Lionbridge

Standards and Paradigms

The difference made by standards oriented processes

Joachim Schurig,
Senior Technical Director Language Technology,
Chair ETSI ISG LIS

How do we produce today?

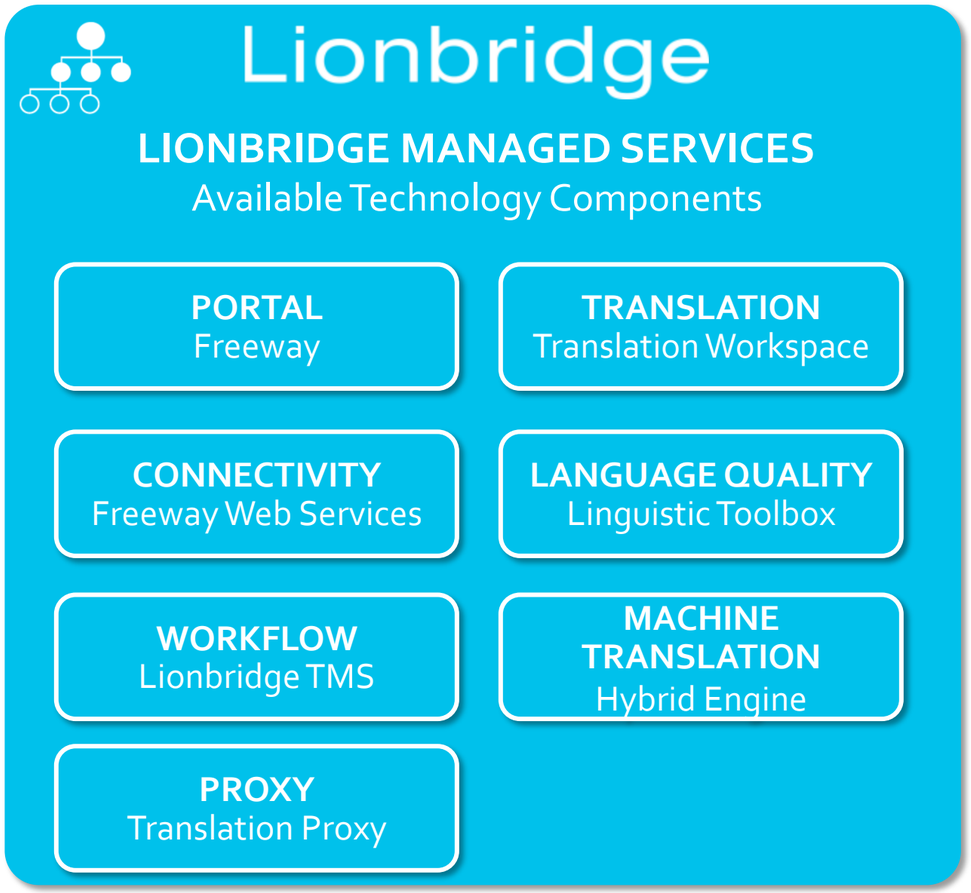
Lionbridge

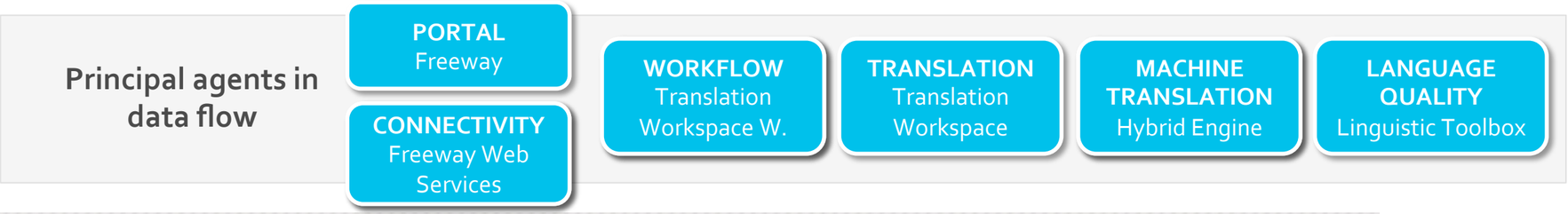
How do we produce today?

Document based approach

Helped by the following translation standards:

- XLIFF 1.2
- TMX
- (ITS 2.0, SRX)





How do we produce today?

