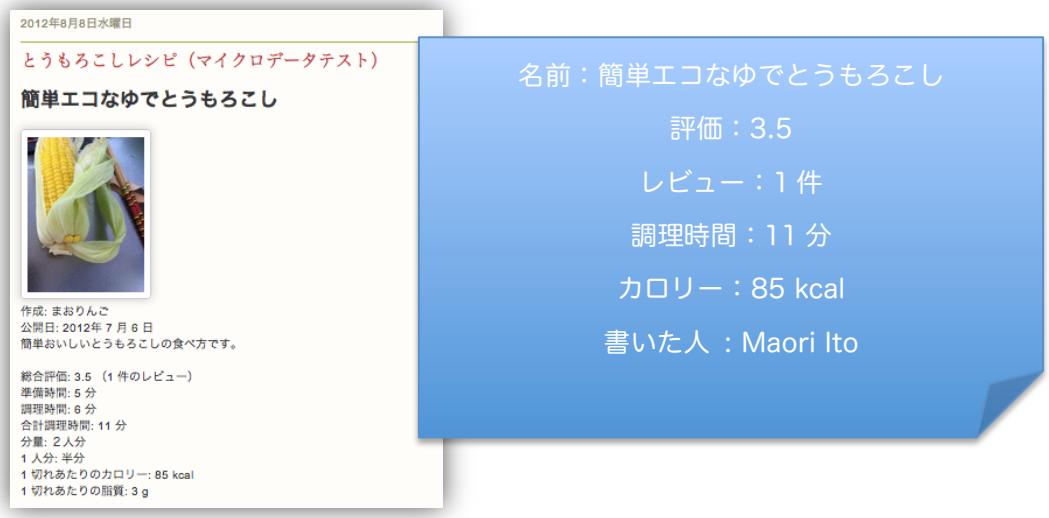
横断検索への Microdata の導入と活用について

2012 年 12 月 19 日（火）

医薬基盤研究所　伊藤真和吏

- Microdata とは
  - HTML にメタデータを埋め込むための記述方法です。
  - Google，Yahoo!，Microsoft Bing 大手検索エンジンサイトは Microdata を利用してより質の高い検索結果を提供するという共同宣言をしています。
- 埋め込むと何が出来るか。
  - イメージとしては，html や database 上の個々のデータにタグ付けをしていき，そのタグ付けを横断検索が拾ってくるイメージです。
    例：とうもろこしレシピ



- Google での検索結果

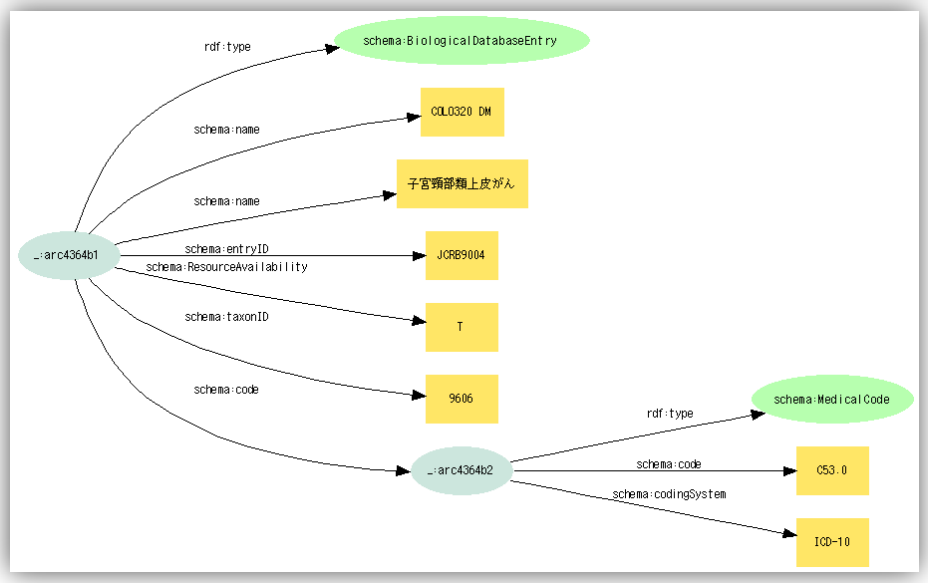
- 具体的な埋め込み方法

Google に取り入れられる Markup の方法：
http://support.google.com/webmasters/bin/answer.py?hl=ja&answer=176035
既に定義されている語彙の一覧：http://schema.org/

