



WHITE PAPER FOR OGC INTERNATIONALIZED TEXT ANALYSIS

TECHNICAL PAPER

CANDIDATE SWG DRAFT

Version: 1.0

Submission Date: 2025-01-25

Approval Date: 2025-01-25

Publication Date: 2025-01-25

Editor: Charles Heazel, Howard Trickey

Notice: This document is not an OGC Standard. This document is an OGC White Paper and is therefore not an official position of the OGC membership. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an OGC Standard. Further, an OGC White Paper should not be referenced as required or mandatory technology in procurements.

License Agreement

Use of this document is subject to the license agreement at <https://www.ogc.org/license>

Copyright notice

Copyright © 2025 Open Geospatial Consortium

To obtain additional rights of use, visit <https://www.ogc.org/legal>

Note

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The Open Geospatial Consortium shall not be held responsible for identifying any or all such patent rights.

Recipients of this document are requested to submit, with their comments, notification of any relevant patent claims or other intellectual property rights of which they may be aware that might be infringed by any implementation of the standard set forth in this document, and to provide supporting documentation.

CONTENTS

- I. ABSTRACTiv
- II. KEYWORDSiv
- III. PREFACEv
- IV. SECURITY CONSIDERATIONSvi
- V. SUBMITTING ORGANIZATIONSvii
- 1. INTERNATIONALIZED TEXT9
 - 1.1. Common Practice9
 - 1.2. International Standards12

LIST OF TABLES

- Table 19
- Table 218

LIST OF FIGURES

- Figure 110
- Figure 210
- Figure 310
- Figure 410
- Figure 511
- Figure 611
- Figure 7 – CharacterString Context.13
- Figure 8 – LanguageString Context.14
- Figure 9 – LanguageSpecificCharacterString Context.15
- Figure 10 – PT_Locale Context.16



ABSTRACT

The Points of Interest (POI) Conceptual Model Standard builds upon the Geospatial Feature and Geometry Standards developed by the Open Geospatial Consortium and ISO Technical Committee 211 (TC211). Its goal is to provide a model using a Unified Modeling Language (UML) object model which any system may adopt for the definition of basic entities, attributes and relations of POI, regardless of encoding language.

In developing the standard, the POI SWG has examined approaches for including in the POI data information in multiple languages without introducing confusion and with the lowest overhead (e.g., computational complexity). Through this report, the SWG members are communicating to the OAB and the geospatial community at large where and how internationalized text is implied in international standards and to make transparent to all concerned the approach the SWG recommends for the POI standard, as well as all other standards that seek to include internationalized text.



KEYWORDS

The following are keywords to be used by search engines and document catalogues.

ogcdoc, OGC document, Internationalization, POI, Points of Interest



PREFACE

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. The Open Geospatial Consortium shall not be held responsible for identifying any or all such patent rights.

Recipients of this document are requested to submit, with their comments, notification of any relevant patent claims or other intellectual property rights of which they may be aware that might be infringed by any implementation of the standard set forth in this document, and to provide supporting documentation.



SECURITY CONSIDERATIONS

No security considerations have been made for this document.



SUBMITTING ORGANIZATIONS

The following organizations submitted this Document to the Open Geospatial Consortium (OGC):

- Digital Flanders
- Google
- Charles Heazel
- Pangaea Innovations
- PEREY Research and Consulting
- US Army Geospatial Center



1

INTERNATIONALIZED TEXT

A POI is a location about which information of general interest is available. A POI can be as simple as a set of coordinates and an identifier, or more complex such as a three-dimensional model of a building with names in various languages, information about open and closed hours, and a civic address. It follows that support for language-specific names, or internationalized text in general, is a requirement for a successful POI Standard.

Internationalization of text is not new. There are a number of techniques in common practice. An understanding of those practices is the first step to developing guidance for internationalization of POIs and potentially OGC Standards in general.

