第 7 章　書誌情報データベース

丸山 岳彦　中村 壮範

7.1　均衡コーパスにおける書誌情報の役割

一般に、均衡コーパスとは、さまざまなメディアやジャンルから抽出されたサンプルの集合体と見なすことができる。ある均衡コーパスがどのようなメディアやジャンルのサンプルを含むかは、そのコーパスがどのような設計に基づいているかに依存するが、どのような設計であっても、そこに含まれている各サンプルの出自が明示されていることが望ましい。均衡コーパスを検索した結果を分析したり解釈したりする際、その結果が幅広いメディアを通して一般的に観察される現象なのか、あるいは（例えば）「雑誌」に特有な現象なのか、といった違いを捉えるためには、各サンプルの出自を表す「書誌情報」が必要不可欠である。

BCCWJ の構築過程においては、サンプリングの作業と並行して、各サンプルの出自を示す「書誌情報データベース」を整備してきた。BCCWJ の利用者は、この書誌情報データベースを参照することにより、BCCWJ を構成するすべてのサンプルの出自と属性を知ることができる。厳密な手順で取得された大量のサンプルを、その書誌情報と関連づけて利用することにより、コーパスの分析結果が現代日本語書き言葉のどの位相に位置づけられるものであるかを明確にすることができるわけである。このような利点は、例えば Web をコーパスとして用いる方法論では得ることのできないものであり、均衡コーパスとしての BCCWJ が持つ意義を最大限に特徴づけるものであると言える。

7.2　書誌情報データベースの構成

BCCWJ-DVD 版で提供される書誌情報データベースは、以下のデータ群から構成される。

- 書誌情報データ（Bibliography.txt）： サンプルを取得した原本に関する情報。
- サンプル情報データ（Sample.txt）： サンプルのID や取得状況に関する情報。
- 人名録データ（Directory.txt）： サンプルの著者や著作権者などに関する情報。
- 記事情報データ（Article.txt）：記事に含まれる文章の著者および初出に関する情報。

以下、各データの構成について概略を示す。詳細は、以下の文献を参照。

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011）『『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用』, 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書（JC-D-10-01）, 特定領域研究「日本語コーパス」データ班.

7.3 「書誌情報データ」（Bibliography.txt）

7.3.1 「書誌情報データ」の概要

書誌情報データ（Bibliography.txt）では、サンプルが抽出された出典元（原本）に関する書誌情報が、表 7-1 に示す 15 列によって表現されている。

表 7-1: 「書誌情報データ」の構成

1	**書誌** ID（Bib_ID）	サンプルを抽出した原本に対して付された ID
2	**タイトル**（Title）	原本のタイトル
3	**副題**（Subtitle）	原本の副題（サブタイトル）
4	**巻号**（Number）	原本の巻号
5	**責任表示**（Bib_author）	原本の責任表示（著者、編者、監修者など）
6	**出版者**（Publisher）	原本の出版者（出版社）
7	**出版年**（Year）	原本の出版年
8	ISBN（ISBN）	原本に付された ISBN（国際標準図書番号）
9	**判型**（Size）	原本のサイズ
10	**ページ数**（Pages）	原本のページ数
11	**ジャンル**(1)（Genre_1）	原本のジャンルに関する情報(1)
12	**ジャンル**(2)（Genre_2）	原本のジャンルに関する情報(2)
13	**ジャンル**(3)（Genre_3）	原本のジャンルに関する情報(3)
14	**ジャンル**(4)（Genre_4）	原本のジャンルに関する情報(4)
15	**責任表示** ID（Bib_author_ID）	原本の責任表示に対応する ID

書誌情報データの例を、表 7-2 に示す。実際には 15 列のタブ区切りテキストだが、ここでは折り返して表示している。「−」が表示されている列は、そのレジスターには情報が付与されないことを示す。

表 7-2: 「書誌情報データ」の例

Register	Bib_ID	Title	Subtitle	Number	Bib_author	Publisher
書籍	BK_20002488	龍臥亭事件	長編推理小説	上	島田荘司|著	光文社
雑誌	PM_00020404	ASAHI パソコン	-	2004 年 2 月 15 日号（通巻 353 号）	-	朝日新聞社
新聞	PN_01030302	朝日新聞	朝刊	2003/3/2	-	朝日新聞社
白書	WR_00000003	わが外交の近況	昭和 51 年版（上）	-	外務省	大蔵省印刷局
教科書	TB_01000009	国語　五上　銀河	-	-	宮地裕|ほか著	光村図書出版
広報紙	PR_14212017	広報あつぎ	-	2008 年 17 号	-	神奈川県厚木市
Yahoo! 知恵袋	YC_00297502	Yahoo!知恵袋	-	-	-	Yahoo!
Yahoo! ブログ	YB_00002691	Yahoo!ブログ	-	-	-	Yahoo!
韻文	VE_00010060	増補版現代短歌全集	紫木蓮まで・風舌	第 17 巻(昭和 55 年～昭和 63 年)	阿木津英|著	筑摩書房
法律	LA_S63HO108	消費税法	-	昭和六十三年十二月三十日法律第百八号	-	-
国会会議録	MD_02010001	国会会議録	-	第 154 回国会	-	-

