JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 18, No. 5, pp. 949-956, Aug. 2017

한국 인터넷신문 HTML 규격 및 시맨틱스 수준 분석

이 병 학[1]
[1]한경대학교 디자인학과

HTML specification and semantics analysis of korean news sites

Byoung-Hak Lee[1]
[1]Department of Design, Hankyong University, Anseong 17579, Korea

[요 약]

오늘날 인터넷 신문들은 대중적인 디지털 콘텐츠로 자리잡고 있다. 인터넷 신문의 시각적 인터페이스는 대동소이하나 그 골조를 이루는 HTML의 수준 및 규격은 천차만별이다. HTML의 가장 기본적인 목적이 다른 컴퓨터도 이해할 수 있도록 문서를 의미론적으로 기술하는 것이기에 HTML5에서 문서의 시맨틱스(semantics)는 더욱 강조되고 있다. 본 연구에서는 글로벌 인터넷 신문 8개의 HTML을 대조군으로 삼아 한국의 110개 인터넷 신문을 분석하여 실질적으로 문서에 사용된 HTML 규격을 점검하고 그 시맨틱스의 수준을 진단하였다. 분석 결과 조사대상인 110개 한국의 인터넷신문 중 68%가 HTML4 규격에 해당하는 것으로 나타났으며, 110개 중 9%에 해당하는 10개의 웹사이트만이 대조군으로 조사한 글로벌 인터넷신문과 동일한 수준의 HTML5 규격으로 작성되었으며 적극적인 시맨틱스를 적용하고 있는 것으로 나타났다. 번역기술이 인공지능으로 인해 한층 개선되고 있는 이 시점에 한국 인터넷신문의 디지털 콘텐츠들이 세계와 소통하기 위해서는 더욱 더 적극적인 시맨틱스를 적용한 HTML 문서 작성 플랫폼으로의 규격 전환이 필요하다.

[Abstract]

Visual interfaces of news sites look similar while their HTML have lots of different specifications and qualities. It's getting more and more significant to describe HTML semantically to make every computer able to understand contents to be shared as HTML5 specification refers. In this study, I have analysed HTML codes of 110 korean news sites in comparison to those of 8 global news sites. As results, 68% of news sites are still described in HTML4 specifications and only 10 out of 110 are in HTML5 specification and as high quality and strong semantics as global news sites. The result shows most korean news sites platforms had not been changed since they developed in mid-2000 and it's needed to be upgraded as language translation technologies are making it possible to share korean digital contents with the rest of world.

색인어 : 뉴스사이트, HTML, HTML5, 시맨틱스, 인터넷신문
Key word : news sites, HTML specification, HTML5, semantics, semantic Web

http://dx.doi.org/10.9728/dcs.2017.18.5.949

Received 16 August 2017; **Revised** 27 August 2017
Accepted 31 August 2017

***Corresponding Author; Byoung-Hak Lee**

Tel: +82-2-2275-4435
E-mail: leebh@hknudesign.kr

Ⅰ. 서 론

1-1 연구동기

HTML(Hyper Text Markup Language)이 처음 고안될 당시의 목적은 과학문서를 공유하고 특히 컴퓨터가 이해할 수 있는 문서를 만드는 것이었다[1]. 하지만 요소를 주축으로 작성되었던 HTML4 규격의 2000년대 초중반 웹사이트들은 기사, 머리말, 제목, 본문, 꼬릿말 등의 구조적 명칭을 임의적인 명칭의 id속성이나 class속성으로 작성하였기에 컴퓨터가 문서구조를 이해하기 어려운 난점이 존재하였다. 이와 관련하여 HTML 개발을 주도했던 팀 버너스리는 줄곧 파악하기 쉬운 '의미론적 웹(semantic web)'을 강조하였으며[2], 지난 2014년 10월 발표된 HTML5 규격에서는 HTML은 시맨틱 웹(semantic web)의 개념을 적극적으로 명시하며 컴퓨터가 이해할 수 있도록 작성함으로써 어떤 플랫폼과 서비스에서도 공유가 가능하도록 의미론적으로 기술하는 것을 중요시하고 있다[3]. 특히 공유가 중요시되는 매체는 인터넷 신문으로 2005년 7월 28일 「신문 등의 자유와 기능보장에 관한 법률 시행령」에 따라 2005년 7월에서 10월까지 176개의 매체가 한국의 인터넷신문으로 등록한 이후 금번 대선에서 2017년 4월 중앙선거관리위원회에서 심의한 인터넷신문의 숫자는 2829개에 달한다[4]. 이렇게 오늘날 급격하게 늘어난 인터넷신문이 하나의 대중매체로 자리잡은 반면 실제로 인터넷신문이 적용하고 있는 HTML 규격 및 시맨틱스 적용 수준에 대해서는 조사된 바 없다. 따라서 현재 한국 인터넷신문의 HTML 규격 파악 및 공유를 위한 의미론적 문서작성을 의미하는 시맨틱스(semantics)가 얼마나 적용되어 있는지 그 수준을 분석하는 작업이 필요하다. 인터넷신문의 기사들이 검색 및 공유되는 경우 비단 사람뿐만 아니라 컴퓨터에게도 문서의 시작과 끝, 머리말, 제목, 단락, 강조, 꼬리말 등 구조가 일관성을 가지고 명확하게 파악될 수 있어야 하기 때문이다. HTML5에서 새롭게 추가된 <header>, <article>, <section>, <figure>, <figcaption>, <footer>와 같이 시맨틱스에 기반한 요소들이 보이는 양상은 앞으로 웹의 스타일링을 담당하는 CSS(Cascading Style Sheet)뿐만 아니라 구조를 기술하는 HTML을 더욱 이해하기 쉽도록 작성해야 함을 시사한다.

