**8.10 Text Analysis**

**[](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html%22%20%5Cl%20%22contents)8.10.1 Definition**

The [Text Analysis](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html%22%20%5Cl%20%22Disambiguation) data category is used by text analysis agents such as named entity recognizers,lexical concept disambiguators, etc. These agents enhance content by suggesting or identifying concepts, identities. Enhancement works via references to resource descriptions such as Internationalized Resource Identifiers (IRIs). Example: A named entity recognizer may provide the information that the string “Dublin” in a certain context denotes a town in Ireland.

While text analysis can be done by humans, this data category is targeted more at software agents.

The information provided by the Text Analysis data category can be used for several purposes, including, but not limited to:

* Informing a human agent such as a translator that a certain fragment of textual content (so-called text analysis target) is subject to follow specific translation rules; examples: proper names, or officially regulated translations.
* Informing software agent such as a content management system about the conceptual type of a textual entity in order to enable special processing; examples: places, personal names, product names, or geographic names, chemical compounds, protein names, are put in a specific index.

The data category provides three pieces of information: confidence, entity type/concept class, entity/concept identifier (see more information in the table below).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Information | Description | Value | Example | Comments |
| **Text analysis confidence** | The confidence of the agent (that produced the annotation)in its own computation | [XML Schema double data type](http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/#double) with the constraining facets [minInclusive](http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/#rf-minInclusive) set to 0 and [maxInclusive](http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/#rf-maxInclusive) set to 1 | 0.5647346 | The confidence value applies to two pieces of information (see following rows in this table). This is opposed to term confidence which is part of the Terminology data category. Term confidence represents the confidence in just a single piece of information: the decision whether something is a term or not (term). term confidence does not relate to the confidence about additional information about the term that can be encoded with termInfoRef. |
| **Entity type / concept class** | The type of entity, or concept class of the text analysis target | IRI | <http://nerd.eurecom.fr/ontology#Place>  |  |
| **Entity / concept identifier** | A unique identifier for the text analysis target | Mode 1: Identifier (string value) of the collection source + identifier of the concept in that collection | “Wordnet3.0" to identify the collection resource; “301467919” to identify a synset in Wordnet3.0 |  |
| Mode 2: Identifier ( IRI) of the text analysis target | http://dbpedia.org/resource/Dubli |   |

Mode 1 and Mode 2 are mutually exclusive. They MUST NOT be used at the same time for the same text analysis target/node.

**Note:**

The use case for Text Analysis is distinct from that for the [Terminology](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#terminology) data category. Text Analysis informs human agents or software agents activities in settings where either explicit terminology information is not (yet) available or would be not appropriate. Example for the latter: for general vocabulary, term-related annotations often are not sensible.

Text Analysis support is achieved by associating a fragment of text with an external web resource that can be interpreted by a language review agent. The agent may for example use the web resource to disambiguate the meaning or lexical choice of the fragment, and thereby contributing to its correct translation. The web resource may as well provide information on appropriate synonyms and example usage. This is for example the case if the web resource is WordNet. In the case of a concept class, the external resource may provide a formalized conceptual definition arranged in a hierarchical framework of related concepts. In the case of a named entity, the external resource may provide a full-fledged description of the associated real world entity.

Extended example: The word 'City' in the fragment 'I am going to the City' may be enhanced by one of the following:

- one of WordNet's synsets that can be represented by 'city'

-an ontological concept of 'City' that could represent a subclass of 'Populated Place' as a concept

- the central area of a particular city - as interpreted as an entity instance (e.g. 'City of London')

**Note:**

A given document fragment can only be annotated once. When support for multiple annotations is necessary – e.g. when all three of the annotations in the extended example above need to be accommodate - NIF 2.0, TEI (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html#NHSO), or other so-called stand-off annotation mechanisms should be considered.

Resources such as DBpedia enable the linking ontological concepts and named entity definitions for same things and in different languages. , This facilitates translation even more (since link traversal does for example allow access to foreign language labels for named entities).

[****](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#contents)**8.10.2 Implementation**

The [Text Analysis](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#Disambiguation) data category can be expressed with global rules, or locally on an individual element. There is no inheritance.

GLOBAL: The TextAnalysisRule element contains the following:

* A required selector attribute that contains an [absolute selector](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#selectors) that selects the nodes to which this rule applies.
* At least one of the following:
	+ A taClassRefPointer attribute that contains a [relative selector](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#selectors) pointing to a node that holds an IRI which implements the Entity type / concept class information.
	+ Exactly one of the following:
		- A taSourcePointer attribute that contains a [relative selector](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#selectors) to a node that holds the identifier of the collection source; and a taIdentPointer attribute that contains a relative selector to a node that holds the identifier of the concept in the collection. See mode 1.
		- A taIdentRefPointer attribute that contains a [relative selector](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#selectors) pointing to a node that holds an IRI that holds the identifier of the text analysis target. See mode 2.