The POI Standard builds on the Geospatial Feature and Geometry Standards developed by the Open Geospatial Consortium and ISO Technical Committee 211 (TC211). It should not be necessary to develop an approach for internationalized text which is specific to POIs. That support should already exist within the foundational standards. This document seeks to provide a window into that support.

1.1. Common Practice

Three general techniques for internationalized text are found in common practice:

Table 1

Implied Language	No language specified for individual text string. There may be a default language for the entire document, or language tags which indicate the language for a scoped section of the document.
Separate Per-Language Attributes	Multiple copies of the same text element that have the language as a suffix (e.g., "name", "name:fr", "name:en", ...) with the unsuffixed version representing use of a default language.
Complex Attributes	In this case text strings are objects, not character strings. Each object has a character string attribute (the text itself) along with a language attribute. Additional attributes, such as character set or country code, can also be provided. If the text is multi-lingual, then the text can be implemented as a sequence of objects, each with content specific to a different language.

1.1.1. Implied Language

Implied Language is the most common approach to internationalized text. When using this technique, all of the text in a document, or section of a document, is in the same language. Support for multiple languages requires duplication of the entire section for each language.

1.1.2. Per-Language Attribute

The per-language attribute approach tags the text string with a language identifier. This identifier can either be a part of the string itself, or an attribute associated with the text string element.

For example, if we wanted to provide the name of the Statue of Liberty in XML using both French and English, then we could use:

A language identifier embedded in the element value

```
<name>"Statue of Liberty:en"</name>  
<name>"Statue de la Liberté:fr"</name>
```

Figure 1

Or provide the language identifier as an attribute of the element

```
<name language="en">"Statue of Liberty"</name>  
<name language="fr">"Statue de la Liberté"</name>
```

Figure 2

1.1.3. Complex Attributes

The Complex Attribute method uses several alternative approaches:

1.1.3.1. Complex Attributes: Sequence-only

Every text string must be a sequence of objects. In the following example, the sequence-only approach is used to provide the name attribute in both French and English.

```
"name": [  
  {"name" : "Statue of Liberty", "language" : "en"},  
  {"name" : "Statue de la Liberté", "language" : "fr"}  
]
```

Figure 3

1.1.3.2. Complex Attributes: Default and Sequence

In this approach the text string is an object that contains a default value and a sequence of language coded values. The language of the default value can be specified by a "language" property in the outermost object of the file.

The following example provides one default, one English, and one French language version of the name attribute.

```
"name" : {  
  "default" : "Statue of Liberty",  
  "names": [  
    "Statue of Liberty",  
    "Statue de la Liberté"  ]  
}
```

```

    {"name" : "Statue of Liberty", "language" : "en"},
    {"name" : "Statue de la Liberté", "language" : "fr"}
  ]
}

```

Figure 4

1.1.3.3. Complex Attributes: Inspire

The Inspire method is based on the PT_Locale concept defined in ISO 19115 and ISO 19139.

This approach is very similar to the Sequence-only method but the elements of the sequence, instead of referring to *languages*, refer to *locales*. This requires definitions of a number of **locales** somewhere at the file scope.

The following is an example of the PT_Local concept adapted to JSON.

```

"locales" : [
  {
    "id" : "locale_en",
    "language_code" : {
      "code_list" : "link:++http://www.loc.gov/standards/iso639-2/"++[],
      "code_list_value" : "en",
      "name" : "English"
    },
    "character_encoding" : {
      "code_list" : "resources/codelist/gmxcodelists.xml#MD_
CharacterSetCode",
      "code_list_value" : "utf8">UTF 8</MD_CharacterSetCode>,
      "name" : "UTF8"
    }
  },
  {
    "id" : "locale_fr",
    "language_code" : {
      "code_list" : "link:++http://www.loc.gov/standards/iso639-2/"++[],
      "code_list_value" : "fr",
      "name" : "French"
    },
    "character_encoding" : {
      "code_list" : "resources/codelist/gmxcodelists.xml#MD_
CharacterSetCode",
      "code_list_value" : "utf8">UTF 8</MD_CharacterSetCode>,
      "name" : "UTF8"
    }
  }
]

```

Figure 5

Then an actual name attribute would look like this:

```

"name": [
  {"name" : "Statue of Liberty", "locale" : "locale_en"},
  {"name" : "Statue de la Liberté", "language" : "locale_fr"}
]

```

Figure 6

One advantage of this approach is that the *Lcales* can be defined once then referenced throughout the document.