Portal/Order System (customer facing)

- Receives any type of content
- Handles quotes and orders
- Passes content on to workflow system

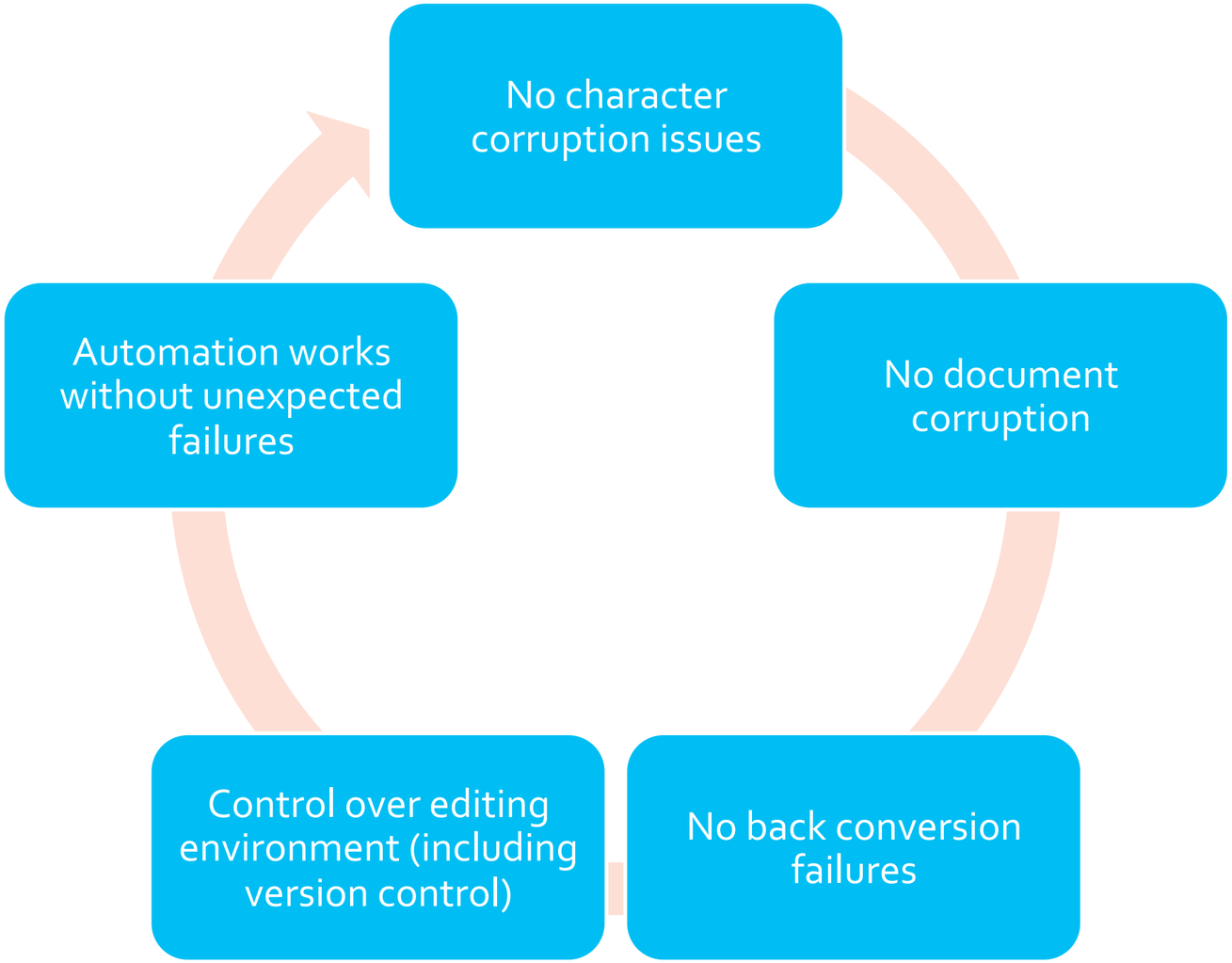
Workflow System (production facing)

- Breaks content down into work items
- Controls conversions to pivotal format (XLIFF 1.2)
- Streams work items through the translation process
- Assigns resources

Translation Memory / Editing environment

- Keeps TUs of every translation ever done
- Real time globally shared TM access

The benefits of using XLIFF 1.2 internally



Case Study: Visual Studio (Orcas)

