Engaging Content
Engaging People

Managing metadata throughout the multilingual content lifecycle

W3C ITS 2.0 in OASIS XLIFF 2.1

Dr. David Filip
OASIS XLIFF OMOS TC Chair
OASIS XLIFF TC Secretary, Editor, Liaison Officer
Spokes Research Fellow
ADAPT Centre
KDEG, Trinity College Dublin

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union
European Union
European Regional Development Fund
SFI Science Foundation Ireland

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

What is XLIFF?

XML Localisation/Localization Interchange File Format

The only **open standard bitext** format

XLIFF 1.2 OASIS Standard since Feb 2008
Superseded by XLIFF 2.0 on 5th August 2014
To be superseded by XLIFF 2.1 by June 2017
XLIFF Version 2.1 opened the 2nd Public Review!

XLIFF Version 2.1 concluded the 1st Public Review
https://multilingual.com/language-in-the-news/xliff-2-1-open-public-review/
The csprd02 version was approved by XLIFF TC on 7th Feb 2017
The official public review link:
http://docs.oasis-open.org/xliff/xliff-core/v2.1/csprd02/xliff-core-v2.1-csprd02.html
Always current Editors' Draft:
http://tools.oasis-open.org/version-control/browse/wsvn/xliff/trunk/xliff-21/xliff-core-v2.1.pdf

The **first "dot" release** after XLIFF 2.0 delivers on the

modularity promise of the XLIFF 2 architecture.

XLIFF 2.1 defines three (3) new namespaces and brings a full native ITS 2.0 capability via its ITS Module without breaking the backwards compatibility with XLIFF 2.0.

XLIFF 2 Core and 7 out of 8 XLIFF 2.0 Modules are unaffected by the 2.1 release.

Apart from a major bugfix for the Change Tracking Module and the brand new ITS Module, XLIFF 2.1 brings Advanced Validation capability. XLIFF 2.1 (and XLIFF 2.0 also) can be now 100% validated with standardized validation artifacts without regress to custom validation code. The expressivity of the validation framework was greatly enhanced by the usage of Schematron and NVDL schema languages on top of XML Schemas (xsd) that were available in XLIFF 2.0.

- Bitext
- Bitext Management

- Open Standards
  - Transparency of development and publishing
  - Availability under Royalty Free (RF) conditions

- Evolution and Adoption of Bitext formats
- **Overview of XLIFF 2.1 ITS module**

Out of scope: Advanced validation, Change Tracking, and what was there already in XLIFF 2.0

- Takeaways

Bitext

A structured (usually mark up language based) artefact that contains aligned source (natural language) and target (natural language) sentences. We consider bitex to be ordered by default (such as in an XLIFF file – defined below, an “unclean” rtf file, or a proprietary database representation). Nevertheless, unordered bitext artefacts like translation memories (TMs) or terminology bases (TBs) can be considered special cases of bitext or bitext aggregates, since the only purpose of TM as an unordered bitext is to enrich ordered bitext, either directly or through training a Machine Translation engine.

(Filip, 2012) (Filip and Ó Conchúir, 2011)

Bitext

A structured (usually mark up language based) artefact that contains **aligned source** (natural language) **and target** (natural language) **sentences**. We consider bitex to be **ordered by default** (such as in an XLIFF file – defined below, an “unclean” rtf file, or a proprietary database representation). Nevertheless, unordered bitext artefacts like translation memories (TMs) or terminology bases (TBs) can be considered special cases of bitext or bitext aggregates, since the only purpose of TM as an unordered bitext is to enrich ordered bitext, either directly or through training a Machine Translation engine.

(Filip, 2012) (Filip and Ó Conchúir, 2011)

Bitext

Tyger Tyger, burning bright,
Tygře, tygře, ohnivou

In the forests of the night;
září svítíš lesní tmou!

What immortal hand or eye,
Kdo ten nesmrtelný byl,

Could frame thy fearful symmetry?
že z ní tvůj souměr sestrojil?

William Blake / Jaroslav Skalický

Bitext

Tyger Tyger, burning bright,	Tygře, tygře, ohnivou
In the forests of the night;	září svítíš lesní tmou!
What immortal hand or eye,	Kdo ten nesmrtelný byl,
Could frame thy fearful symmetry?	že z ní tvůj souměr sestrojil?
Willian Blake	Jaroslav Skalický

Bitext

<source>Tyger Tyger, burning bright, </source>
<target>Tygře, tygře, ohnivou </target>

<source>In the forests of the night; </source>
<target>září svítíš lesní tmou! </target>

<source>What immortal hand or eye, </source>
<target>Kdo ten nesmrtelný byl, </target>

<source>Could frame thy fearful symmetry? </source>
<target>že z ní tvůj souměr sestrojil? </target>

Bitext


```
msgid "Tyger Tyger, burning bright, "
msgstr "Tygře, tygře, ohnivou "

msgid "In the forests of the night; "
msgstr "září svítíš lesní tmou! "

msgid "What immortal hand or eye, "
msgstr "Kdo ten nesmrtelný byl, "

msgid "Could frame thy fearful symmetry? "
msgstr "že z ní tvůj souměr sestrojil? "
```

Bitext

<source xml:lang="EN">Tyger Tyger, burning bright, </source>
<target xml:lang="CS">Tygře, tygře, ohnivou </target>

<source xml:lang="EN">In the forests of the night; </source>
<target xml:lang="CS">září svítíš lesní tmou! </target>

<source xml:lang="EN">What immortal hand or eye, </source>
<target xml:lang="CS">Kdo ten nesmrtelný byl, </target>

