

网页头部的声明应该用 lang="zh" 还是 lang="zh-cn" ?

关注问题

写回答

1 条评论

分享

邀请回答

举报 ...

21 个回答

默认排序



贺师俊

JavaScript、前端开发、编程 等 4 个话题的优秀回答者

收录于 编辑推荐 · 639 人赞同了该回答

把我在另一个回答里的评论整理一下搬运过来：

lang属性的取值应该遵循 [BCP 47 - Tags for Identifying Languages](#)。

单一的 zh 和 zh-CN 均属于废弃用法。

问题主要在于，zh 现在不是语言code了，而是macrolang，能作为语言code的是cmn（国语）、yue（粤语）、wuu（吴语）等。我通常建议写成 zh-cmn 而不是光写 cmn，主要是考虑兼容性（至少可匹配 zh），有不少软件和框架还没有据此更新。

zh-CN 的问题还在于，其实多数情况下标记的是简体中文，但是不恰当的使用了地区，这导致同样用简体中文的 zh-SG（新加坡）等无法匹配。更典型的是 zh-TW 和 zh-HK。所以其实应该使用 zh-Hans / zh-Hant 来表示简体和繁体。那么完整的写法就是 zh-cmn-Hans，表示简体中文书写的普通话/国语。一般而言没有必要加地区代码，除非要表示地区特异性，一般是词汇不一样（比如维基百科的大陆简体和新马简体）。

所以回答题主，两种都不对。

如何标记的例子：

1. 简体中文页面：html lang=zh-cmn-Hans
2. 繁体中文页面：html lang=zh-cmn-Hant
3. 英语页面：html lang=en
4. 《回来》的音频，以国语演唱：audio lang=zh-cmn
5. 《海阔天空》的音频，以粤语演唱：audio lang=yue
6. 《进来白相相》的音频，以沪语演唱：audio lang=wuu

注意上述因为都是音频，所以不应加上 Hans/Hant 标记。

7. 《最终信仰》的音频，双语演唱，因此audio不需要标记lang（如果一定需要标记可以用 lang=mul），但是每个段落的歌词可以分别用 p lang=en 和 p lang=zh-cmn 标记（根据简体或是繁体可用 zh-cmn-Hans 或 zh-cmn-Hant）。
8. 《忐忑》的音频，没有歌词，因此不需要标记lang（如果一定需要标记可用 audio lang=zxx）。

需要加地区代码的情况一般比较少，除非为了强调不同地区汉语使用差异。比如：

```
<p lang="zh-cmn-Hans">
```

```
<b lang="zh-cmn-Hans-CN">菠萝</b>和<b lang="zh-cmn-Hant-TW">鳳梨</b>其实是同一种水果。只是大陆和台湾称谓不同，且新马一带的称谓也是不同的，称之为<b lang="zh-cmn-Hans-SG">黄梨</b>。
```

```
</p>
```

【参考菠萝和凤梨有区别吗？】

当然，由于历史原因，有时候不得不继续使用zh-CN。比如中文维基百科，沿用了传统的zh-CN/zh-HK/zh-SG/zh-TW（按照标准应该使用 zh-cmn-Hans-CN、zh-cmn-Hant-HK、zh-cmn-Hans-SG、zh-cmn-Hant-TW）。这时候，合理的软件行为，是将 zh-CN 等转化为 zh-cmn-Hans（即转化为最常见的误用所对应的实际标准写法）。

实际上，各相关标准，也存在一定的滞后。例如CSS的:lang选择器，不支持选择仅仅简体/繁体中文（而不管是cmn或是yue或是min等汉语方言）。理想情况是CSS3对:lang选择器的语法进行升级，即BCP 47中的高级匹配算法，支持:lang(*-Hans)这样的写法。

更新：[Selectors Level 4](#) 已经加入了对BCP 47高级匹配算法的支持，即支持:lang(*-Hans)的写法。（感谢 @一丝 君。）

编辑于 2014-05-05



猎奇怪蜀黍

本质骑士认证

302 人赞同了该回答

zh 是中文，代表的早宏语言（Macrolanguage），zh 单独由表示由立整体，可以早方言，立言立，简繁体等混合内容，比音大陆地区大部分人都能认识不少

639

40 条评论

分享

收藏

感谢

话，局限性就体现出来了，没法表达繁体普通话，这时为了精准性，应该用独立语种替换，包括但不限于普通话和七大方言：

cmn 普通话（官话、国语）
wuu 吴语（江浙话、上海话）、czh 徽语（徽州话、严州话、吴语-徽严片）
hak 客家语
yue 粤语（广东话）
nan 闽南语（福建话、台语）、cpx 莆仙话（莆田话、兴化语）、cdo 闽东语、mnp 闽北语、zco 闽中语
gan 赣语（江西话）
hsn 湘语（湖南话）
c jy 晋语（山西话、陕北话）