For an example, see [Example 59](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#EX-disambiguation-html5-rdfa-companion-document).

LOCAL: The following local markup is available for the [Text Analysis](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#Disambiguation) data category:

* An optional taConfidence attribute that implements the text analysis confidence.
* At least one of the following:
	+ A taClassRef attribute that holds an IRI which implements the Entity type / concept class information.
	+ Exactly one of the following:
		- A taSource attribute that holds the identifier of the collection source; and a taIdent attribute that holds the identifier of the concept in the collection. See mode 1.
		- A taIdentRef attribute that holds the identifier of the text analysis target. See mode 2.

Any node selected by the [Text Analysis](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#Disambiguation) data category with the taConfidence attribute specified [MUST](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#rfc2119) be contained in an element with the annotatorsRef (or in HTML its-annotators-ref) attribute specified for the [Text Analysis](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#Disambiguation) data category. For more information, see [Section 5.8: ITS Tools Annotation](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#its-tool-annotation).

Example 57: Local mixed usage of Usage of taClassRef and taIdentRef in HTML.

**<!DOCTYPE html>**

**<html** lang="en" its-annotators-ref="text-analysis|http://enrycher.ijs.si"**>**

 **<head>**

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

 **<title>**Text Analysis: Local Test**</title>**

 **</head>**

 **<body>**

 **<p><span**

 its-ta-confidence="0.7"

 its-ta-class-ref="http://nerd.eurecom.fr/ontology#Place"

 its-ta-ident-ref="http://dbpedia.org/resource/Dublin"

 **>**Dublin**</span>**

 is the **<span**

 its-ta-source="Wordnet3.0"

 its-ta-ident="301467919"

 its-ta-confidence="0.5"

 **>**capital**</span>** of Ireland.**</p>**

 **</body>**

**</html>**

[Source file: [examples/html5/EX-Text Analysis-html5-local-1.html](http://www.w3.org/International/multilingualweb/lt/drafts/its20/examples/html5/EX-disambiguation-html5-local-1.html)]

**Note:**

For expressing Entity type / concept class information, implementers are encouraged to use an existing repository of entity types such as the Named Entity Recognition and Disambiguation [[NERD]](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html%22%20%5Cl%20%22nerd) ontology. Of course this requires that the repository satisfies the constraints imposed by the text analysis data category (e.g. use of IRIs).

Various target types can be expressed via Entity type / concept class: types of entities, types of lexical concepts, or ontology concepts. While a relationship between these types may exist, this specification does not prescribe a way of automatically inferring a one target type from another.

**Note:**

Text Analysis is primarily intended for textual content. Nevertheless, the data category can also be used in multi-media contexts. Example: objects on an image could be annotated with DBpedia IRIs.

When serializing the ITS [Text Analysis](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#Disambiguation) data category markup in HTML, the preferred way is to serialize in RDFa Lite or Microdata due to the existing search and crawling infrastructure that is able to consume these formats.

Example 58: Local mixed usage of taClassRefPointer, taIdentRefPointer in HTML+RDFa Lite.

See [Example 59](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#EX-disambiguation-html5-rdfa-companion-document) for the companion document with the mapping data.

**<!DOCTYPE html>**

**<html** lang=en**>**

 **<head>**

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

 **<link** href=EX-text analysis-html5-rdfa.xml rel=its-rules**>**

 **<title>**Entity: Local Test**</title>**

 **</head>**

 **<body>**

 **<p><span** property="http://xmlns.com/foaf/0.1/name"

 about="http://dbpedia.org/resource/Dublin"

 typeof="http:/nerd.eurecom.fr/ontology#Place"**>**Dublin**</span>** is

 the capital of Ireland.**</p>**

 **</body>**

**</html>**

[Source file: [examples/html5/EX-Text Analysis-html5-rdfa.html](http://www.w3.org/International/multilingualweb/lt/drafts/its20/examples/html5/EX-disambiguation-html5-rdfa.html)]

Example 59: Companion document, having the mapping data for [Example 58](http://www.w3.org/International/multilingualweb/lt/drafts/its20/its20.html#EX-disambiguation-html5-rdfa).

**<its:rules** xmlns:its="http://www.w3.org/2005/11/its" version="2.0"**>**

 **<its:textAnalysisRule** selector="//\*[@typeof and @about]"

 taClassRefPointer="@typeof" taIdentRefPointer="@about" **/>**

**</its:rules>**

[Source file: [examples/html5/EX-Text Analysis-html5-rdfa.xml](http://www.w3.org/International/multilingualweb/lt/drafts/its20/examples/html5/EX-disambiguation-html5-rdfa.xml)]

### Best practices/Hints [separate document]

In order to accommodate work with existing annotation APIs, we provide some example mappings.

For instance, looking at NERD and ITS2.0, we have the following correspondence:

* label corresponds to the text node in the HTML/XML document
* nerdType corresponds to taClassRef
* confidence corresponds to taConfidence

extractor corresponds to annotatorsRef