

## 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



名前: 簡単エコなゆでとうもろこし  
評価: 3.5  
レビュー: 1件  
調理時間: 11分  
カロリー: 85 kcal  
書いた人: Maori Ito

- Google での検索結果



- 具体的な埋め込み方法

```
1 <div itemscope itemtype="http://data-vocabulary.org/Recipe" >  語の宣言: Recipe
2 <h1 itemprop="name">簡単エコなゆでとうもろこし</h1> name: 簡単エコなゆでとうもろこし
3 <span itemprop="rating">3.5</span> rating: 3.5
4 (<span itemprop="count">1</span> 件のレビュー) <br> count: 1 totalTime: 11分
5 合計調理時間: <time datetime="PT11M" itemprop="totalTime">11分</time><br>
6 1切れあたりのカロリー: <span itemprop="calories">85 kcal</span> calories: 85 kcal
7 </div>
```

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)

Caucasian Human colonの検索結果: 2 hits 表示オプション: 同義語展開あり (デフォルト) ↓  
ファセットによる絞り込み: 生物資源

**案1:画像の表示**

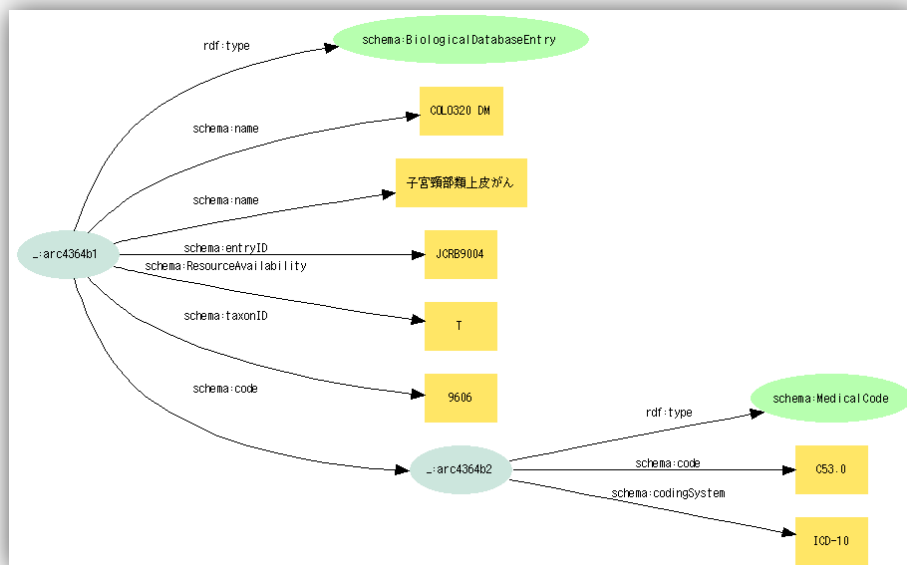
COLO320 DM - JCRB細胞バンク  
<http://cellbank.nibio.go.jp/> JCRB細胞バンク  
生物資源 | ヒト, 動物 (ヒト以外) | 細胞・組織 - ID: JCRB0225  
Species: Human - Disease: 子宮頸部の悪性新生物 - 譲渡: 可  
JCRB0225 COLO320 DM Human Homo sapiens Female 50 year-old  
colon carcinoma of the sigmoid colon colon, adenocarcinoma a round shaped colon  
adenocarcinoma tumor Quinn,LA. Caucasian Human c..... **分譲の可否**

**疾患名の表示**

HT-29-Luc - JCRB細胞バンク  
<http://cellbank.nibio.go.jp/> JCRB細胞バンク  
生物資源 | ヒト, 動物 (ヒト以外) | 細胞・組織 - ID: JCRB1383 - Species: Human  
Disease: 子宮頸部の悪性新生物 - 譲渡: 不可 -Image  
JCRB1383 HT-29-Luc Human Homo sapiens female  
pMSCV-luc transfectee!! line. epithelial-like trans  
stably expressing cell line (HT-29; Hum..... **案2:画像のポップアップ**

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

```
maori-itos-macbook-2:- Maori$ php ~/Desktop/abmicro/urlmicro.php sample3.html
@name=COLO320 DM
@name=子宮頸部類上皮がん
@entryID=JCRB9004
@ResourceAvailability=T
@taxonID=9606
@MedicalCode_code=C53.0
@MedicalCode_codingSystem=ICD-10
```

- ◇ 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。