Natural Language Processing (NLP) in Real-World Multilingual Production

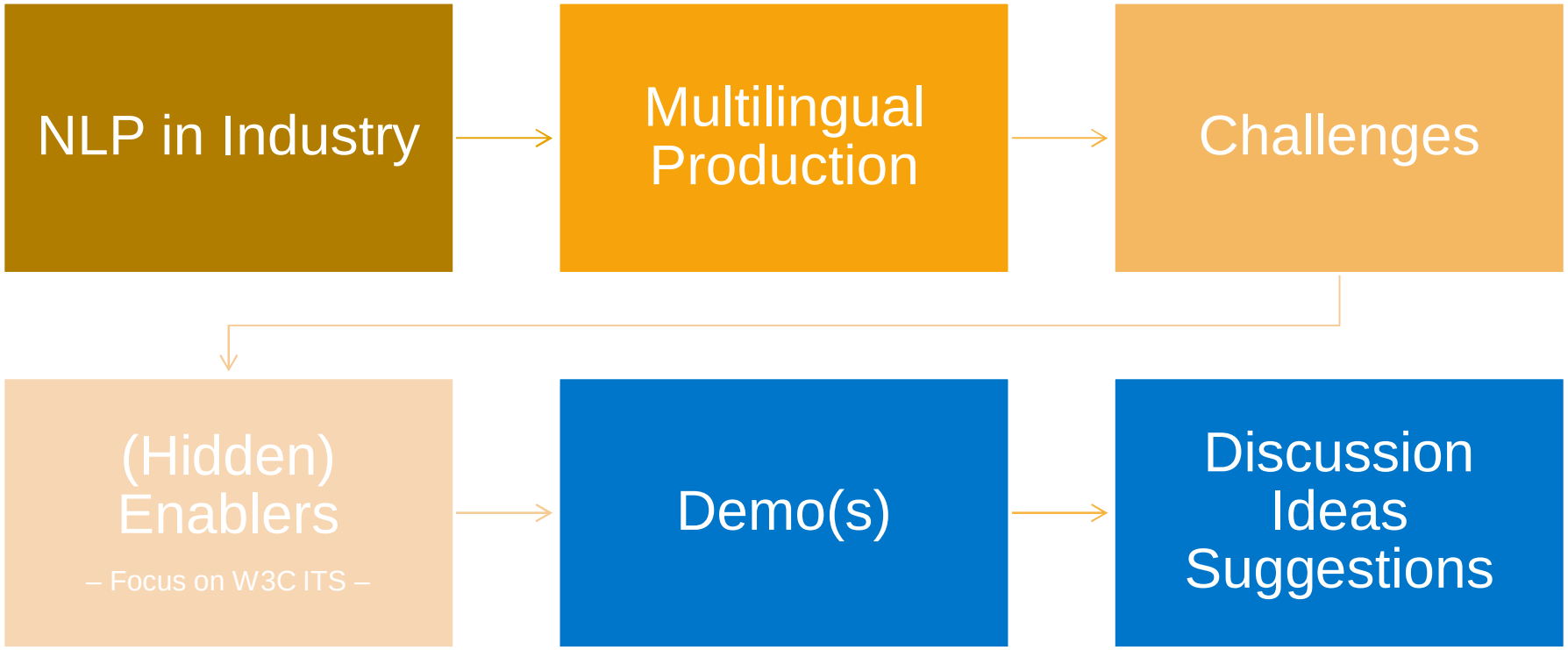
Christian Lieske (Globalization Services, SAP AG)

– A Personal View –

Grammatical Framework Summer School (August 2013)

This presentation is **purely personal** — my employer does not have responsibility for any information contained here.

Overview

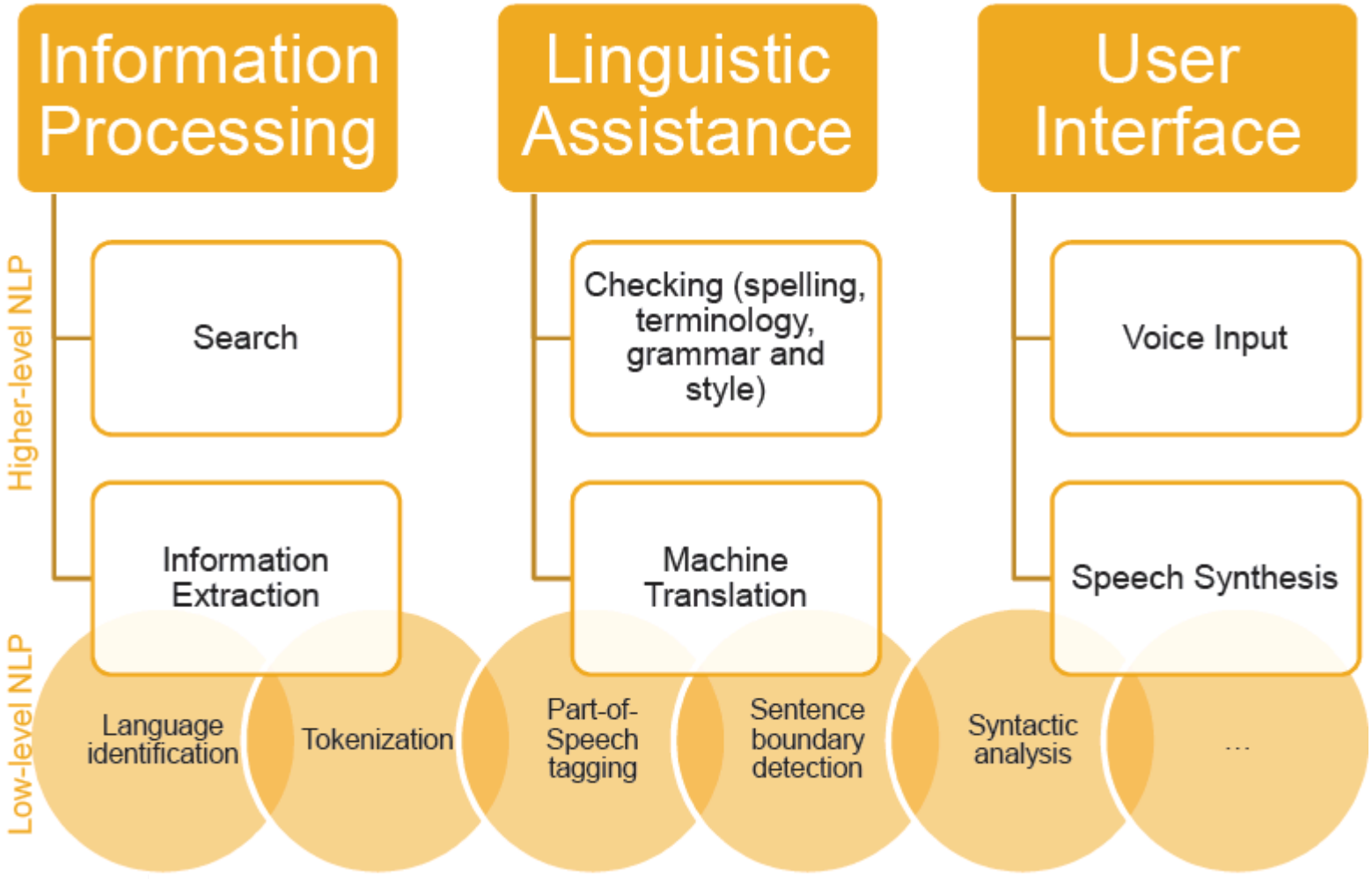

NLP in Industry
Multilingual Production
Challenges
(Hidden) Enablers
– Focus on W3C ITS –
Demo(s)
Discussion Ideas Suggestions