이와 관련한 선행연구들로 HTML에서 본문과 같은 특정 내용을 추출하는 알고리즘 및 방법론에 관한 연구들을 들 수 있는데 이러한 연구들의 기본적인 목적이 HTML4 규격의 자의적 한계성을 극복하려는 목적임을 감안한다면, 시맨틱스가 적용된 HTML5 기반의 웹에서는 <article>과 같은 시맨틱스 요소를 통해 훨씬 더 쉽고 효율적으로 원하는 내용을 추출하는 것이 가능하다[5][6].

1-2 HTML의 버전 별 특성

W3C(World Wide Web Consortium)에 수록된 첫 공식 HTML 기술문서는 1997년 1월에 작성된 3.2버전이다. 이후 1999년 HTML 4.01, 2000년 XHTML 1.0을 거쳐 2014년 10월 HTML5가 등장했다. XHTML이 XML과 HTML을 통합하기 위한 규격이라는 점에서 사실상 HTML4와 별다른 차이가 없음을 감안할 때 HTML4 이후 HTML5가 등장하기까지 15년의 세월이 걸렸기에 가장 오랜 기간 동안 보급된 버전은 HTML4다. 본 연구의 결과에서도 HTML4 규격으로 작성된 인터넷신문이 가장 많은 것으로 나타났다.

1) HTML3

1996년 CSS Level 1이 공식 발표되긴 했지만, 기술이 보급되는 속도를 감안할 때 21세기 초반의 HTML3는 별도의 CSS(Cascading Style Sheet)가 포함되지 않은 채 작성되는 것이 일반적이었고 HTML 자체가 과학 문서를 표준화하는데 1차적 목적이 있었기에 HTML만으로는 그리드나 다단편집과 같은 시각적 구조를 표시하기 어려웠다. 따라서 웹사이트 전체를 <table>요소를 사용하여 하나의 거대한 표로 작성하는 편법이 보편화되었다. <table>요소를 사용한 HTML은 당시의 한계를 극복하고 원하는 시각적 결과물을 얻는 데는 용이했으나 경직된 구조적 특성으로 인해 다양한 크기의 화면에 유동적으로 대처할 수 없었으며 스타일을 적용할 수 없어 일일이 코드를 수정해야하는 번거로움이 있었다. 또한 편법을 사용한 탓에 문서의 개념적 구조를 다시 하나의 표를 만들어 작성함으로써 학습이 힘들었고 이러한 연유로 당시 HTML3를 사용했던 세대가 현재의 HTML5을 받아들이기 어렵게 만드는 걸림돌이 되고 있다.

2) HTML4

요소만을 사용하고 class, id 속성을 적극적으로 활용하여 구조 설계를 맥락에 맞게 직관적으로 단순화시킨 것이 HTML4로 작성된 문서의 특징이다. <font>요소를 통해 HTML 문서에 포함되어 혼란을 가중시켰던 타이포그래피 속성들이 CSS로 이식되면서 웹의 구조와 스타일링은 HTML과 CSS로 명확하게 이원화되었다[5]. 목차, 머리말, 본문, 날개, 꼬리말과 같이 문서의 구조는 예나 지금이나 커다란 변화가 없기에 HTML4에 머물러 있던 15년의 세월 동안 웹폰트, 반응형 디자인 등을 통해 많은 변화를 보인 것은 CSS다.

HTML4에서는 요소 하나만을 사용하여 구조를 작성하기 때문에 요소를 식별하기 위하여 ' '와 같이 'class', 'id' 속성이 추가되었다. 그런데 이 속성들의 명칭이 별도 규칙이 있는 것이 아니라 작성자에 따라 자의적이기에 일관된 해석이 어렵고 혼란을 가중시킨다. '중앙일보', '연합뉴스', '다음뉴스'의 HTML을 살펴보면 동일한 구성임에도 불구하고 그 명칭은 서로 다르게 나타난다.