1.1.3.4. Complex Attributes: Choice of Implied or Complex

The “Choice” approach allows the text string to be one of two types: a simple string with an Implied Language or one of the other Complex Attribute approaches. If the value is only a simple string, then it is assumed to be using the Implied Language approach. Otherwise, the text string will be structured using one of the Complex Attribute approaches.

The advantage of this approach is that the simple Implied Language approach can be used when appropriate. The Complex Attribute approaches can be limited to those cases where the increased complexity is needed. The result is a file that has the level of complexity that it needs, and no more.

1.2. International Standards

This section discusses how the Standards developed by ISO TC211 and the OGC support internationalized text and the common practices described above.

1.2.1. Conceptual Schema

ISO 19103 defines the conceptual schema language (CSL) for developing computer interpretable models, or schemas, of geographic information. This includes definition of the primitive data types used in OGC and ISO TC211 developed standards. `CharacterString` is one of those primitive data types.

ISO/TS 19103:2005 — Geographic information — Conceptual schema language was the first version of ISO 19103. This version defined the `CharacterString` class as a primitive type with attributes for `CharacterSetCode`, `maxLength`, `size`, and `elements` (characters).

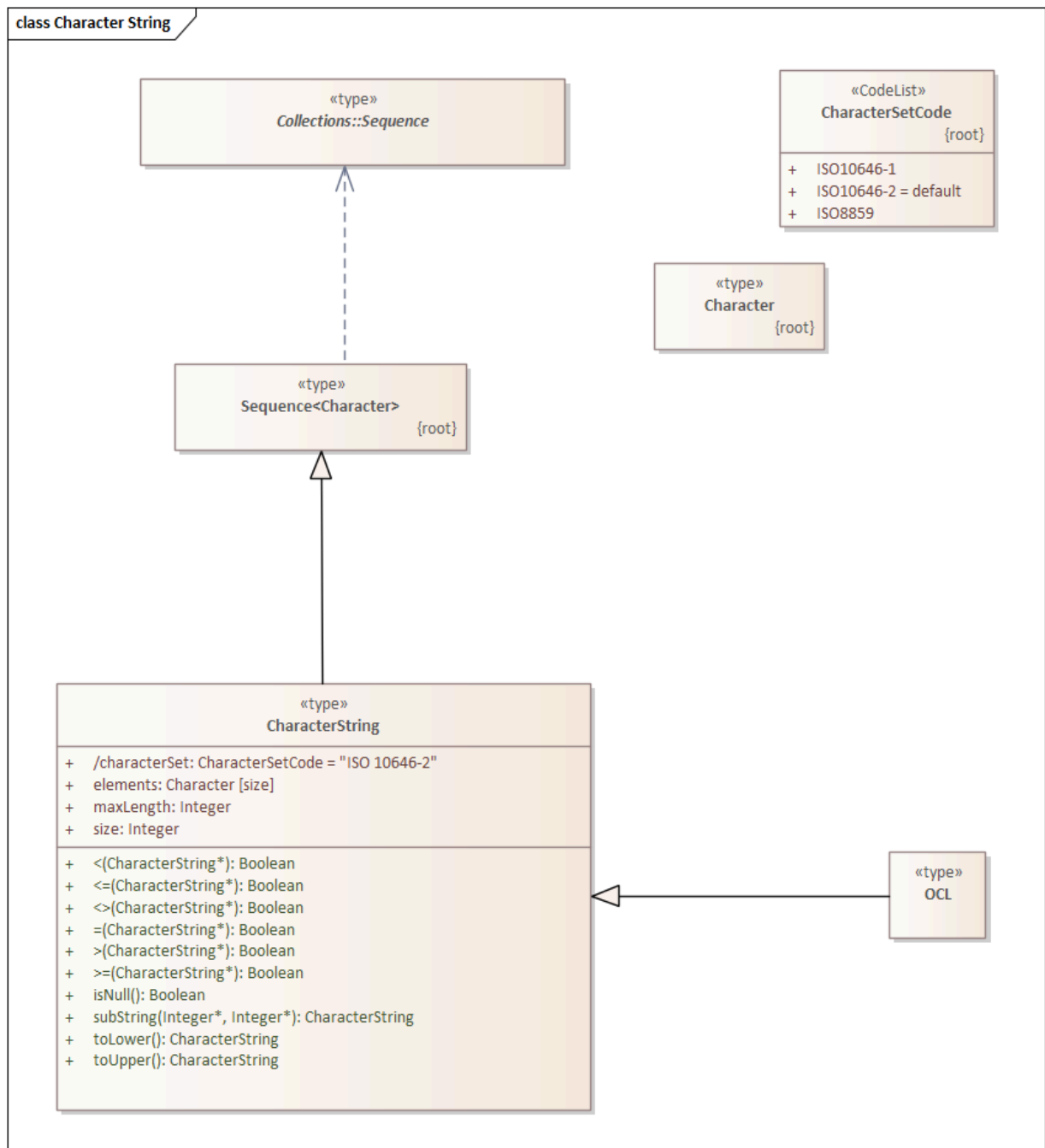


Figure 7 – CharacterString Context.

ISO 19103 was updated by ISO 19103:2015 – Geographic information – Conceptual schema language.

In this version, the CharacterString class has the same information content as in 19103:2005. However, this class is modeled as an interface instead of a type. It also adds a normative Annex C “Data types – extension types”.

Section C.2 of this annex addresses “Cultural and Linguistic Adaptability” which includes a new “LanguageString” class. LanguageString is a subclass of CharacterString. This subclass

adds a languageCode whose values come from ISO 639. As a result, a LanguageString is a CharacterString with a associated language code.

