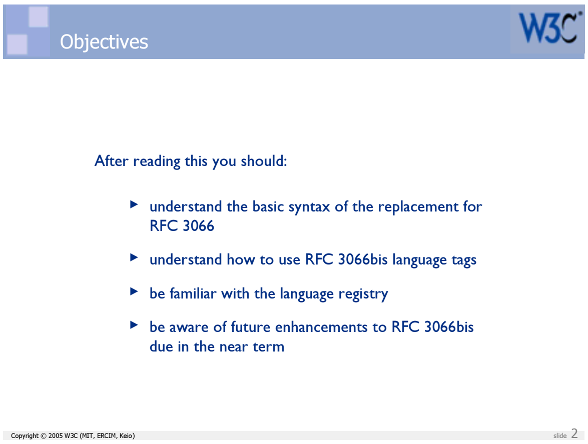
Understanding the new Language Tags

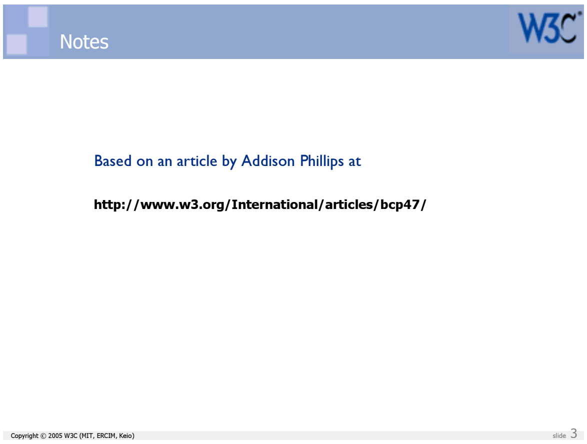
Richard Ishida
W3C Internationalization Activity Lead



Objectives

After reading this you should:

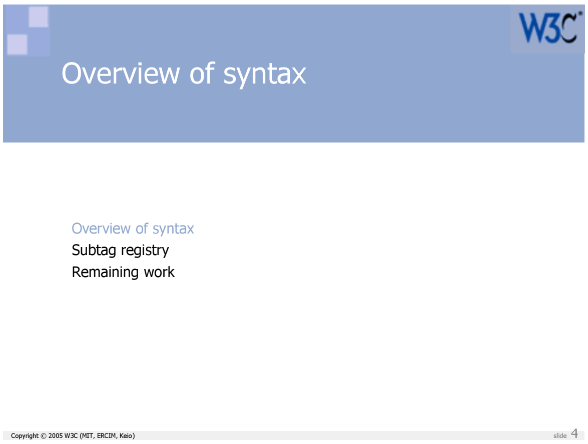
- understand the basic syntax of the replacement for RFC 3066
- understand how to use RFC 3066bis language tags
- be familiar with the language registry
- be aware of future enhancements to RFC 3066bis due in the near term



Notes

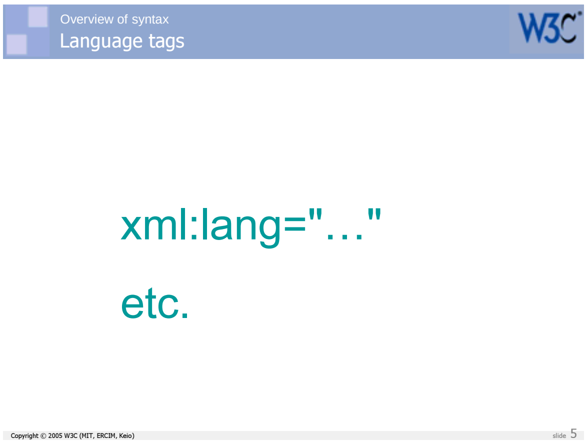
Based on an article by Addison Phillips at