<source xml:lang="EN">Could frame thy fearful symmetry? </source>
<target xml:lang="CS">že z ní tvůj souměr sestrojil? </target>

Bitext


```
<unit id=1>
   <segment>
      <source xml:lang="EN">Tyger Tyger, burning bright, </source>
      <target xml:lang="CS">Tygře, tygře, ohnivou </target>
   </segment>
   <segment>
      <source xml:lang="EN">In the forests of the night; </source>
      <target xml:lang="CS">září svítíš lesní tmou! </target>
   </segment>
   <segment>
      <source xml:lang="EN">What immortal hand or eye, </source>
      <target xml:lang="CS">Kdo ten nesmrtelný byl, </target>
   </segment>
   <segment>
      <source xml:lang="EN">Could frame thy fearful symmetry? </source>
      <target xml:lang="CS">že z ní tvůj souměr sestrojil? </target>
   </segment>
</unit>
```

Bitext management

[Bitext Management is a g]roup of processes that consist of high and low level manipulation of ordered and/or unordered bitext artefacts. Agents can be both human and machine. Usually the end purpose of Bitext Management is to create target (natural language) content from source (natural language) content, typically via other enriching Bitext Transforms, so that Bitext Management Processes are usually enclosed within a bracket of source content extraction and target content re-import.

(Filip, 2012) (Filip and Ó Conchúir, 2011)

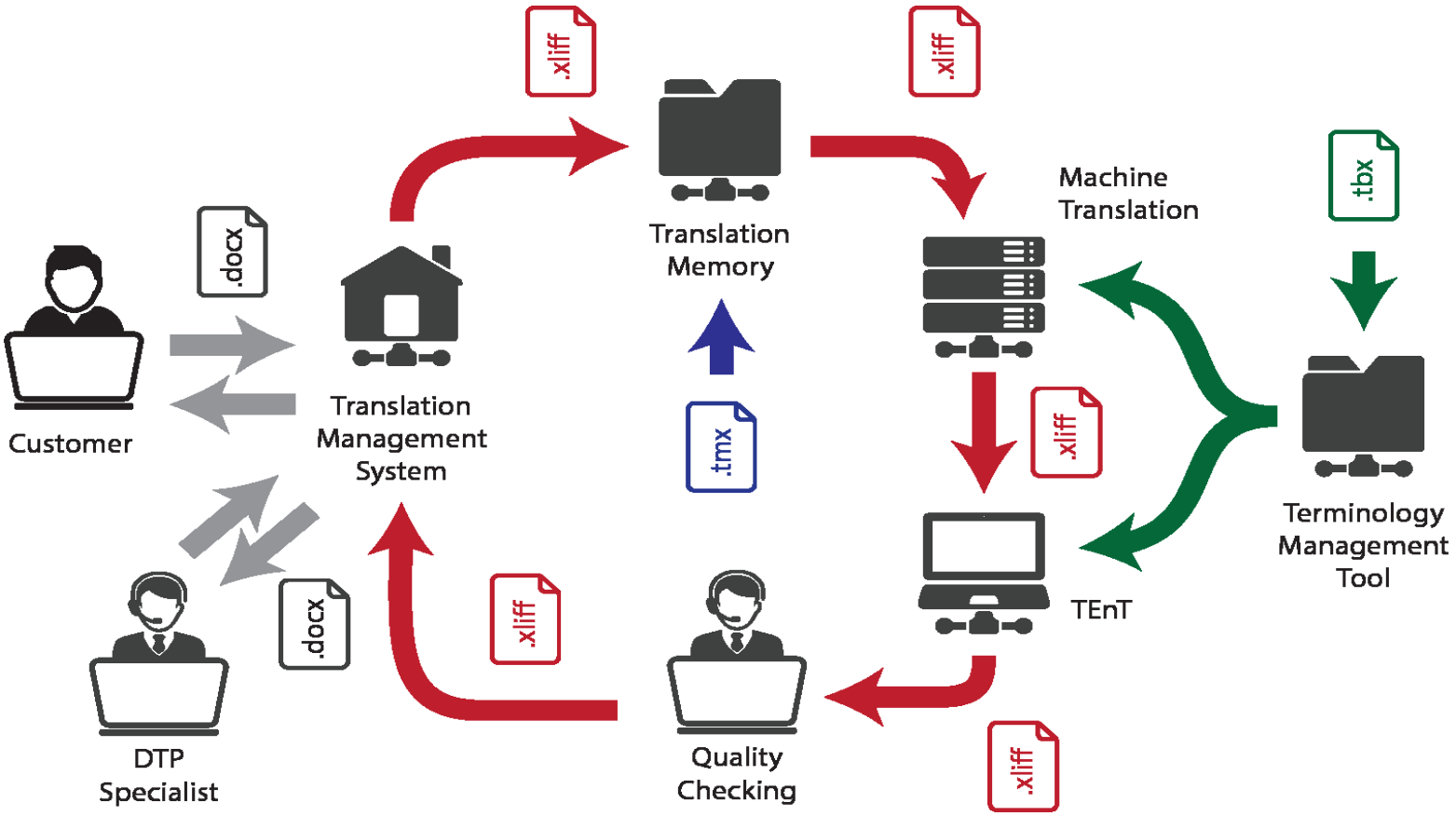
Bitext management

[Bitext Management is a **g]roup of processes** that consist of high and low level **manipulation of ordered and/or unordered bitext artefacts**. Agents can be both human and machine. Usually the end purpose of Bitext Management is to create target (natural language) content from source (natural language) content, typically via other enriching Bitext Transforms, so that Bitext Management Processes are usually enclosed within a bracket of source content extraction and target content re-import.