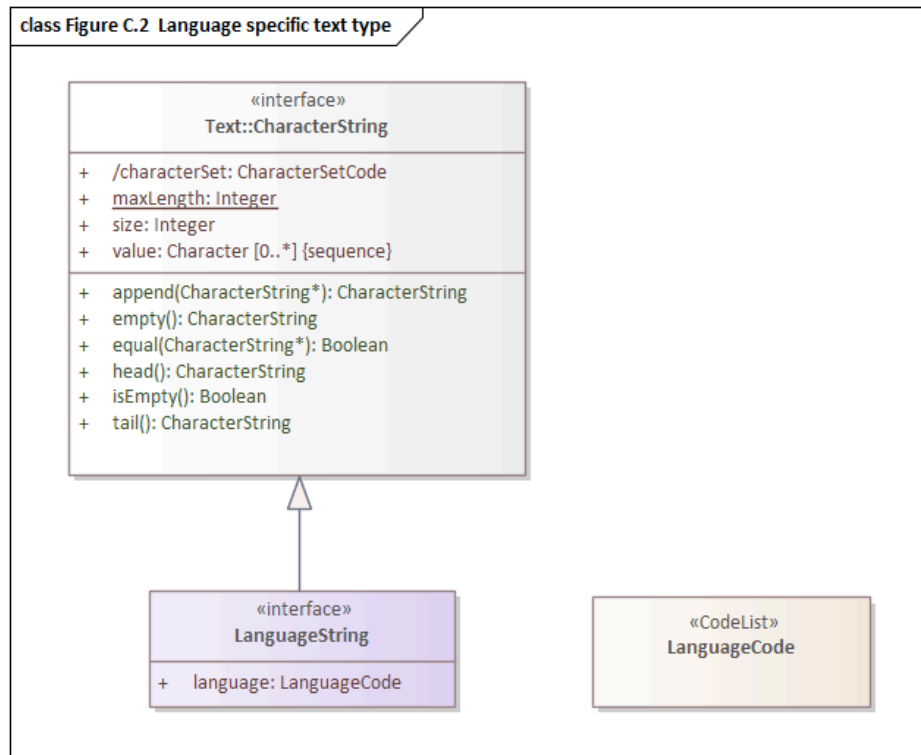


Figure 8 — LanguageString Context.

Since each LanguageString has multiple mandatory attributes, this is an example of the “Complex Attribute” approach.

1.2.2. Location-Based Services

The Location-Based Services family of standards was developed to extend Web-based location services to the bandwidth limited cell phones of the time. This is arguably the first appearance of a language-specific text field in an ISO TC211 Standard.