http://www.w3.org/International/articles/bcp47/



Overview of syntax

Overview of syntax
Subtag registry
Remaining work

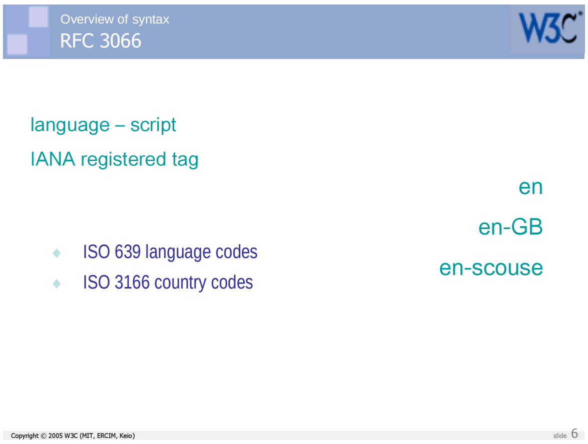


Overview of syntax

Language tags

xml:lang="…"

etc.



Overview of syntax

RFC 3066

language – script

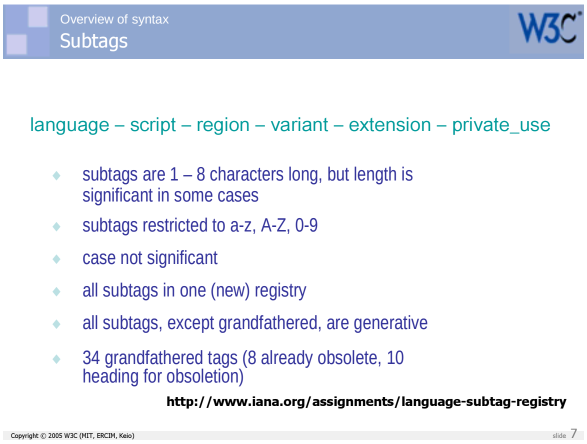
IANA registered tag

- ISO 639 language codes
- ISO 3166 country codes

en

en-GB

en-scouse



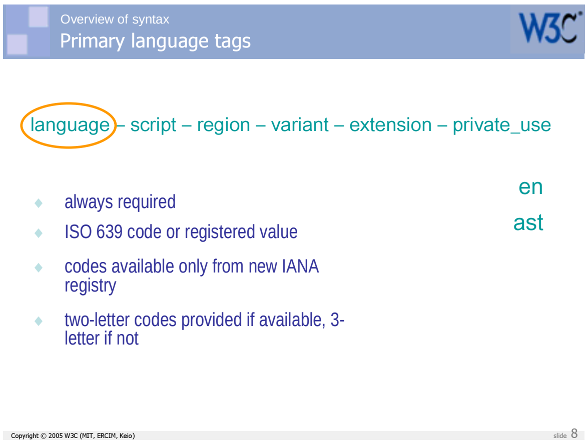
Overview of syntax

Subtags

language – script – region – variant – extension – private_use

- subtags are 1 – 8 characters long, but length is significant in some cases
- subtags restricted to a-z, A-Z, 0-9
- case not significant
- all subtags in one (new) registry
- all subtags, except grandfathered, are generative
- 34 grandfathered tags (8 already obsolete, 10 heading for obsoletion)

http://www.iana.org/assignments/language-subtag-registry



Overview of syntax

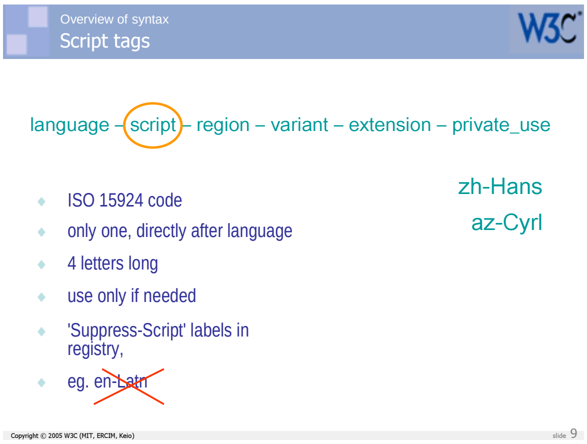
Primary language tags

language – script – region – variant – extension – private_use

- always required
- ISO 639 code or registered value
- codes available only from new IANA registry
- two-letter codes provided if available, 3-letter if not