语言文字标签书写顺序：

language-extlang-script-region-variant-extension-privateuse
语言文字种类-扩展语言文字种类-书写格式-国家和地区-变体-扩展-私有

语言文字标签示例：

language:
fr 法语
es 西班牙语

language-extlang 或 language:
zh-lzh 或 lzh 中文（文言文）
sgn-csl 或 csl 手语（中国大陆）

language-region:
zh-CN 中文（简体，中国大陆）
en-US 英语（美国）

language-script:
cmn-Hans 普通话（简体）
yue-Hant 粤语（繁体）

language-script-region:
cmn-Hans-CN 普通话（简体，中国大陆）
cmn-Hant-TW 普通话（繁体，台湾）

language-script-variant:
zh-Latn-pinyin 汉语拼音（例：nǐ，拉丁字母形式）
zh-Bopo-pinyin 注音拼音（例：ㄋㄧˇ，罗马字形式，中国大陆1958年以前使用，台湾目前使用）

附注：

1. 以下所有 zh 开头写法已于 2009 年废弃，**不推荐使用**，请直接去掉 zh- 前缀并使用 cmn、wuu、yue、gan 等（上述标记已由 2005 年的 extlang 升级到 2009 年的 language，维基百科吴、粤、赣等语言版本均已使用最新的标准写法）代替即可：

zh-Hans, zh-Hans-CN, zh-cmn, zh-cmn-Hans, zh-cmn-Hant, zh-wuu, zh-yue, zh-gan

废弃可能原因：在语言学的分类上，中国语言学者多认为吴语、粤语、闽语等是汉语的方言，而西方学者多认为这些语言是一门和汉语同级关系的单独语种。

2. 以下两种写法均正确，后者描述更精准，但目前浏览器和操作系统都只支持前者，使用新标准可能会造成无法匹配浏览器用户定义字体、网页翻译、程序语言自动切换等功能，前端和码农为了兼容性推荐使用前者：

zh-CN 中文（简体，中国大陆）对应 cmn-Hans-CN 普通话（简体，中国大陆）
zh-SG 中文（简体，新加坡） 对应 cmn-Hans-SG 普通话（简体，新加坡）
zh-HK 中文（繁体，香港） 对应 cmn-Hant-HK 普通话（繁体，香港）
zh-MO 中文（繁体，澳门） 对应 cmn-Hant-MO 普通话（繁体，澳门）
zh-TW 中文（繁体，台湾） 对应 cmn-Hant-TW 普通话（繁体，台湾）

3. 同是简体中文页面，默认或书面语均标记为 cmn 普通话，全文使用地方方言表达时，使用相应语言文字标签标记，例：

“你” “我们” 用 cmn-Hans 普通话（简体）
“依” “啊啦” 用 wuu-Hans 吴语（简体）
“你” “我哋” 用 yue-Hans 粤语（简体）

4. 一般情况不用 region 选项，但如果要针对特定地区特定用语的友好性，则应设置区域，例：

“消息” “黄梨” “自行车” 用 cmn-Hans-SG 普通话（简体，新加坡）
“消息” “菠萝” “自行车” 用 cmn-Hans-CN 普通话（简体，中国大陆）
“消息” “菠蘿” “自行車” 用 cmn-Hant-CN 普通话（繁体，中国大陆）
“訊息” “菠蘿” “單車” 用 cmn-Hant-HK 普通话（繁体，香港）
“訊息” “鳳梨” “腳踏車” 用 cmn-Hant-TW 普通话（繁体，台湾）

“analyse” “color” “elevator” 用 en-US 英语（美国）
“analyze” “colour” “lift” 用 en-GB 英语（英国）

参考：

1. W3C [Language tags in HTML and XML](#)
2. IETF [BCP 47 - Tags for Identifying Languages](#)
3. IANA [Language Subtag Registry](#)

编辑于 2016-05-20

▲ 302 ▼ 21 条评论 分享 ★ 收藏 ♥ 感谢 ...