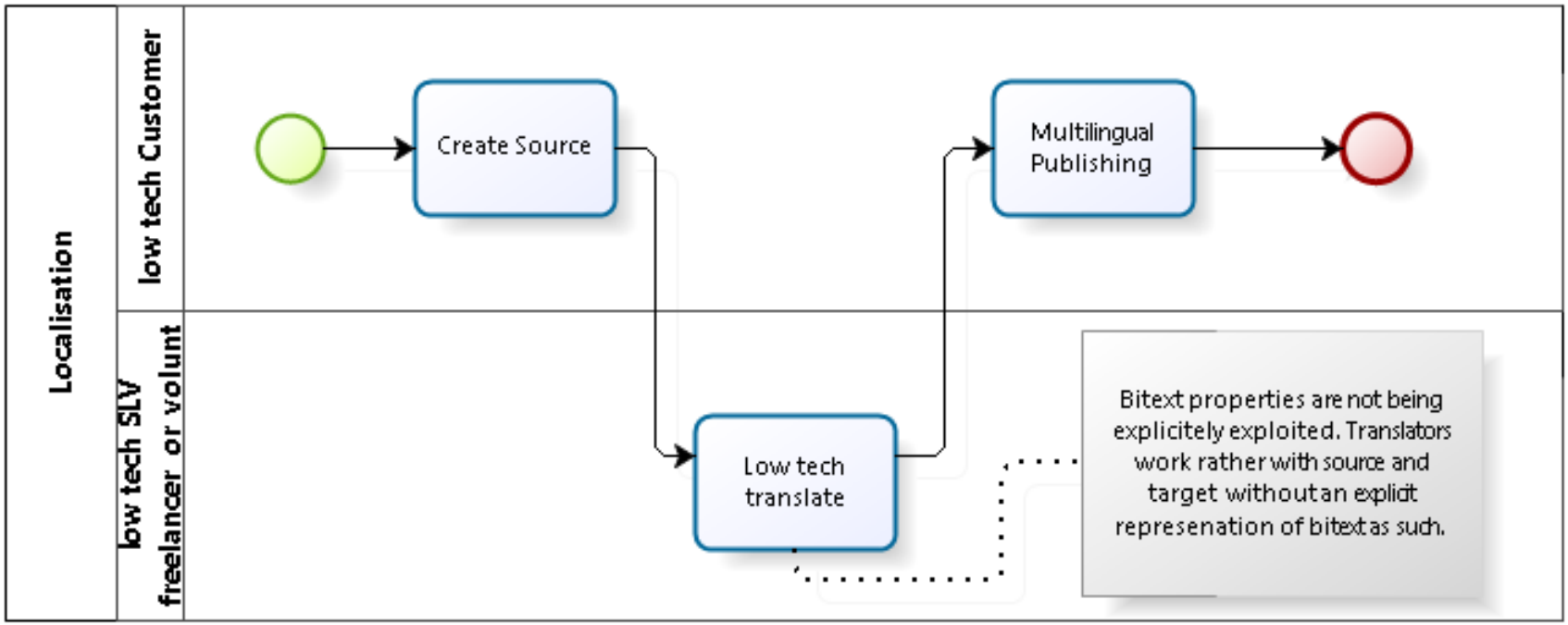
(Filip, 2012) (Filip and Ó Conchúir, 2011)

