# Terminology use case for LT-Web metadata

Terminology identification (extraction):

* is **language and domain specific** (it is not advised to mix multiple languages and multiple domains together in one pass through the data).
* works best if performed on **larger text spans** (the identification phase and not the recognition phase).

Therefore, in the Terminology use case we will model term annotation on a single ITS enriched document as a term annotation task in multiple dimensions:

**Dimension 1**: language (the question, however remains how to treat different locales – as different languages or different )

**Dimension 2**: subject field (or domain in ITS terminology)

The Terminology Annotation Web Service will thus split the content in multiple dimensions, analyse (identify and recognise terms) the data separately and combine all separate analyses into one.

Apart from the Language Information and Domain data categories, the fully implemented version of the Terminology use case will feature also:

**Element Within Text** – in order to separate paragraphs where necessary and restrict overlapping annotations.

**Terminology** – in order to identify existing terminology markup and restrict overlapping annotations.

**Locale Filter** – in order to identify fragments that should be ignored by terminology annotation.

## Annotation of terms in HTML5 (with ITS tags)

Terms will be annotated only with local tags/attributes.

The annotation service will create the following types of annotation:

1. Only identified terms:
	1. If previous annotation has been present:

*note that termConfidence is not currently in ITS V2.0*

**<p>**We need a new **<b** its-term="yes" its-term-confidence="0.9"**>**motherboard**</b></p>**

* 1. If no previous annotation has been present:

**<p>**We need a new **<span** its-term="yes" its-term-confidence="0.9"**>**motherboard**</span></p>**

1. Terms recognised using term bases:
2. If previous annotation has been present:

**<p>**We need a new **<b** its-term="yes" its-term-info-ref="http://www.eurotermbank.com/information%20technology%20and%20data%20processing/motherboard"**>**motherboard**</b></p>**

1. If no previous annotation has been present:

**<p>**We need a new **<span** its-term="yes" its-term-info-ref="http://www.eurotermbank.com/information%20technology%20and%20data%20processing/motherboard"**>**motherboard**</span></p>**

Information about the annotation tool will be added with an *AgentInfo* (?) rule (?) or attribute (?).

Should it be something like:

**<p>**We need a new **<span** its-term="yes" its-term-info-ref="http://www.eurotermbank.com/information%20technology%20and%20data%20processing/motherboard" its-**agent-info-ref="http://www.tilde.com/termBaseAnnotator.asmx">**motherboard**</span></p>**

External rules will be supported only if the paths will be accessible. **For a discussion** – should processing break down if external rules are missing or should the processing be continued ignoring the fact of external rule absence?

# Questions:

1. For MT systems the “translate” data category defines what should be analysed and what not. **Should there be a similar category for other linguistic tasks?** (or more precisely, how should one know what to ignore and what not for other kinds of analysis – Named entity recognition and term identification/recognition)
2. For overriding, imagine the following example:

**<html>**

 **<head>**

 **<meta** charset**=**"utf-8"**>**

 **<title>**The Origins of Modern Novel**</title>**

 **<its:rules** version="2.0"**>**

 **<its:translateRule** translate="no" selector="//span"**/>**

 **</its:rules>**

 **</head>**

 **<body>**

 **<div** translate="yes"**>**

 **<p>**It would certainly be quite a **<span>**faux pas**</span>** to start a dissertation on the origin of modern novel without mentioning the **<b>**Epic of Gilgamesh**</b>**...**</p>**

 **</div>**

 **</body>**

**</html>**

Just for clarification: **should the <span> be translated?** That is, are rules always less important than local attributes even if they can address precise target elements? **According to “*5.5 Precedence between Selections*” – yes!**

1. **ITS V2.0 currently specifies Terminology not to be inheritable.** What if the situation is as follows?

**<p>**We need a new **<b** its-term="yes" its-term-confidence="0.9"**><i>**central**</i>** **<u>**processing**</u>** **<i>**unit**</i></b></p>**

1. This is sort of a comment. In cases of overlapping HTML5 mark-up or any other ITS related mark-up, identified possible terms will be ignored and not marked.

See an example below.

**<p>**We need a **<b>**new central**</b>** **<u>**processing unit**</u>**

**This is a content creation issue and won’t be resolved by automated methods.** For what we know, the different elements could be positioned in two different places on a screen using CSS – combining them might corrupt the design.

1. Regarding ***AgentInfo*** (or *ProcessInfo* (?), or *ToolInfo* (?)) …

The real need for the annotations of terminology are to identify who identified which terms in a document (Was it a human? Was it an expert? Was it a term-base based annotation tool? Was it an unguided terminology annotation tool? Who or what exactly?). If there is a sophisticated translation provenance, it should also be able to identify these annotators.

*Wouldn’t it be reasonable to make the* ***translation agent provenance more general*** *so that it can be also used to identify subparts, including “Named entity recognition agent provenance” or “Term identification/recognition agent provenance”?*

1. **Language Information** according to ITS can be en-US, en-GB, etc., which makes sense, however, existing terminology resources rarely distinguish the geographical differences (geographical usage in TBX). Thus, for term identification and recognition fall-back to simple English will most probably need to be used.