(1) 중앙일보[7]

```
<div id="doc">
        <div id="gnb"></div>
        <div id="head"></div>
        <div id="nav"></div>
        <div id="body">...기사 본문...</div>
```

```
        <div id="foot">...</div>
</div>
```

(2) 연합뉴스[8]

```
<div id="wrap">
        <div id="header"></div>
        <div id="content">...기사 본문...</div>
        <div id="footer"></div>
</div>
```

(3) 다음뉴스[9]

```
<div id="kakaoWrap">
        <div id="kakaohead"></div>
        <div id="kakaoContent">...기사 본문...</div>
        <div id="kakaoFoot"></div>
</div>
```

3) HTML5

본 연구에서 주목한 HTML5의 특징은 시맨틱스다. HTML5는 시맨틱스 도입을 통해 HTML4의 단점인 태그의 복잡성을 극복하고 본래 HTML의 의미를 충실히 따르고자 했다. 다음은 규격별 HTML 문서에서 설명하는 HTML의 의미다.

- 플랫폼 간 이동이 자유로운 하이퍼텍스트 문서[11][12]
- 모든 컴퓨터가 이해할 수 있는 표준 퍼블리싱 언어[7]
- 의미론적으로 작성되는 과학문서를 위한 언어[3]

여기서 HTML의 가장 주요한 기능은 어떠한 상황에서건 문서의 구조와 내용이 해석되어야 함을 알 수 있다. 처음 HTML을 고안한 팀 버너스리는 1998년「The Semantic Web」이라는 기사를 통해 다음과 같이 의미론적 웹을 설명한다.

"시맨틱 웹(semantic web)은 별개의 웹이 아니라 체계적으로 정의된 의미의 정보가 담긴 현재 웹의 개선안이다. 시맨틱 웹을 통해 컴퓨터와 사람들은 협동할 수 있다. 현재 쓰고 있는 웹의 구조에 시맨틱 웹을 더하려는 움직임은 이미 시작되었으며, 머지않아 컴퓨터들은 지금은 단순히 표시하는 것에 그치고 있는 정보들을 "이해"하는 것으로 그 기능이 향상될 것이다[2]."

따라서 HTML5에서 추가된 새로운 요소들(elements)의 역할은 HTML을 더욱 복잡하고 어렵게 만드는 것이 아니라 HTML 문서를 더욱 쉽고 명확하게 파악할 수 있는 의미론적 특성인 시맨틱스(semantics)를 강화한 것이다. 영역 구분에 있어 확실한 의미를 나타내는 <section>, <article>, <main>, <aside>, <header>, <footer>, <nav>와 같은 요소들이 그 의미가 자의적이었던 를 대신할 수 있도록 새롭게 추가되었다(HTML5 3.2). 본문 작성에 있어서도 이미지와 캡션을 합쳐 <figure>라는 하나의 구조체로 만들었고, 캡션을 위하여 <figcaption>이라는 전용 요소를 추가하였다[3]. 이러한 요소들의 이름이 강조(emphasis)를 의미하는 처럼 축약되지 않고 실제 단어를 그대로 사용한 점 또한 새로 추가된 요소들이 보이는 공통적인 특징이다. 다음 뉴욕 타임즈와 가디언지의 HTML에서 일관성 있는 시맨틱스의 적용이 드러난다.

(1) 뉴욕타임즈[13]

```
<div id="shell">
        <header></header>
        <nav></nav>
        <main>
                <article>...기사...</article>
                <aside></aside>
                <section></section>
        </main>
        <section></section>
        <footer></footer>
</div>
```

(2) 가디언[14]

```
<header>
        <nav></nav>
</header>
<div class="l-side-margins">
        <article>...기사 본문...</article>
</div>
<footer>
        <aside>
                <section></section>
        </aside>
</footer>
```

II. 분석대상 및 연구방법

2-1 분석대상

HTML 규격은 W3C의 권장사항(Recommendation)이기에 반드시 규격을 따라야 할 필요는 없다. 따라서 현재 HTML5 규격이 실제로 바람직하게 적용되고 있는 수준을 파악하기 위하여 글로벌 언론사인 뉴욕타임즈(New York Times), 가디언(Guardian), 르몽드(Le Monde), 텔레그래프(Telegraph), CNN, 내셔널 지오그래픽(National Geographic), 워싱턴 포스트(Washington Post), 로이터(Reuter)의 웹사이트를 대조군으로 선정하여 2017년 5월 각 언론사에서 운영중인 인터넷신문의 기사 HTML을 분석하였다.