XLIFF Roundtrip example

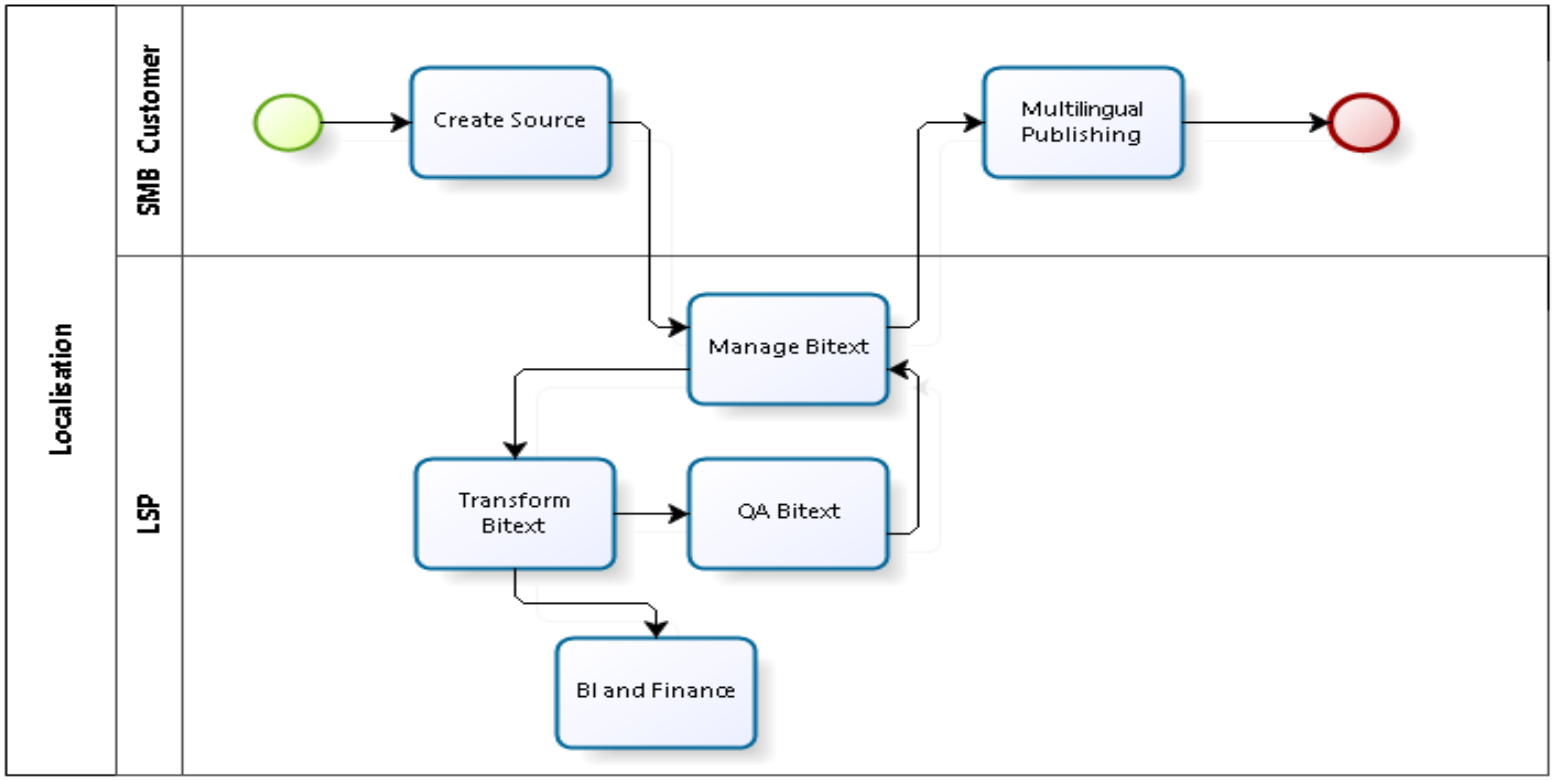
Courtesy of Arle Lommel and Alan Melby



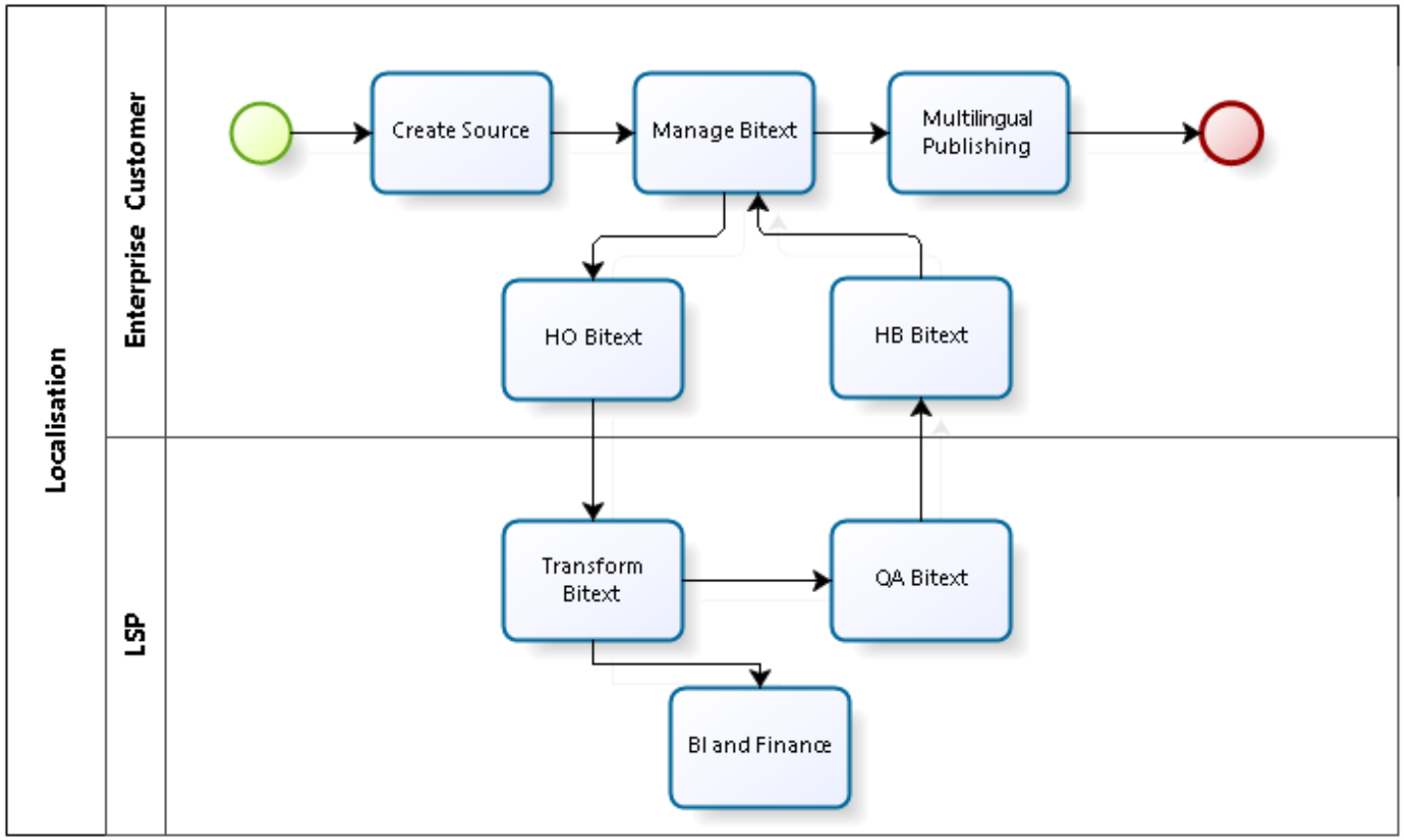
Bitext management?