收起 ^



skydiver
程序员

6 人赞同了该回答

两种写法都可以，看你的需求选择。参考 [W3C: Language information and text direction](#)：

Briefly, language codes consist of a primary code and a possibly empty series of subcodes:

```
language-code = primary-code ( "-" subcode )*
```

Here are some sample language codes:

- "en": English
- "en-US": the U.S. version of English.
- "en-cockney": the Cockney version of English.
- "i-navajo": the Navajo language spoken by some Native Americans.
- "x-klingon": The primary tag "x" indicates an experimental language tag

编辑于 2013-02-23

▲ 6 ▼ 添加评论 分享 ★ 收藏 ♥ 感谢 ...



sinkcup

找个小女盆友，养她读到美少女博士٩_٩

62 人赞同了该回答

“实际上它是一个刮胡刀”——长文考证，结束此题。

▲ 639 40 条评论 ★ 收藏 ♥ 感谢

既然是说网页HTML，那就先找到HTML的制定组织，即W3C (w3.org)。

然后看官方文档：

最新的HTML5的：3 Semantics, structure, and APIs of HTML documents

HTML 4.01的：[Basic HTML data types](#)

HTML5官方文档节选：

The lang attribute (in no namespace) specifies the primary language for the element's contents and for any of the element's attributes that contain text. Its value must be a valid BCP 47 language tag, or the empty string. Setting the attribute to the empty string indicates that the primary language is unknown. [\[BCP47\]](#)

HTML 4.01官方文档节选：

The value of attributes whose type is a language code (`%LanguageCode` in the DTD) refers to a language code as specified by [\[RFC1766\]](#), section 2. For information on specifying language codes in HTML, please consult the section on [language codes](#).

现在这个年代，讨论HTML5即可，那就看看“BCP 47”是什么。W3C官方有讲解：[Choosing a language tag](#)，这篇文章详细的讲解了如何选择语文代码，仔细读一遍，此问题即有了答案。

节选如下：

Language tag syntax is defined by the IETF's [BCP 47](#). In the past it was necessary to consult lists of codes in various ISO standards to find the right subtags, but now you only need to look in the [IANA Language Subtag Registry](#). We will describe the new registry below.

BCP stands for 'Best Current Practice', and is a persistent name for a series of RFCs whose numbers change as they are updated. The latest RFC describing language tag syntax is [RFC 5646, Tags for the Identification of Languages](#), and it obsoletes the older RFCs [46463066](#) and [1766](#).

翻译大意就是：

语言文字代码按照“BCP 47”执行。过去需要查看很多ISO文档，拼起来代码，现在省事了，直接看IANA整理出来的即可。

BCP是“最佳实践”的意思，是由一系列会更新的RFC文档组成。“BCP 47”即“第47号最佳实践”规定了语言文字代码，目前最新文档为RFC 5646，取代4646、3066和1766。

可以看到这解决了HTML 4.01中写死RFC 1766导致无法更新的问题。

然后文中详细的讲解了langtag如何拼接，和BCP 47文档中一样，格式为：

```
langtag      = language
               [ "-" script ]
               [ "-" region ]
               *( "-" variant )
               *( "-" extension )
               [ "-" privateuse ]

language      = 2*3ALPHA                ; shortest ISO 639 code
               [ "-" extlang ]           ; sometimes followed by
                                       ; extended language subtags
               / 4ALPHA                  ; or reserved for future use
               / 5*8ALPHA                ; or registered language subtag

extlang       = 3ALPHA                  ; selected ISO 639 codes
               *2( "-" 3ALPHA )         ; permanently reserved

script        = 4ALPHA                  ; ISO 15924 code

region        = 2ALPHA                  ; ISO 3166-1 code
               / 3DIGIT                  ; UN M.49 code

variant       = 5*8alphanum             ; registered variants
               / (DIGIT 3alphanum)
```

简单的说就是：

1：语种，到IA

1次。

▲ 639

● 40 条评论

★ 收藏

♥ 感谢

2：书写格式，到IANA文档中搜 “Type: script” 。此项可省略，出现0或1次。
3：区域，到IANA文档中搜 “Type: region” 。此项可省略，出现0或1次。
4：声音变种（方言）或文字变种。此项可省略，出现0或多次。
等等。

打开IANA的文档：[iana.org/assignments/la...](http://iana.org/assignments/language-subtags/)

