Inline Markup

David Filip
ADAPT Centre @ TCD
@merzbauer

MomenT Workshop #MTSummit2019
2019-08-19

Convenor, JTC 1/SG 1 Open Source Software

Convenor, JTC 1/SC 42/WG 3 Trustworthiness of AI

SC 42 representative on JTC 1/SG 5 Trustworthiness

National mirror chair, NSAI TC 02/SC 18 AI | Head of the Irish national delegation, ISO/IEC JTC 1/SC 42 Artificial Intelligence

NSAI expert on CEN/CENELEC FG AI

Chair & Editor, OASIS XLIFF OMOS TC | Secretary & Lead Editor, OASIS XLIFF TC

SC 42 liaison to SC 38

NSAI expert to ISO/IEC JTC 1/SC 38 Cloud Computing,

SC 42 liaison to ISO TC 37 and vice versa

NSAI expert to ISO TC 37/SC 3 Terminology management, SC 4 Language resources, SC 5 Language technology

Moravia Research Fellow
ADAPT Centre, KDEG, Trinity College Dublin

Agenda

Markup?

Structural vs Inline

Bitext vs Parallel Data

Interoperability

Parsing an object model vs Regular expessions

Markup?

Distinguishable from text

Semantic vs formatting

Is JSON markup?

….

….

No, but..

JSON, no mixing character data with markup

```
<source>Eat <ph id="1" equiv="[number]"/> eggs for <mrk
id="2">breakfast</mrk>. </source>


{
"source" : [
{"text" : "Eat "} ,
{"kind" : "ph" , "id" : "1", "equiv" : "[number]"} ,
{"text" : " eggs for "} , {"kind" : "sm" , "id" : "2"} ,
{"text" : "breakfast"} ,
{"kind" : "em" , "id" : "3" , "startRef" : "2"} ,
{"text" : ". "}
]
}
```

Structural vs Inline

HTML

<body>, <p>,

, , <i>, <br/>, <var>, <a>

DocBook

<article>, <section>, <para>

<glossterm>, <olink>, <ulink>

XLIFF

<file>, <group>, <unit>,

<sc/>, <ec/>, <pc>, <ph>, <sm/>, , <mrk>

Bitext vs Parallel Data

bi-text [psycholinguistic] -> bitext [markup artifact]

Bitext is segmented, ordered, and *always* aligned

… cleaning data ??

… Yes, but…

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,

In the forests of the night;

What immortal hand or eye,

Could frame thy fearful symmetry?

Willian Blake

Tygře, tygře, ohnivou

září svítíš lesní tmou!

Kdo ten nesmrtelný byl,

že z ní tvůj souměr sestrojil?

