vistatec/JliffGraphTools. But the TC is aware of at least two early implementations that cannot be publicly disclosed at the moment. Seems that the capability of JLIFF to create fragments compatible with the XLIFF 2 data model is exactly what implementers need for real time API based interoperability.

Importantly, the JLIFF format will be reused by the GALA TAPICC Track 2, when specifying their real time translation API (JRARTEBU – JLIFF REST API for Real Time Exchange of Bitext Units).

XLIFF OMOS (Object Model and Other Serializations) TC has been chartered in December 2015 and started to develop a JSON serialization in parallel to creation of the abstract object model for the XLIFF 2 family of standards that it is developing as its first priority. The objective is to create such a JSON serialization of XLIFF and XLIFF fragment data categories that would allow lossless interchange between JLIFF and XLIFF based tool stacks. The development of the JLIFF specification and artifacts is being conducted on this GitHub repository: https://github.com/oasis-tcs/xliff-omos-jliff. The repository is public but the contributors need to be or become OASIS XLIFF OMOS TC members. Anyone can raise issues or comments though via the associated "Issues" or "Wiki" https://github.com/oasis-tcs/xliff-omos-jliff/wiki, as well as minor bug fixes via pull requests. It's up to the repository maintainers if those pull requests will be merged or not (the usual GitHub collaboration workflow).

Short Description

JLIFF aims to be compliant with the Abstract Object Model defined in UML diagrams and a prose specification as XLIFF OM. JLIFF is currently available as a JSON Schema that is reasonably stable but not yet officially published by OASIS.

JLIFF's main intended use case is the real time interoperability of TMS and CAT tools that support the XLIFF 2 data model. Importantly and unlike XLIFF, JLIFF can support exchange of fragments at unit, group, and file levels. For full XLIFF interoperability, it also supports exchange of whole XLIFF Document JLIFF equivalents, but it's not the main use case. It is important to stress that in order to preserve data integrity the lowest interchange level is that of logically separate bitext units and not arbitrary segments.

LDML

See UTS #35

MQM
Multidimensional Quality Metrics

Standard Short Name

MQM

Owner

ASTM
W3C

IPR Mode

RAND at ASTM

No IPR Mode under W3C, as it is not being developed under the Recommendation Track, just republishing the ASTM maintained structure of data categories

Current version

1.0, December 30, 2015; http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

This URL leads to the latest published version http://www.qt21.eu/mqm-definition/

Current Work In Progress

The W3C MQM Community Group is being reactivated to publish updates from ASTM F43 https://www.astm.org/DATABASE.CART/WORKITEMS/WK46396.htm

Short Description

Multidimensional Quality Metrics (MQM) is primarily a set of error data categories that can be used in Language Quality Assessment especially in bilingual translation scenarios but also in monolingual review or authoring scenarios. The aim of MQM is to cover the whole logical space of possible errors and non-conformities related to language and locale specific presentation of content of any kind.

Since it is impossible (and even if possible not likely very useful) to use all the error data categories from all levels and sub-levels. Profiling is a key method defined by MQM, so that in any particular project or job a particular profile or subset of relevant error data categories has to be defined based on stakeholder requirements and expectations. Thus MQM concentrates on quality in the sense of fitness for purpose. It specifically rejects the notion of some general or abstract quality without defining the purpose, related requirements, and a resulting relevant set of possible error types. Also notable is that MQM doesn't define any severity levels. There is an option for implementers to predefine severity levels and scoring methods based on any particular MQM subset.

Currently MQM is a proper superset of TQF. MQM and TQF had been developed separately by the QT21 EU project (coordinated by DFKI) http://www.qt21.eu/quality-metrics/ and TAUS respectively. The EC funders effectively forced these two projects to reconcile their data models.

For effective exchange of MQM/TQF metadata within projects and jobs, inline capturing and encoding of the recorded errors is critical. This needs to be implemented as the ITS Localization Quality Issue data category in native formats or via the same data category implemented as a XLIFF Version 2.1 Module in bitext. The inline encoding of MQM follows this mapping guidance:
https://www.w3.org/TR/its20/#lqissue-typevalues,
https://www.w3.org/International/its/wiki/Localization_quality_types_mappings, and
http://www.qt21.eu/mqm-definition/definition-2015-12-30.html#relationship-its.

OLIF
Open Lexicon Interchange Format

Standard Short Name

OLIF

Owner

OLIF Consortium

IPR Mode

Unclear. The specifications and schemas are available for free, but no IPR mode seems to have been specified.

Current version

v.2.1 (v.2)

Current Work In Progress

v.3 is in Beta since 2008, no current work seems to be under way.

Short Description

OLIF is a stable and relatively widely used lexicon interchange format it has a rich metadata structure and allows for exchange of complex lexicon entries for various purposes such as terminology management, machine translation etc. OLIF had been designed for use in both monolingual and multilingual context via cross-linking of "mono" elements.

TAPICC (Translation API Cases and Classes)

Standard Short Name

TAPICC

Owner

GALA TAPPIC Group

IPR Mode

The Steering Committee originally targeted Non-Assertion (RF) but instead opted to be organized as an Open Source project licensed under the BSD 3 Clause license for code and Creative Commons 2.0 BY for documentation. GALA decide not to mimic an SDO IPR Mode but to donate stable deliverables to OASIS XLIFF or OASIS XLIFF OMOS Technical Committees, where these would be maintained under RF or Non-Assertion IPR modes respectively.