실험군의 선정에 있어서 중앙선거관리위원회에 심의를 요청한 2829개 인터넷신문의 HTML을 전수조사하기에는 본 분석이 일부 정성적 방법을 포함하고 있기에 상대적으로 그 표집이 방대하였다. 또한 인터넷신문이 대중매체라는 점에서 그 대중성을 고려하지 않을 수 없기에 본 분석에서는 유의미한 결과를 도출하기 위하여 대형 포털사이트인 네이버와 다음에 기사를 제공하는 108개 인터넷신문 및 SNS를 통해 빈번하게 공유되는 인터넷신문인 'ㅍㅍㅅㅅ'와 '허핑턴포스트 코리아'로 구성된 총 110개 인터넷신문을 실험군으로 선정하여 2017년 5월에 퍼블리싱된 기사의 HTML을 분석하였다. 대조군과 실험군

및 조사결과는 다음 구글 공유 스프레드시트 문서를 참고 https://docs.google.com/spreadsheets/d/1BE7ZMnzVoLkkDF82MrVOxqj7fbFiGTm2HeiUGgF8gOs/edit#gid=0 [15].

표 1. 조사항목
Table 1. Investigating items

Declaration	Open Graph	Layout	Sectioning Elements	Phrasing Elements
HTML5, HTML4, HTML3, n/a	O, X	semantics, , <table>	<header>, <nav>, <main>, <article>, <section>, <aside>, <footer>	<h1~h6>, <p>, <figure>, , , <b>,

2-2 연구방법

각 HTML의 분석은 크롬(Chrome) 웹브라우저에서 제공하는 '개발자 도구'를 사용하였으며, HTML5 규격 상세문서를 토대로 시맨틱스와 관련된 요소를 선정하였다. 조사항목은 다음과 같다.

(1) 선언한 HTML규격과 실제적으로 작성된 HTML규격의 차이를 분석하기 위하여 각 인터넷신문이 선언한 HTML규격을 조사하였다.

(2) 기사를 공유하기 위한 정보를 담고 있는 Open Graph의 작성 여부를 조사하였다.

(3) 실질적인 HTML 규격을 파악하기 위해 <body>의 전체 영역을 구분하는 레이아웃 방식을 HTML5의 시맨틱스 방식, HTML4의 방식, HTML3의 <table>방식으로 구분하였다.

(4) 영역 구분(sectioning)에서의 시맨틱스 적용 수준을 판단하기 위하여 <header>, <nav>, <main>, <article>, <aside>, <footer>요소의 사용 여부를 조사하였다. 이 중 <header>, <nav>, <footer>는 문서에 필수적으로 포함되는 요소들이며 <main>, <article>, <aside>, <section>의 경우 콘텐츠에 따라 선택적으로 포함되는 요소다. 대조군의 경우 최소한 4종의 영역 구분 시맨틱스 요소를 사용한 것으로 나타났다.

(5) 본문 시맨틱스의 수준을 판단할 수 있는 4가지 기준을 정하여 차례대로 대상을 좁혀나가는 방식으로 본문 시맨틱스의 수준을 분석하였다. 4가지 기준은 다음과 같다. 첫째, 본문의 단락구분 여부. 둘째, 단락구분에 문단을 의미하는 <p>요소의 사용여부. 셋째, 제목을 의미하는 <h1>~<h6>요소의 사용여부. 넷째, 그 외 도판을 의미하는 <figure>요소 및 강조를 의미하는 , 요소의 사용여부. 단락구분에
요소를 사용하는 것은 현재 규격에서 권장하지 않고 있다.

Ⅲ. 조사결과

먼저 대조군으로 삼은 8개의 글로벌 인터넷 신문에 적용된 HTML 특성을 우선적으로 분석하고 그에 비추어 110개 한국 인터넷 신문의 HTML 조사 결과는 5가지 항목별로 나누어 정리하였다.

3-1 글로벌 인터넷 신문의 HTML

표 2. 글로벌 인터넷 신문 조사결과
Table 2. HTML of Global Internet journal

Declaration	Open Graph	Layout	Sectioning Elements	Phrasing Elements
HTML5	O(7), X(1)	Semantics	At last 4 semeantic elements used in sectioning	<h1>~<h6>(all), <p>(all), <figure>(6), , (3)

대조군으로 삼은 인터넷 신문의 경우, 공히 HTML5 규격의 시맨틱스를 적극적으로 적용하고 있는 것으로 나타났다. 영역 구분에 사용된 시맨틱 요소의 경우 모든 신문이 <head> 요소와 <footer>요소를 공통으로 사용하고 있었으며, 이외에 최소 4개의 요소를 사용하고 있다. 본문 작성의 경우 모든 신문이 기본적으로 <h1>~<h6>과 <p>요소의 권장사항을 준수하고 있었으며, 75%에 해당하는 6개의 신문이 HTML5 규격에서 새롭게 정의된 도판 요소인 <figure>태그를 적용하고 있는 것으로 나타났다. 또한 강조의 스타일링에 <b>또는 <i>요소를 지양하고 과 을 사용하고 있는 등 글로벌 인터넷 신문의 경우 전반적으로 HTML5 규격의 권장사항을 적극적으로 준수하고 있는 것으로 나타났다.

