「読売新聞記事データ集」tag/データ構成

1. ファイルについて

・文字コードはSJISです。
・DOS用の改行コード(CR, LF)ではなく、UNIX用の改行コード(LFのみ)を使用しております。

2. データ形式

・形式は“＼tag＼データ”
・tag/データともにすべて全角ですが、一部、半角文字が含まれている場合があります。
・ファイルは月単位に分かれています。

3. tag/データ構成

データ項目	TAG	データ内容/備考
ID番号	ID	00000010から10up ★ユニーク
記事ID	C0	例) 20020819TYM01AA　001 ★ユニーク
掲載年月日	YF	例) 20020819
頁	Y0	例) 01
面種	Y1	データ内容は下記参照
面名	Y2	例) 一面
版	Y3	例) 140
段	Y4	例) 07
写真・図表有無	YE	「写真」「表」「写真・表」の3パターン
本文文字数	Y5	例) 1557
記事分類コード	Y6	データ内容は下記参照
タイトル	T1	記事見出し
キーワード	KB	表記(漢字)
キーワード	AB	ヨミ(カナ)
記事本文	T2	

4. データ内容

1) 記事ID(tag.C0)
・データ内容は以下の通りです。
掲載年月日(8桁)＋紙誌名(3桁)＋頁(2桁)＋面種(2桁)＋スペース(全角4文字分)＋ページ内記事通番(3桁)

2) キーワード(tag.KB)
・記事の主題に見合った主要なキーワードの先頭に「＄」を付与しています。

3) 写真・図表有無(tag.YE)
・写真や図表を含む記事にのみtag/データを付与しています。

4) 紙誌名(tag.C0内)
・コード内容は以下の通りです。

T	東京本社	O	大阪本社
W	西部本社	C	中部支社
Y	読売		
M	朝刊	E	夕刊

東京本社の場合、以下のようなパターンがあります。
TYM　東京本社・読売新聞・朝刊
TYE　東京本社・読売新聞・夕刊
TMn　東京本社・朝刊・別刷りn部(nが2ならば朝刊別刷り2部)
TEn　東京本社・夕刊・別刷りn部
大阪本社＝TがOになります
西部本社＝TがWになります
中部支社＝TがCになります(中部は夕刊はありません)

5) 面種(tag.Y1)＊主なもの

AA	一面
AB	二面
AC	三面
AJ	２社面
AK	社会面
UA	一面(夕刊)
UB	二面(夕刊)
UG	２社面(夕刊)
UH	社会面(夕刊)

6) 版(tag.Y3)

・朝夕刊の「版建て」に対応したコードが付与されています。

010～100＝夕刊

110～200＝朝刊

例えば夕刊の2版は「020」、朝刊の13版は「130」となります。

・3桁目にはKとSが付与されることがあります。

Kは選挙などで締切時間が延長された「改造版」のことです。

Sは朝刊と夕刊が配達される地域のみ掲載される紙面(版)のことです。

7) 記事分類コード(tag.Y6)

Z01	政治	X01	社会	U02	健康
Z02	右翼・左翼	X02	市民運動	U03	衣
Z03	選挙	X03	社会保障	U04	食
Z04	行政	X04	環境	U05	住
Z05	地方自治	X05	婦人	U06	余暇
Z06	司法	X06	子供	U07	行事
Z07	警察	X07	中高年	T01	犯罪・事件
Z08	日本外交	X08	勲章・賞	T02	事故
Z09	軍事	X09	労働	T03	災害
Z10	戦争	X10	教育	S01	科学
Y01	経済	W01	スポーツ	S02	宇宙
Y02	財政	W02	巨人軍	S03	地球
Y03	金融	V01	文化	S04	理工学
Y04	企業	V02	学術	S05	生命工学
Y05	中小企業	V03	美術	S06	動植物
Y06	技術	V04	映像	R01	国際
Y07	情報	V05	文学	R02	アジア　太平洋
Y08	サービス	V06	音楽	R03	南北アメリカ
Y09	貿易	V07	演劇	R04	西欧
Y10	国土・都市計画	V08	芸能	R05	旧ソ連・東欧
Y11	鉱工業	V09	舞踊	R06	中東
Y12	資源・エネルギー	V10	宗教	R07	アフリカ
Y13	農林水産	U01	生活	Q01	皇室

以上