Current version

XLIFF 2 Extraction and Merging Best Practice (XLIFF EMBP), Version 1.0

https://galaglobal.github.io/TAPICC/T1/WG3/prd01/XLIFF-EM-BP-V1.0-prd01.xhtml

https://galaglobal.github.io/TAPICC/T1/WG3/XLIFF-EM-BP-V1.0-LP.xhtml returns the latest published GALA TAPICC version.

Several other deliverables have not yet reached stability but engaged in a number of public consultations via GALA webinars etc.

Current Work In Progress

The XLIFF EBMP, Version 1.0 was contributed by GALA to the OASIS XLIFF TC for publication as a TC note.

TAPICC is in a staged development phase where the Track 1 chartered on February 08 2017 has 4 active Working Groups. T1/WG1 was chartered to develop a consensual set of business metadata that will enable efficient payload exchange. T1/WG 2 was chartered to specify what kind of payload will be allowed to be exchanged via the TAPICC API and how to resolve potential conflicts between payload encoded metadata and the business level imposed metadata. T1/WG3 was chartered to work on best practices of creating and consuming XLIFF 2 including documenting public libraries and services. T1/WG4 was chartered to design an actual RESTful API that will work with data models specified by the other T1 Working Groups, most importantly the T1/WG1. The joint purpose of the T1 groups is to enable asynchronous API exchange within the complex Localization supply chain, while relying on an XLIFF 2 based canonical data model.

In October and November 2018, TAPICC chartered and launched a call for volunteers for T2/WG1 (widely publicized in tcworld 2018 in Stuttgart on November 14, 2018). T2/WG1 - JLIFF REST API for Real Time Exchange of Bitext Units (JRARTEBU) has currently been assigned 1 Work Package: WP1 Unit Exchange. Batch or buffered exchange is out of scope of WP1 although not out of scope of Track 2 or T2/WG1. SOAP, any other protocols or bindings are out of scope of T2/WG1 although not out of scope of TAPICC Track 2. Even based on the name of T2/WG1, it is clear and TAPICC Track 2 is going to reuse the OASIS XLIFF OMOS TC specified JLIFF format, the JSON serialization of the XLIFF 2 based object model. This synchronicity of data models will ensure semantic and behavioral interoperability between the asynchronous T1 and real-time transactional T2 bitext interchange.

Short Description

FEISGILTT Montreal hosted on 26th October the very first TAPICC Symposium where the stakeholders from the GALA and XLIFF communities discussed the new effort in the context of payload exchange formats such as XLIFF and TBX, as well as business oriented Translation project data exchange initiatives such as Linport, TIPP and DERCOM COTI.

The first TAPICC Symposium brought a four way categorization of use cases that were considered in scope of the Translation API (TAPI) standardization. 1) Exchange of a payload blackbox accompanied with rich business or project metadata. In this area, TAPICC wants to abstract common business metadata used in other initiatives past and current including Linport, TIPP, COTI, XLIFF 1.2 project group, OASIS Translation Services TC (closed) and so on. 2) Bidirectional real time exchange of XLIFF unit data in arbitrary serializations as well as 3) Enriching of XLIFF units in arbitrary serializations with terminology, translation suggestions, text analysis, process and quality assurance metadata etc. 4) exchange of data for layout representation purposes during the bitext management process.

Scenarios 2) and 3) deemed to be dependent on Work in Progress within OASIS XLIFF OMOS TC and hence put on the back burner at TAPICC until XLIFF OMOS TC produces the required abstract data models and JSON (JLIFF) and other non-XML serializations. Although scenario 4) was deemed in scope of the TAPICC project, neither GALA nor XLIFF OMOS TC have resources to tackle the scenario at the moment, therefore volunteers to engage in the API specification for this scenario are sought. The groundwork for this has been done at OASIS XLIFF TC and the provisions of the XLIFF 2 resource data module can be used to address this. Work on non-XML payload exchange would have a dependency on the XLIFF OMOS work in progress for scenarios 2) and 3).

SOAP (Simple Object Access Protocol)

Standard Short Name

SOAP

Owner/ Maintainer

W3C XML Protocol Working Group (closed) / no current maintainer

The SOAP 1.2 multipart recommendation was produced by the XML Protocol Working Group at W3C, the group was closed on 10th July 2009.

IPR Mode

RF

Current version

1.2 (Second Edition), published as W3C Recommendation on 27th April 2017.

Current Work In Progress

N/A, the protocol as such is not being developed or maintained at W3C or elsewhere. We list SOAP in this reader because it was resurrected in our industry by the COTI level 3 conformance requirement to implement a SOAP based web service automation.

Short Description

SOAP (the expansion of the acronym is not in use since at least 2007) is an XML based web services protocol. Version 1.0 was submitted to IETF in autumn 1999 as an Internet Draft but never reached an RFC status, so it actually hadn't become a standard at IETF. SOAP 1.1 had Note status at W3C and the only SOAP version that ever reached the standard status (W3C Recommendation) is SOAP 1.2.

The SOAP protocol consists of 5 layers: message format, transfer protocol bindings, message processing models, message exchange patterns, and extensibility. Unlike REST that is strictly HTTP based, SOAP is protocol neutral and can work over several lower level protocols such as HTTP (Hypertext Transfer

Protocol), SMTP (Simple Mail Transfer Protocol), TCP (Transmission Control Protocol), UDP (User Datagram Protocol, or JMS (Java Message Service).