NLP in Industry

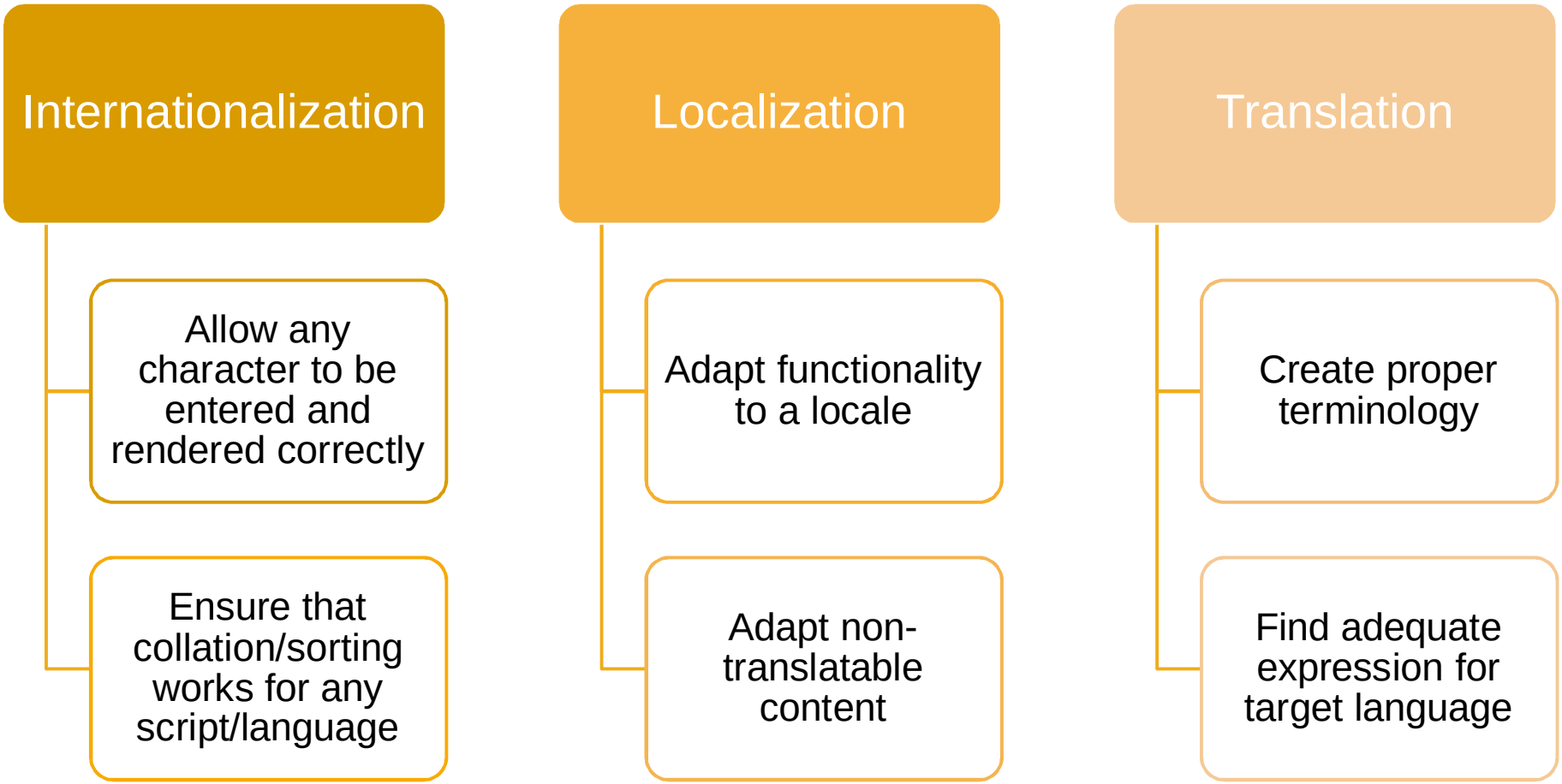
Part of Solution or Application

(Multilingual) Production

Part of Solution of Application



Multilingual Production – Globalization Tripod



Globalization Size, Impact, and Prospects*

82 %
of online shops only in one language

1/3
goes to the translator

2/3
of consumers prefer e-shop in own language

1.8 million
pages translated

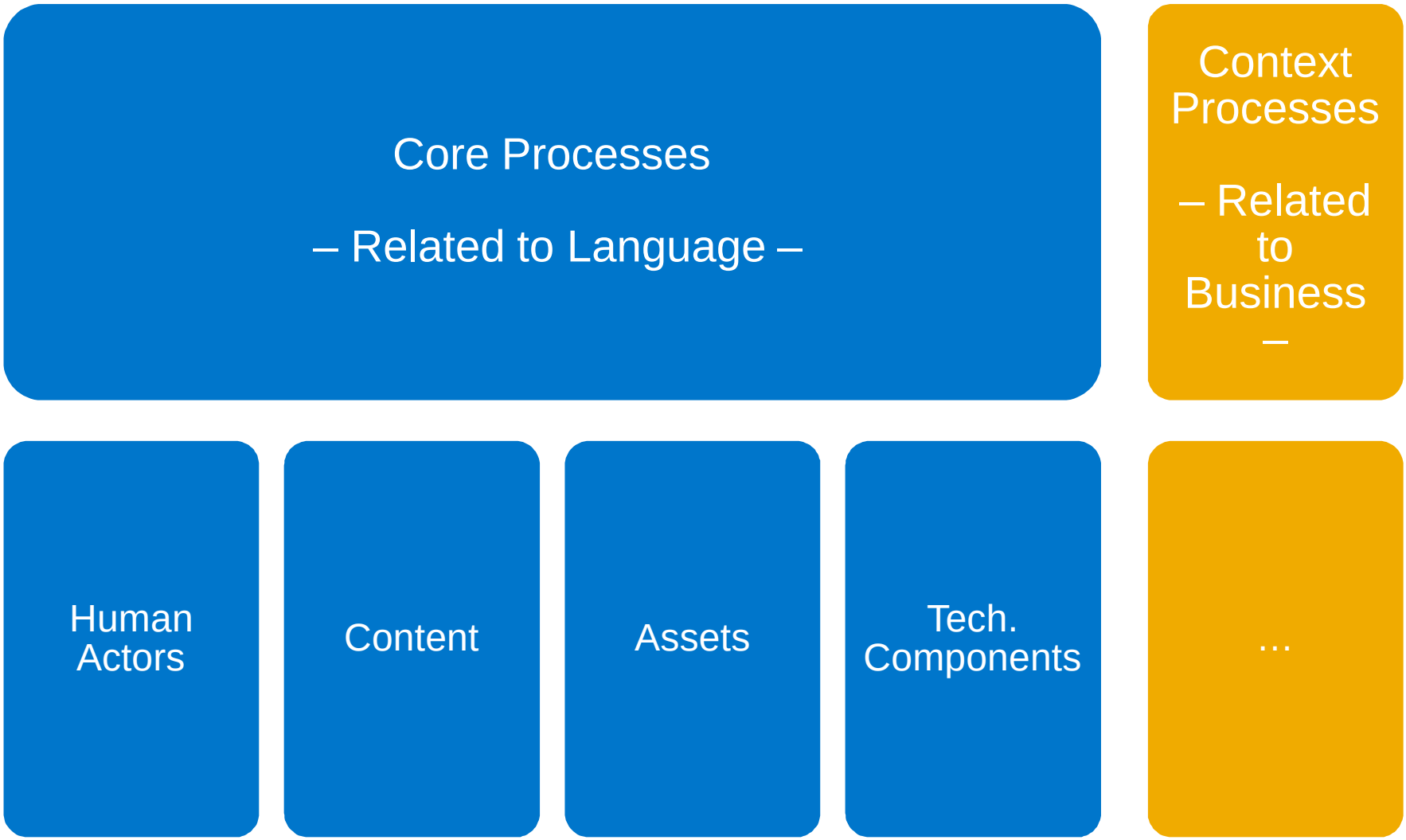
202 million
words translated

$ 6.5 billion
revenues for language services market

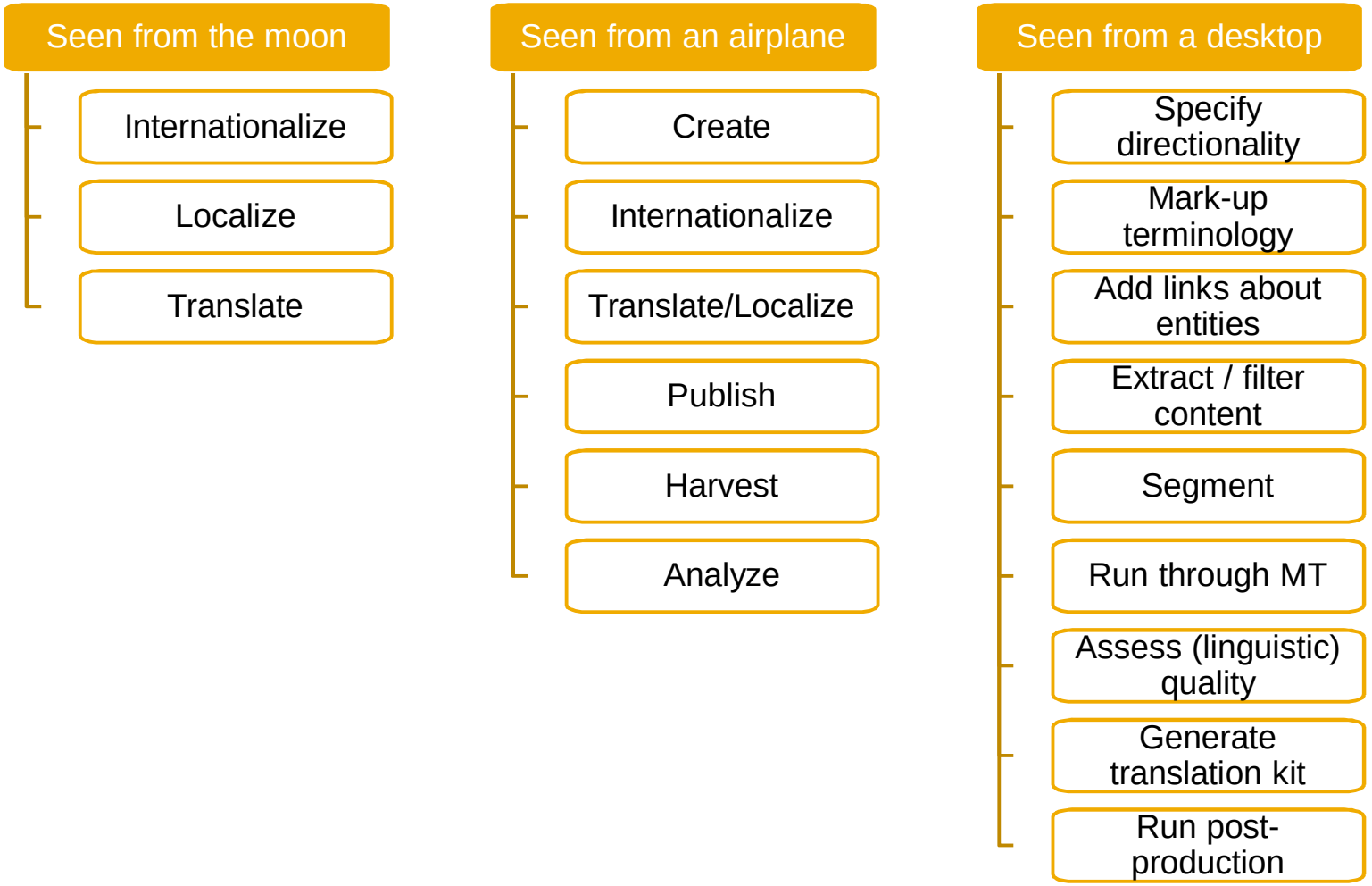
4500/$ 450 million
employees/revenue for large Language Service Provider