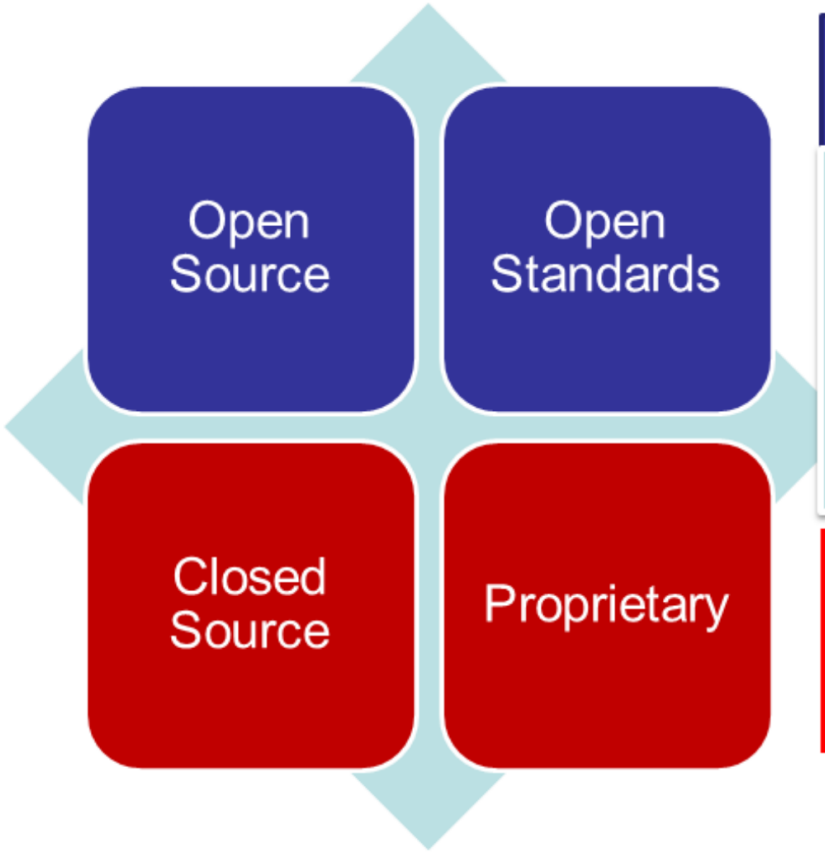
Bitext management – SMBs don't manage bitext