ISO 19132:2007 “Location-based Services Reference Model” introduces the LanguageSpecificCharacterString class. This is a subclass of the CharacterString class from 19103:2005. Its sole function is to add a LanguageCode attribute to CharacterString.

Since LanguageSpecificCharacterString is a subclass of CharacterString, any CharacterString can (in principle) be a LanguageSpecificCharacterString.

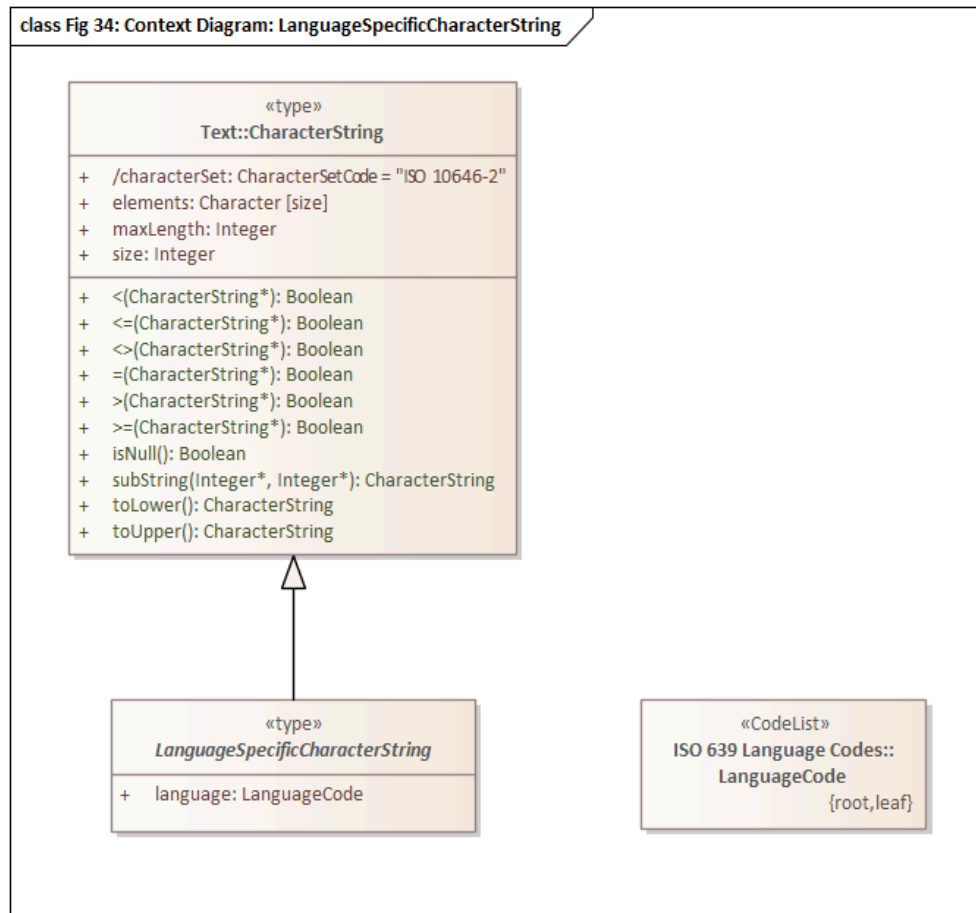


Figure 9 — LanguageSpecificCharacterString Context.