en

ast



Overview of syntax

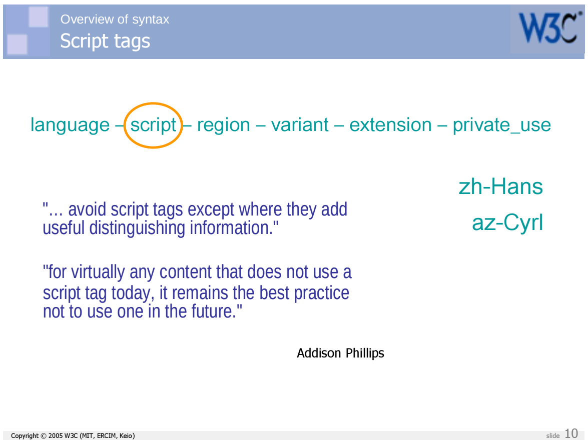
Script tags

language – script – region – variant – extension – private_use

- ISO 15924 code
- only one, directly after language
- 4 letters long
- use only if needed
- 'Suppress-Script' labels in registry,
- eg. ~~en-Latn~~

zh-Hans

az-Cyrl



Overview of syntax

Script tags

language – script – region – variant – extension – private_use

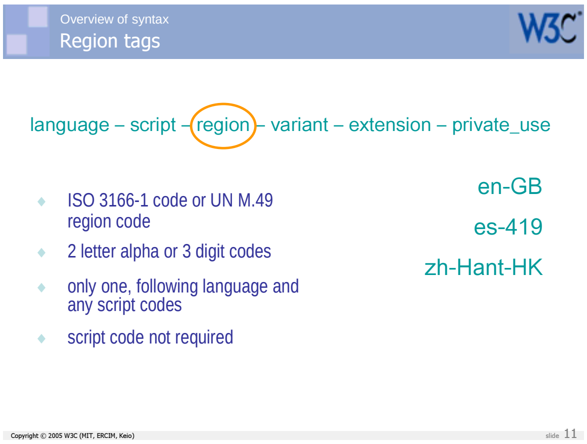
"... avoid script tags except where they add useful distinguishing information."

"for virtually any content that does not use a script tag today, it remains the best practice not to use one in the future."

Addison Phillips

zh-Hans

az-Cyrl



Overview of syntax

Region tags

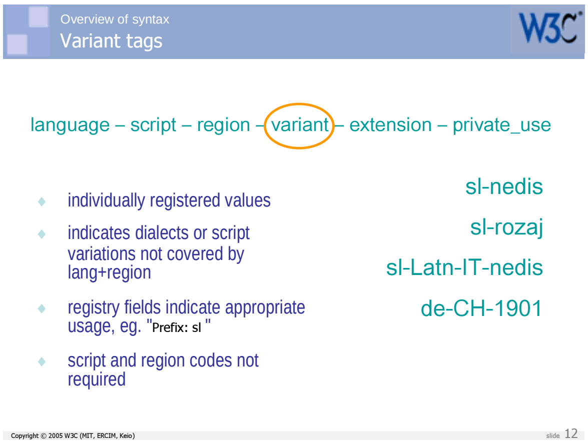
language – script – region – variant – extension – private_use

- ISO 3166-1 code or UN M.49 region code
- 2 letter alpha or 3 digit codes
- only one, following language and any script codes
- script code not required

en-GB

es-419

zh-Hant-HK



Overview of syntax

Variant tags

language – script – region – variant – extension – private_use

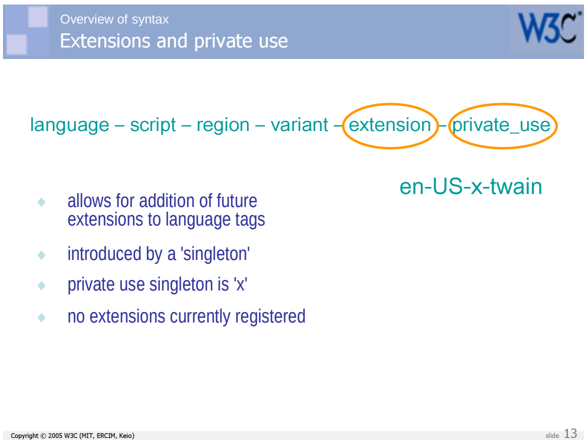
- individually registered values
- indicates dialects or script variations not covered by lang+region
- registry fields indicate appropriate usage, eg. "Prefix: sl "
- script and region codes not required