3-2 한국 인터넷 신문의 HTML

(1)HTML 규격 선언 및 Open Graph

절반에 해당하는 인터넷 신문이 HTML5로 규격을 선언하고 있었다. XHTML이 기본적으로 HTML4에 해당함을 고려하면 HTML5와 HTML4가 각각 절반에 가까운 분포를 보였으며, 아무런 선언도 하지 않은 경우는 2%에 불과했다. Open Graph의 경우 90%에 이르는 대부분의 인터넷신문이 <head>요소에 오픈그래프를 포함하며 나머지 10%는 없거나 부족하게 작성되어 있었다.

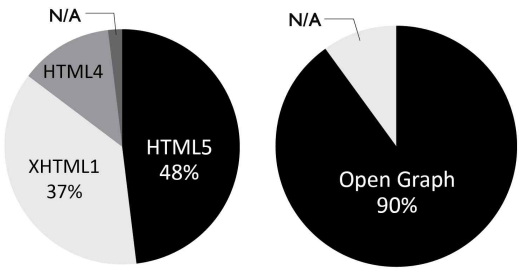


그림 1. HTML 규격 선언 및 Open Graph 사용여부
Fig. 1. HTML declaration and Open Graph

(2) 영역구분 방식 및 시맨틱 영역 요소 활용

선언부의 규격보다도 HTML 규격을 더욱 정확하게 판단할 수 있는 영역구분 방식을 분석하였다. <table>요소를 사용한 HTML3, 요소를 사용한 HTML4, 시맨틱스를 적용한 HTML5로 나누어 영역을 구분하는 방식에 따라 HTML 규격을 조사한 결과 시맨틱 영역 요소를 사용한 HTML5 규격의 인터넷 신문은 18%에 불과하였으며, 이 중 충분히 시맨틱스가 적용된 인터넷 신문은 전체 110개 조사대상 가운데 16개인 14.5%로 나타났다. 뉴스토마토, 디스패치, 맥스무비, 법률신문, 블로터, MBC, 서울경제, 시사저널, 전자신문, 조선일보, 코리아헤럴드, 텐아시아, 프프ㅅㅅ, 한국경제, 한겨레21, 허핑턴포스트가 이에 포함되었다.

표 3. 영역구분 방식
Table 3. Sectioning HTML

HTML5	HTML4	HTML3
By Semantics Elements	By	By <table>
20 (18%)	75 (68%)	15 (14%)

표 4. HTML5에서의 시맨틱 영역 요소 활용 정도
Table 4. Quality of Sectioning semantics in HTML5

Strong Semantics		Weak Semantics
all	4 ~ 6 elements	under 3 elements
4 (20%)	12 (60%)	4 (20%)
Sisa Journal, EtNews, 10asia, 프프ㅅㅅ	News Tomato, Dispatch, Max Movie, Lawtimes, Bloter, Se daily, Chosun.com, Korea Herald, Hani21, Hankyung, Huppington Post Korea, MBC	OSEN, TV Chosun, Hankook Ilbo, Health Chosun

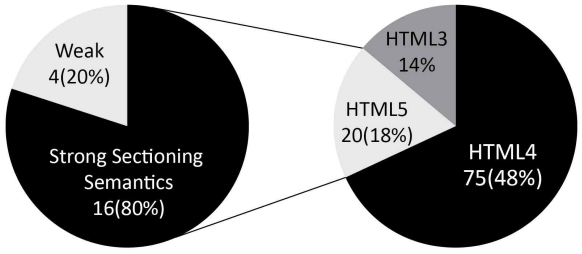


그림 2. 영역구분방식에 의한 HTML 규격 조사결과 및 시맨틱 요소 사용의 적극성
Fig. 2. HTML specification by sectioning elements and Quality of semantics

(3) 본문 단락구분 방식 및 시맨틱 요소 활용

본문의 HTML에서 가장 주요한 분석 기준은 본문의 단락구분 여부다. HTML4 규격이후 HTML에서는
요소에 의한 단순개행보다는 문단을 의미하는 <p>태그의 사용을 권장하고 있다. 따라서 본문의 경우 단락구분 방식, 단락구분 요소, 제목요소, 도판 및 강조요소의 순서로 대상을 좁혀나가며 본문에 적용된 시맨틱스의 수준을 분석하였다. 결과 연합뉴스, 프프ㅅㅅ, 다음뉴스, KBS WORLD, 경향신문, 전자신문, 텐아시아, 맥스무비, 허핑턴포스트, 헬스조선의 기사 본문에 적절한 수준의 시맨틱스가 적용된 것으로 나타났다.