*Numbers not current

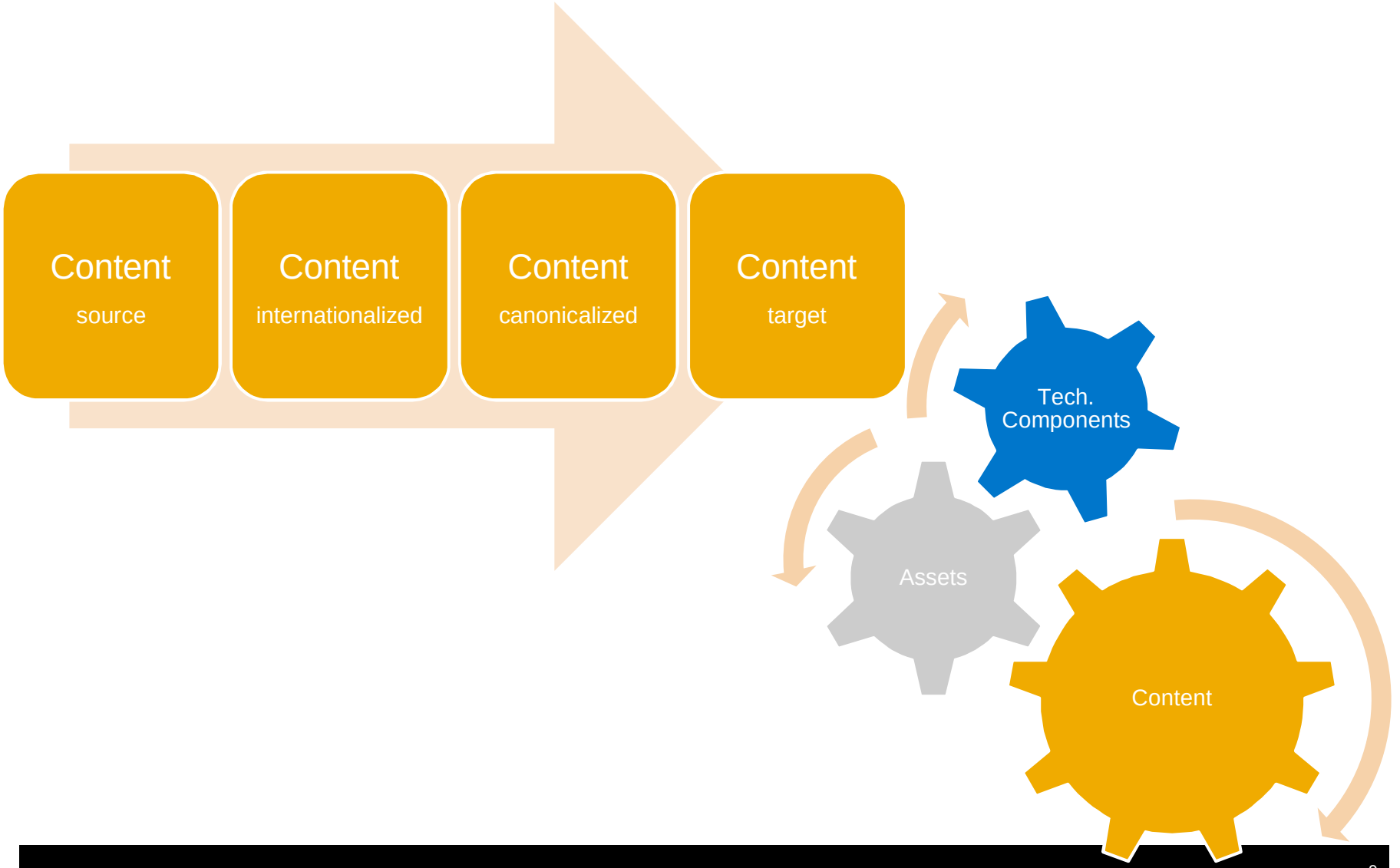
Production‘s Core and Context



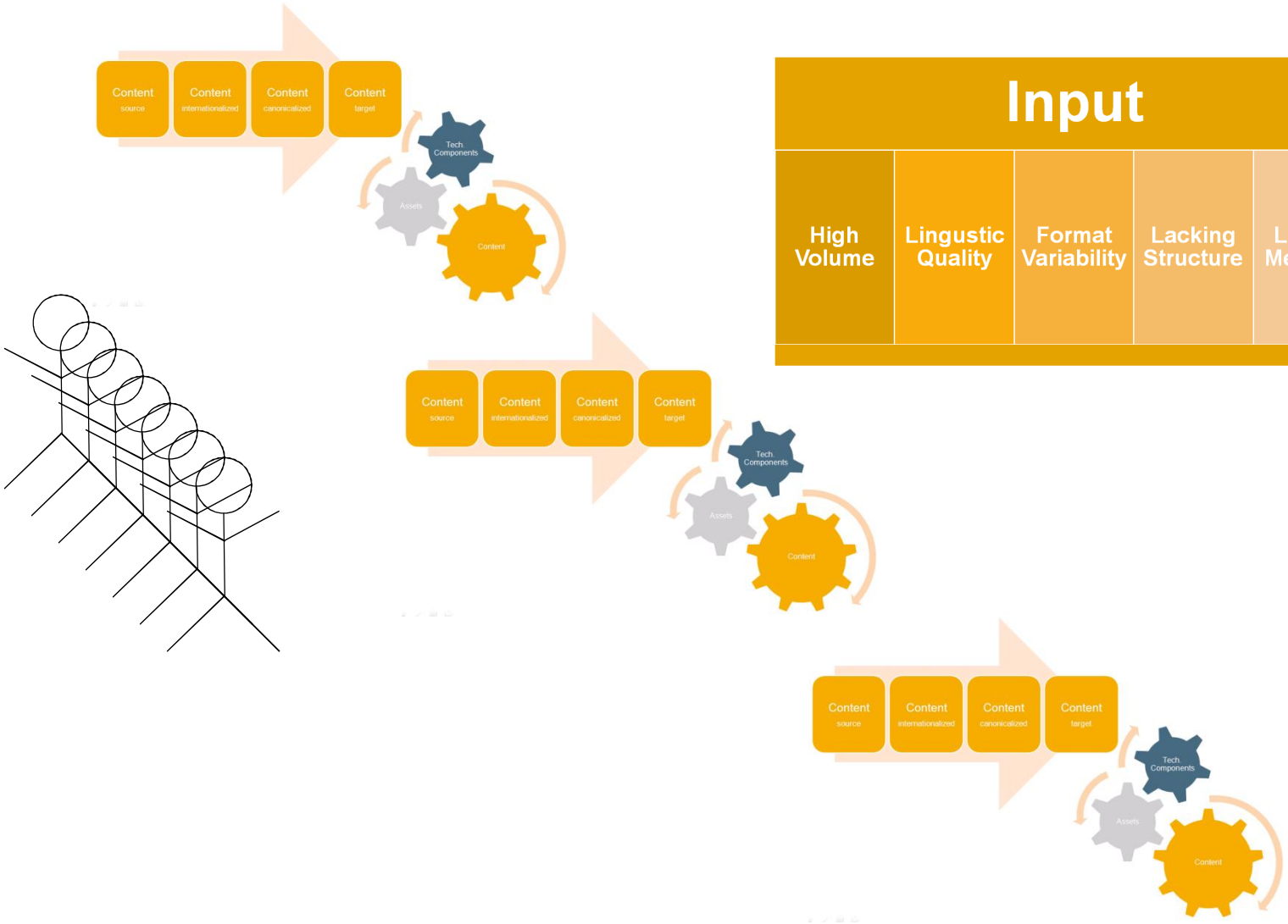
Multilingual Production – Challenges (1/4)