sl-nedis

sl-rozaj

sl-Latn-IT-nedis

de-CH-1901



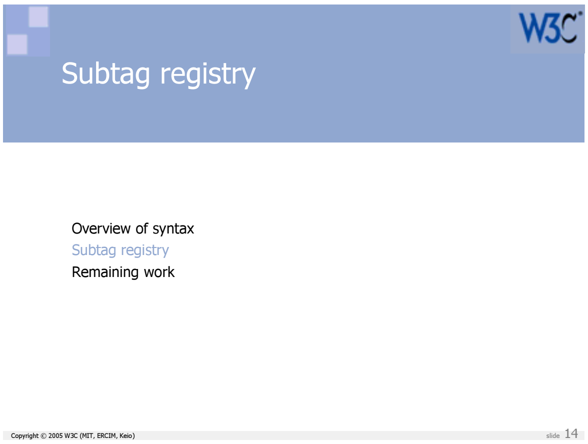
Overview of syntax

Extensions and private use

language – script – region – variant – extension – private_use

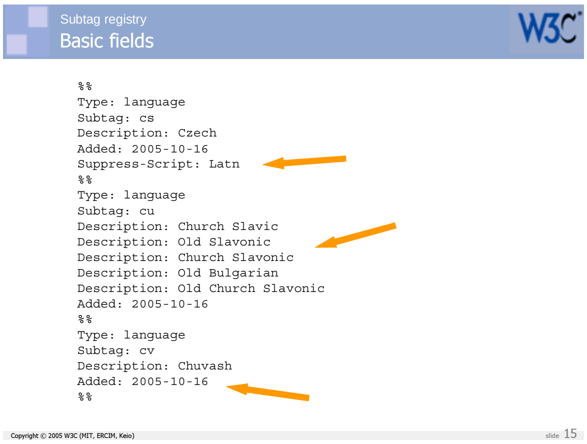
en-US-x-twain

- allows for addition of future extensions to language tags
- introduced by a 'singleton'
- private use singleton is 'x'
- no extensions currently registered



Subtag registry

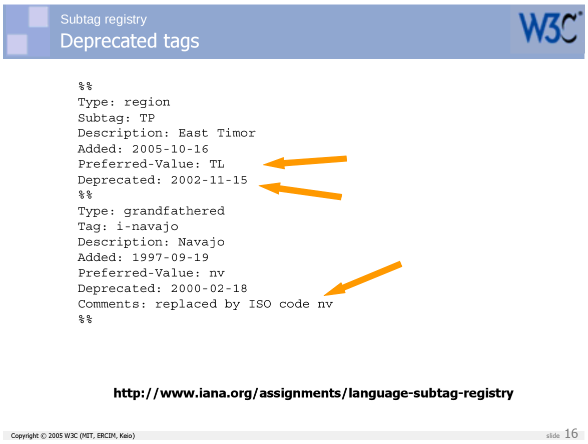
Overview of syntax
Subtag registry
Remaining work



Subtag registry

Basic fields

```
%%
Type: language
Subtag: cs
Description: Czech
Added: 2005-10-16
Suppress-Script: Latn
%%
Type: language
Subtag: cu
Description: Church Slavic
Description: Old Slavonic
Description: Church Slavonic
Description: Old Bulgarian
Description: Old Church Slavonic
Added: 2005-10-16
%%
Type: language
Subtag: cv
Description: Chuvash
Added: 2005-10-16
%%
```