표 5. 본문 시맨틱스 조사결과
Table 5. Quality of Phrasing Semantics

paragraph tagging	block markup				line break
paragraph elements	<p>				
heading	<h1>~<h6>		n/a		
figure, emphasis	<figure>, ,	n/a			
	10	17	8	7	68
phrasing semantics	strong	weak			bad
도수 (%)	10 (9%)	32 (29%)			68 (62%)

Ⅳ. 종합 분석

전체 분석 결과를 종합한 결과 110개의 한국 인터넷 신문을 HTML 시맨틱스 요소 적용 여부 및 레이아웃 방식에 따라 우선 HTML5, HTML4, HTML3로 분류하였으며 HTML5로 분류된 그룹은 다시 대조군인 글로벌 인터넷 신문의 HTML에 비추어 적극적으로 시맨틱스를 적용한 HTML5그룹과 그렇지 않은 그룹으로 세분화하고 HTML4로 분류된 그룹에 대해서는 본문의 단락구분에 있어
요소를 사용하여 단순개행을 한 경우와 혹은 <p>요소를 사용하여 블록지정을 한 경우로 세분화하였다. 최종적으로 국내 인터넷 신문의 HTML 수준을 총 5가지 수준의 그룹으로 나누었고 각 수준을 분석하였다.

표 6. Group A. HTML5 - 강한 시맨틱스
Table 6. Group A. HTML5 - Strong Semantics

- More than 4 sectioning semantics elements - Recommended phrasing semantics.
Dispatch, Max Movie, Lawtimes, Bloter, Sisa Journal, Etnews, 10asia, 프프ㅅㅅ, Han 21, Huppington post korea (10)

A그룹은 글로벌 인터넷신문과 대등한 수준의 HTML5 규격을 갖고 있는 한국 인터넷 신문들이다. '법률신문', '전자신문', '디스패치' 등 종합일간지가 아닌 특정 카테고리의 기사를 전문적으로 취급하는 중소규모의 인터넷 신문들로 구성되었으며, 종합일간지 중에서는 '한겨레21'이 유일하게 포함되었다. 따라서 기존의 기사량이 방대한 대규모의 종합일간지의 경우 업데이트 및 플랫폼 업그레이드가 쉽지 않은 것으로 판단된다. 특히 포털 사이트에 콘텐츠를 노출하지 않고 SNS을 통해서만 공유되는 '프프ㅅㅅ'과 '허핑턴포스트'가 상대적으로 높은 수준의 HTML을 갖추고 있으며, 공유에 적합한 것으로 나타났다. 이와 관련하여 기사가 대중들에게 확산되는 속도가 빨라야 하

는 대중문화 관련 인터넷 신문들인 '디스패치', '맥스무비', '텐아시아'의 수준이 높은 이유도 같은 맥락에서 판단할 수 있다. '맥스무비'의 경우 모바일 전용의 레이아웃을 갖추고 있는 것으로 나타났으며, '블로터'의 경우 워드프레스를 사용하였다.

표 7. Group B. HTML5 - 약한 시맨틱스
Table 7. Group B. HTML5 - Weak Semantics

- Less than 3 semantics elements - Don't use <h1>~<h6> or use   on paragraphs
News tomato, Se daily, Chosun.com, Korea Herald, Hankyung, Hankookilbo, Health Chosun, MBC, OSEN, TV Chosun (10)

B그룹은 HTML5 권장사항을 일부 적용하긴 하였으나, 전체적으로 적용하였다고 보기 어려운 약한 시맨틱스로 판단되는 인터넷 신문이다. 'MBC', '코리아헤럴드', '한국경제', '서울경제', '조선일보', 'TV조선'의 경우 여전히 본문의 단락구분에
요소를 사용하고 있는데, 이는 앞서 언급한 바와 같이 대규모의 인터넷신문의 경우에 기본 기사들의 HTML을 완벽하게 새로운 규격으로 전환하기 어려움을 시사한다. 반면 영역 구분의 경우에는 HTML 소스의 1회 수정으로 다소 간편하게 수정할 수 있기 때문에, B그룹의 경우 HTML5 규격의 중요성을 인지하고 대응한 것으로 판단된다.

표 8. Group C. HTML4 - <p> 단락구분
Table 8. Group C. HTML4 - <p> for paragraphs

- for sectioning - <p> or for paragraph
EBS, KBS WORLD, Zdnet korea, Game Mecca, Kyunghyang, Kihoilbo, New Daily, Daum News, The Scoop, Donga Science, Money S, Mediatoday, Vop.co.kr, Science Times, Sport Seoul, Elle, Yonhap Newd, Wiki Tree, Joongboo Ilbo, Channel Yes, Hani.co.kr, Green Post Korea (22)

C그룹은 를 사용하여 영역을 구분한 전형적인 HTML4 규격의 인터넷신문들이며 본문의 단락구분에 <p>요소를 사용하였기에 HTML5 규격의 중요성을 인지하기만 한다면 손쉽게 전환할 수 있는 인터넷 신문들이다. 대부분 대규모의 종합일간지들로 구성되어 있으며, 초기 개발 시 권장되는 HTML4 규격을 준수하여 작성된 것으로 판단된다. C그룹의 경우 HTML4 권장사항을 잘 따랐기에 A그룹 수준으로의 전환이 용이하다.