搜索结果如下：

```
%%
Type: language
Subtag: zh
Description: Chinese
Added: 2005-10-16
Scope: macrolanguage
%%
Type: language
Subtag: cmn
Description: Mandarin Chinese
Added: 2009-07-29
Macrolanguage: zh
%%
Type: language
Subtag: yue
Description: Yue Chinese
Description: Cantonese
Added: 2009-07-29
Macrolanguage: zh
%%
Type: language
Subtag: nan
Description: Min Nan Chinese
Added: 2009-07-29
Macrolanguage: zh
%%
Type: language
Subtag: lzh
Description: Literary Chinese
Added: 2009-07-29
Macrolanguage: zh

%%
Type: extlang
Subtag: cmn
Description: Mandarin Chinese
Added: 2009-07-29
Preferred-Value: cmn
Prefix: zh
Macrolanguage: zh
%%
Type: extlang
Subtag: yue
Description: Yue Chinese
Description: Cantonese
Added: 2009-07-29
Preferred-Value: yue
Prefix: zh
Macrolanguage: zh

%%
Type: script
Subtag: Hans
Description: Simplified Chinese
Added: 2005
```

```
%%
Type: script
Subtag: Hant
Description: Han (Traditional variant)
Added: 2005-10-16
%%
Type: script
Subtag: Latn
Description: Latin
Added: 2005-10-16

%%
Type: region
Subtag: CN
Description: China
Added: 2005-10-16
%%
Type: region
Subtag: HK
Description: Hong Kong
Added: 2005-10-16
%%
Type: region
Subtag: TW
Description: Taiwan, Province of China
Added: 2005-10-16

%%
Type: redundant
Tag: zh-Hans-TW
Description: Taiwan Chinese in simplified script
Added: 2005-04-11

%%
Type: redundant
Tag: zh-cmn-Hans
Description: Mandarin Chinese (Simplified)
Added: 2005-07-15
Deprecated: 2009-07-29
Preferred-Value: cmn-Hans
%%
Type: redundant
Tag: zh-cmn-Hant
Description: Mandarin Chinese (Traditional)
Added: 2005-07-15
Deprecated: 2009-07-29
Preferred-Value: cmn-Hant
%%
Type: redundant
Tag: zh-wuu
Description: Shanghaiese or Wu
Added: 1999-12-18
Deprecated: 2009-07-29
Preferred-Value: wuu
%%
Type: redundant
Tag: zh-yue
Description: Cantonese
Added: 1999-12-18
Deprecated: 2009-07-29
Preferred-Value: yue
```

可以发现zh、cmn、yue、nan是平级的，都是language，都可以作为第1位代码。那zh又作为macrolanguage是什么意思？

ISO的解释如下
在2007年的国

▲ 639

💬 40 条评论

★ 收藏

♥ 感谢

北方话，cpx – 莆仙语，czh – 徽语，czo – 闽中语，gan – 赣语，hak – 客家话，hsn – 湘语，mnp – 闽北语，nan – 闽南语，wu – 吴语，yue – 粤语。

BCP 47的解释如下：[BCP 47 - Tags for Identifying Languages](#)

For example, the macro language Chinese ('zh') encompasses a number of languages. For compatibility reasons, each of these languages has both a primary and extended language subtag in the registry. A few selected examples of these include Gan Chinese ('gan'), Cantonese Chinese ('yue'), and Mandarin Chinese ('cmn'). Each is encompassed by the macro language 'zh' (Chinese). Therefore, they each have the prefix "zh" in their registry records. Thus, Gan Chinese is represented with tags beginning "zh-gan" or "gan", Cantonese with tags beginning either "yue" or "zh-yue", and Mandarin Chinese with "zh-cmn" or "cmn". The language subtag 'zh' can still be used without an extended language subtag to label a resource as some unspecified variety of Chinese, while the primary language subtag ('gan', 'yue', 'cmn') is preferred to using the extended language form ("zh-gan", "zh-yue", "zh-cmn").

W3C解释如下：[Choosing a language tag](#)

As we recommended for the collection subtags mentioned above, in most cases you should try to use the more specific subtags, but there are a small number of important exceptions. These are situations where you should continue using a macrolanguage subtag for reasons of backward compatibility.

For example, although BCP 47 explains that zh (the macrolanguage subtag for Chinese) doesn't actually specify which of the many, sometimes mutually unintelligible, dialects of Chinese is actually meant by this subtag, in practice convention overwhelmingly associates the macrolanguage subtag with the predominant language among the encompassed subtags - in this case, cmn (Mandarin Chinese). If your application identified Mandarin Chinese in the past using the language tag zh-CN (Chinese as used in Mainland China), or even just zh, you can continue to use zh in this way. Using cmn or cmn-CN may cause serious compatibility problems if the software or users expect a tag such as zh.

If, on the other hand, you are using zh to refer to another Chinese dialect such as Hakka, you should use the language subtag hak instead.

大意就是：推荐使用language（具体语种，比如cmn、yue），而不是macrolanguage（大语种），如果是为了兼容老设备，才使用macrolanguage（比如zh-cmn、zh-yue）。

从IANA的文档中还能发现，“zh-Hant”、“zh-yue”和“zh-cmn-Hans”都是“Deprecated”，是不让用了吗？再仔细一看提示“Type: redundant”，查查这是什么意思，BCP 47解释如下：

Many of these registered tags were made redundant by the advent of either RFC 4646 or this document. A redundant tag is a grandfathered registration whose individual subtags appear with the same semantic meaning in the registry. For example, the tag "zh-Hant" (Traditional Chinese) can now be composed from the subtags 'zh' (Chinese) and 'Hant' (Han script traditional variant). These redundant tags are maintained in the registry as records of type 'redundant', mostly as a matter of historical curiosity.