Bitext management – corporations do manage bitext

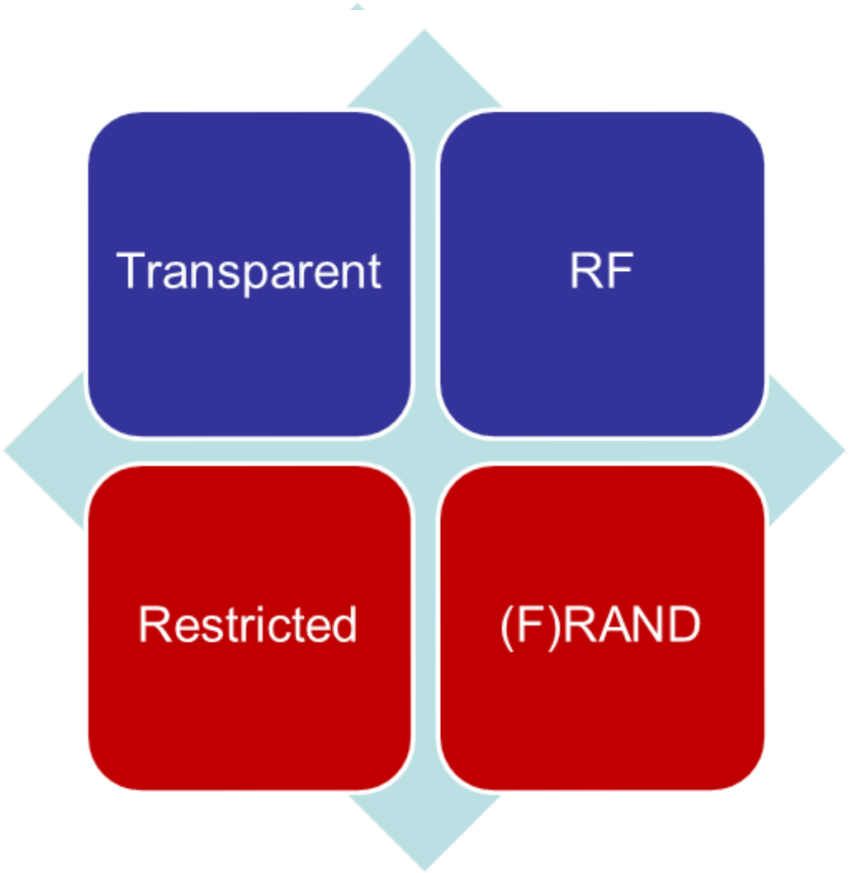


Open Standards



1. open-open
2. closed-open
3. open-proprietary
4. closed-proprietary

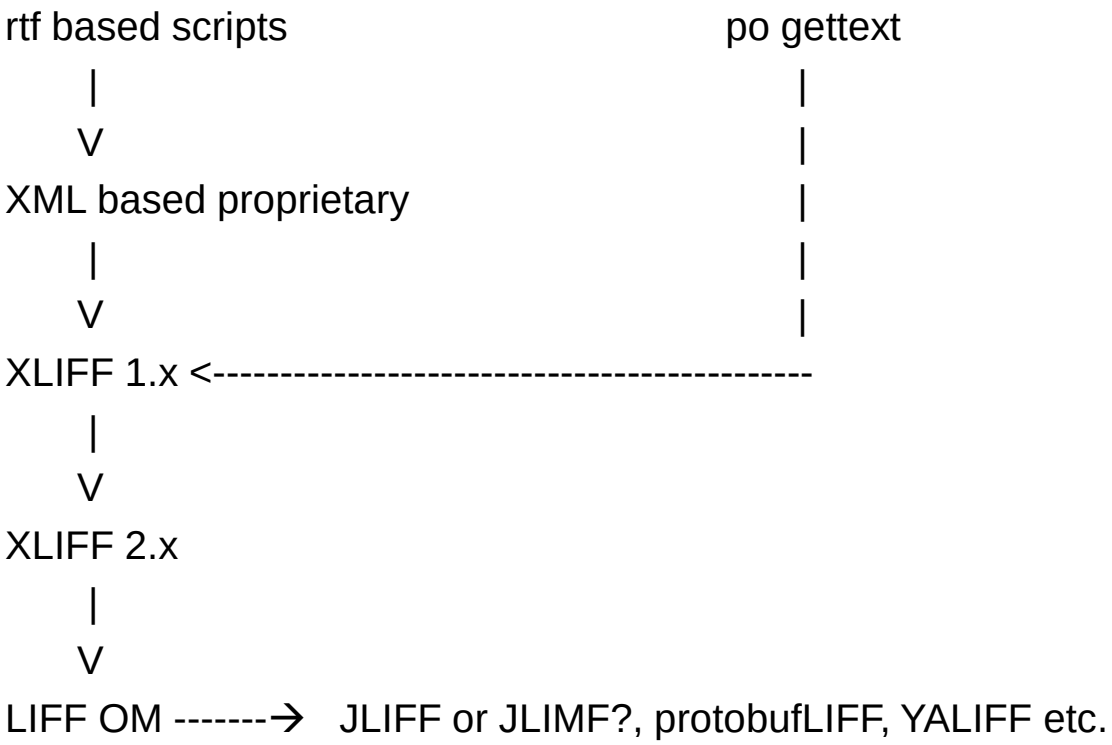
Open Standards



Open

Proprietary

Evolution and adoption of bitext formats

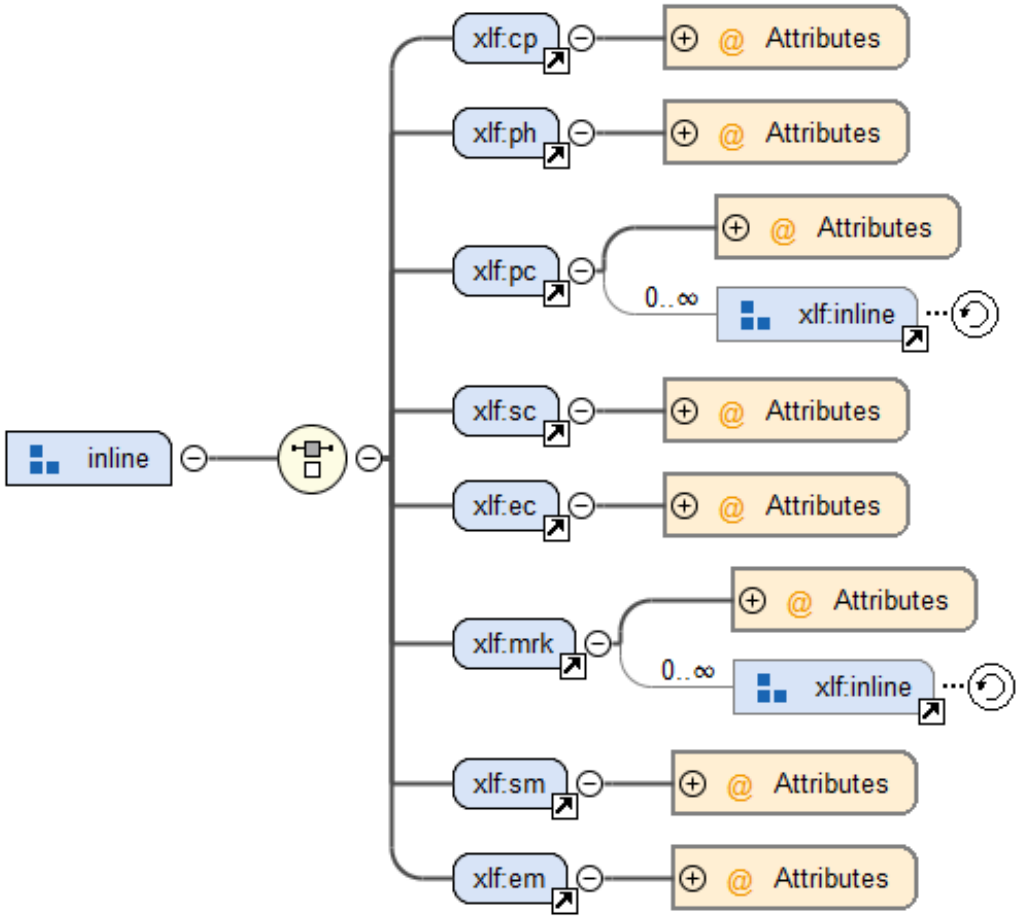


XLIFF 2 Overview - Agents

Agent

- **Writer**
  - **Extractor**,
  - **Enricher**, e.g. reviewer workbench, TM server, MT broker etc.
  - **Modifier**, e.g. a translation editor
- **Merger**
- Etc. e.g. validator