표 9. Group D. HTML4 -
 단락구분
Table 9. Group D. HTML4 -
 for paragraphs

- for sectioning -   for paragraphs - don't use <h1>~<h6> for heading
CEO Score Daily, EBN, It Chosun, Ize, JTBC, KBS, MBN, SBS, SBS sports, Ytn Science, Kyeonggi.com, Kmib.co.kr, Kookbang Ilbo, Naver News, Next Daily, Nocut News, Nongmin News, News1, Newsis, Dailian, Donga Ilbo, Digital Times, MyDaily, MK.co.kr, MT.co.kr, Bridge News, Seoul.co.kr, Segye.com, Star News, Sports Donga, Sport World, Sports Chosun, Sportal Korea, Sisa In, Asise, Inews24, Aju News, Able News, Xports News, Ohmynews, Edaily, Incheon Ilbo, Isplus, Ilyo.co.kr, Chosun Biz, Joongang Daily, Joongang Ilbo, Korea Times, Financial News, Pressian, WowTV, Herald Biz (53)

D그룹은 C그룹과 같이 전형적인 HTML4 규격의 인터넷 신문들이며 C그룹과 합쳐 전체에서 68%에 해당하여 가장 많은 분포를 보이는 그룹으로 한국 인터넷 신문들은 대부분은 HTML4에 여전히 머물러있는 것으로 판단된다. D그룹은 C그룹에 비해 기사의 본문 단락 구별에
을 사용함으로써 상위 규격인 HTML5 규격으로의 전환이 쉽지 않을 것으로 예상된다. 기사 본문의 HTML은 기사 작성 시에 DB에 본문과 함께 저장되기 때문이다. HTML5 시맨틱스의 중요성을 인식한다면 적어도 영역구분에 있어서는 손쉽게 B그룹 수준으로 규격 전환이 가능하다.

표 10. Group E. HTML3
Table 10. Group E. HTML3

- no declaration - <table> for sectioning - some are made by frameset which isn't recommended at the moment - some used <font> element whis ins't recommended at the moment
News Culture, Newsen, Design Jungle, Le Monde Diplomatique, MK economy, Munhwa.com, Business Post, Hankooki.com, Smart PC Love, Siminilbo, Women News, Yonhap Infomax, JoyNews24, Computer World, TV daily (15)

E그룹의 경우 현재 사용이 지양된 <table> 요소를 사용하여 영역을 구분하였고, HTML 규격을 선언하지 않았다는 점에서 극초기의 HTML3 규격을 그대로 유지하고 있는 것으로 나타나 상위규격으로의 전환이 시급한 그룹이다. '디자인정글'의 경우 텍스트를 이미지에 합성한 이미지를 제목으로 사용하고 있는데 이러한 방식은 컴퓨터에게 전혀 정보를 전달할 수 없다는 치명적인 방식이다. '매경이코노미'와 '여성신문'의 경우엔 현재 사용하지 않는 프레임 구조로 작성되어 있다.

Ⅴ. 결 론

인터넷 신문을 통해 대부분의 뉴스 콘텐츠가 대중들에게 확산된다는 점 및 인공지능 번역 서비스의 기술 수준이 급속도로 향상되고 있다는 점을 감안한다면 전 세계로 한국의 뉴스 콘텐츠를 확산하기 위해서 상위 규격에 대한 관심 및 W3C에서 권장하는 규격에 관심을 가지고 HTML을 작성할 필요가 있다.

HTML5로 분류된 전체 20개(18%)의 HTML5 규격을 시맨틱스 적용 정도에 따라 다시 두 그룹으로 세분화 한 결과 9%에 해당하는 10종의 인터넷 신문만이 대조군으로 설정한 글로벌 인터넷 신문과 동등한 수준의 HTML을 갖추고 있었으며 그 다음으로 전체 110개 조사대상 중 75개(68%)로 가장 많은 수를 차지한 것은 영역구분에 시맨틱요소가 아닌 요소를 사용한 HTML4 규격이었다. 이 중 본문 단락구분을 여전히 단순 줄바꿈에 해당하는
요소로 작성한 인터넷신문이 53개(48%)로 가장 비중이 높았다. 마지막으로 전체의 14%를 차지하는 15개의 인터넷신문이 HTML3 및 그 이전 규격의 공통적 관행인 전체를 하나의 표로 레이아웃하는 <table>요소를 사용하고 있기에 HTML5 규격으로의 개선이 시급한 것으로 판단된다. \