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>

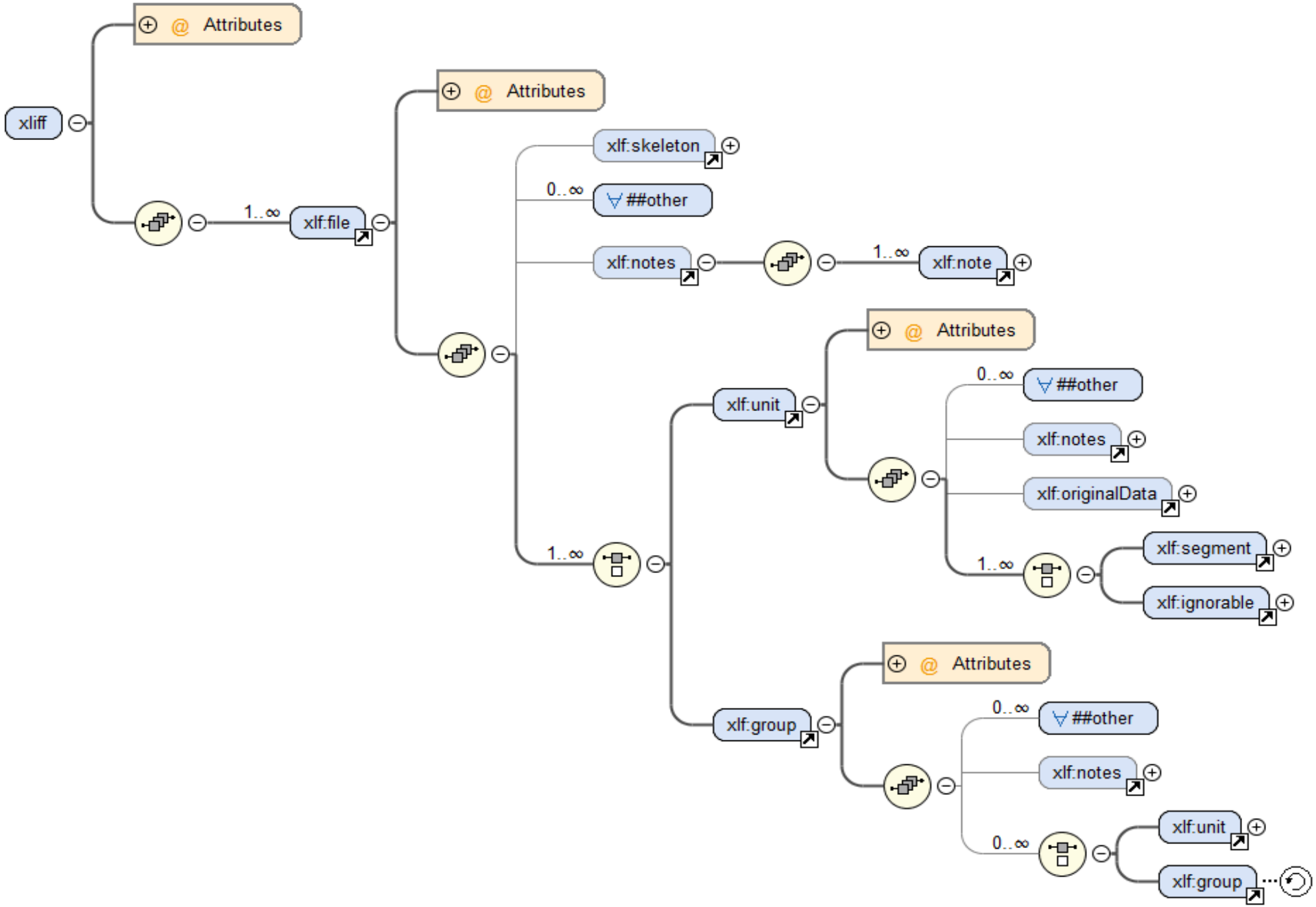
<source xml:lang="EN">In the forests of the night; </source>
<target xml:lang="CS">září