TBX

Standard Short Name

TBX, ISO 30042

Owner

ISO TC 37/SC 3

IPR Mode

FRAND (within ETSI)

RAND (within ISO)

BSD 3 clause licensed open source project at LTAC/TerminOrgs

Current version

ISO 30042:2008

TBX-Basic 3.1 (the version maintained by LTAC/TerminOrgs, last updated Sep 12 2014) This is the latest version of TBX-Basic compatible with ISO 30042:2008.

Although the revised (2nd Edition) ISO 30042 hasn't been published yet – it will hopefully become ISO 30042:2019, as it currently is ISO FDIS 30042. It is in a sense the current version as no further technical work is possible on this version since August 2018 and it should appear within a month of this Issue of Multilingual being shipped.

Although the underlying ISO Standard has not yet been officially published. LTAC has already released industry oriented dialects compliant with ISO FDIS 30042 (and thus hopefully ISO 30042:2019) in the very near future. These dialects are available through GitHub.io webpages. Notably TBX-Basic Dialect Version 1.0 (2018-06-12) https://ltac-global.github.io/TBX-Basic_dialect/. LTAC also published a robust developer guide to the public dialects compliant with the new modular TBX design http://www.tbxinfo.net/tbx-downloads/.

Current Work In Progress

Although the ISO DIS 30042 ballot held in 2018 didn't fail per se, the Draft international Standard (DIS) received a large number of substantial comments that lead to technical changes. It wasn't surprising as it was clear even during the TC 37/SC 3 meeting in June 2017 in Vienna that the Committee Draft (CD) was far from stable and should have gone for 2nd CD rather than for a DIS ballot. Per ISO publishing procedures, a DIS cannot proceed to publication as an International Standard in case the DIS ballot and the following Comment Resolution Meeting (CRM) necessitates technical changes. Therefore, SC 3 spent a better part of Q3 2018 preparing the Final Draft International Standard (FDIS) text. The FDIS ballot launched in November 2018 and should yield the officially published International Standard (IS) in February 2019.

The most important data model change to happen in ISO 30042:2019 is making the inline data model compliant with the XLIFF 2 inline data model. This was implemented by the LTAC/Terminorgs TBX Steering Committee and TC 37/SC 3 with input from XLIFF and XLIFF OMOS TCs. The modular public dialects (Min, Basic, and Linguist) compatible with the 2nd edition of ISO 30042 were already released, also with the new inline data model.

The second ISO edition of TBX accumulated quite a number of breaking changes. But most of those changes are dealing with modernization of the XML tooling. Apart from the inline data model change and related introduction of explicit directionality support, it is important to stress that TBX joined other modern day standards in leaving its former monolithic design for a new modular design.

2nd ISO edition TBX specifies a non-negotiable core and prescribes that all compliant dialects must include that core. The standard then specifies a modularity mechanism. Outside of ISO, LTAC made sure that the most common dialects are extended from the core using a telescoping principle. The simplest possible dialect is TBX-Core, the TBX-Min is composed from the Core and Min modules. TBX-Basic is TBX-Min plus the Basic module, and TBX-Linguist is TBX-Basic plus the Linguist module. The stakeholders have high hopes that this will vastly improve the so called “blind” or plug and play interoperability. The second ISO TBX edition also defines TBX agents and provide a compliance clause targeting the document compliance as well as specific agents’ compliance. Notably, this edition introduces namespace based modules, albeit only in the DCT (data category as tag name) style, which is a compromise that should not disturb those who are not worried with the DCA (data category as attribute name) style. And vice versa, implementers who are afraid of implementing namespaces, can stick to the old DCA style. The good thing is that both DCA and DCT are extended from the same core and modules specified in either DCA or DCT are semantically equivalent (based on the same sets of additional data categories). DCT is easier to validate and it’s also easier to filter out unsupported modules (as those are in different namespaces) in the DCT style.

ISO 30042:2008 was identical with TBX 2.0 developed at LISA OSCAR. Despite some effort to schedule joint work with ISO, ETSI ISG LIS didn’t manage to renew work on the TBX work item inherited from LISA OSCAR. Instead, LTAC Global and Terminorgs, took to publishing TBX dialects or (also called) industry specifications as well as other Terminology management related resources based on the current and planned ISO versions of the TBX (Default) standard. GALA and ttt.org still host the latest (albeit stale) LISA OSCAR version including the legacy TBX Basic.

Work in liaison with OASIS XLIFF TC and later with OASIS XLIFF OMOS TC resulted in development and testing of a TBX-Basic to XLIFF 2 with Glossary Module mapping. This mapping has been actually developed and described in a FEISGILTT paper that later appeared in Localisation Focus https://www.localisation.ie/sites/default/files/publications/LFV14-1-4.pdf.

The mapping upgraded to the modular 2nd ISO edition based TBX-Basic (that is semantically equivalent with the old TBX-Basic) is now being specified on the standards track within OASIS XLIFF OMOS TC. See the latest editor's draft at https://tools.oasis-open.org/version-control/browse/wsvn/xliff-omos/trunk/XLIFF-TBX/xliff-tbx-v1.0.pdf.

Short Description

TermBase eXchange, is a family of XML based Terminology markup languages that should allow for lossless exchange of terminology related data and metadata. So far two more lightweight versions known as TBX-Basic (published in 2008 by LISA Terminology Special Interest Group) and TBX-Min (published in 2013 by LTAC Global) have been developed. TBX-Min should be targeting the use case of exchanging terminology with translators in the form of simple (mappable onto UTX) glossaries. However, TBX-Basic is more suitable for mapping between TBX and XLIFF 2.0 with glossary module.