（表 7-2 続き）

Register	Year	ISBN	Size	Pages	Genre_1	Genre_2	Genre_3	Genre_4	Bib_author_ID
書籍	1999	4334728898	16cm	577	9 文学	913	0193	-	00122924
雑誌	2003	-	A4 変型判	128	工業	電気機／電子	コンピュータ／情報処理	月２回刊	-
新聞	2003	-	ブランケット判	37	全国紙	-	-	-	-
白書	1976	-	-	-	外交	-	-	-	-
教科書	2006	-	-	-	国語	小	5	-	0045734
広報紙	2008	-	-	-	関東地方	神奈川県	-	-	-
Yahoo! 知恵袋	2005	-	-	-	子育てと学校	子育て, 出産	子育ての悩み	-	-
Yahoo! ブログ	2008	-	-	-	家庭と住まい	住まい	ガーデニング	-	-
韻文	2002	4480138374	23cm	500	短歌	-	-	-	00110019
法律	1988	-	-	-	23_国税	-	-	-	-
国会会議録	2002	-	-	-	衆議院	常任委員会	環境委員会	-	-

7.3.2 書誌 ID

書誌 ID（Bib_ID）列は、サンプルを取得した原本に対して一意に付された ID を表す。

「書籍」レジスターの書誌 ID

例：BK_20208020　→『うたかたの月』

1・2 桁目 「BK」「書籍（Book)」であることを表す。

3 桁目 「_」区切り記号。

4～11 桁目 原本に付された一意の ID。

「雑誌」レジスターの書誌 ID

例：PM_00030103　→『アサヒカメラ』、2001 年 3 号

1・2 桁目 「PM」「雑誌（Magazine)」であることを表す。

3 桁目 「_」区切り記号。

4～7 桁目 同一タイトルの雑誌に付された一意の ID。

※ 「0003」は『アサヒカメラ』に付与されたID。

8～9 桁目 発行年（2001 年から 2005 年の下 2 桁、01～05)。

10～11 桁目 その発行年における号数。

「新聞」レジスターの書誌 ID

例： PN_01010202　→ 朝日新聞・朝刊（0101)、2 月 2 日発行

1・2 桁目 「PN」「新聞（Newspaper)」であることを表す。

3 桁目 「_」区切り記号。

4～5 桁目 新聞タイトル・朝夕刊の別を表す ID。（7.A.3 を参照）

6～7 桁目 発行年（2001 年から 2005 年の下 2 桁、01～05)。

8～11 桁目 発行日（1 月 1 日から 12 月 31 日、0101～1231)。

「白書」レジスターの書誌 ID

例： WR_00000001　→『エネルギー白書』2004 年版

1・2 桁目 「WR」「白書」であることを表す。

3 桁目 「_」区切り記号。

4～11 桁目 原本に付された一意の ID。

「教科書」レジスターの書誌 ID

例： TB_01000001　→『こくご　一上　かざぐるま』

1・2 桁目 「TB」「教科書（TextBook)」であることを表す。

3 桁目 「_」区切り記号。

4 桁目 教科。

「0」= 国語　「3」= 社会　「6」= 芸術　「9」= 生活

「1」= 数学　「4」= 外国語　「7」= 保健体育

「2」= 理科　「5」= 技術家庭　「8」= 情報

5 桁目 学校。

「1」= 小学校 「2」= 中学校 「3」= 高校

6〜11 桁目 教科・学校ごとに分類された教科書の通し番号。

「広報紙」レジスターの書誌 ID

例： PR_14212017 → 『広報あつぎ』2008 年 17 号

1・2 桁目 「PR」「広報紙（Public Relations)」であることを表す。

3 桁目 「_」区切り記号。

4〜8 桁目 自治体に付された一意の ID。（7.A.6 を参照）

9〜11 桁目 その自治体における号数。

「Yahoo!知恵袋」レジスターの書誌 ID

例：YC_00297787 → 小カテゴリ「政治、社会問題」