HTML 규격은 규칙이 아니라 권장사항이기에 어떤 규격을 따르던지 간에 적어도 시각적으로는 동일한 화면을 구현할 수 있다. 하지만 스마트폰을 위주로 화면의 크기가 줄어들면서 점차 중요시되는 것은 풍부한 디자인보다는 컴퓨터에게도 해석될 수 있는 의미의 전달이다. 특히 인터넷신문과 같은 대중매체일수록 공유를 통해 확산될 가능성이 높기 때문에, 검색과 공유를 위해 보다 해석되기 용이한 HTML을 작성하는 것이 필요하다. 분석결과 네이버나 포털에 기사를 제공하지 않고 SNS 통해 주로 공유되는 '프프ㅅㅅ'와 '허핑턴포스트'가 HTML5의 시맨틱스를 적극적으로 사용하고 있다는 점이 이를 반증한다. 하지만 현재 한국 인터넷 신문의 주를 이루는 규격은 HTML4(68%)로 나타났고 HTML5 규격을 적극적으로 적용한 경우는 9%에 불과했다. HTML5 규격이 3년 전에 발표되었음을 감안할 때 인터넷신문의 경우 더욱 시맨틱스를 적극적으로 적용하려는 노력이 필요하다. 다수의 HTML4 및 HTML3 규격이 여전히 존재하는 점은 앞으로 관련교과교육에서도 장애가 될 것이다.

참고문헌

[1] W3C(World Wide Web Consortium). What is HTML? [internet]. Available: https://www.w3.org/TR/1999/REC-html401-19991224/intro/intro.html#h-2.2.

[2] Berners-Lee, Tim(2001, May). The Semantic Web. Scientific American. com, [internet]. Available: https://www.scientificamerican.com/article/the-semantic-web/.

[3] W3C. HTML 5 [internet]. Available: https://www.w3.org/TR/2014/REC-html5-20141028/.

[4] National Election Commission. The definition of internet journals [internet]. Available: http://www.nec.go.kr/portal/knowLaw/quanDetailView.do?contId=201202150112&contSid=0001&quanId=201203038058.

[5] Hyun-Gee Jeon and Chan KOH, "Text Extraction Algorithm using the HTML Logical Structure Analysis", *The Journal of Digital Contents Society*, Vol. 16, No. 3, pp. 445-455, June 2015.

[6] Jeff P., Dan R., "Extracting Article Text from the Web with Maximum Subsequence Segmentation," The 18th international conference on World wide web, pp.971-980, 2009.

[7] W3C. HTML 4 [internet]. Available: https://www.w3.org/TR/1999/REC-html401-19991224/intro/intro.html#h-2.3.2

[8] Joongang-Il-Bo. HTML source [internet]. Available: http://news.joins.com/article/21557874?cloc=joongang|home|newslist1

[9] Yonhap News. HTML source [internet]. Available: http://www.yonhapnews.co.kr/politics/2017/05/10/0501000000AKR20170510072400001.HTML?template=2085

[10] Daum News. HTML source [internet]. Available: http://v.media.daum.net/v/20170514094213264

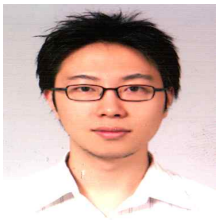
[11] Berners-Lee, Tim. Hypertext Markup Language - 2.0 [internet]. Available: https://www.w3.org/MarkUp/html-spec/html-spec_toc.html

[12] Raggett, Dave. HTML 3.2 Reference Specification [internet]. Available: https://www.w3.org/TR/REC-html32-19970114

[13] New York Times. HTML source [internet]. Available: https://www.nytimes.com/2017/05/09/opinion/an-agenda-for-south-koreas-new-leader.html?action=click&pgtype=Homepage&clickSource=story-heading&module=opinion-c-col-right-region&region=opinion-c-col-right-region&WT.nav=opinion-c-col-right-region&_r=0

[14] The Guardian. HTML source [internet]. Available: https://www.theguardian.com/world/2017/may/09/moon-jae-in-the-south-korean-pragmatist-who-would-be-presidentc

[15] Byoung Hak, Lee. Analysis of korean internet journalism HTML specification and quality of semantics research result [internet]. Available: https://docs.google.com/spreadsheets/d/1BE7ZMnzVoLkkDF82MrVOxqj7fbFiGTm2HeiUGgF8gOs/edit#gid=0



이병학(Byoung-Hak Lee)

2010년 : 서울대학교 대학원 (공예·디자인학 석사)
2013년 : 서울대학교 대학원 (디자인학 박사)

2016년～현 재: 한경대학교 디자인학과 조교수
※관심분야 : 정보인터랙션(Infomative interaction), 타이포그래피(typography) 등