Multilingual Production – Challenges (2/4)



Multilingual Production – Challenges (3/4)



Multilingual Production – Challenges (4/4)

Anyone, anything (proprietary, XML ...), anytime

Scaling, consistency, compliance …

Coupling

- Object Linking and Embedding, HTTP, Web Services, ...
- Libraries/Application Programming Interfaces/Software Development Kits
- Orchestration (e.g. synchronization of calls, and "bus-like" integration or annotation framework)

http://www.dagstuhl.de/mat/Files/12/12362/12362.LieskeChristian.Slides.pdf

Multilingual Content Processing for Multilingual Production

```
<p>We are presenting on the <a hrif="http://www.tekom.de/">tcworld conference 2011</a>
<img src="tcWorld2011.jpg" width="450" height="174" alt="The tcWorld 2011 Logo"/>.</p>
```

Content is more than natural language text

Quality, cost, and delivery count

Often more than just linguistic stuff is in the mix (Natural Language Processing vs. Text Technology)

Sample natural language questions/tasks

Is there existing or new terminology?

Are spelling, grammar, and style alright?

Can I recycle an existing translation?

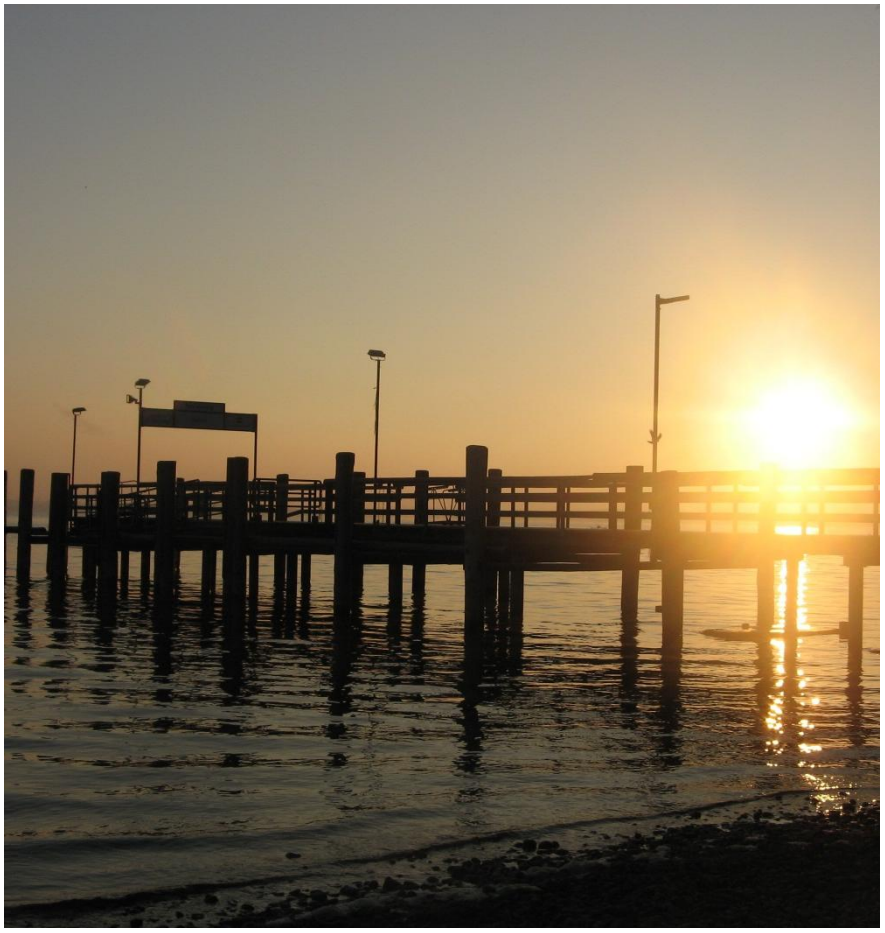
```
<p>Text technology provides answers not only related to characters, but also to other areas:</p>
<ul>
  <li>Content (e.g. HTML, XML, XML-based vocabularies like DocBook or DITA, ...)</li>
  <li>Metadata (e.g. RDF)</li>
  <li>Filters to go from general XML to XLIFF via ITS</li>
</ul>
```

```
<p>テキストテクノロジーは文字だけではなく、下記の他の分野にも基本になる。</p>
```

Enablers – Overview (1/2)

You need Natural Language Processing (NLP)/Language Technology (LT) for natural language.

Text Technology is the base for solid, sustainable NLP/LT in real world deployment scenarios.



Enablers – Overview (2/2)

Best Practices and Standardization

Computer-Assisted Linguistic Quality Support

Computer-Assisted Linguistic Assistance

i. Needs assets
ii. Creates assets
iii. Relates to Natural Language Processing

Text Technology Standards for Universal Coverage (1/2)

Think about content with the world in mind

- Can I encode all characters?
- Can my HTML display the content properly?
- Can I translate efficiently?

Only world-ready NLP/LT is solid and sustainable

Unicode standard

- Allows for content creation and processing in a wide range of languages.
- Applied in many contexts (XML, HTML, multilingual Web addresses like http://ja.wikipedia.org/wiki/東京 , etc.)

Unicode support should be considered as a key feature of any NLP/LT offering.

Text Technology Standards for Universal Coverage (2/2)

Content formats (e.g. HTML, XML, XML-based vocabularies like DocBook or DITA, …)

Metadata (e.g. Resource Description Framework)

Filters, e.g. to go from general XML to XLIFF (XML Localization Interchange File Format) based on W3C Internationalization Tag Set (ITS)

…

Standards for Universal Efficiency and Effectiveness

Assets

- Terminology – TermBase eXchange (TBX)
- Former Translations – Translation Memory eXchange (TMX)

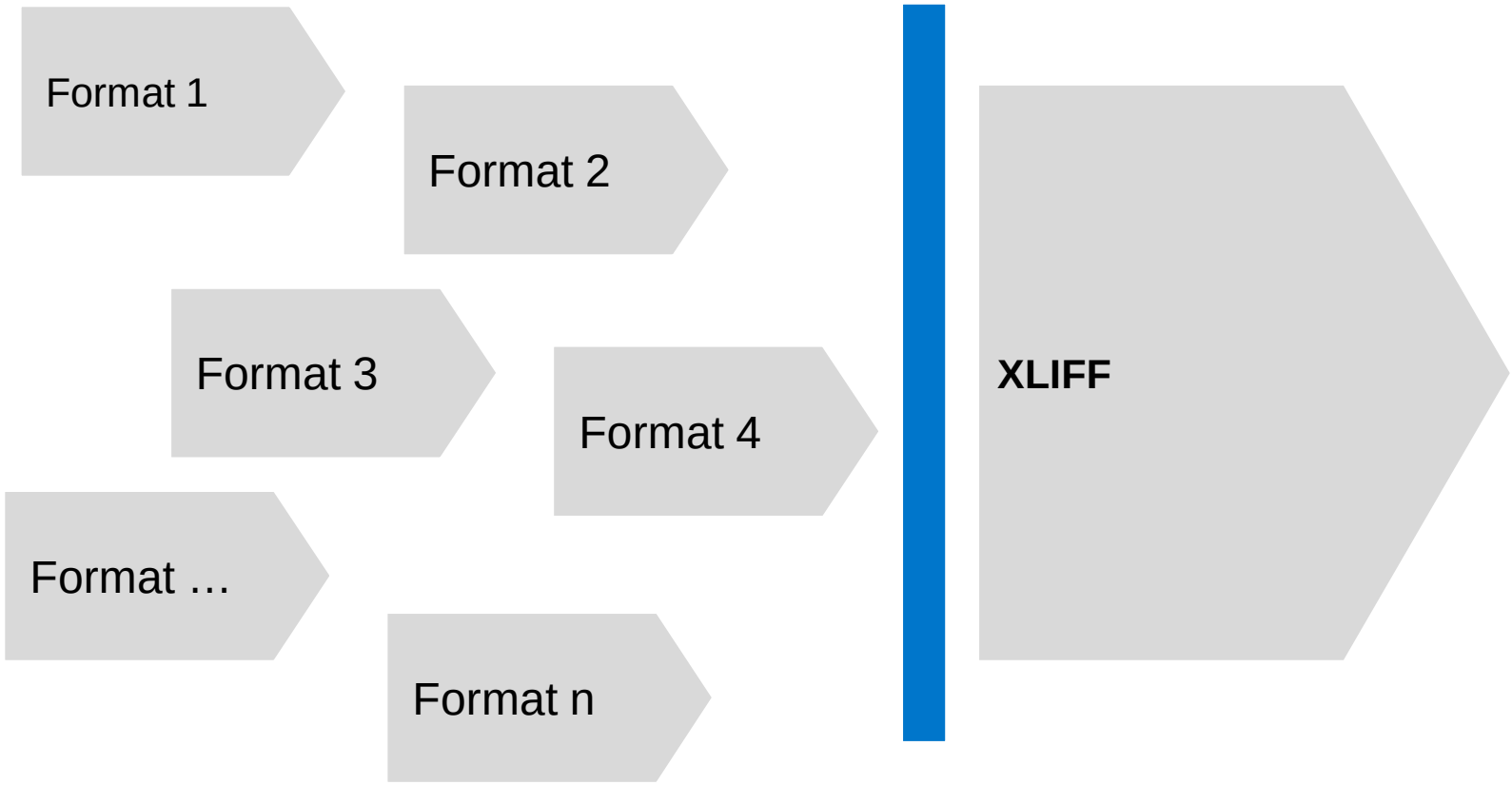
Canonicalized Content

- XML Localization Interchange File Format (XLIFF)