The resulting class is almost identical to the LanguageString class defined in ISO 19103:2015.

1.2.3. Metadata

There are two series of TC211 Standards for Metadata. The ISO 19115 series provides the conceptual models. The ISO 19139 series defines XML schema for the ISO 19115 Standards.

1.2.3.1. ISO 19139

The ISO 19139:2007 “Metadata XML Implementation” Standard defines the XML schema for ISO 19115:2003. ISO 19139 also addresses internationalization, but it uses a different approach from that used in ISO 19132.

ISO 19139 introduces the LocalizedCharacterString class. This class is a realization of the CharacterString class from ISO 19103:2005. LocalizedCharacterString adds to CharacterString an association with the new PT_Locale class. PT_Locale includes attributes for CharacterSetCode, Country, and LanguageCode. Only the LanguageCode is required.

Note that CharacterSetCode is already defined by CharacterString so it is redundant in PT_Locale.

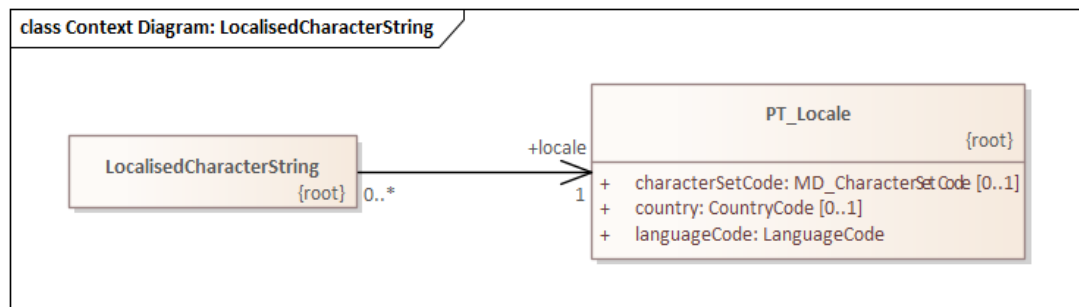


Figure 10 – PT_Locale Context.

1.2.3.2. ISO 19115

ISO 19139:2007 defines the XML schema for ISO 19115:2003. However, the PT_Locale class does not appear in ISO 19115:2003. So initially PT_Locale was only defined for XML metadata encodings.

This limitation was partially lifted by ISO 19115-1:2014. This update to ISO 19115 moves PT_Locale up to the conceptual level. It is no longer limited to XML encodings. The XML encoding standard for 19115-1 is ISO 19115-3:2016. This Standard continues the support the PT_Locale from ISO 19139:2007.

At this time XML is the only standardized encoding for ISO 19115. Until other encodings are defined, PT_Locale will continue to be a concept for use in XML metadata encodings.

PT_Locale is an example of the “Complex Attribute” approach. Its adoption by Inspire is discussed above.

1.2.4. OGC Web APIs

“OGC API – Common – Part 1: Core” provides a brief discussion of text internationalization in section 8.6. It includes a recommendation that “For encodings that support string internationalization, the server SHOULD include information about the language for each string value that includes linguistic text.”

The text internationalization techniques it describes are:

1. Use of the HTTP Accept-Language header to convey the language desired by the requestor. This approach assumes that the server has multiple copies of the resource, each in a different language, or that the server is capable of generating a language-specific version based on the Accept-Language header value.
2. Use of the hreflang attribute of the Link class. The Link class is based on RFC 8288 (Web Linking). It provides the information needed to properly invoke an HTTP request. This includes the URL, format, and language of the target resource.