1・2 桁目 「YC」「Yahoo!知恵袋（Yahoo! Chiebukuro)」であることを表す。

3 桁目 「_」区切り記号。

4〜11 桁目 小カテゴリごとに付された一意の ID。（7.A.7 を参照）

「Yahoo!ブログ」レジスターの書誌 ID

例：YB_00000075 → 小カテゴリ「インテリア」

1・2 桁目 「YB」「Yahoo!ブログ（Yahoo! Blog)」であることを表す。

3 桁目 「_」区切り記号。

4〜11 桁目 小カテゴリごとに付された一意の ID。（7.A.8 を参照）

「韻文」レジスターの書誌 ID

例：VE_89028672 → 『稲垣足穂詩集』

1・2 桁目 「VE」「韻文（Verse)」であることを表す。

3 桁目 「_」区切り記号。

4〜11 桁目 原本に付された一意の ID。

「法律」レジスターの書誌 ID

例：LA_S54HO004 →「民事執行法」（昭和五十四年三月三十日法律第四号）

1・2 桁目 「LA」「法律（Law)」であることを表す。

3 桁目 「_」区切り記号。

4〜6 桁目 法律の公布年（S54 → 昭和 54 年)。

7〜8 桁目 「法律（HO)」であることを表す。

4〜11 桁目 その年における法令番号。

「国会会議録」レジスターの書誌 ID

例：MD_79050005 → 国会会議録（1979 年第 91 回国会、参議院、常任委員会、外務委員会）

1・2 桁目 「MD」「国会会議録（Minutes of the Diet)」であることを表す。

3 桁目 「_」区切り記号。

4～5 桁目 会議の開催年。

6～7 桁目 開催院および会議種別。

「01」= 衆議院・常任委員会　「05」= 参議院・常任委員会

「02」= 衆議院・特別委員会　「06」= 参議院・特別委員会

「03」= 衆議院・本会議　「07」= 参議院・本会議

「04」= 衆議院・その他　「08」= 参議院・その他

8～11 桁目 会議種別ごとの会議に付された一意の ID。

7.3.3 タイトル

タイトル（Title）列は、原本のタイトルを表す。

例 「ファンの心をときめかせた世界の映画ベストセレクション」（書籍）

「塩狩峠; 道ありき」（書籍）

「週刊朝日」（雑誌）

「北海道新聞」（新聞）

「情報通信白書」（白書）

「こくご　一上　かざぐるま」（教科書）

「広報あげお」（広報紙）

「Yahoo!知恵袋」（Yahoo!知恵袋）

「Yahoo!ブログ」（Yahoo!ブログ）

「谷川俊太郎詩集」（韻文）

「民事保全法」（法律）

「国会会議録」（国会会議録）

7.3.4 副題

副題（Subtitle）列は、原本の副題・サブタイトルを表す。

例 「伝説の呼び屋・永島達司の生涯」（書籍）

「朝刊」（新聞）

「平成 4 年版」（白書）

「サラダ記念日」（韻文）

7.3.5 巻号

巻号（Number）列は、原本の巻号・巻次に関する情報を表す。

例 「第 6 巻」（書籍）

「3（神の星編）」（書籍）

「2002 年 4 月 15 日号（第 15 巻第 16 号、通巻 750 号）」（雑誌）

「サンデー毎日臨時増刊（第 80 巻第 49 号、通巻 4467 号）」（雑誌）

「2001/10/24」（新聞）
「2008 年 12 号」（広報紙）
「第 17 巻（55 年～昭和 63 年）」（韻文）
「平成元年六月二十八日法律第五十八号」（法律）
「第 154 回国会」（国会会議録）

7.3.6 責任表示

責任表示（Bib_author）列は、原本の責任表示（著者、編者、監修者など）の情報を表す。

例　「司馬遼太郎 | 著」（書籍）
「七田眞、七田厚 | 著」（書籍）
「高橋貞巳 | 監修；三菱総合研究所 | 著」（書籍）
「カフカ | 著；池内紀 | 訳」（書籍）
「ロナルド・A.　モース | 編著；日下公人 | 監修；時事通信社外信部 | ほか訳」（書籍）
「経済産業省；厚生労働省；文部科学省」（白書）
「宮地裕 | ほか著」（教科書）