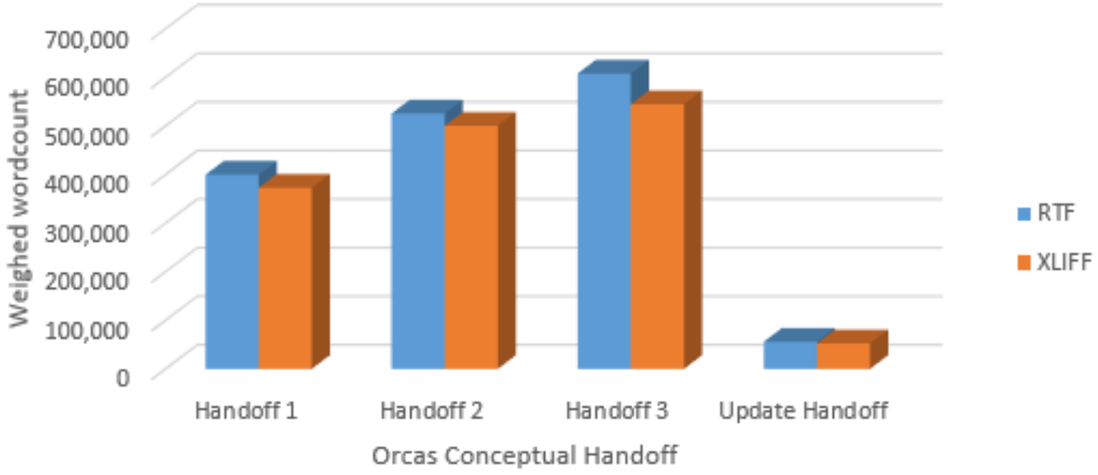
- Orcas was the first big project translated using the XLIFF file format and Xliff Editor
- During the project we tracked the wordcount differences between RTF and XLIFF
- The results confirmed the expectations of reduced cost and reduced cycle times.
- The tracked data helped us to convince the client about the benefit of switching to the XLIFF file format.

Wordcount reduction

- Weighed Wordcounts using XLIFF file format lower by **~10%**

Improved Productivity

- XLIFF file format allowed a reduction in translation time
- Improved productivity by avoiding format issues.



Lionbridge

What about the promise of interchangeability of XLIFF 1.2?

It was just that – a promise.

- In practice, working with 3rd party XLIFF 1.x documents was causing more difficulties than parsing plain XML or binary formats (where we knew at least what to expect)

Reasons for bad interchange:

- Erroneous implementations
- Ambiguous and incomplete specification

The benefits still let XLIFF 1.2 fly!

Adoption in Lionbridge initially happened by word-of-mouth of success stories

Heavily campaigning XLIFF transition since 2010

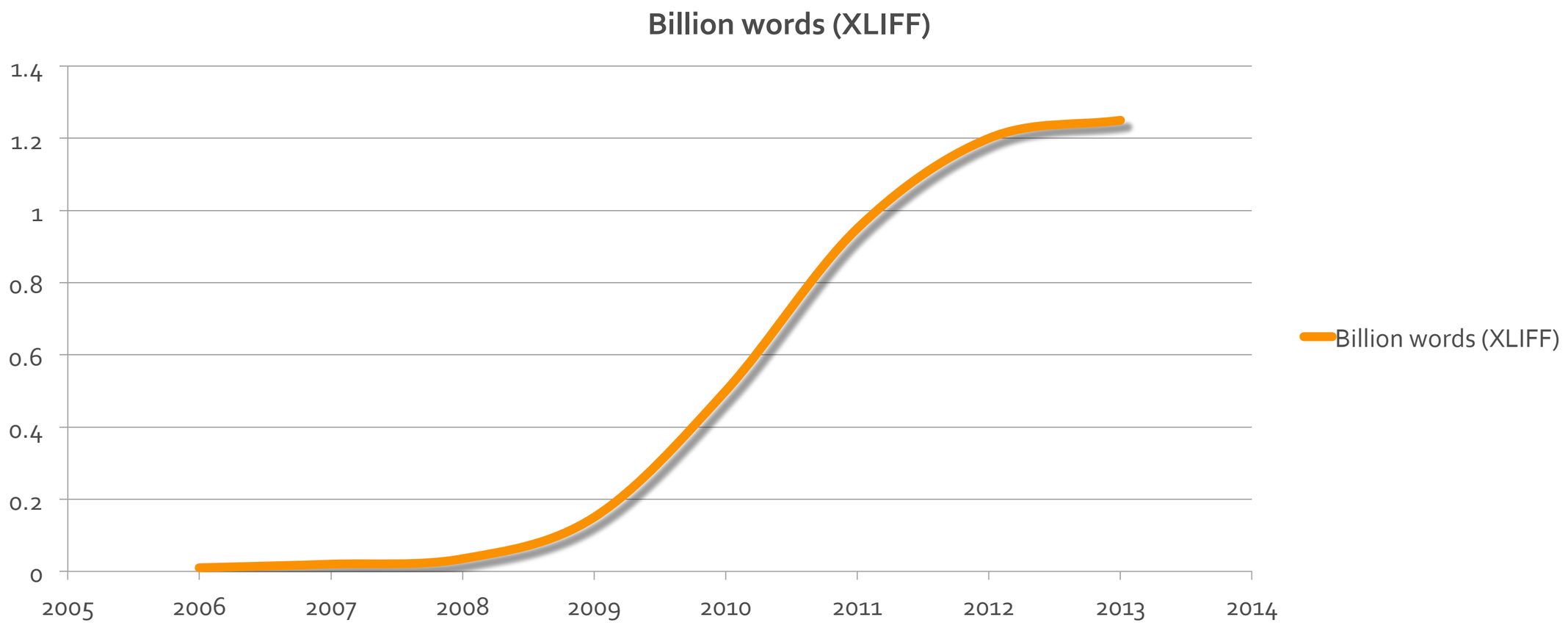
Early 2012 showed 85% transition to XLIFF, which slowly converged to 95% nowadays