TBX has been criticized for industry disconnect, for being too heavy on one hand and being too restrictive and not very well suitable for MT training on the other. One of the reasons might be that TBX is supposed to be both a representation and exchange format for terminology but it has been struggling to define a minimum set of terminology metadata suitable for practical interchange in localization context. Some industry implementations can be hardly considered in the spirit of the standard, not enforcing inclusion of even very basic metadata such as part of speech or not being structurally compliant with any of the predefined data structures.

TMX

Standard Short Name

TMX

Owner

ETSI ISG LIS (heir designate of LISA OSCAR portfolio)

IPR Mode

RF (the LISA published versions)

FRAND (ETSI ISG LIS)

Current version

1.4b (LISA numbering)

ETSI GS LIS 002 V1.4.2 (2013-02)

Current WIP

There is no current work in progress, as the ETSI ISG LIS group was closed. Unfortunately OASIS and ETSI were not able to agree on transferring the TMX IP to OASIS XLIFF OMOS TC that has been chartered to take over the ownership of TMX.

Based on published executive summary from November 2011, ETSI ISG LIS wished to coordinate TMX 2.0 development with XLIFF 2.0 definition of inline markup codes. During the existence of ETSI ISG LIS there was no technical development on the TMX front. The group only republished the latest LISA version on the ETSI Group Specification template. ETSI ISG LIS Chair made a public consultation of a possible related work item that ETSI ISG LIS would like to undertake, standardization of fuzzy matching (similarity calculations for text segments) at FEISGILTT (June) 2013 in London. At the time of ETSI closing ISG LIS (August, September 2015), there were informal talks about OASIS XLIFF OMOS TC taking over the TMX ownership, maintenance and development. The XLIFF OMOS TC was chartered and scoped to be able to take over TMX ownership. However, formal IP transfer between ETSI and OASIS has not been negotiated to the present day. Because of the legal matters being stuck, delegates of the June 2016 FEISGILTT and XLIFF Symposium discussed developing an XLIFF 2 profile with mandatory usage of the Translation Candidates module as a maintainable replacement of the TMX functionality. In 2017 Andrzej Zydron proposed developing a Note within OASIS XLIFF TC that would describe how to use an XLIFF 2 profile instead of the obsolete and not maintained TMX format. This work hasn’t however progressed due to lack of interest and technical consensus.

Short Description

Translation Memory eXchange has been arguably the most important and most widely implemented localization standard format. TMX is a simple XML vocabulary that was designed to provide lossless translation memory exchange. However, there are several obstacles that prevented TMX from reaching the set goal. Level 1 implementations are too low a common denominator to actually secure lossless interoperability, because of segmentation differences (that should be in theory addressed by SRX) and because of absence of inline markup on level 1. Level 2 stipulates lossless exchange of native inline codes that are however ignored by many tools and encoded as abstract placeholders.

Unfortunately, TMX is now too far behind industry developments in CAT tools (OSCAR had been failing to produce a major new version since 2004), it will continue to be important for some time as a legacy format, mainly for collecting MT training corpora from legacy tools and repositories that are for some reason or other unable to produce XLIFF files.

TQF

See MQM.

SRX
Segmentation Rules eXchange

Standard Short Name

SRX

Owner

ETSI ISG LIS (heir designate of LISA OSCAR portfolio)

ISO TC 37/SC 4 was intended as a co-owner of SRX. No memorandum of understanding between either LISA or ETSI ISG LIS was signed. This work item expired at SC 4 and was deleted on July 22, 2013 according to the ISO standards publishing policy.

Unicode ULI TC

ULI TC agreed to host the current SRX version (after ETSI ISG LIS disbanded) as a publicly available specification. It is however unclear if ULI TC would be legally allowed to produce a new version without formally negotiating IP transfer from ETSI.

IPR Mode

RF for the LISA version

FRAND in ETSI (but wasn't re-published)

Current version

2.0 (released by LISA OSCAR on April 7, 2008)

Current WIP

co-publication attempt with ISO TC37/SC 4 as ISO CD 24621 that successfully passed the Committee Draft (CD) ballot, but no work progressed and the project was deleted on July 22, 2013.

No current work in progress due to ETSI ISG LIS dissolution. Based on published executive summary from November 2011, ETSI ISG LIS scheduled an SRX meeting for March 2012. This meeting happened behind the closed door. There are or should be dependencies with ULI Work in Progress on UAX #29 segmentation behavior modifications. SRX has potential as an XML based exchange vehicle for segmentation rules, because there is no ultimate finite solution to the segmentation issue in natural languages, at least not with overloaded full stop sign. ULI TC started in 2018 a note on segmentation and wordcounting that would build on the principles of UAX #29.

Short Description

Segmentation Rules eXchange. It is an XML vocabulary that is facilitating exchange of segmentation rules between TMX compliant systems (i.e. it assumes only TMX-legal inline codes). SRX relationship to Unicode is not a transparent one. SRX can be considered incomplete from the engineering point of view. However, its proclaimed goal was not to provide a set of segmentation rules for a number of languages, but rather to provide a mechanism to exchange the rules to improve TMX interoperability. It is a known issue of TMX that it often fails to guarantee its targeted lossless transfer of Translation Memory data due to segmentation differences (chief among other issues). The current SRX incarnation works on a closed world assumption, i.e. it recreates (and adapts) UAX #29 rules. UAX #29 is referenced and its study encouraged but the relationship is currently not a maintainable linkage.

