راهنمای برچسب های پیکره همشهری

پیکره همشهری در قالب XML نگهداری میشود. جدول ۱ برچسب های استفاده شده در پیکره همشهری را نشان می دهد. شایان ذکر است نسخه ۲ پیکره تعداد برچسب های بیشتری نسبت به نسخه ۱ دارد که در جدول مشخص شده اند.

در نسخه ۲ پیکره همشهری برچسب ORIGINALFILE مسیر نسبی[1] صفحه اینترنتی که خبر از آن استخراج شده است را نشان می دهد. این برچسب بنا به درخواست برگزار کنندگان همایش CLEF از نسخه ۱ پیکره جهت استفاده در آزمایشات Persian@CLEF2008 و Persian@CLEF2009 حذف گردید.

جدول ۱: مشخصات برچسب های نسخه ۱ و ۲ پیکره همشهری در قالب CLEF

برچسب	توضیح	نسخه ۱	نسخه ۲
HAMSHAHRI2	هر فایل نسخه ۲ پیکره با این برچسب شروع می شود که کلیه اخبار یک روز را در خود جای می دهد		✓
HAMSHAHRI	هر فایل نسخه ۱ پیکره با این برچسب شروع می شود که کلیه اخبار یک روز را در خود جای می دهد	✓	
COPYRIGHT	متن حقوق محفوظ را در خود جای می دهد (در تصویر بالا خلاصه شده است).		✓
DOC	شروع یک خبر را مشخص می کند.	✓	✓
DOCID	شماره منحصر به فرد خبر را در کل پیکره مشخص می کند.	✓	✓
DOCNO	کاملاً مشابه DOCID می باشد.	✓	✓
ORIGINALFILE	نام و مسیر نسبی صفحه اینترنتی خبر را در سایت همشهری مشخص می کند.		✓
ISSUE	مشخصات انتشار خبر را مشخص می کند. این رشته مستقیماً از صفحه اینترنتی خبر استخراج شده و پردازشی روی آن انجام نشده است.	✓	✓
DATE	تاریخ خبر را مشخص می کند. هر خبر دارای دو برچسب DATE می باشد که یکی تاریخ شمسی و دیگری تاریخ میلادی انتشار خبر را در خود جای می دهد.	✓	✓
CAT	طبقه بندی موضوعی خبر را مشخص می کند. هر خبر دارای دو برچسب CAT می باشد که موضوع خبر را به دو زبان فارسی و انگلیسی مشخص می کند.	✓	✓
TITLE	عنوان خبر		✓
TEXT	متن خبر	✓	✓
IMAGE	تصاویر را مشخص می کند. این برچسب در هر کجای متن خبر می تواند ظاهر شود. ویژگی		✓

[1] Relative Path

caption عنوان تصویر (در صورت وجود) را در خود دارد.

به همین ترتیب برچسب های IMAGE نیز از نسخه ۱ پیکره حذف گردیدند. دلیل ذکر شده از طرف دست اندرکاران برگزار کننده همایشCLEF آن بود که یک پیکره باید خود-محتوی[۲] باشد و کلیه اطلاعات مورد نیاز برای استفاده از آن در خود پیکره وجود داشته باشد. از آنجایی که برچسب های IMAGE تنها پیوندی به تصاویر را نشان می دهد برای حفظ خود-محتوی بودن پیکره لازم بود کل تصاویر نیز بین شرکت کنندگان توزیع شود. با توجه به حجم بالای تصاویر (نزدیک به ۲ گیگا بایت) و بی استفاده بودن آن در خط پی گیری بازیابی تک منظوره تصمیم بر آن شد که این برچسب نیز از نسخه ۱ حذف گردد. برچسب COPYRIGHT نیز اخیراً به نسخه ۲ پیکره اضافه گردیده است.

نکته: خواننده گرامی توجه داشته باشد که نسخه ۱ و ۲ پیکره در دو اجرای متفاوت ساخته شده اند و شماره منحصر به فرد یک سند که با DOCID مشخص می شود در دو نسخه پیکره متفاوت می باشد.