No vendors ever rejected working in XLIFF

No language preference on XLIFF adoption

Remaining 5% RTF due to Office centric, complex, legacy workflows

And we do more than 1,2 billion words in XLIFF 1.2 per year!



Lionbridge

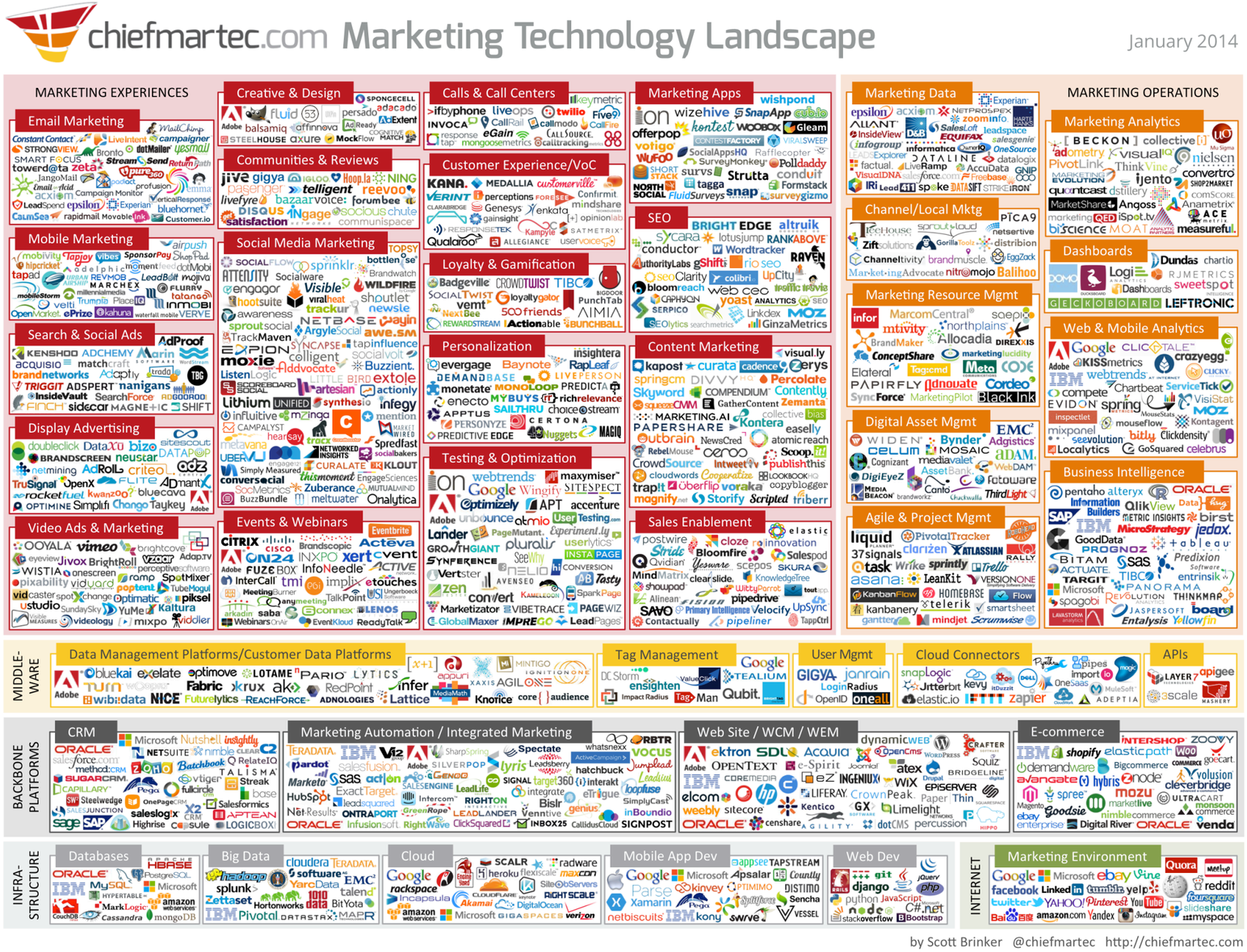
Standardization v1.0 – representing the traditional process