svítíš lesní tmou! </target>

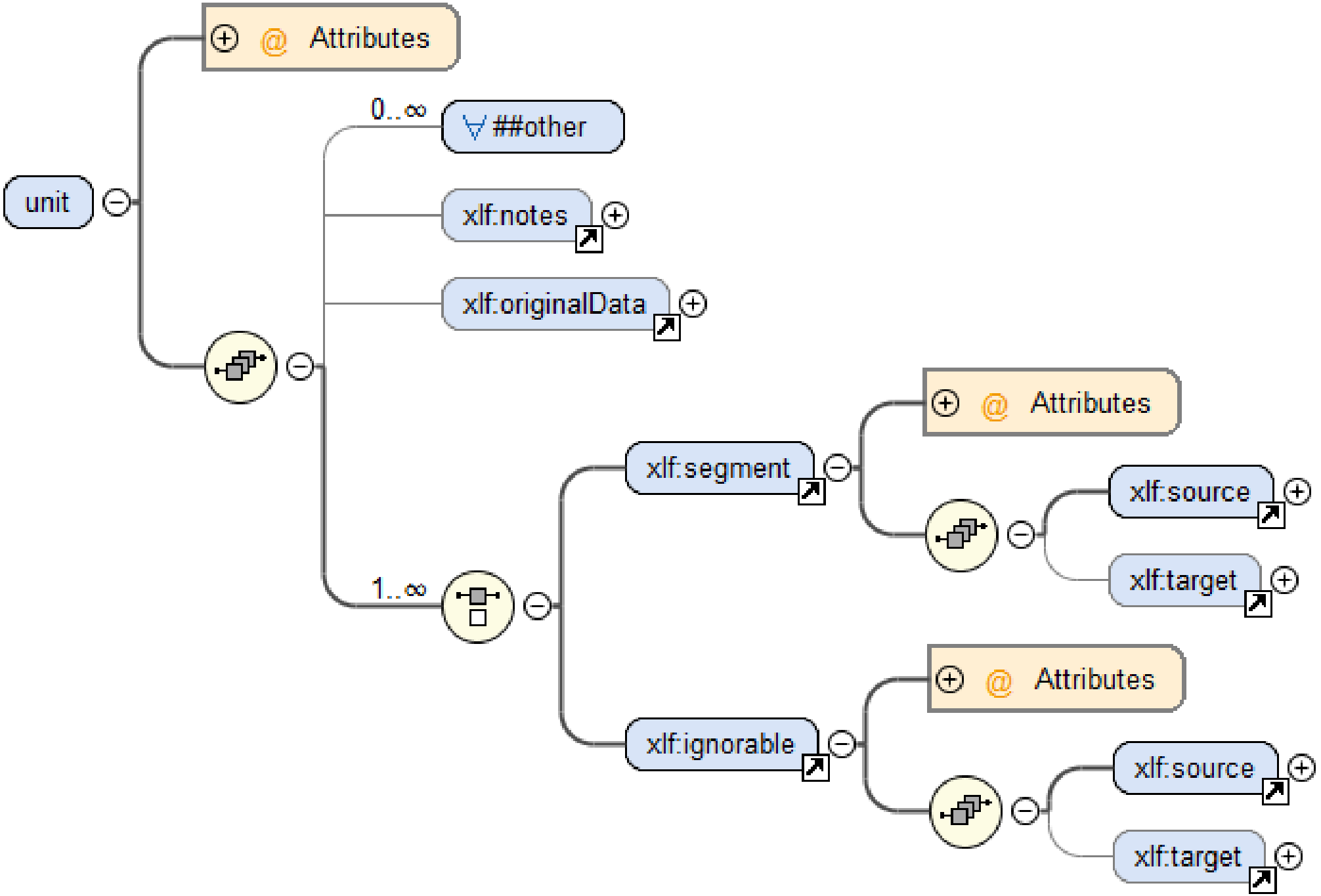
<source xml:lang="EN">What immortal hand or eye, </source>
<target xml:lang="CS">Kdo ten nesmrtelný byl, </target>