大意是：“zh”和“Hant”已经被收录了，拼起来就是“zh-Hant”，所以重复了，不应该作为单独条目收录了，所以废弃了。而拼起来的“zh-Hant”正常使用。[@猎奇怪蜀黍](#)

回到提主的问题：网页应该用“zh”还是“zh-cn”？

答：都不推荐，二选一的话那“zh-cn”好点，因为zh表示的范围太大了。

原因：

- 1、不推荐使用zh，而推荐使用cmn、yue。
- 2、如果前面用精确了，一般用不到CN这种区域代码

为什么不推荐使用zh-CN？

请看下面例子：

“佢係一个须刨嚟嘅”是汉字，和“它是一个刮胡刀”一起用“zh-CN”表示，有几个人理解？明明是相互隔离的两个语种，用同一种代码表示，显然不合理。

推荐这些：

文字代码：

cmn-Hans：普通话用规范汉字写下来，比如“它是一个刮胡刀”，为兼容时，可使用“zh-cmn-Hans”。

yue-Hans：粤语用规范汉字写下来，比如“佢係一个须刨嚟嘅”，没必要为兼容而使用“zh-yue-Hans”，因为看得懂“zh-CN”的人，看不懂这些字。

yue-Hant：粤语用國字写下来，比如“佢係一個鬚刨嚟嘅”。

lzh-Hans：文言文用规范汉字写下来，比如“驴不胜怒，蹄之”。

lzh-Hant：文言文用國字写下来，比如“駱不勝怒，蹄之”。

声音（语言）代码：
cmn：普通话或国语
cmn-CN：普通话
cmn-TW：國語
yue：粤语，比如粤语歌曲、电影的粤语音轨

另附我的一篇文章：《语言和文字的区别》

很多人没有理清一个基本概念：“语言”和“文字”是不同的东西。

想想小学“语文”课本，而不是“语言”课本，为什么？

因为“语言”是指发音，“文字”是单独的事物。

看看法律，就明白了：

《中华人民共和国国家通用语言文字法》第二条：本法所称的国家通用语言文字是普通话和规范汉字。

《中华人民共和国宪法》第一百二十一条：民族自治地方的自治机关在执行职务的时候，依照本民族自治地方自治条例的规定，使用当地通用的一种或者几种语言文字。

《中华人民共和国香港特别行政区基本法》第九条：香港特别行政区的行政机关、立法机关和司法机关，除使用中文外，还可使用英文，英文也是正式语文。

台湾：《常用國字標準字體表》、《次常用國字標準字體表》與《罕用字體表》《語言文字基本法》等。

而大学课本就不严谨了，“C语言编程”是错误的，应该叫“C文字编程”，其发音是“英语”，不存在单独的“C发音”。

所以“language”是“语文”的意思，英文代码也要分两种：语言代码和文字代码，即language tags分为：spoken language tags 和 written language tags。

常见语言：

普通话

粤语

英语

常见文字：

规范汉字（大陆法律规定如此，香港法律规定为“中文”）

國字（台湾法律规定如此）

英文

大部分网站是纯文字的，所以只牵扯到文字代码。

而现在各种形态都出现了，声音代码也需要理清楚。比如：

电影（VCD/DVD年代就有这个问题了）：声音有“普通话”、“粤语”等多种，字幕也有“规范汉字”、“英文”等多种。

游戏：比如星际争霸2的剧情任务，配音有多种，字幕和界面也有多种。

豆瓣电台：文字是“规范汉字”，声音有多种。

不过很遗憾，程序员很少有能把语言和文字分清楚的，做出来的程序凶神恶煞.....

如图：“中文”这种“声音”.....俺真没听过.....

编辑于 2016-04-04

▲ 62 ▼ ● 9 条评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...

收起 ^



管斌瑞
画网页的

7 人赞同了该回答

目前，最新的语言标签标记法的国际标准是IETF的BCP 47(Best Current Practice)，也就是RFC 5646取代了之前的RFC 4646，3066和1766.

RFC 5646和之前版本的区别：

- 所有子标签都在IANA registry查询。
- 所有子标签有固定的位置和长度。
- 描述更具灵活性。

RFC 5646规定语言标签按以下形式排列：

language-extlang-script-region-variant-extension-privateuse

所有语言标签都是大小写无关的，但是依照惯例 language 标签小写，region 标签大写，script 标签首字大写。

所有标签都可以在 [IANA registry](#) 查询，获取更多信息可以阅读[Choosing a language tag](#)。

以下链接包含更多的介绍：

- [Language tags in HTML and XML](#)
- [语种名称代码——阮一峰](#)

编辑于 2015-09-13

▲ 7 ▼ ● 添加评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...