- This was an example for an *internal translation process* which profits from standardization
- Pivotal document format is XLIFF 1.2, and TM data are exchanged with TMX 1.4.2. (TMX 1.4.2, not 1.4b? Yes: http://uri.etsi.org/lis/002/v1.4.2/gs_LIS002v010402p.pdf)
- Internal standardization allows to streamline and automate processes as long as all tools are controlled by the same authority
- Benefits can be huge – our lower limit acceptance level for a new client came down from several thousand US$ to $100 due to reduced transaction cost!
- But benefits from XLIFF 1.2 end when interoperating with 3rd party software
- Similarly, TMX shows its limitations when exchanging with 3rd party software

Standardization v1.0 – representing the traditional process

- So this represents the traditional workflow, where we get files/documents from the customer, and create a TM on our side.
- Our industry's standards suite has been focused on this traditional workflow, too:
  - TMX: document/TM centric
  - SRX: document/TM centric in the way that it assumes there is more than one correct segmentation
  - TBX: document centric in the way that it uses a document to transport the complete set of terms
  - XLIFF: document centric as it *is* the document

Now what about the CMS systems?



- Starting in the mid-90s, CMS systems have by now become the ubiquitous way of managing content electronically
- Over the past few years, CMS system count has ten-fold from the hundreds to the thousands
- Unfortunately, a large part of the CMS systems do not specifically support translation processes (although there are great exceptions)

Copy and Paste?

- Accessing content data for translation in CMS systems is often difficult
- Sometimes one gets the impression that content is held hostage
- We see many situations where data is transported via copy and paste
- Automation requires Connector functionality, which interfaces between CMS and translation service
- Each CMS requires its own Connector implementation
- There are > 1200 CMS on the market in 2014!

And the segmentation?

- Even with a successful source/target roundtrip process, CMS systems lose the association between source and target segments – because their content block size typically is larger than one sentence.
- When target content now gets modified by a *subject matter expert*, it is extremely difficult to represent the modifications back in a sentence segmented translation memory.
- This is a major issue, as it makes the information flow unidirectional, forbids in-context reviews, or creates an insane amount of manual adjustment.
- As an industry, we need a technical solution for this problem.
- As an economy, we need a technical solution for this problem – as it unproportionally increases cost of business for multilingual regions.

Is this the last millennium's Translation Model?

- Is the translation model with a TM external to the content data still the right one?
- Should we not see the CMS as one of the TMs, including versioning for old translations?
- Should the CMS not offer online APIs to push new translatable content and allow queries for old translations?
- Is variable segmentation an idea of the past, and ULI's universal segmentation initiative the future? We interconnect already today up to hundred TMs for one large query base, different segmentation rules only split the content into silos.

We need a new type of standards!

- Opening interoperability: tighter specification!
- Getting away from the document centric model!
- Focusing on services and protocols!
- Including the king of content of today – the CMS
- Need incentives and motivation for CMSs to adapt to multilingualism

- All content and support data of today is dynamic, and should not be forced anymore into a document container

Whish List:

- XLIFF 2.0 adopted by all content creators for a translation interface
- Interoperability part of standards conformance tests
- Standardized connector protocol for CMS systems (COTI as a blueprint for the orders, Linport for the content?)
- CMS maintaining sentence segmentation/association – can it be supported by some formalism?
- ITS 2.0 alongside all exchanged content, persistently stored in CMS
- Standard protocol definitions for TM query, Terminology, Content/Segment exchange
- Hard competition on a unified market! We need the best products and services possible to bridge the language gap!

Lionbridge

Thank you!

Joachim Schurig,
Senior Technical Director Language Technology,
Chair ETSI ISG LIS

Email: joachim.schurig@lionbridge.com
Mobile: +33 6 14 50 27 75

www.lionbridge.com
http://blog.lionbridge.com
http://twitter.com/Lionbridge
http://www.facebook.com/L1onbridge