7.3.7 出版者

出版者（Publisher）列は、原本の出版者（出版社）を表す。

例　「岩波書店」（書籍）
「日本図書刊行会；近代文芸社（発売）」（書籍）
「マガジンハウス」（雑誌）
「株式会社朝日新聞社」（新聞）
「大蔵省印刷局」（白書）
「光村図書出版株式会社」（教科書）
「北海道札幌市東区」（広報紙）
「Yahoo!」（Yahoo!知恵袋、Yahoo!ブログ）
「筑摩書房」（韻文）

7.3.8 出版年

出版年（Year）列は、4 桁の数字で、原本が出版された年を表す。

※　「Yahoo!知恵袋」と「Yahoo!ブログ」の場合、それぞれ「2005」「2008」の一通りとなる。質問や記事が実際に投稿された日時は、サンプル情報データ（Sample.txt）の「タイムスタンプ（Timestamp）」列を参照のこと。

※　「法律」の場合、法律が公布された年を表す。

※　「国会会議録」の場合、会議が開催された年を表す。

7.3.9 ISBN

ISBN（ISBN）列は、原本に付された ISBN（国際標準図書番号）を表す（10 桁）。

7.3.10 判型

判型（Size）列は、原本の大きさを表す。

7.3.11 ページ数

ページ数（Pages）列は、原本の総ページ数を表す。

7.3.12 ジャンル(1)～(4)

ジャンル(1)～(4)（Genre_1～Genre_4）列は、原本のジャンルに関連した情報を表す。レジスターごとに取るジャンル情報の種類を、表 7-3 に示す。

表 7-3: ジャンル情報の種類

レジスター	ジャンル(1)	ジャンル(2)	ジャンル(3)	ジャンル(4)
書籍	NDC(1 桁)＋分類名	NDC(3 桁)	C コード	
雑誌	大ジャンル名	中ジャンル名	小ジャンル名	刊行形態
新聞	配達エリア			
白書	ジャンル名			
教科書	教科名	学校種	学年	
広報紙	地域	都道府県名		
Yahoo!**知恵袋**	大カテゴリ名	中カテゴリ名	小カテゴリ名	
Yahoo!**ブログ**	大カテゴリ名	中カテゴリ名	小カテゴリ名	
韻文	韻文種別			
法律	ジャンル名			
国会会議録	開催院	会議種別	委員会名	

※ ジャンル情報の詳細については、付録 7-A を参照。

7.3.13 責任表示 ID

責任表示 ID（Bib_author_ID）列は、責任表示（Bib_author）列に記載されている人名・組織名などに対して付された ID を表す。記載されている ID は、人名録データ(Directory.txt)の「人名 ID （Directory_ID)」列に記載された ID に対応している。

例 「00685074」（書籍）
「00254659；00184422」（書籍）
「00113880；00166885；00124738」（教科書）
「00037561」（韻文）

7.4 「サンプル情報データ」（Sample.txt）

7.4.1 「サンプル情報データ」の概要

サンプル情報データ（Sample.txt）では、BCCWJ に収録された各サンプルの ID や抽出状況に関する情報が、表 7-4 に示す 5 列によって表現されている。

表 7-4: 「サンプル情報データ」の構成

1	**サンプル ID**（Sample_ID）サンプルに対して一意に付された ID
2	**書誌 ID**（Bib_ID）サンプルを抽出した原本に対して付された ID
3	**サンプル抽出基準点ページ**（Sampling_page）サンプル抽出基準点を取得したページ
4	**サンプル抽出基準点座標**（Sampling_point）サンプル抽出基準点を取得した交点
5	**投稿日時**（Timestamp）Yahoo!知恵袋の質問、Yahoo!ブログの記事の投稿日時

サンプル情報データの例を、表 7-5 に示す。

表 7-5: 「サンプル情報データ」の例

レジスター	Sample_ID	Bib_ID	Sampling_page	Sampling_point	Timestamp
出版・書籍	PB10_00047	BK_20205918	163	5D	-
雑誌	PM11_00053	PM_10550109	76	9F	-
新聞	PN1a_00013	PN_01010225	4	6C	-
図書館・書籍	LBa1_00004	BK_86049602	230	2H	-
白書	OW6X_00009	WR_00000066	285	4C	-
教科書	OT01_00008	TB_01000002	31	8A	-
広報紙	OP00_00001	PR_01103001	-	-	-
ベストセラー	OB0X_00001	BK_75079014	358	4D	-
Yahoo!知恵袋	OC01_00001	YC_00297514	-	-	2004/4/29 18:35
Yahoo!ブログ	OY01_00005	YB_00010571	-	-	2008/6/24 21:39
韻文	OV0X_00001	VE_00010001	-	-	-
法律	OL3X_00072	LA_H01HO058	-	-	-
国会会議録	OM11_00001	MD_80010001	-	-	-