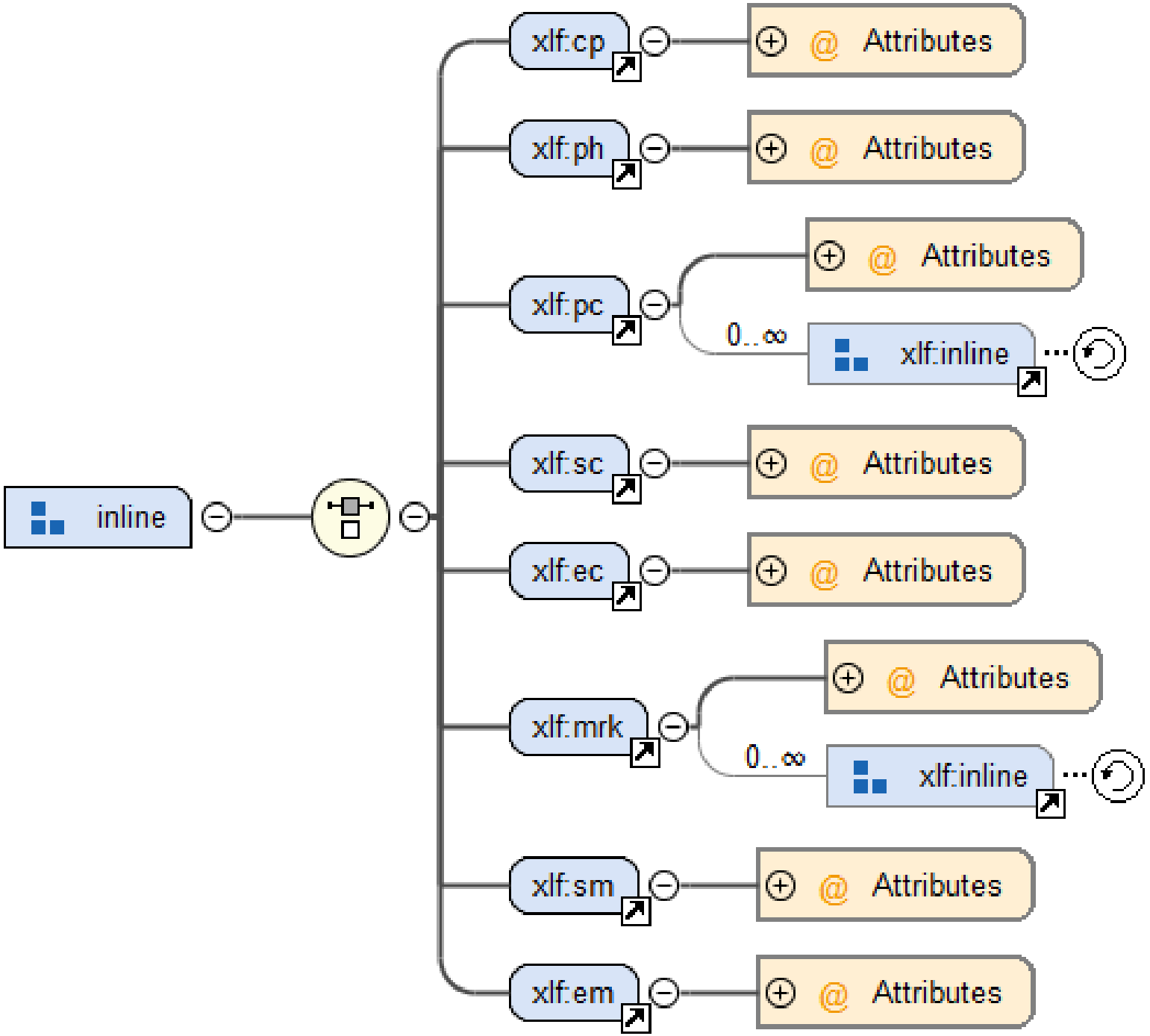
<source xml:lang="EN">Could frame thy fearful symmetry? </source>
<target xml:lang="CS">že z ní tvůj souměr sestrojil? </target>

```
<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>
```