3. The built-in capabilities of JSON-LD.

The same content appears in “OGC API — Features — Part 1: Core” Section 7.10.

Since most OGC API standards build off of API-Common or API-Features, they all inherit the same recommended approaches.

The Accept-Language header is an example of the “Implied Language” approach. The Link class is an example of the “Complex Attribute” approach. These two techniques can be used together.

1.2.5. JSON-LD

One of the capabilities that JSON-LD adds to JSON is the ability to annotate strings with their language. The capabilities provided include:

- language-tagged string type: A language-tagged string type consists of a string and a non-empty language tag. This is implemented using the `rdf:langString` type.
- `@language` keyword: The `@language` keyword is used to specify the language for a particular string value or the default language for use within a JSON-LD document. When used within a context, it specifies the default language to be used within the scope of that context. If used within the top-level context, it defined the default language for the whole document unless overridden by a lower level context.

Text internationalization is addressed in section 4.2.4 of the JSON-LD 1.1 Standard (<https://www.w3.org/TR/json-ld/#string-internationalization>)

Language Maps: Allows multiple values differing in their associated language to be indexed by language tag.

Language Maps are addressed in section 9.8 of the JSON-LD 1.1 Standard (<https://www.w3.org/TR/json-ld/#language-maps>)

The net result is that all three common approaches can be used in a JSON document.

1.2.6. Conclusion

All standards which build on the ISO 19103:2015 Standard have the option to use the “`LanguageString`” class for text data. This class allows the association of a language identifier with the text string. In some encoding languages, addition of this attribute to a `CharacterString` is easily accomplished. Others may require definition of a `LanguageString` class to support the additional attribute.

The `PT_Locale` concept is useful, particularly if the locality information includes more than just the language code. Use of `PT_Locale` establishes a dependency on the ISO 19115 metadata standards. However, there are many additional data types defined in these Standards which deserve reuse. Such a dependency may be a good practice.

JSON-LD provides powerful text internationalization capabilities, which unfortunately are only applicable to that encoding language.

Here is a table of some pros and cons of the various methods discussed.

Table 2

METHOD	PROS	CONS
Implied Language	Simple. Easy to write and use.	The only way to handle multiple languages is to provide a choice of multiple files, or duplicate sections within one file. Keeping the content in sync across sections is subject to error.
Separate Per-Language Attributes	Familiar to OpenStreetMap users. Handles the one-language case well. Compact.	Conceptually unclear to have multiple versions (languages) of an attribute at the same level as other attributes. Needs more processing to gather together all the versions in the internal data format.
Complex Attributes: Sequence Only	Moderately simple to read and process.	Bulkier for one-language case. Not clear what the default value is.
Complex Attributes: Default and Sequence	Handles one-language case moderately well. Can tell what value to use by default.	Bulkier than most other options, and still not ideal for one-language case.
Complex Attributes: Inspire	Closest to “standards compliant”. Allows specification of character encoding too, and in a less-verbose way than if done per attribute.	Complex to read and write (the locales part). Need to process metadata in another part of the file and connect to each POI: a POI feature would not be standalone. Extra complexity of indirection is only useful if multiple character encodings are needed.
Complex Attributes: Choice of Implied or Complex	Handles one-language case well. Limits complexity to those attributes where it is needed.	Need to support all internationalization approaches and select between them when processing.

It is recommended that implementors choose the simplest approach which meets the needs of their content. A document that is in a single language can rely on the HTTP Accept-Language header to address all internationalization requirements. Mixed languages in a single document will require a more complex approach.

The selection of approach will also depend on the capabilities of the encoding technology. For example, the JSON-LD techniques described above are not supported for JSON encodings. In another example, the per-language attribute approach is common practice in XML, but difficult with JSON.

Ultimately the selection of approach to internationalization will depend on these two factors.