张国舜

7 人赞同了该回答

如果仔细看语言tag (bcp47) 的BNF语法定义，其实@李萧 与 @管斌瑞 的回答才是靠谱的！

bcp47中语言tag的BNF语法定义在的第三、四页：

```
Language-Tag = langtag           ; normal language tags
               / privateuse       ; private use tag
               / grandfathered    ; grandfathered tags
```

规范的语言tag必须是符合3种情况其中的一种。第二种的私有定义我们就不讨论了，而第三种为了兼容和过渡，比如zh-xiang，为什么说zh-xiang特例，后面会说到。

我们只关注第一种情况：看看langtag的定义：

```
langtag       = language
               [ "-" script]
               [ "-" region]
               *( "-" variant)
               *( "-" extension)
               [ "-" privateuse]

language      = 2*3ALPHA          ; shortest ISO 639 code
               [ "-" extlang]     ; sometimes followed by
                                   ; extended language subtags
               / 4ALPHA           ; or reserved for future use
               / 5*8ALPHA         ; or registered language subtag

extlang       = 3ALPHA            ; selected ISO 639 codes
               *2( "-" 3ALPHA)    ; permanently reserved
```

script

▲ 639 ▼ ● 40 条评论 ★ 收藏 ♥ 感谢

```
region          = 2ALPHA                ; ISO 3166-1 code
                  / 3DIGIT              ; UN M.49 code

variant         = 5*8alphanum           ; registered variants
                  / (DIGIT 3alphanum)

extension       = singleton 1*("-" (2*8alphanum))
```

“langtag”定义中很明确：“language”是必须的，后面“script”，“region”，“variant”都为可选，且必须使用-分隔，“variant”可以出现0次或无数次，而“script”等只能出现0或1次。

这个最重要的“language”的定义有三种情况：使用最短ISO 639代码2个或3个字符（注意这个最短！！）加可选的“extlang”（3个字符长度的ISO 639-3代码）。language必须优先考虑ISO 639-1中的代码，如果没有则选择639-2中三位的，如果有extlang，那extlang必须使用ISO 639-3中的代码和扩展预留。

另外两种“language”的定义，一个是必须4个字符长度保留，另一个是5到8个字符的subtag。subtag其实就是一个符合language的定义，因为extlang的扩展可以出现0到2次，所以subtag的字符长度可能是5-8之间。

这里稍微岔开一下说说ISO639的问题，ISO639规范下有6个部分，639-1是2字符编码，639-2是3字符编码，相对ISO639-1来说另一种表示方法，他们都是属于ISO639的一部分，不存在新旧替换的问题。用途不一，比如说ISO 639-2存在两种编码，B类用于书籍著作等，T类则用于技术。接下来就是639-3，其实是ISO639-2的超集。剩下的4、5、6都是实现指导和族的分类编码等，干其他用的，和bcf47没太大关系。

关于 @贺师俊 的回答中，他回答的重点不在bcf47上，而是在说ISO 639。我觉得他错误理解ISO639的几个组成部分之间的关系。准确说各部各有用处，没有新旧替换的概念！zh和zho在639-3中分类属于宏语言（macrolanguage，其实就像香蕉、苹果都叫水果一样，水果就属于macro）而yue，cmn则是独立语言（individual）。非bcf47的用途上说，他的用法没错。但在bcp47中，他的说法则不符合shortest ISO 639 code这一要求。所以“language”定义的要优先考虑ISO639-1的编码，如果不存在则使用ISO639-2T。需要extlang，则从ISO639-3中individual选取，因为macrolanguage的编码在639-1和639-2中已经有了。因此zh,zh-yue, ar,ar-arb都是符合规范的“language”定义的。

所以说“zh-xiang”并不符合“language”的语义定义，它定义于“grandfathered”中为了兼容和过度。

接着“language”后都是可选的，但必须保证先后次序！“script”是描述使用“脚本”，按我意思理解就是使用那种文字表述方式，这个值必须使用ISO 15924的编码，4个字符长度。获取ios系统的语言环境就可以见到这种language+script方式。比如简体中文就是zh-Hans,繁体中文就是zh-Hant

我们如果是在ios中获取locale信息，返回的就是language+script+region，region的定义使用ISO 3166-1的2个大写字母编码或者UN M.49三位数字编码。比如US,TW,CN。当然我们如果不用script那就是我们最最常见的了：zh-TW,zh-CN,en-US,en-GB。基本大部分操作系统和浏览器依然沿用这种language+region的方式。其实形式是非常符合规范的，没有听说要替换！剩下就是可以出现0次或n次的和变体和扩展，这个除非你针对非常专业语言分类才用得上。对于技术上说，网页和操作系统，基本也就是language+script+region就足够了！