According to former ULI Chair Helena Chapman, SRX developers should take UAX #29 (and its ICU implementation) for granted, and use SRX only for exchange of rules that differ from the standard UAX #29 behavior. ULI has been collecting natural language exceptions to UAX#29 for several major languages and included these in the CLDR release cycle. CLDR however uses LDML as its description language not SRX.

UAX #9
Unicode Bidirectional Algorithm

Standard Short Name

UAX #9

Owner

Unicode Consortium

IPR Mode

RAND

Current version

Revision 39 released for Unicode 11.0.0 on May 09, 2018. This URL http://www.unicode.org/reports/tr9/ always points to the current official version.

Current Work In Progress

UAX #9 is being constantly revised to be up to date with the current Unicode release. This URL http://www.unicode.org/reports/tr9/proposed.html always point to the latest proposed version (if such a proposal exists). However, so far no revision has been proposes for Unicode 12.0 (to appear in May or June 2019, Unicode 12 Beta was authorized on September 21, 2018). Major changes happened between Revision 27 and Revision 29. Revision 29 (for Unicode 8) sent ripples that profoundly influenced handling of bidirectional text in the world of markup languages. It led to introduction of new direction handling elements and attributes in HTML and XML vocabularies. Because directionality handling in HTML and several XML vocabularies was in flux in 2013, ITS 2.0 does not contain normative provisions for directionality handling. XLIFF 2 does contain up to date directionality markup and a valid guidance how to combine it with directionality control characters if necessary when modifying segmentation in compliance with the Revision 35 of UAX #9.

Short Description

Default text flow of Arabic and Hebrew scripts is Right to Left. However text written in these scripts often contains portions with left to right directionality (such as numbers, Latin script based quotes, names of companies or products etc.). That is why such text is called bidirectional (BiDi). Many material characters (i.e. characters with natural language semantics) have strong directionality properties but there are also characters with weak directionality behavior and neutral characters whose directionality depends on context. In practice, (normally invisible) control characters (markers) need to be used in order to encode BiDi in plain text (that is by nature sequential). Simply said, UAX #9 is a detailed normative account of Unicode BiDi behavior (mainly) in plain text.

In theory (according to UTR #20 that is a joint Technical Report of Unicode and Note of W3C Internationalization Core WG) the control and stateful characters that Unicode Bidirectional Algorithm makes use of (apart from material characters with natural semantics that have a natural directionality behavior) to explicitly set text flow direction should not be used within markup context (after crossing the plain text/mark up boundary); The bidirectional flow control characters should be replaced with appropriate markup controlling the text flow (markup that implements the ITS directionality data category). In practice many CAT tools do not rely on directionality markup and apply UAX #9 in full (including the control characters) even in structured and markup environments. Speculating about the reasons, we can identify two main ones. 1) UAX #9 has a long tradition (since Unicode 2 in 1996), 2) There is a standardization gap, as Unicode control and stateful characters are not provided with clear processing requirements to be applied on entering markup environments. UTR #20 unfortunately provides only an abstract guidance rather than clear and unambiguous processing requirements. While XLIFF 2.0 has its own directionality attributes it does not have attributes corresponding to inline bidirectional overrides or embeddings, so these are allowed as UAX #9 control characters if needed.

UAX #29
Unicode Text Segmentation

Standard Short Name

UAX #29

Owner

Unicode Consortium

IPR Mode

RAND

Current version

Revision 33, released for Unicode 11.0.0 on May 22, 2018. This URL http://www.unicode.org/reports/tr29/ always points to the current official version.

Current Work In Progress

UAX #29 is being constantly revised to be up to date with the current Unicode release. This URL http://www.unicode.org/reports/tr29/proposed.html always point to the latest proposed version as long as one exists. The Revision 34 for Unicode 12.0.0 has not been proposed yet at the time of writing.

ULI exceptions use the UAX#29 behavior as the baseline. UAX #29 contains an informative pointer to CLDR released segmentation exceptions developed by ULI that can be used to modify the locale independent baseline segmentation behavior described in UAX #29.

Short Description

Unicode Annex Nr. 29: Unicode Text Segmentation is the key normative source of segmentation rules. Apart sentence boundaries (that are most relevant for CAT tools interoperability); it defines more basic grapheme cluster and word boundaries. The segmentation rules are given in (more or less) natural language (with many specific terms and operators) as an inductive succession of rules; the specification states itself that the same set of rules can be given using regular expressions. Unfortunately no finite set of regular expression based rules can ensure 100% successful sentence segmentation of English text (as an example of a Western alphabetical language written with Latin script that uses the full stop as the sentence boundary). The main reason for this is the semantic ambiguity (overload) of the full stop character ".". Apart from closing sentences, the same character is being used for a few more distinct but not disjoint purposes (closing abbreviations, decimal points and so on). UAX #29 rules do give proper segmentation in English in more than 80% but far less than 100% cases. Interestingly, in Hebrew this problem virtually does not exist, as Hebrew does not overload the full stop with the abbreviation function.

Although UAX #29 cannot possibly achieve completeness, it is still beneficial to implement it (e.g. via ICU) as the basic set of rules, and apply more fine grained exception rules on top of it. The ULI TC does not plan to influnce the default UAX #29 segmentation behavior, but it started releasing the locale specific segmentation exceptions as part of CLDR.

Unicode Standard

Standard Short Name

Unicode

Owner

Unicode Consortium

IPR Mode

RAND

Current version

Unicode 11.0.0 published on June 05 2018. This URL http://www.unicode.org/versions/latest/ always points to the current official version.