NL(P)-related Resource Descriptions

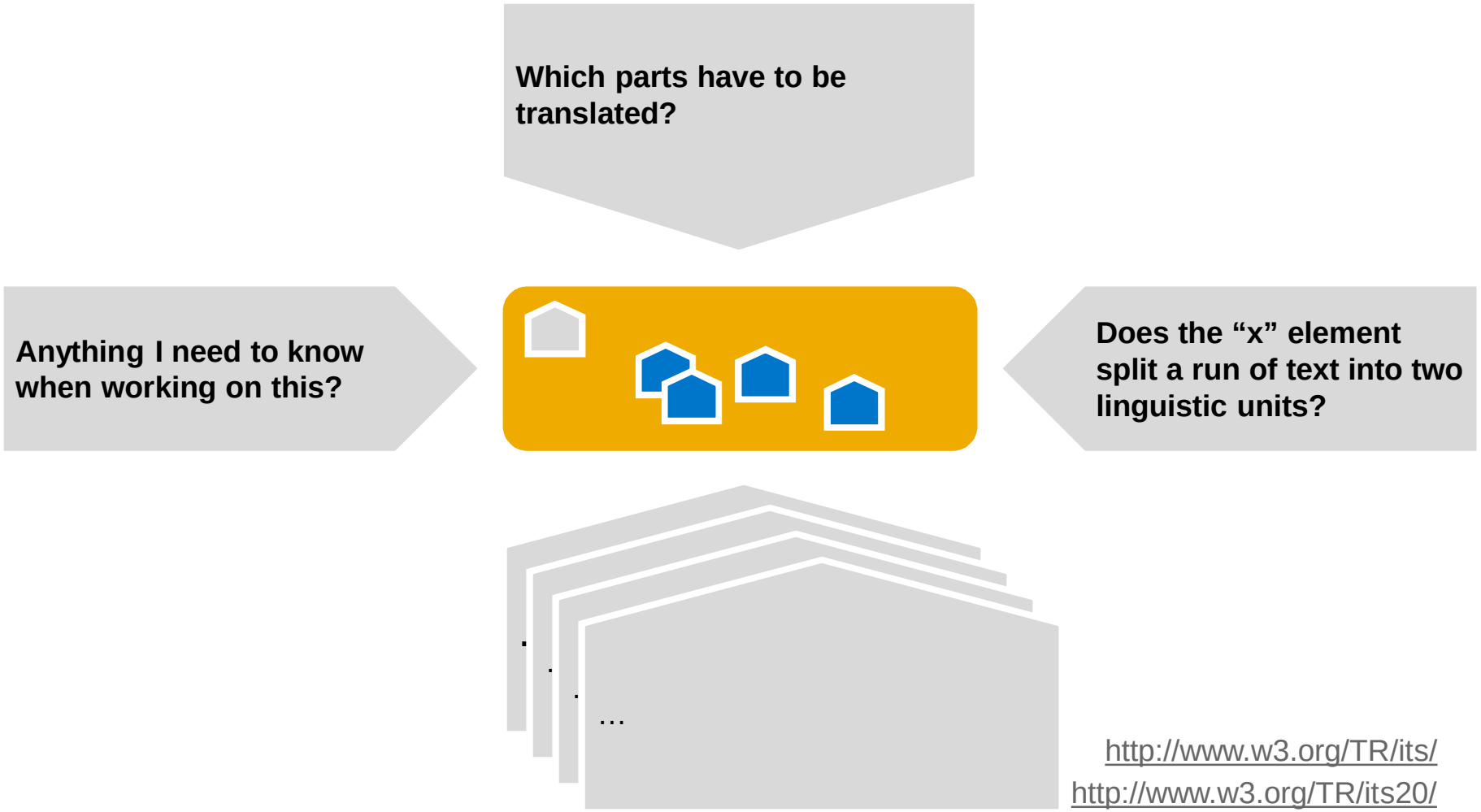
- Internationalization Tag Set (ITS) 1.0 and 2.0

Enablers – Canonicalized Content



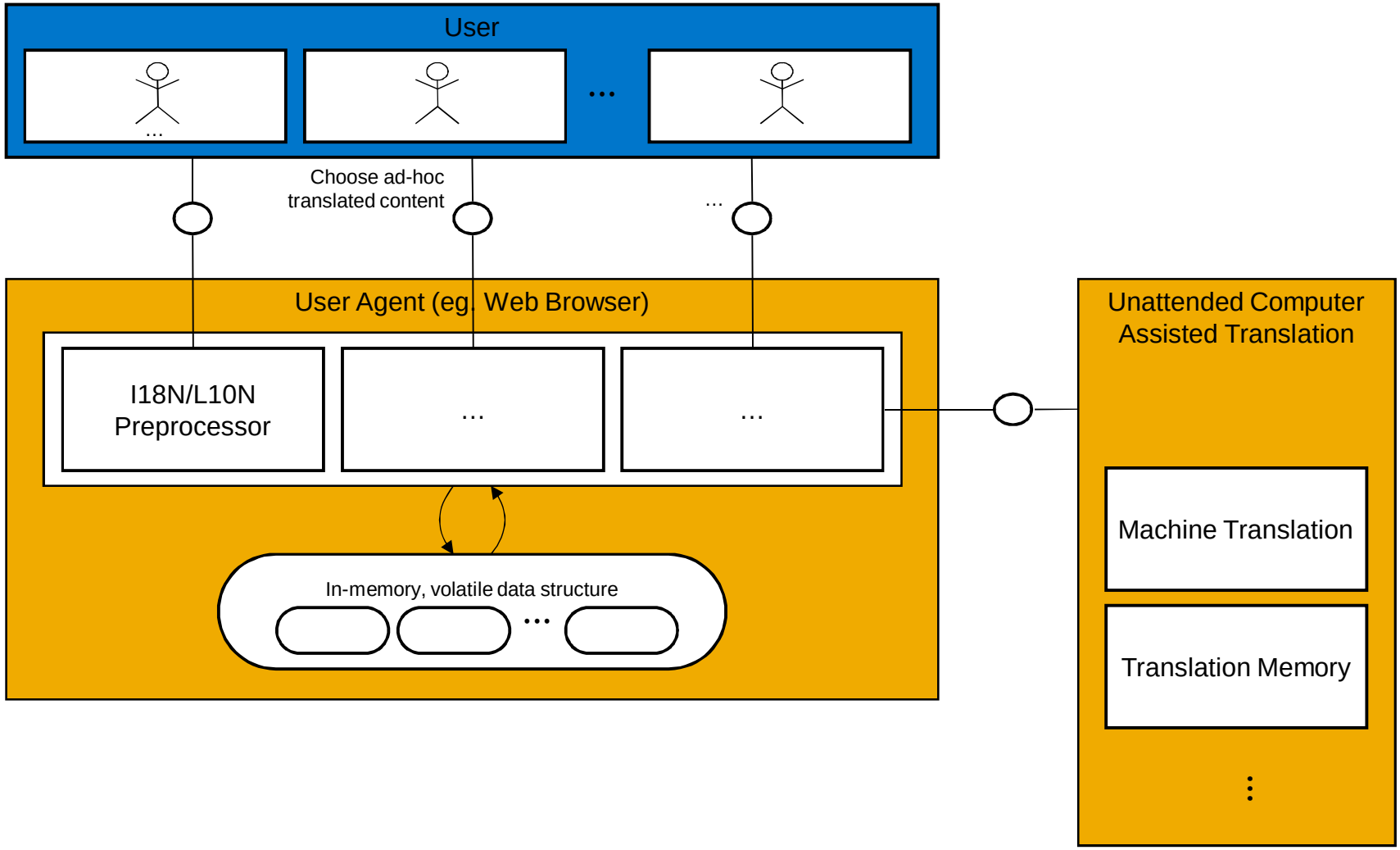
http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html