接下来回答题主的问题，究竟用zh,还是zh-CN。我的意见是语言声明写不写用处不大，你要关心的是你采用的什么字符集而已，如果是utf8编码，你页面就是写泰语、日语、韩语和汉语，照样都能显示！除非你用的特殊字体，访问者电脑上没这些字体。

如果你想做网站的自动多国语言支持，浏览器请求头中Accept-Language中告诉你用户正在使用哪种语言，你可以根据请求中的语言标记来安排服务器返回什么语言的内容给他。

linux\windows\mac的应用程序实现的自动多语言支持也是这样，去访问系统或提供给你的locale信息，这个信息都是符合bcp47的。开发程序时，就不要文字内容写带代码中，而是提供语言包，程序根据获取到的当前的locale信息去判断应该使用那个语言包。只不过有些系统提供的是language+region,如linux，android等。有些是用language+script+region如苹果的系统。

编辑于 2015-07-14

▲ 7 ▼ ● 2 条评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...

收起 ^



程博
基层劳动者

1 人赞同了该回答

参照 google，这个是简体版

发布于 2017-11-16

▲ 1 ▼ ● 添加评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...

怪棋

▲ 639

● 40 条评论

★ 收藏 ♥ 感谢



FSE, 微信公众号: aisinme

8 人赞同了该回答

```
<html xmlns="http://www.w3.org/1999/xhtml" lang="zh-CN" xml:lang="zh-CN">
```

应该是 zh-cn 的，对于目前在互联网使用的中文一般有三种：

简体中文（中国）zh-cn
繁体中文（台湾地区）zh-tw
繁体中文（香港）zh-hk

编辑于 2013-02-23

▲ 8 ▼ ● 3 条评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...



李萧

来不及洞察和理解这个鬼魅的世界

17 人赞同了该回答

目前，语言的标签表示法的国际标准是RFC 4646，名称是《Tags for Identifying Languages》。
简单说，这个文件规定，一种语言的标签应该按照如下方式排列：

language-script-region-variant-extension-privateuse

1. language：这部分就是ISO 639规定的代码，比如中文是zh。
2. script：表示变体，比如简体汉字是zh-Hans，繁体汉字是zh-Hant。
3. region：表示语言使用的地理区域，比如zh-Hans-CN就是中国大陆使用的简体中文。
4. variant：表示方言。
5. extension-privateus：表示扩展用途和私有标识。

一般约定，language标签全部小写，region标签全部大写，script标签只有首字母大写。不同标签之间用连字号-链接。

下面列出一些与中文有关的语言标签。

zh-Hans 简体中文
zh-Hans-CN 大陆地区使用的简体中文
zh-Hans-HK 香港地区使用的简体中文
zh-Hans-MO 澳门使用的简体中文
zh-Hans-SG 新加坡使用的简体中文
zh-Hans-TW 台湾使用的简体中文
zh-Hant 繁体中文
zh-Hant-CN 大陆地区使用的繁体中文
zh-Hant-HK 香港地区使用的繁体中文
zh-Hant-MO 澳门使用的繁体中文
zh-Hant-SG 新加坡使用的繁体中文
zh-Hant-TW 台湾使用的繁体中文

此外，还有一些目前仍在使用，但因不符合规范，将被逐步替代（grandfathered）的标签。

zh-hakka 客家话
zh-cmn 普通话
zh-cmn-Hans 简体普通话
zh-cmn-Hant 繁体普通话
zh-gan 江西话
zh-guoyu 国语
zh-min 福建话
zh-min-nan 闽南话
zh-wuu 吴语（上海话）
zh-xiang 湖南话
zh-yue 粤语

【参考】：[语种的名称代码](#)

发布于 2013-02

▲ 639 ▼ ● 40 条评论 ★ 收藏 ♥ 感谢



Lee William

你必须非常努力,才能看起来毫不费力

6 人赞同了该回答

虽然各位说了很多....

先抬头看下

比如 知乎当前这个页面用的是"lang=zh-CN" =.=

发布于 2016-06-29

▲ 6 ▼ 7 条评论 分享 收藏 感谢 ...



橘子叶上的太阳

5 人赞同了该回答

看完也没明白到底哪种是最标准的，也是醉了！！

编辑于 2017-01-16

▲ 5 ▼ 2 条评论 分享 收藏 感谢 ...