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.

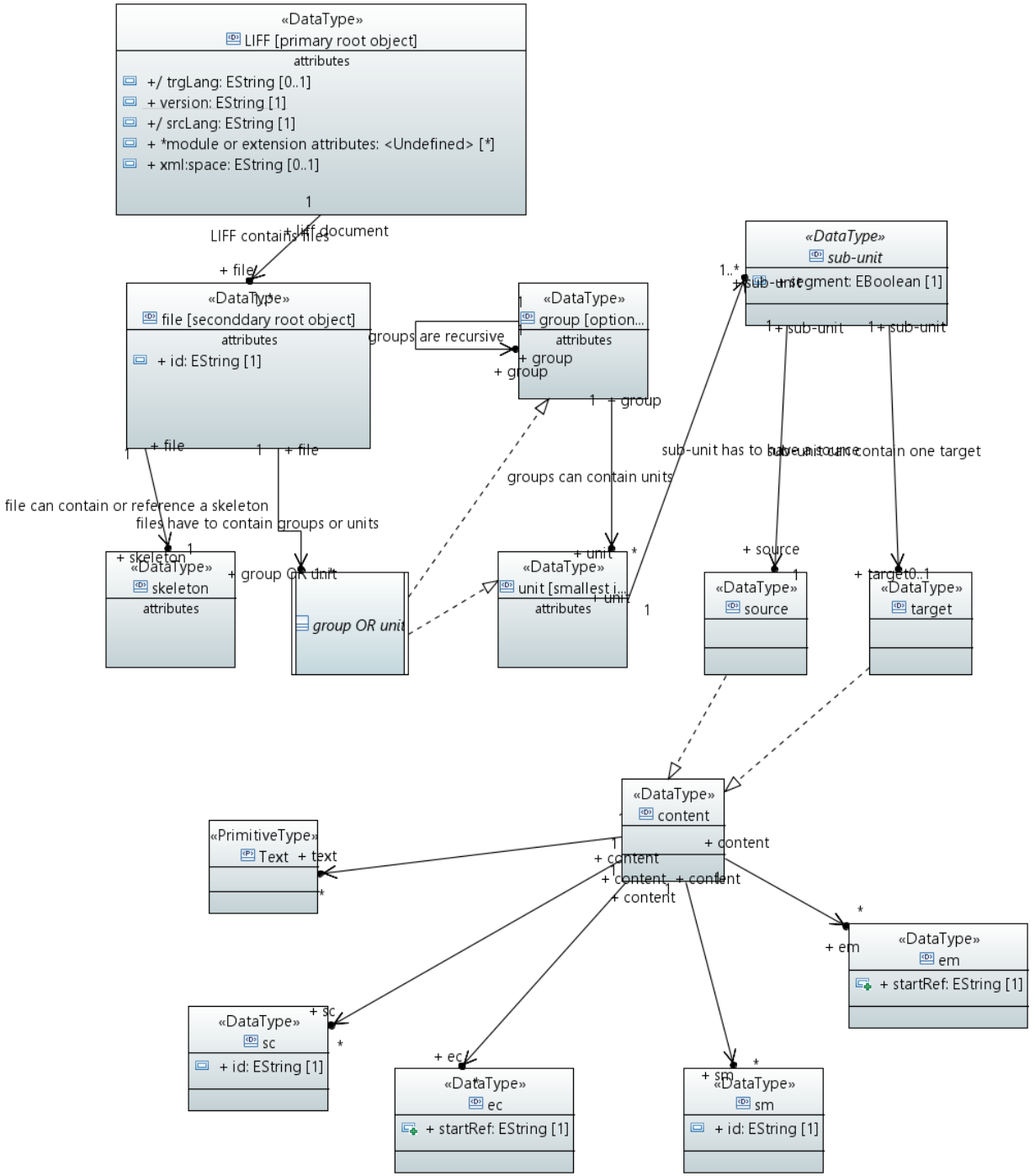

xliff
@ Attributes
1..∞
xlf:file
@ Attributes
xlf:skeleton
0..∞
##other
xlf:notes
1..∞
xlf:note
1..∞
xlf:unit
@ Attributes
0..∞
##other
xlf:notes
xlf:originalData
1..∞
xlf:segment
xlf:ignorable
xlf:group
@ Attributes
0..∞
##other
xlf:notes
0..∞
xlf:unit
xlf:group


unit
@ Attributes
0..∞
##other
xlf:notes
xlf:originalData
1..∞
xlf:segment
@ Attributes
xlf:source
xlf:target
xlf:ignorable
@ Attributes
xlf:source
xlf:target

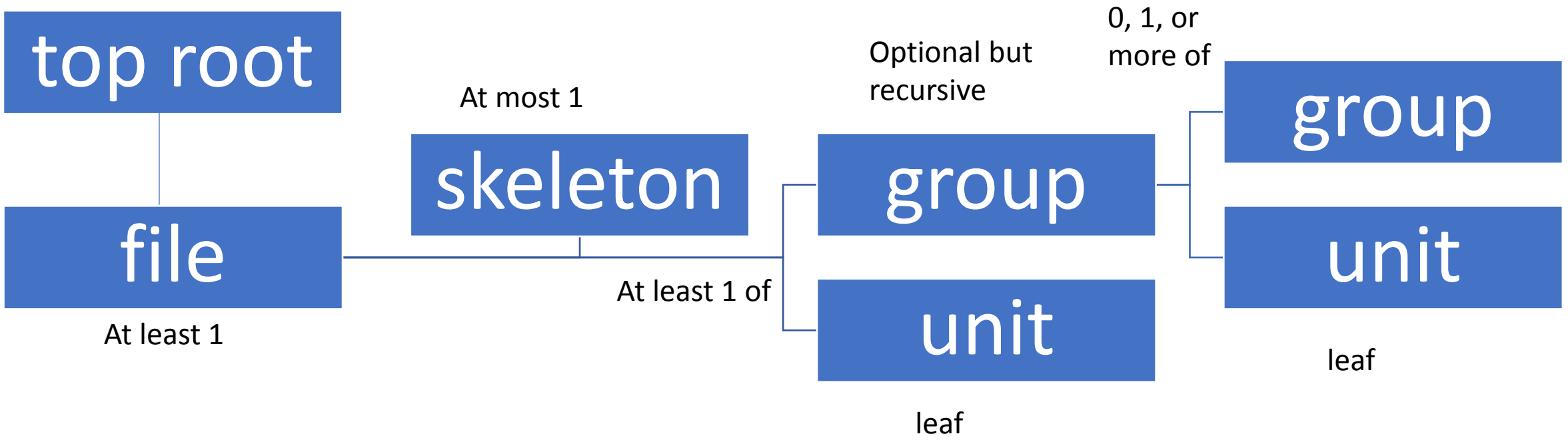

inline
xlf:cp
@ Attributes
xlf:ph
@ Attributes
xlf:pc
@ Attributes
0..∞
xlf:inline
xlf:sc
@ Attributes
xlf:ec
@ Attributes
xlf:mrk
@ Attributes
0..∞
xlf:inline
xlf:sm
@ Attributes
xlf:em
@ Attributes