XLIFF 2 data model – Inline



All **Modules** other than the **Core** are **OPTIONAL** sets of features

- Core **<- ITS**
  - Translation Candidates Module 2.0 **+ ITS**
  - Glossary Module 2.0
  - Format Style Module 2.0
  - Metadata Module 2.0
  - Resource Data Module 2.0 **<- ITS**
  - Change Tracking Module ~~2.0~~ -> 2.1
  - Size and Length Restriction Module 2.0 **<- ITS**
  - Validation Module 2.0
  - **ITS Module 2.1**

ITS 2.0	XLIFF	Other standards
Translate	Translate Annotation	
Localization Note	Note	
Terminology	Term Annotation	TBX
Directionality	Directionality	UAX #9
		locale specific layouts for HTML
Language Information	srcLang, trgLang, xml:lang, itsm:lang	BCP47
Elements Within Text	Subflows mechanism	
Domain	Itsm:domains	
Text Analysis	ITSM Text Analytics	TBX
Locale Filter	Extraction / ITSM mechanism	
Provenance	ctr: module + ITSM Provenance	Dublin Core
External Resource	res: module	IANA media types
Target Pointer	<source> & <target> siblings	
ID value	XLIFF specific ID and FRAGID	xml:id
Preserve Space	xml:space	xml:space
Localization Quality Issue	ITSM LQI	MQM
Localization Quality Rating	ITSM LQR	
MT Confidence	mtc: module / itsm:mtConfidence	TMX
Allowed Characters	ITSM Allowed Characters	regular expressions
Storage Size	slr: module	Unicode code points, fonts etc.

Example 1. Translate expressed locally in HTML

```
<!DOCTYPE html>
<html>
    <head>
        <meta charset=utf-8>
        <title>Translate flag test: Default</title>
    </head>
    <body>
        <p>The <span translate=no>World Wide Web Consortium</span> is making the World Wide Web worldwide!</p>
    </body>
</html>
```

Example 1. Translate expressed locally in HTML

```
<!DOCTYPE html>
<html>
    <head>
        <meta charset=utf-8>
        <title>Translate flag test: Default</title>
    </head>
    <body>
        <p>The <span translate=no>World Wide Web Consortium</span> is making the World Wide Web worldwide!</p>
    </body>
</html>
```

Example 2. Translate expressed locally in XML

```
<messages its:version="2.0" xmlns:its="http://www.w3.org/2005/11/its">
    <msg num="123">Click Resume Button on Status Display or
<panelmsg  its:translate=“no">CONTINUE</panelmsg> Button on
printer panel</msg>
</messages>
```

Example 3. Translate expressed globally in XML

```
<its:rules version="2.0" xmlns:its="http://www.w3.org/2005/11/its">
    <its:translateRule translate="no" selector="//code"/>
</its:rules>
```

Example 4. XLIFF Core @translate on a structural leaf element

```
<unit id='1' translate="yes">
    <segment>
        <source>Translatable text</source>
    </segment>
</unit>
<unit id='2' translate="no">
    <segment>
        <source>Non-translatable text</source>
    </segment>
</unit>
```

The above could be an *Extraction* of the following HTML snippet:

```
<p translate='yes'>Translatable text</p>
<p translate='no'>Non-translatable text</p>
```

The same snippet:

Translatable text

Non-translatable text

Example 5. XLIFF representing ITS Translate by Extraction behavior w/o explicit metadata

```
<unit id='1'>
    <segment>
        <source>Translatable text</source>
    </segment>
</unit>
```

The same snippet:

** Translatable text **

Non-translatable text

Example 5. XLIFF representing ITS Translate by Extraction behavior w/o explicit metadata

```
<unit id='1'>
    <segment>
        <source>Translatable text</source>
    </segment>
</unit>
```

Source metadata that inform Extraction behavior

- Translate http://www.w3.org/TR/its20/#trans-datacat
- Locale Filter http://www.w3.org/TR/its20/#LocaleFilter
- External Resource http://www.w3.org/TR/its20/#externalresource
- Elements Within Text http://www.w3.org/TR/its20/#elements-within-text

Source metadata that inform Extraction behavior

- Translate http://www.w3.org/TR/its20/#trans-datacat
- Locale Filter http://www.w3.org/TR/its20/#LocaleFilter
- External Resource http://www.w3.org/TR/its20/#externalresource
- Elements Within Text http://www.w3.org/TR/its20/#elements-within-text

Other metadata that inform localization behavior

- Language Information http://www.w3.org/TR/its20/#language-information
- Target Pointer http://www.w3.org/TR/its20/#target-pointer
- Localization Note http://www.w3.org/TR/its20/#locNote-datacat
- Directionality http://www.w3.org/TR/its20/#directionality
- Preserve Space http://www.w3.org/TR/its20/#preservespace
- ID Value http://www.w3.org/TR/its20/#idvalue
- Allowed Characters http://www.w3.org/TR/its20/#allowedchars
- Storage Size http://www.w3.org/TR/its20/#storagesize

Subject Matter related datacats

- Terminology http://www.w3.org/TR/its20/#terminology
- Text Analysis http://www.w3.org/TR/its20/#textanalysis
- Domain https://www.w3.org/TR/its20/#domain