Current Work In Progress

As of Unicode 7.0.0 UTC started an aggressive yearly schedule for major releases, which means that Unicode 8.0.0 appeared in June 2015, Unicode 9.0.0 in June 2016, Unicode 10.0.0 in June 2017, Unicode 11.0.0 was released last June and Unicode 12.0.0 is on track for June 2019.

Unicode 11 added 7 new scripts, 66 new emoji characters, after adding 684 characters in this release the new total of Unicode characters stands at 137,374. Version 10 added 4 scripts, Version 9 added 6 scripts, Version 8 added 6 scripts, version 7 added 23 scripts!

UTC continues adding individual characters, character groups and whole new scripts as per worldwide communities' requirements, mainly for marginalized and historic languages coverage; Relatively heavy editorial reshuffle in 7.0 enables the upcoming rapid cycle for major releases. Unicode core had to adapt to changes introduced in the significant changes in Revision 29 or the Bidirectional Algorithm (See UAX #9). Changes in Unicode 11 brought updates to UTS #10 Unicode Collation Algorithm, UTS #39 Unicode Security Mechanisms, UTS #46, and UTS #51 Unicode Emoji.

Short Description

Unicode is the core standard that allows the humanity to encode all written human languages for computer use. Hundreds of thousands of characters covering alphabetic, syllabic, ideographic and other types of scripts are ordered in planes along with punctuation and control and private use characters. Unicode standard has been published since October 1991.

UTR #20 [Withdrawn]
Unicode in XML and other Markup Languages

Standard Short Name

Unicode-XML

Owner

W3C (Internationalization Core WG)

Unicode Consortium [UTR #20 has been withdrawn at Unicode in March 2016, maintenance of the document transferred to W3C]

IPR Mode

RF (W3C)

Current version

Although this s was last released as a W3C Working Group Note on July 13 2017, the status is still withdrawn, and the material is still mostly stale and dated as it had been developed against Unicode Unicode 6.2.0 (!).

Current Work In Progress

This important note is still in need of a major rewrite. Handling of directionality overrides and isolates in XHTML has been described recently.

The newest developments of this important note can be followed on the GitHub Editor's draft http://w3c.github.io/unicode-xml/.

Short Description

Unicode in XML and other Markup Languages. Unicode, as its main target is plain text, contains many control, formatting, or otherwise stateful characters. This document that had been first published jointly by Unicode and W3C (as a UTR and a WG Note respectively) gives a normative overview and general guidelines on, which characters can and should, and which must not or should not be used in markup context. In general any Unicode character that is XML illegal, stateful, or would require additional metadata for interpretation should come with a markup handling/replacement recommendation, or processing requirement. Authoring tools, XML editors and browsers are generally encouraged to ignore inappropriate or deprecated Unicode characters, so their preservation on crossing of plain text/markup boundary will often lead to harmful loss of data or metadata. In general, plain text is linear and requires special control characters or specific application behavior to encode metadata and/or styling information that can be handled with structured markup in XML or HTML environments. It is important to keep separated 1) methods suitable for linear and point-like handling of text from 2) markup methods that benefit from structure and inheritance.

UTS #18
Unicode Regular Expressions

Standard Short Name
UTS #18

Owner
Unicode Consortium

IPR Mode
RAND

Current version
Version 19, Revision 19, released on October 18, 2016. This URL http://www.unicode.org/reports/tr18/ always points to the current official version.

Current Work In Progress
Once a new draft will be proposed, this URL http://www.unicode.org/reports/tr18/proposed.html will point to it. No new revisions have been proposed for Unicode 10, 11, or 12 at the time of writing.

Short Description
Unicode Regular Expressions gives general guidelines for regular expression engines how to comply with the Unicode standard. Three levels are specified, of which two are default (one of them the minimum feasible for programmers, the other more end user friendly) and the finest is language specific. Revision 19 has Unicode 9 as its baseline so that it covers all the new scripts, characters and emojis.

UTS #35
Unicode Locale Data Markup Language (LDML)

Standard Short Name

LDML, UTS #35

Owner

Unicode Consortium

IPR Mode

RAND

Current version

Version 34, Revision 53, released on October 10, 2018

Current WIP

Since the latest Version and Revision is for CLDR Version 34, a new release can be expected in about a year for CLDR Version 35.

Short Description

Unicode Technical Standard #35: Unicode Locale Data Markup Language (LDML) specifies an XML vocabulary for encoding locale specific generic data categories (dates, amounts, decimals, units of measure, currency symbols etc.). Its main purpose is to enable the creation and maintenance of CLDR but is also used directly in programming frameworks such as .NET.

UTX
Universal Terminology Exchange

Standard Short Name

UTX

Owner

AAMT

IPR Mode

The documentation is available under Creative Commons 4.0 BY. No IPR mode specified.

Current version

UTX 1.2 Minimal Specification, February 18, 2018. There are some header changes and the 1.2 version supports multiple sub-glossaries as part of one glossary. Editing in spreadsheet software became easier.

Current WIP

N/A, the latest developments consisted in the development of TBX-Min mapping by LTAC Global (see TBX).

Short Description

UTX is a simple bilingual glossary format that was originally targeting Machine Translation training. It is a simple tab delimited format, an XML version also exist incited by criticism of the localization standardization community. UTX embedding has been specified for XLIFF:doc. TBX-Min mapping was proposed by LTAC Global.