Parsing an Object Model vs Regular Expressions

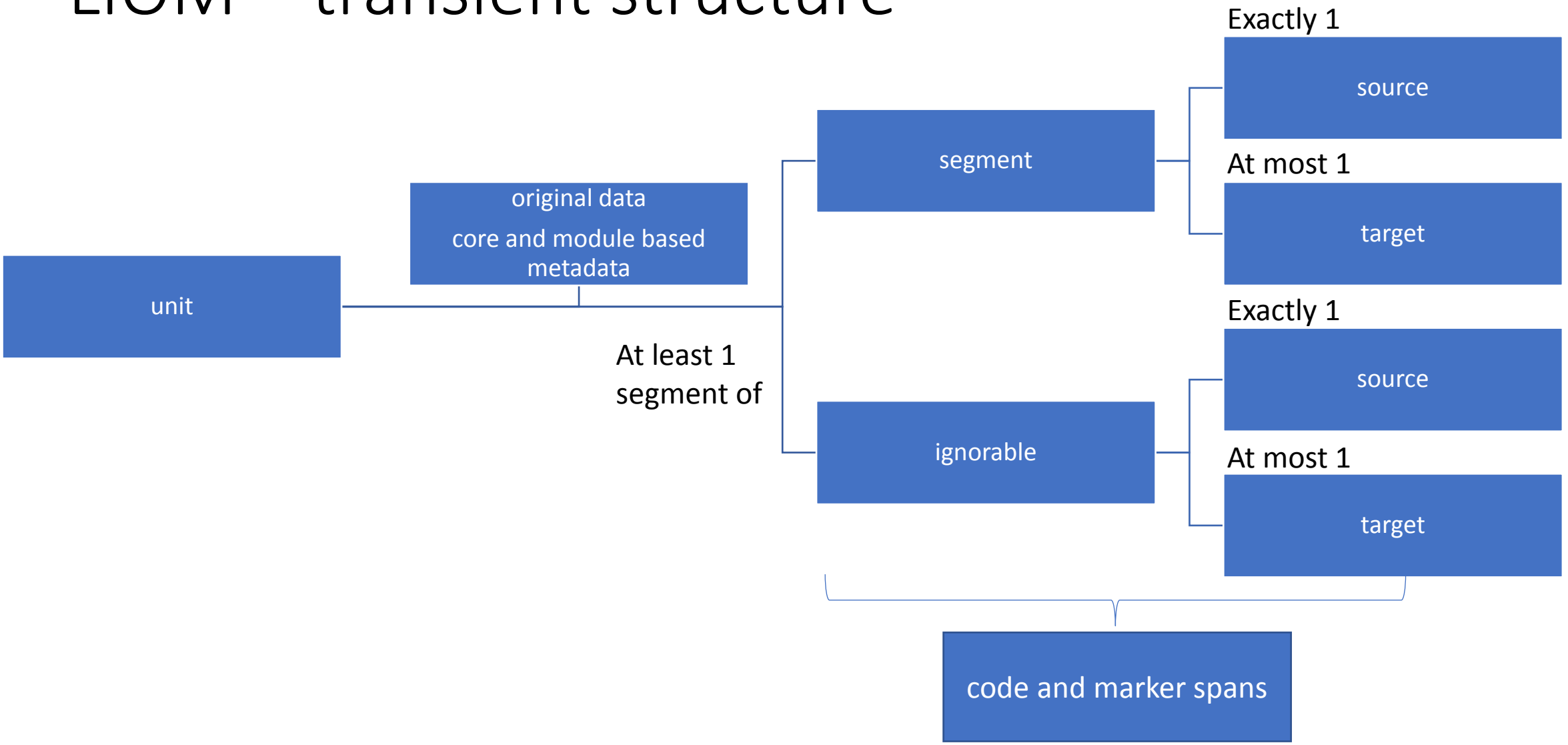
LIOM


«DataType»
LIFF [primary root object]
attributes
+/ trgLang: EString [0..1]
+ version: EString [1]
+/ srcLang: EString [1]
+ *module or extension attributes: <Undefined> [*]
+ xml:space: EString [0..1]
1
LIFF contains files
+ liff document
+ file
«DataType»
file [seconddary root object]
attributes
+ id: EString [1]
groups are recursive
«DataType»
group [option...
attributes
+ group
+ group
«DataType»
sub-unit
+ segment: EBoolean [1]
1..*
+ sub-unit
1 + sub-unit
1 + sub-unit
sub-unit has to have a source
sub-unit can contain one target
groups can contain units
file can contain or reference a skeleton
files have to contain groups or units
+ skeleton
«DataType»
skeleton
attributes
+ group OR unit
group OR unit
+ unit
«DataType»
unit [smallest i...
attributes
+ unit
+ source
«DataType»
source
+ target 0..1
«DataType»
target
«DataType»
content
+ content
+ content
+ content
+ content
«PrimitiveType»
Text
+ text
+ em
«DataType»
em
+ startRef: EString [1]
+ sc
«DataType»
sc
+ id: EString [1]
+ ec
«DataType»
ec
+ startRef: EString [1]
+ sm
«DataType»
sm
+ id: EString [1]