Enablers – (Non-intrusive) Universal NLP-related Resource Descriptions

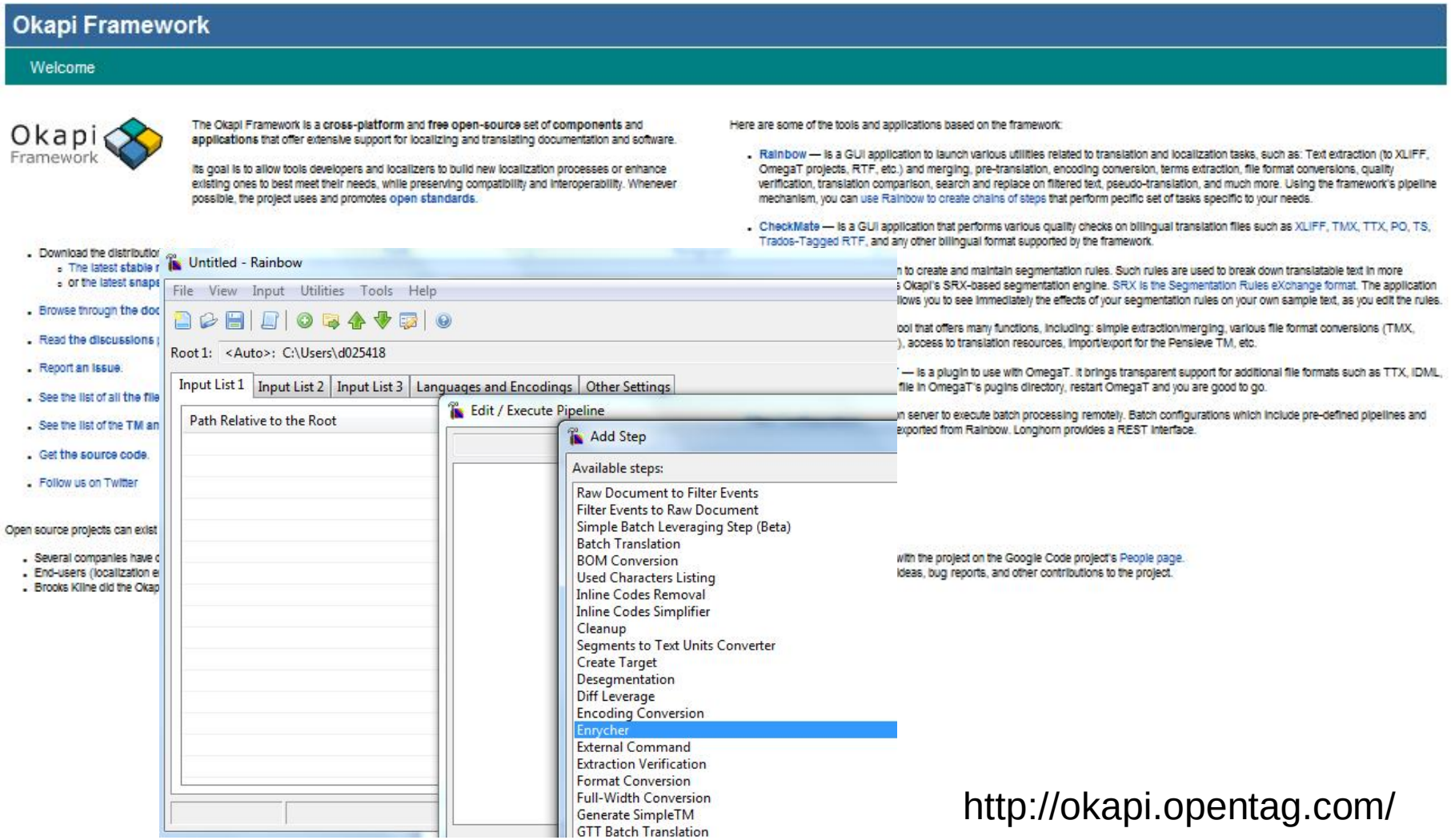


http://www.w3.org/TR/its/
http://www.w3.org/TR/its20/

Standards-based Scenario in Main Web Stack

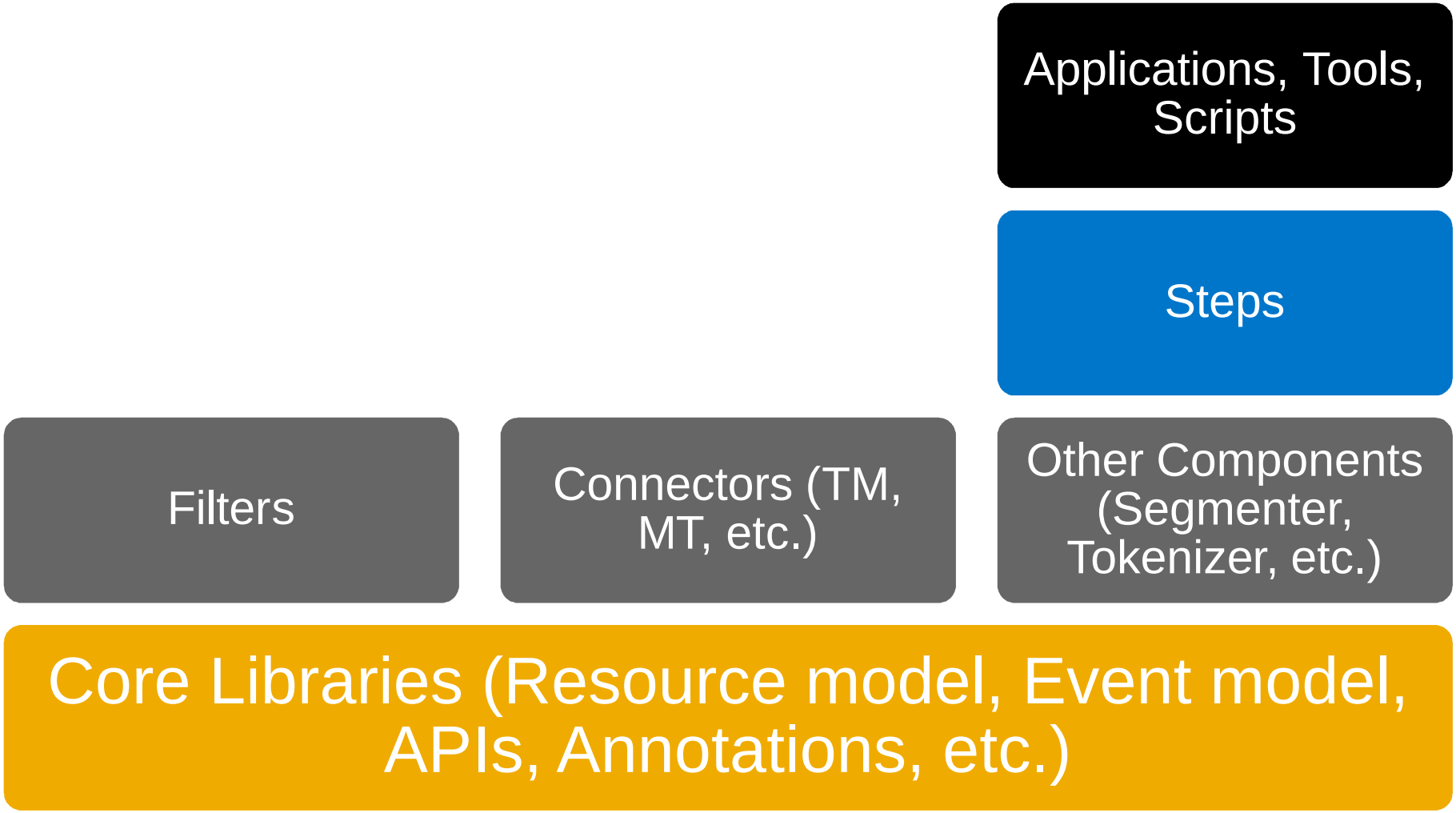


Standards-based Scenario – OKAPI / RAINBOW / CheckMate (1/2)