Although rather minimal, this tiny terminology exchange standard specifies the part of speech field as mandatory and also provides an optional field for term status tracking with predefined values.

XLIFF

Standard Short Name

XLIFF

Owner

OASIS XLIFF TC

IPR Mode

RF on RAND terms in OASIS (freely available)
RAND in ISO (sold)

Current version

XLIFF Version 2.1, published as OASIS Standard on February 13, 2018.
application/xliff+xml (.xlf) registered as a media type on the IANA Standards tree on April 6, 2018
https://www.iana.org/assignments/media-types/application/xliff+xml
ISO 21720:2017 - XLIFF (XML Localisation interchange file format) is identical to the OASIS XLIFF Version 2.0 from August 05, 2014.

Current Work In Progress

After successfully delivering the ITS Module and advanced validation capabilities as part of the 2.1 version, the TC works on reformatting XLIFF Version 2.1 in ISO style, so that it can been again submitted for publishing in ISO TC 37/SC 5.

The TC also started publicly tracking feature development for the Version 2.2 at https://wiki.oasis-open.org/xliff/FeatureTracking.

The biggest and most interesting feature proposed for Version 2.2 is Rendering Requirements. A detailed proposal of XLIFF Rendering Requirements was presented at the 40th edition of the ASLING

Translating and the Computer conference on November 16, 2018. The idea is to give possibly normative guidance to CAT tool developers on how to display the XLIFF bitext and its rich metadata, but also to lower the technology complexity of rendering bitext and make interactive rendering of XLIFF possible in Web browsers. Other proposed features include, and attempt to reintroduce the Change Tracking Module that was deprecated between 2.0 and 2.1, specification of semantic domains for placeholders, roundtrip of segmentation metadata.

The XLIFF TC also wrestles with the idea how to use XLIFF for Translation Memory Exchange since TMX is in legal limbo and cannot be maintained or further developed. Three technical options are under consideration: 1) informatively describe a an XLIFF 2.0 profile that only uses XLIFF 2 Core and the Translation Candidates Module 2) describe how to use XLIFF Core only storage as TM 3) develop a new grammar that would reuse the XLIFF 2 vocabularies but define a different structure that would be more suitable for bulk exchange of Translation Memories, also in multilingual scenarios.

Work on TBX Basic mapping to and from XLIFF 2.0 and higher was transferred to the XLIFF OMOS TC.

Short Description

In February 2018, XLIFF TC delivered the first “dot” release of XLIFF 2, which is backwards compatible with XLIFF 2.0. XLIFF 2.1 provides full ITS 2.0 support and advanced validation support. The 2.1 Version also deprecated the Change Tracking Module and provided a number bug fixes in both core and modules. Interestingly and usefully, XLIFF 2.0 Documents are forwards compatible with XLIFF Version 2.1 advanced validation artifacts.

The approved Advanced Validation feature has been successfully added, it is however not a module in the same sense as the other 9 XLIFF 2 modules. Advanced Validation has a very small footprint in the XLIFF prose specification; instead, it provides new declarative validation artifacts that provide vastly more automated validation capability than the XML Schema files that were available in previous versions of XLIFF. As of XLIFF Version 2.1, XLIFF 2 provides also Schematron and NVDL based validation. These new validation artifacts build on the structural validation available through the existing XML Schemas, as they can express additional constraints, functional dependencies and processing requirements. With Schematron and NVDL, XLIFF 2 provides standardized artifacts to automatically validate very nearly all of its prose requirements and not only its structural integrity, as was the case with XLIFF 2.0 and earlier. This brings more transparence to verifying co-occurrence constraints, id uniqueness constrains and so on.

The ITS Module in XLIFF 2.1 specifies support for six (6) ITS data categories from scratch. Allowed Characters, Domain, Locale Filter, Localization Quality Issue, Localization Quality Rating, and Text Analysis are fully defined in the XLIFF 2.1 ITS Module. Four (4) data categories were already fully available as XLIFF 2.0 Core or Module features, hence XLIFF 2.1 only provides informative guidance on how to map Localization Note, Preserve Space, Translate and External Resource onto existing XLIFF elements and attributes. Five (5) ITS data categories have partial overlap with XLIFF 2.0 features, hence XLIFF 2.1 specifies the additional elements and attributes needed to fully express Language Information, MT Confidence, Provenance, and Terminology. Storage Size could be fully expressed within the Size and

Length Restriction Module, but the extended profile needed to do that has not been specified for XLIFF 2.1 (this could make it into XLIFF 2.2). Five (5) categories don't represent metadata after XLIFF Extraction, some of them help inform the extraction in various ways.

In August 2014, OASIS XLIFF arrived at a major new release after 6 years. XLIFF 2.0 is not backwards compatible with XLIFF 1.2, yet it draws heavily from XLIFF 1.2 and lessons learnt related to its implementation and adoption. Most importantly the XLIFF 2 architecture allows for producing a series of XLIFF 2 standards, XLIFF 2.x backwards compatible "dot" releases.

XLIFF 2 Core covers about 20% of XLIFF 1.2, because in 1.2 everything was core and hence the core was too big to be implemented across industry. Yet XLIFF (2.0 and) 2.1 provides a powerful array of advanced functionality through nine (9) modules of various extent and complexity. Modules are optional yet standard and protected throughout the roundtrip.