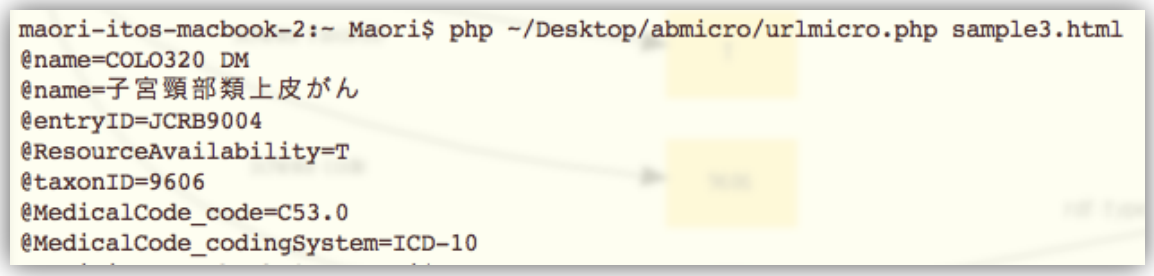
- ライフサイエンス分野の横断検索の例（イメージ）

- 具体的な埋め込み方法（JCRB 細胞バンクより抜粋,
  語彙<BiologicalDatabaseEntry>は提案中）とそのグラフ化

```
<div itemscope itemtype="http://schema.org/BiologicalDatabaseEntry">
<span itemprop="name">COLO320 DM</span>
    <span itemprop="entryID">JCRB9004</span>
    <meta itemprop="ResourceAvailability" content="T"/>
    <meta itemprop="taxonID" content="9606" />human<br>
    <span itemprop="name">子宮頸部類上皮がん</span>
        <span itemprop="code" itemscope itemtype="http://schema.org/MedicalCode">
        <meta itemprop="code" content="C53.0"/>
        <meta itemprop="codingSystem" content="ICD-10"/>
        </span>
</span>
</div>
```

- クローリングの方法
  - マイクロデータを横断検索用に抽出する専用のクローラーを作成
  - 上記の html をクローリングすると以下のようにデータを取得。



    - 神崎正英さんの Microdata Parser and RDF Extractor for ARC2:www.kanzaki.com/works/2012/pub/1007-microdata2rdf.html を横断検索用にアレンジしたものです。
    - @語彙_プロパティの順で出力され，BiologicalDatabaseEntry のみ語彙を省略しています。
    - urlmicro.php と Microdata.php さえあれば取得可能です。
  - 実装方法（追記するだけなら以下の一文で挿入できます）
    - 例：Perl
    - print OUT `php urlmicro.php $url`;
    - 例：PHP
    - echo `php urlmicro.php $url`;
- 相談したいこと
  - BiologicalDatabaseEntry の語彙は提案中でして
    - http://www.w3.org/wiki/WebSchemas/BioDatabases
  - どういった書き方が適切か，議論に参加していただけると嬉しいです。
    - 12 月 19 日（水）〜21 日（金）の BH12.12 で初版は決めてしまおうと考えています。
  - 細胞バンク以外の具体例を一つ一つ増やしていきたいと考えています。
  - 有用な具体例のアイデアを募集しています。
    - 例：データ提供者，臓器，論文の ID，構造式等の表示

- データを繋げるという意味での Microdata
  - グラフにもあるように Microdata は書き方によって，特定のデータがある性質を持ち，入れ子にして書き込むことも出来るので，RDF にも落とし込める性質を持っています。複数の Microdata を持っているデータが共通の表記をしていた場合に繋いでいくということも可能です。そして，RDF とは異なり，データの提供者が実装しやすく既存の DB と共存できるということが，Microdata の強みだと考えています。ご興味ある方にはそれらについても一緒に調査にご協力いただけると幸いです。