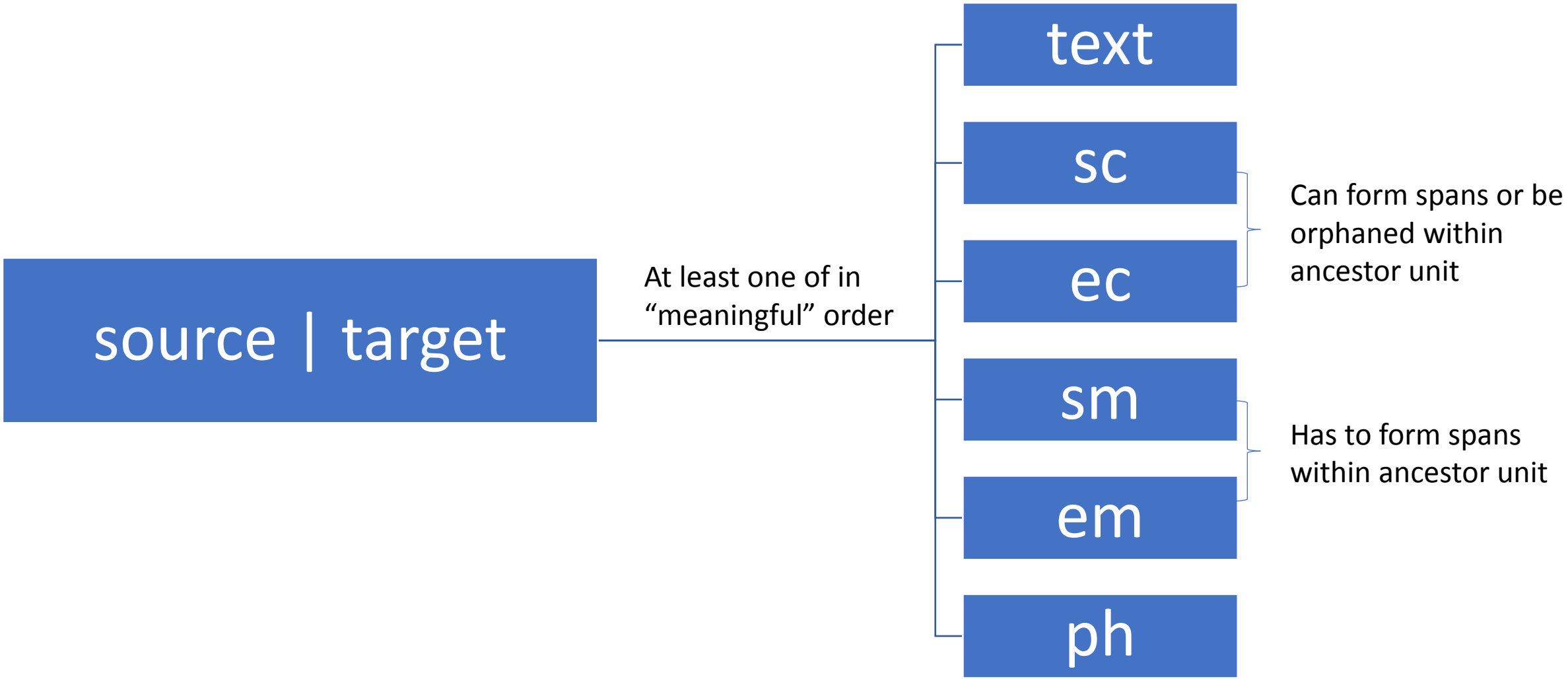
LIOM – static structure



LIOM – transient structure



LIOM – inline content model



XLIFF Top Level Element

```
<xliff xmlns="urn:oasis:names:tc:xliff:document:2.0"
xmlns:uext1="http://example.com/userextension/1.0"
xmlns:uext2="http://example.com/userextension/2.0"
version="2.1" srcLang="en" trgLang="fr">
 <file … >
    <group … > /arbitrary group depth including 0/
       <unit … > [ … /truncated payload structure / … ]
       </unit>
    </group>
 <file>
</xliff>
```

JLIFF Anonymous Top Level Object

```
{
 "jliff": "2.1",
 "@context": {
 "uext1": "http://example.com/userextension/1.0",
 "uext2": "http://example.com/userextension/2.0"
 },
 "srcLang": "en",
 "trgLang": "fr",
 "files | subfiles | sugbroups | subunits": [ … /truncated payload structure / … ]
}
```

Thanks a million!