Subject Matter related datacats

- Terminology http://www.w3.org/TR/its20/#terminology
- Text Analysis http://www.w3.org/TR/its20/#textanalysis
- Domain https://www.w3.org/TR/its20/#domain

Metadata that are produced during or by localization transformations of content

- MT Confidence http://www.w3.org/TR/its20/#mtconfidence
- Localization Quality Issue http://www.w3.org/TR/its20/#lqissue
- Localization Quality Rating http://www.w3.org/TR/its20/#lqrating
- Provenance http://www.w3.org/TR/its20/#provenance

Already in

- Translate http://www.w3.org/TR/its20/#trans-datacat
- Preserve Space http://www.w3.org/TR/its20/#preservespace
- External Resource http://www.w3.org/TR/its20/#externalresource

---

- Localization Note https://www.w3.org/TR/its20/#locNote-datacat

Implemented from scrcatch

- Allowed Characters http://www.w3.org/TR/its20/#allowedchars
- Domain https://www.w3.org/TR/its20/#domain
- Locale Filter http://www.w3.org/TR/its20/#LocaleFilter
- Localization Quality Issue http://www.w3.org/TR/its20/#lqissue
- Localization Quality Rating http://www.w3.org/TR/its20/#lqrating
- Text Analysis http://www.w3.org/TR/its20/#textanalysis

Partial Overlap

- Language Information http://www.w3.org/TR/its20/#language-information
- MT Confidence http://www.w3.org/TR/its20/#mtconfidence
- Provenance http://www.w3.org/TR/its20/#provenance
- Terminology http://www.w3.org/TR/its20/#terminology
- Storage Size http://www.w3.org/TR/its20/#storagesize

Not represented

- Directionality http://www.w3.org/TR/its20/#directionality
- Elements Within Text http://www.w3.org/TR/its20/#elements-within-text
- ID Value http://www.w3.org/TR/its20/#idvalue
- Locale Filter http://www.w3.org/TR/its20/#LocaleFilter
- Target Pointer http://www.w3.org/TR/its20/#target-pointer

Scope and Pointer Issues

XLIFF 2.1 First Public Review Draft used only the itsm namespace

`urn:oasis:names:tc:xliff:itsm:2.1`

This made many ITS data categories inaccessible by generic ITS Processors because of the lack of global pointers in the ITS 2.0 spec.

Thus XLIFF TC attempted to reuse the W3C its namespace

`https://www.w3.org/2005/11/its/`

Using only the W3C namespace would howver break the solutions for **Language Information** and **Domain**, the current public review draft uses both namespaces to maximize expressivity and accessibility..

Scope and Pointer Issues

Still there is the fundamental issue that generic ITS processors cannot access pseudo-spans that are necessary in XLIFF inline data model

Nevertheless, ITS Processors can easily implement an additional capability to detect spans like this one

```
<sm id="1"/>span of text<em startRef="1"/>
```

without going into any more XLIFF specific features.

Scope and Pointer Issues

Also existing and overlapping features are not accessible in cases, where ITS 2.0 lacks global pointers.

It is again relatively easy and straightforward to introduce these as extensions via the W3C ITS Interest Group (IG).

Takeaways 1

- XLIFF 2.1 gives guidance how to roundtrip each of the 19 ITS 2.0 datacats
- All of the ITS module's based metadata is accessible by ITS Processors, except for the pseudo-span issue..
  - ITS Processors can easily implement an additional capability to detect spans like this one

```
<sm id="1"/>span of text<em startRef="1"/>
```

  without going into any more XLIFF specific features.

Takeaways 2

- The release of a technically stable public review draft of XLIFF 2.1 constitutes another important step in harmonization of Internationalization and Localization standards based at OASIS, W3C, Unicode Consortium and elsewhere
- Early adopters of XLIFF 2.1 should subscribe to the XLIFF TC Comment List https://www.oasis-open.org/committees/comments/index.php?wg_abbrev=xliff to be notified on further progress of the review drafts towards the official publication as an OASIS Standard hopefully in summer 2017..
- All issues raised through the comment list are publicly solved on the XLIFF TC JIRA project https://issues.oasis-open.org/browse/XLIFF/?selectedTab=com.atlassian.jira.jira-projects-plugin:issues-panel

Thanks a million for your attention

david.filip@adaptcentre.ie

@merzbauer

XLIFF 2 Resources

https://www.oasis-open.org/committees/**xliff-omos/**

https://github.com/oasis-tcs/xliff-omos-om

https://github.com/oasis-tcs/xliff-omos-jliff

https://tools.oasis-open.org/version-control/browse/wsvn/xliff-omos/trunk/XLIFF-TBX/xliff-tbx-v1.0.pdf

https://www.oasis-open.org/committees/**xliff/**

http://docs.oasis-open.org/xliff/xliff-core/v2.1/xliff-core-v2.1.html

https://www.oasis-open.org/committees/comments/index.php?wg_abbrev=xliff

https://issues.oasis-open.org/browse/XLIFF/

Presentations from 7th XLIFF Symposium at 5th FEISGILTT at #LocWorld31 Dublin June 7-8, 2016

http://locworld.com/feisgiltt2016-cfp/

Presentations from 6th XLIFF Symposium at 4th FEISGILTT at #LocWorld28 Berlin June 2-3, 2015
http://locworld.com/feisgiltt-program/

FEISGILTT Localisation Focus Volumes

http://www.localisation.ie/locfocus/issues/14/1

http://www.localisation.ie/locfocus/issues/12/1