六十六

左翼程序员联盟

3 人赞同了该回答

前排大佬们回答得真实多啊，各种引用，还是最底下的回答靠谱

逼乎就是逼乎

发布于 2017-06-02

▲ 3 ▼ 添加评论 分享 收藏 感谢 ...



匿名用户

1 人赞同了该回答

就上面对汉语的划分，个人有点看法。

把粤语、客家话等列为汉语方言的各语种，提高到和汉语等同地位其实是混淆了汉语和普通话。

所谓语言（个人认为更应该成为语种），应该是文和言的组合。“文”就是书写、记载所使用的字；“言”就是发音，怎么说出来。事实上根本不存在汉语单独一个语言，汉语，究其本意就是汉族使用的语言。粤语、客家话、赣语等等都是汉族人使用的语言，就是汉语的一个组成部分。所以汉语更应该是属于一个宏语种（macrolang）。就像通常所说的英语。

事实上，造成这种误解，主要就是把普通话简单的等同于汉语了，其实大家都明白，普通话实际上就是在北京地方发音的基础上配合简体汉字制定出来的一种语言方案，说到底，普通话和粤语、长沙话、客家话等等一样都是汉语在流通过程中，随着历史的变迁，逐渐形成了各有特色的“方言”。准确的说，不应该叫“方言”，尤其粤语，因为在粤语中使用的部分字和普通话中使用的字都有所差异，而且就使用的广泛性，都已经超出了普通方言的范畴。就和普通话一样，只不过，普通话是以国家的名义的推广，减少民族间的交流阻碍，建立一种统一的语言环境。而粤语、客家话等则是随着使用人群的自然流动，自发的传播开的，就其使用的范围来说，说是“方言”都有点名不符实了。

发布于 2016-05-11

▲ 1 ▼ 添加评论 分享 收藏 感谢 ...



李克振

前端工程师

在赞同最多 @贺师俊 的回答中有一句“单一的 zh 和 zh-CN 均属于废弃用法。”我不知道是出处何在，但是我觉得是不妥当的会误导很多人，zh，zh-cn，cmn是适合不同场合用的，关于题主的问题，如果是在网页中推荐zh，因为阅读的人只要是中国人实际是无所谓什么汉字的，当然你的网站有特别要求除外

编辑于 2017-12-14

▲ 0 ▼ 添加评论 分享 收藏 感谢 ...



Rosev

未来的

▲ 639

40 条评论

收藏

感谢

w3c官网是这样说的：

发布于 2017-05-04

▲ 0 ▼ ● 添加评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...



随便看看

看了半天，我想说查看当前网页的源代码 lang="zh-CN"

发布于 2016-12-26

▲ 0 ▼ ● 添加评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...



xgqfrms

software engineer

ISO 639-1 Language Codes

ISO 639-1 defines abbreviations for languages.

In HTML and XHTML they can be used in the lang and xml:lang attributes.

See also: next reference for country codes.

w3schools.com/tags/ref_...

编辑于 2016-10-21

▲ 0 ▼ ● 2 条评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...



黄赫哲

好奇怪 而且很不理解 简体字的页面如果用zh-cmn-Hant-HK的话是不是就变成繁体字了？

发布于 2016-06-03

▲ 0 ▼ ● 1 条评论 ➦ 分享 ★ 收藏 ♥ 感谢 ...

▲ 639 ● 40 条评论 ★ 收藏 ♥ 感谢



匿名用户

3 人赞同了该回答

[HTML 语言代码参考手册](#)

搞不懂各位大神在说些什么

发布于 2016-12-10

▲ 3 ▼ ● 添加评论 ↗ 分享 ★ 收藏 ♥ 感谢 ...



匿名用户

1 人赞同了该回答

值 涵义

1. zh-cmn 普通话
2. zh-cmn-Hant
3. 普通话(繁体)
4. zh-cmn-Hans 普通话(简体)
5. zh-cmn-Hant-HK 普通话（繁体，香港地区）
6. zh-cmn-Hans-CN 普通话（简体，中国大陆）
7. zh-gan 赣语
8. zh-hak 客家话
9. zh-yue 粤语
10. zh-hsn 长沙话
11. zh-yue-Hant-HK 粤语（繁体，香港地区）

摘录源：[Understanding the New Language Tags](#)

不会英语，只会谷歌翻译。。欢迎勘误

编辑于 2014-07-23

▲ 1 ▼ ● 添加评论 ↗ 分享 ★ 收藏 ♥ 感谢 ...



雅努斯

1 人赞同了该回答

推荐《[HTTP权威指南](#)》

发布于 2015-09-14

▲ 1 ▼ ● 添加评论 ↗ 分享 ★ 收藏 ♥ 感谢 ...

[写回答](#)

4 个回答被折叠（为什么？）