شکل ۲ و ۳ نمونه ای از اسناد پیکره همشهری نسخه یک و دو را در قالب برچسب های استاندارد CLEF نشان میدهند.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<DOC>
  <DOCID>H-750402-1S1</DOCID>
  <DOCNO>H-750402-1S1</DOCNO>
  <DATE>1996-06-22</DATE>
  <CAT xml:lang="fa">هنر و ادب</CAT>
  <CAT xml:lang="en">Literature and Art</CAT>
  <TEXT>
    هنر طریق از گروهی زندگی در جاودانگی
    طباطبایی احمد هنری آثار نمایشگاه به نگاهی
    . . .
  </TEXT>
</DOC>
```

شکل۲: یک نمونه ای از اسناد مجموعه همشهری نسخه یک

[2] Self-Contained

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<!DOCTYPE HAMSHAHRI2 SYSTEM "hamshahri.dtd">

<HAMSHAHRI2>
<COPYRIGHT>
<![CDATA[Copyright Notice …]]>
</COPYRIGHT>
<DOC>
<DOCID>HAM2-831014-001</DOCID>
<DOCNO>HAM2-831014-001</DOCNO>
<ORIGINALFILE>/1383/831014/irshahr/_armansh.htm</ORIGINALFILE>
<ISSUE>دوشنبه 14 دی 1383 - سیزدهم سال - شماره - 3601</ISSUE>
<DATE calender="Western">2005-01-03</DATE>
<DATE calender="Persian">1383/10/14</DATE>
<CAT xml:lang="fa">شهری.گوناگون</CAT>
<CAT xml:lang="en">Miscellaneous.Urban</CAT>
<TITLE>
<![CDATA[کلانشهرها اجلاس در رسمی وغیر رسمی های حرف]]>
</TITLE>
<TEXT>
<IMAGE caption="گفت عکس فاش را آن توان نمی گاهی که است مهم آنقدر ها حرف:
مختاریان هادی ">/1383/831014/irshahr/027012.jpg</IMAGE>
<![CDATA[
مشایخی مهرداد
یکدیگر با کلانشهرها ریلی سیستم وضعیت اینکه به اشاره با اللهی فتح مهندس
فاز تبریز و شیراز در است، طراحی مرحله در اهواز و کرج مترو :افزود است، متفاوت
در وی.است اول بخش اندازی راه مرحله در نیز مشهد مترو و شده آغاز مترو ساخت اول
:گفت و کرد اشاره زمینه این در متفاوت های دیدگاه به تاکسیرانی ناوگان خصوص
… و نوسازی به معتقد دیدگاه این بیشتر
]]>
</TEXT>
<DOC>
```

شکل ۲: نمونه ای از اسناد پیکره همشهری ۲

راهنمای DTD

برای اعتبار سنجی برچسب ها در فایل های XML از فایلهای جانبی DTD[۳] استفاده می شود. فایل های DTD به یک زبان فرمال[۴] نوشته می شوند و دقیقاً مشخص می کنند چه عناصر و نهاد هایی ممکن است در یک فایل XML ظاهر شوند و چه صفت[۵] هایی داشته باشند. شکل های ۳ و ۴ به ترتیب DTD های مورد استفاده برای نسخه ۱ و تصویر ۲ پیکره همشهری را نشان می دهند.

[3] Document Type Definition
[4] Formal
[5] Attribute

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!ELEMENT HAMSHAHRI        (DOC)+>
<!ELEMENT DOC
       (DOCID,DOCNO,DATE,CAT,CAT,TEXT)>
<!ELEMENT DOCID            (#PCDATA)>
<!ELEMENT DOCNO            (#PCDATA)>
<!ELEMENT DATE             (#PCDATA)>
<!ELEMENT CAT              (#PCDATA)>
<!ATTLIST CAT xml:lang (fa|en) #REQUIRED>
<!ELEMENT TEXT             (#PCDATA)>
```

شکل ۳: DTD پیکره همشهری نسخه ۱

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!ELEMENT HAMSHAHRI2 (COPYRIGHT, DOC+)>
<!ELEMENT DOC              (DOCID,  DOCNO,  ORIGINALFILE,  ISSUE,  DATE,  CAT,
CAT, TITLE, TEXT)>
<!ELEMENT DOCID            (#PCDATA)>
<!ELEMENT DOCNO            (#PCDATA)>
<!ELEMENT ORIGINALFILE     (#PCDATA)>
<!ELEMENT ISSUE            (#PCDATA)>
<!ELEMENT DATE             (#PCDATA)>
<!ATTLIST DATE calendar (Persian | Western) #REQUIRED>
<!ELEMENT CAT              (#PCDATA)>
<!ATTLIST CAT xml:lang (en | fa) #REQUIRED>
<!ELEMENT TITLE            (#PCDATA)>
<!ELEMENT TEXT             (#PCDATA)>
<!ATTLIST IMAGE caption    CDATA #IMPLIED>
```

شکل ٤: DTD پیکره همشهری نسخه ۲