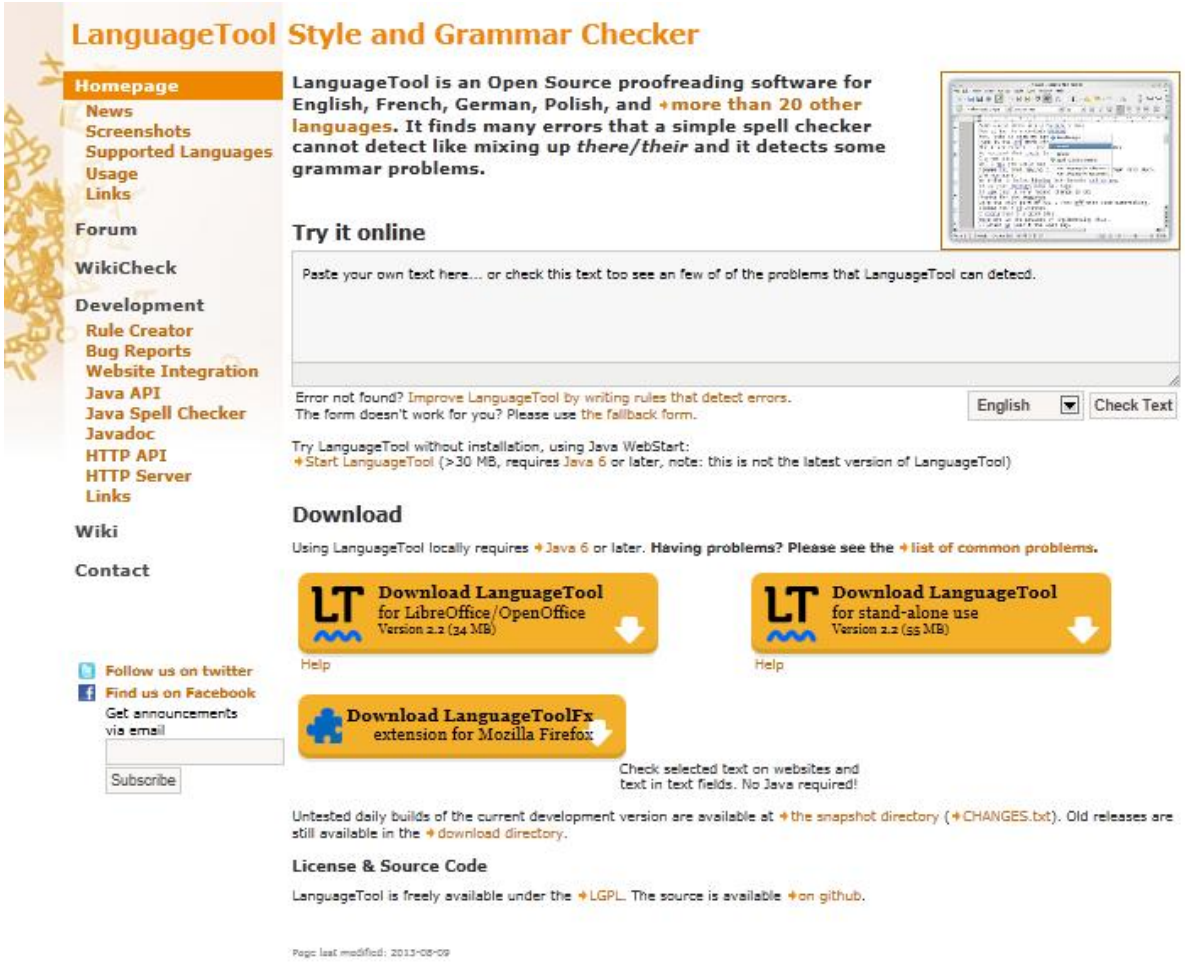


http://okapi.opentag.com/

Standards-based Scenario – OKAPI / RAINBOW / CheckMate (2/2)



Standards-based Scenario – LanguageTool (1/5)



http://www.languagetool.org/

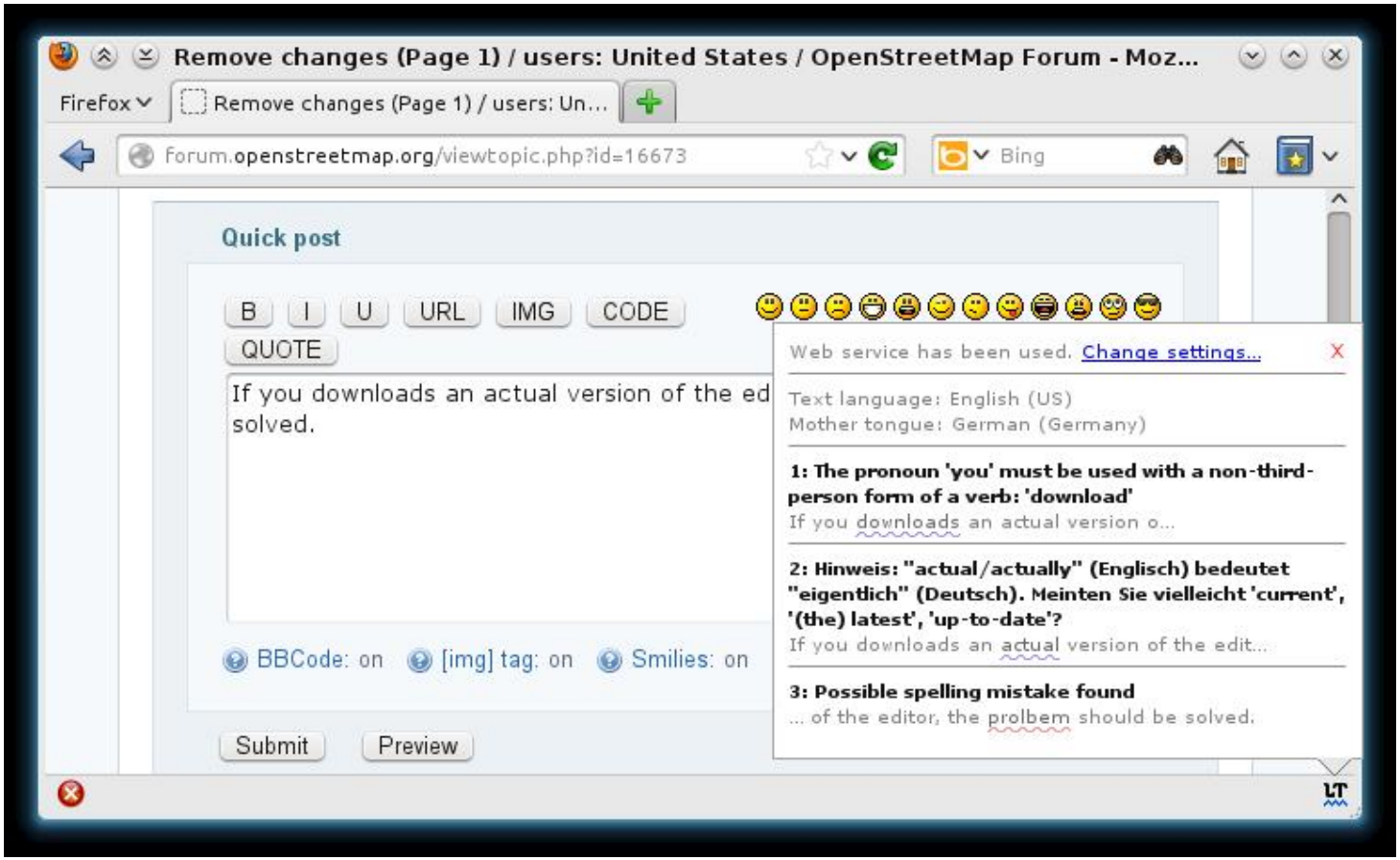
Standards-based Scenario – LanguageTool (2/5)

```
<rule id="GENITIV-ARTIKEL">
<pattern>
  <token postag_regexp="yes"
postag="SUB:.*"/>

  <token postag_regexp="yes"
postag="ART:(DEF|IND):GEN:.*" skip="-
1"/>

  <token postag_regexp="yes"
postag="SUB:GEN:.*"/>
</pattern>

<message>Genitiv gefunden:
&quot;<match no="2"/>&quot;
Vermeiden Sie den Genitiv.</message>

</rule>
```

```
 <rule id="GENITIV-
POSSESSIVPRONOMEN">
<pattern>
  <token postag_regexp="yes"
postag="SUB:.*"/>

        <token postag_regexp="yes"
postag="PRO:POS:GEN:.*" skip="-1"/>

        <token postag_regexp="yes"
postag="SUB:GEN:.*"/>

      </pattern>

      <message>Genitiv gefunden:
&quot;<match no="2"/>&quot; Vermeiden
Sie den Genitiv.</message>

    </rule>
```

Courtesy of Annika Nietzio

Standards-based Scenario – LanguageTool (3/5)