XLIFF 2.0 and 2.1 were created and approved by a representative group of industry and academic standardizers: big enterprise translation buyers (such as IBM, Microsoft, and Oracle), Large LSPs (such as SDL and Lionbridge) and Toolmakers (apart from SDL and Lionbridge, also Multicorpora, and ENLASO) Industry Associations (PSBT, GALA, TAUS), Academics (LRC at the University of Limerick) and Individuals (notably Chair Bryan Schnabel). Many more took interest in the final OASIS organizational approval rounds (July/August 2014 and January/February 2018) and during the extensive public reviews (organized between March 2013 and April 2014 for Version 2.0 and between October 2016 and October 2017 for Version 2.1).

XLIFF 2 is a lean modular standard that allows for plug and play interoperability in the areas of inline markup, segmentation, glossary exchange, translation and review, engineering QA etc., effectively catering for end to end mash up of best of breed solutions throughout the content value chain, not only the localization roundtrip.

The translation candidates module allows for local inclusion of translation candidates coming from various sources including Translation memories, Machine Translation services, and various crowdsourcing scenarios. The glossary module allows for local inclusion of relevant terminology in both source and target languages, it supports both monolingual and bilingual scenarios, including terminology lifecycle management scenarios, one-to-one mapping with TBX Basic and UTX is being specified in the XLIFF OMOS TC. Formatting style module provides two attributes for embedding HTML encoded preview information that can be used by agents for on the fly preview generation to provide context for human translators. Metadata module is a non-namespaced private extensibility mechanism. Custom metadata can be grouped and presented as key-value pairs, which is a predictable way facilitating interoperable display. Resource data module is to replace the binary module capability known from 1.2. Resources can be provided for localization with external media type specific editors or simply as reference. Change tracking module allow to log change history including the provenance of changes. Size and length restriction module, provides a generalized way for specifying text size/volume restrictions. This module can cater for simple use cases such as fitting a database field based on Unicode codepoint or byte count, up to very complex scenarios, such as fitting a specific hardware display restrictions using a specific font.

Validation module provides a mechanism for specifying simple localization rules to be checked with relation to source and/or target. You can for instance make sure that a brand name will be included in the target text, or on the other hand make sure that a deprecated, offensive or legally inadmissible term will NOT be included. For Modules and functionality changed or added in XLIFF Version 2.1 look at the Current Work in Progress.

XLIFF Object Model

Standard Short Name

XLIFF OM

Owner

OASIS XLIFF OMOS TC

IPR Mode

Non-Assertion (RF)

Current version

Pre-release

Current Work In Progress

XLIFF OMOS (Object Model and Other Serializations) TC has been chartered in December 2015 to address the industry need of an abstract object model for the XLIFF 2 family of standards. The objective is to provide a serialization independent description of XLIFF and XLIFF fragment data categories that would allow development of non-XML serializations of the same data model. The goal is to make sure that arbitrary serializations of localization interchange data will be interoperable with XLIFF 2, vice versa, and among each other. The Object Model is being developed on this GitHub Repository: https://github.com/oasis-tcs/xliff-omos-om. The repository is public but the contributors need to be or become OASIS XLIFF OMOS TC members. Anyone can raise issues or comments though via the associated "Issues" or "Wiki", as well as minor bug fixes via pull requests. It's up to the repository maintainers if those pull requests will be merged or not (the usual GitHub collaboration workflow). UML Diagram development on that repository is powered by the open source Papyrus project.

Short Description

Abstract Object Model defined in UML diagrams and a prose specification. Definition of an XLIFF equivalent data model but also of interoperable file, sub-file, unit, sub-unit and sub-segment fragments. This requires detailed specification of data integrity dependencies that was not needed in XLIFF 2 that is always being exchanged as a complete XML file.

The main goal is to be able to transfer the XLIFF Inline data model into arbitrary XML and non-XML serializations. As such this is groundwork on an interchange data model that is needed to ground any messaging, bus or webservices architecture or a standard Translation API (see TAPICC).

xml:tm

Standard Short Name

xml:tm

Owner

ETSI ISG LIS (heir designate of LISA OSCAR portfolio)

IPR Mode

RF (LISA published versions)

FRAND

Current version

Last version by LISA, 1.0, released on February 26, 2007.

Current WIP

Andrzej Zydroń has exposed the 2.0 version for “public comment” on the webpage of his company XTM International. However, the document appears to be a final ISG LIS specification with publication date August 19, 2012. Its ETSI status is unclear from publicly available sources and the specification is not available on the ETSI download portal.

Short Description

xml:tm is a so called namespace application, which means it is NOT designed to form independent documents that could exist on its own. Instead it is designed to be injected as a relatively heavy explicit internationalization apparatus (following other more established element standards) into any well formed XML document containing human readable language. xml:tm is based on a smart idea how to make XML content intelligent and aware of its various versions, be it authoring or translation versions. Unfortunately, it is hardly possible to call this specification a standard due to a very low number (2) of implementations. The standard is actually being pushed by one company only without wider industry consensus. The specification was indeed developed by XTM’s (then XML-INTL) Andrzej Zydroń and donated to LISA that published it as an OSCAR standard in early 2007. The failure of this specification to become an actual standard should be a memento of the importance of broad consensus building while creating industry standards. OAXAL committee specification (Andrzej is the Chair) works with xml:tm as the key standard of the open authoring and localization architecture, which might sound strange as xml:tm implementation is only required for the highest (level 1 - complete) of OAXAL conformance. OKAPI framework is for instance only OAXAL level 2 compliant as it does not implement xml:tm. It seems unclear how the architecture can exist without its key component on levels 2 and 3.