Q&A

@merzbauer
@DavidFatDavidF
david.filip@adaptcentre.ie

Resources

Recent papers

Filip, D., Ritchie, P., van Engelen, R., 2019. JLIFF, Creating a JSON Serialization of OASIS XLIFF, in: XML Prague 2019 - Conference Proceedings, XML Prague. Presented at the XML Prague 2019, University of Economics, Prague, Prague, pp. 295–321.

http://archive.xmlprague.cz/2019/files/xmlprague-2019-proceedings.pdf#page=307

Filip, D., Husarčík, J., 2018. Modification and Rendering in Context of a Comprehensive Stand-ards Based L10n Architecture. Proceedings ASLING Translating and the Computer, Translating and the Computer 40, 95–112.

https://www.asling.org/tc40/wp-content/uploads/TC40-Proceedings.pdf#page=103

XLIFF, LIOM, and JLIFF resources (1)

[ITS20] D. Filip, S. McCance, D. Lewis, C. Lieske, A. Lommel, J. Kosek, F. Sasaki, Y. Savourel, Eds.: Internationalization Tag Set (ITS) Version 2.0. W3C Recommendation, 29 October 2013. W3C. http://www.w3.org/TR/its20/
[JGT] P. Ritchie, JLIFF Graph Tools. Vistatec, 2019. https://github.com/vistatec/JliffGraphTools/commit/74ffde990d8dd6d6d5d3f80d78e76ea8b0dc8736
[JLIFF] D. Filip and R. van Engelen, JLIFF Version 1.0 [wd01]. OASIS, 2018. https://github.com/oasis-tcs/xliff-omos-jliff/commit/7e63e0d766bb7394f9dccaa93d7fa54bf1a394d3
[JLIFFSchema] R. van Engelen, JLIFF Version 1.0, JSON Schema [wd01]. OASIS, 2018. https://github.com/oasis-tcs/xliff-omos-jliff/commit/2ed3b57f38548600f1261995c466499ad0ade224/>
[JSON-LD] M. Sporny, G. Kellogg, M. Lanthaler, Eds. JSON-LD 1.0, A JSON-based Serialization for Linked Data W3C Recommendation 16 January 2014. https://www.w3.org/TR/2014/REC-json-ld-20140116//>
[L10nStandards] D. Filip: Localization Standards Reader 4.0 [v4.0.1], Multilingual, vol. 30, no. 1, pp. 59–73, Jan/Feb-2019. https://magazine.multilingual.com/issue/jan-feb-2019dm/localization-standards-reader-4-0/
[LIOM] D. Filip, XLIFF 2 Object Model Version 1.0 [wd01]. OASIS, 2018. https://github.com/oasis-tcs/xliff-omos-om/commit/030828c327998e7c305d9be48d7dbe49c8ddf202/>

XLIFF, LIOM, and JLIFF resources (2)

[XLIFF20] T. Comerford, D. Filip, R. M. Raya, and Y. Savourel, Eds.: XLIFF Version 2.0. OASIS Standard, 05 August 2014. OASIS. http://docs.oasis-open.org/xliff/xliff-core/v2.0/os/xliff-core-v2.0-os.html
[XLIFF21] D. Filip, T. Comerford, S. Saadatfar, F. Sasaki, and Y. Savourel, Eds.: XLIFF Version 2.1. OASIS Standard, 13 February 2018. OASIS http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html
[XLIFFEMBP] D. Filip and J. Husarčík, Eds., XLIFF 2 Extraction and Merging Best Practice, Version 1.0. Globalization and Localization Association (GALA) TAPICC, 2018. https://galaglobal.github.io/TAPICC/T1/WG3/rs01/XLIFF-EM-BP-V1.0-rs01.xhtml/>
[XLIFFglsTBXBasic] J. Hayes, S. E. Wright, D. Filip, A. Melby, and D. Reineke, Interoperability of XLIFF 2.0 Glossary Module and TBX-Basic, Localisation Focus, vol. 14, no. 1, pp. 43–50, Apr. 2015.https://www.localisation.ie/resources/publications/2015/260
[XLIFFRender] D. Filip and J. Husarčík, Modification and Rendering in Context of a Comprehensive Standards Based L10n Architecture, Proceedings ASLING Translating and the Computer, vol. 40, pp. 95–112, Nov. 2018. https://www.asling.org/tc40/wp-content/uploads/TC40-Proceedings.pdf

XLIFF 2 Compliance

Low level open source implementations

- Okapi XLIFF Toolkit
  https://bitbucket.org/okapiframework/xliff-toolkit
- Microsoft XLIFF 2 Object Model
  https://github.com/Microsoft/XLIFF2-Object-Model

Other implementers:

SDL[all products] , Lionbridge, Memsource, Ocelot, DITA XLIFF Toolkit, Xmarker, Globalsight, XTM, Multilizer, AEM at al.