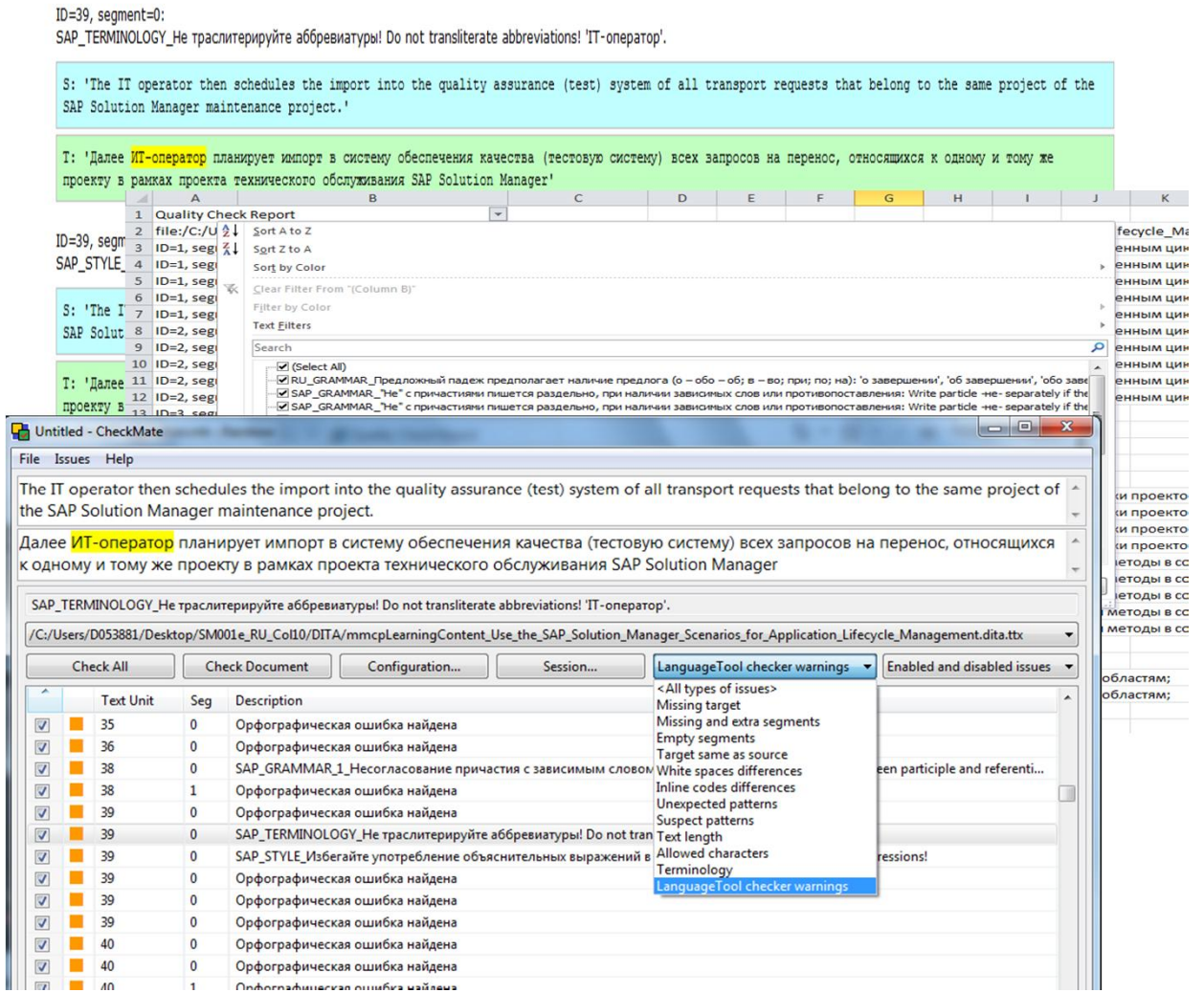


https://addons.mozilla.org/de/firefox/addon/languagetoolfx/

Standards-based Scenario – LanguageTool (4/5)

http://www.languagetool.org/de/leichte-sprache/

Standards-based Scenario – LanguageTool (5/5)



Multilingual content processing needs help

“Which `data` elements need to be processed by NLP?”

<rsrc id="123"> ...

- <data type="text">images/cancel.gif</data>
- <data type="position">12,20</data>
- <data type="text“>**Cancel**</data>
- <data type="position">60,40</data>
- <data type="text“>**Number of files:** </data>

</rsrc>

ITS 2.0 – The help

Comprehensive	• Supports internationalization, translation, localization and other aspects of the multilingual content production cycle
Standardized	• Building on W3C ITS 1.0
Meta data	• data categories, values etc.

ITS 2.0 Basic principles

Say important things

About specific content

In a standard way

- With agreed upon syntax and values

1. Say important things: ITS 2.0 “data categories”

Translate, Localization Note, Terminology, Directionality, Language Information, Elements Within Text, Domain, Text Analysis, Locale Filter, Provenance, External Resource, Target Pointer, Id Value, Preserve Space, Localization Quality Issue, Localization Quality Rating, MT Confidence, Allowed Characters, Storage Size

Definition in prose

Selection of content via two approaches

2. About specific content: Content selection approaches

ITS selection can be compared to

- global = “style” element
- local = “style” attribute

Selection global	• XPath to select markup nodes
Selection local	• ITS local attributes

```
<rsrc ...>
<its:rules xmlns:its="http://www.w3.org/2005/11/its"
version="2.0">
        <its:translateRule selector="//data" translate="no"/>
</its:rules>
    •<data type="text" its:translate="yes">Cancel</data>
    •<data type="position">60,40</data> ... </rsrc>
```

3. In a standard way (1/2)

Pre-defined (if appl.) meta data values	• “Translate”: “yes” or “no”
Specific defaults (if appl.)	• Elements: translate “yes”, attributes: translate “no”
Specific HTML5 behaviour	• E.g. “alt” attribute default “yes”

3. In a standard way (2/2)

Independent /orthogonal

- Powerful (e.g. easy combination)
- Dublin Core, xml

Strict conformance clauses

- Supported ITS 2.0 data categories
- Supported selection mechanism (local / global) and type of content (HTML / XML)

Why ITS 2.0? (1/2)

ITS 1.0 = simplified view of multilingual content production

Too limited for comprehensive automated content processing/usage scenarios (see http://www.w3.org/TR/mlw-metadata-us-impl/ for various ITS 2.0 usage scenario descriptions)

Example gap: too few data categories

Why ITS 2.0? (2/2)

Coverage for additional types of content: HTML5

- Bridge to Web & app content
- Accommodate relevant HTML5 markup (e.g. HTML5 “translate” attribute behaviour)

Easy mapping/conversion to other formats

- XML Localization Information Markup (XLIFF; status: informal mapping, under discussion) = workflows
- Natural Language Processing Interchange Format (NIF) = bridge to the Semantic Web and Natural Language Processing

Example: MT Confidence

Score from machine translation engine

Example for new ITS capability: Tool traceability

```
<!DOCTYPE html> ...
<body its-annotators-ref="mt-confidence|file:///tools.xml#T1">
 <p>
  <span its-mt-confidence="0.8982">Dublin is the capital of Ireland.</span></p>
</body></html>
```

Example: Locale Filter

Content relevant only for a specific locale

```
<!DOCTYPE html> ...
<div its-locale-filter-list="*-ca">
  <p>Text for Canadian locales.</p>
</div>
<div its-locale-filter-list="*-ca" its-locale-filter-type="exclude">
 <p>Text for non-Canadian locales.</p>
</div> ...
```

Example: Localization Quality Issue

For quality assessment

```
<!DOCTYPE html> ... <
 its-loc-quality-issue-comment="should be 'quality'"
 its-loc-quality-issue-profile-
ref=http://example.org/qaMovel/v1
 its-loc-quality-issue-severity=50
 its-loc-quality-issue-type=spelling>qulaity
```

Demo(s)

1. **„Filtering“ with Okapi Rainbow (built-in filter)**
2. **„Filtering“ with Okapi Rainbow (custom filter configuration based on W3C Internationalization Tag Set)**
3. **Browser-based demo of LanguageTool**

Discussion/Ideas/Suggestions

1. GF for checking term entry conventions
2. GF for term lookup
3. GF in automated pre-editing for MT
4. GF in automated post-editing for MT
5. GF and markup
6. GF from Okapi (e.g. one or more steps)
7. Okapi from GF (e.g. as pre- and post-processor)
8. GF and LanguageTool
9. GF and TermBase eXchange (TBX)
10. GF and Translation Memory eXchange (TMX)
11. GF and Pseudo-translation

Thank You!

More information on W3C ITS:

http://www.w3.org/TR/its/
http://www.w3.org/TR/its20/

http://www.w3.org/International/its/ig/
http://lists.w3.org/Archives/Public/public-i18n-its-ig (public list, free to subscribe)

Contact:

Christian Lieske
christian.lieske@sap.com
www.sap.com

Disclaimer

All product and service names mentioned and associated logos displayed are the trademarks of their respective companies. Data contained in this document serves informational purposes only. National product specifications may vary.

This document may contain only intended strategies, developments, and is not intended to be binding upon the authors or their employers to any particular course of business, product strategy, and/or development. The authors or their employers assume no responsibility for errors or omissions in this document. The authors or their employers do not warrant the accuracy or completeness of the information, text, graphics, links, or other items contained within this material. This document is provided without a warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability, fitness for a particular purpose, or non-infringement.

The authors or their employers shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials. This limitation shall not apply in cases of intent or gross negligence.

The authors have no control over the information that you may access through the use of hot links contained in these materials and does not endorse your use of third-party Web pages nor provide any warranty whatsoever relating to third-party Web pages.