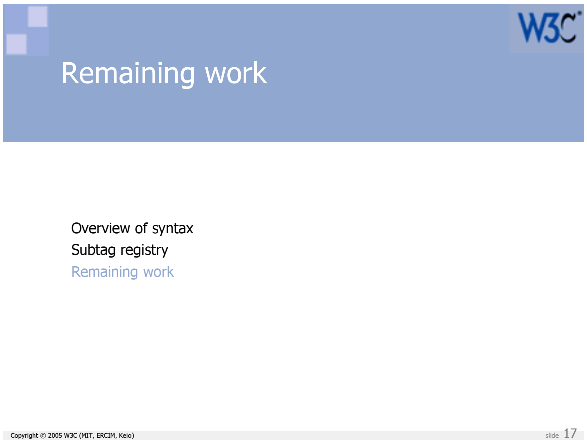


Subtag registry

Deprecated tags

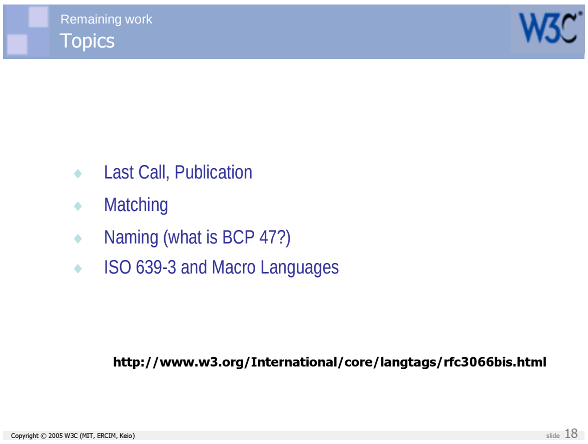
```
%%
Type: region
Subtag: TP
Description: East Timor
Added: 2005-10-16
Preferred-Value: TL
Deprecated: 2002-11-15
%%
Type: grandfathered
Tag: i-navajo
Description: Navajo
Added: 1997-09-19
Preferred-Value: nv
Deprecated: 2000-02-18
Comments: replaced by ISO code nv
%%
```

http://www.iana.org/assignments/language-subtag-registry



Remaining work

Overview of syntax
Subtag registry
Remaining work

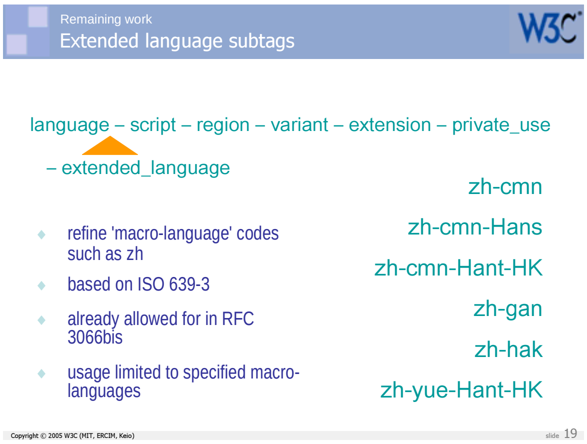


Remaining work

Topics

- Last Call, Publication
- Matching
- Naming (what is BCP 47?)
- ISO 639-3 and Macro Languages

http://www.w3.org/International/core/langtags/rfc3066bis.html



Remaining work

Extended language subtags

language – script – region – variant – extension – private_use

– extended_language

- refine 'macro-language' codes such as zh
- based on ISO 639-3
- already allowed for in RFC 3066bis
- usage limited to specified macro-languages

zh-cmn

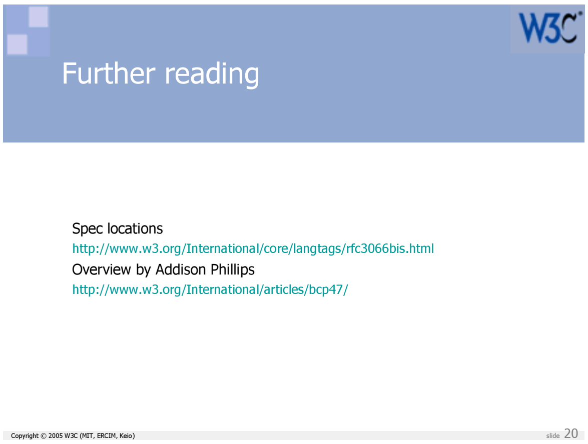
zh-cmn-Hans

zh-cmn-Hant-HK

zh-gan

zh-hak

zh-yue-Hant-HK



Further reading

Spec locations

http://www.w3.org/International/core/langtags/rfc3066bis.html

Overview by Addison Phillips

http://www.w3.org/International/articles/bcp47/