7.4.2 サンプル ID

サンプル ID（Sample_ID）列は、各サンプルに対して一意に付された ID を表す。

- 出版サブコーパス「書籍」レジスターのサンプル ID

  例： PB10_00001

  1 桁目 「P」出版サブコーパス（Publication）に所属することを表す。

  2 桁目 「B」書籍（Book）のサンプルであることを表す。

  3 桁目 「1～5」出版年を表す。

  「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年

  「2」＝ 2002 年　「4」＝ 2004 年

  4 桁目 「0～9,n」当該書籍に付された NDC（日本十進分類法）の第 1 次区分を表す。

  「0」＝総記　　「4」＝自然科学　　「8」＝言語

  「1」＝哲学　　「5」＝技術・工学　「9」＝文学

  「2」＝歴史　　「6」＝産業　　「n」＝分類なし

  「3」＝社会科学　「7」＝芸術・美術

  5 桁目 「_」区切り記号。

  6～10 桁目 各出版年・各 NDC におけるサンプルの取得順位を表す。

- 出版サブコーパス「雑誌」レジスターのサンプル ID

  例： PM11_00002

  1 桁目 「P」出版サブコーパス（Publication）に所属することを表す。

  2 桁目 「M」雑誌（Magazine）のサンプルであることを表す。

  3 桁目 「1～5」出版年を表す。

  「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年

  「2」＝ 2002 年　「4」＝ 2004 年

  4 桁目 「1～6」当該雑誌に付されたジャンルを表す。

  「1」＝総合　　「4」＝産業

  「2」＝教育・学芸　　「5」＝工業

  「3」＝政治・経済・商業　「6」＝厚生・医療

  5 桁目 「_」区切り記号。

  6～10 桁目 各雑誌タイトル・各出版年におけるサンプルの取得順位を表す。

- 出版サブコーパス「新聞」レジスターのサンプル ID

  例： PN1a_00001

  1 桁目 「P」出版サブコーパス（Publication）に所属することを表す。

  2 桁目 「N」新聞（Newspaper）のサンプルであることを表す。

  3 桁目 「1～5」出版年を表す。

  「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年

「2」＝ 2002 年　「4」＝ 2004 年

4 桁目 「a～o」新聞タイトルを表す。

「a」＝朝日新聞　「f」＝中日新聞　「k」＝神戸新聞
「b」＝毎日新聞　「g」＝西日本新聞　「l」＝中国新聞
「c」＝読売新聞　「h」＝河北新報　「m」＝高知新聞
「d」＝産経新聞　「i」＝新潟日報　「o」＝琉球新報
「e」＝北海道新聞　「j」＝京都新聞

5 桁目 「_」区切り記号。

6～10 桁目 各新聞タイトル・各出版年におけるサンプルの取得順位を表す。

- 図書館サブコーパス「書籍」レジスターのサンプル ID

例： LBa0_00002

1 桁目 「L」図書館サブコーパス（Library）に所属することを表す。

2 桁目 「B」書籍（Book）のサンプルであることを表す。

3 桁目 「a～t」出版年を表す。

「a」＝ 1986 年　「h」＝ 1993 年　「o」＝ 2000 年
「b」＝ 1987 年　「i」＝ 1994 年　「p」＝ 2001 年
「c」＝ 1988 年　「j」＝ 1995 年　「q」＝ 2002 年
「d」＝ 1989 年　「k」＝ 1996 年　「r」＝ 2003 年
「e」＝ 1990 年　「l」＝ 1997 年　「s」＝ 2004 年
「f」＝ 1991 年　「m」＝ 1998 年　「t」＝ 2005 年
「g」＝ 1992 年　「n」＝ 1999 年

4 桁目 「0～9,n」当該書籍に付された NDC（日本十進分類法）の第 1 次区分を表す。

「0」＝総記　「4」＝自然科学　「8」＝言語
「1」＝哲学　「5」＝技術・工学　「9」＝文学
「2」＝歴史　「6」＝産業　「n」＝分類なし
「3」＝社会科学　「7」＝芸術・美術

5 桁目 「_」区切り記号。

6～10 桁目 各出版年・各 NDC におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「白書」レジスターのサンプル ID

例：OW1X_00000

1 桁目 「O」特定目的サブコーパスに所属することを表す。

2 桁目 「W」白書のサンプルであることを表す。

3 桁目 「1～6」出版時期を表す。

「1」＝第 1 期（1976～1980 年）　「4」＝第 4 期（1991～1995 年）
「2」＝第 2 期（1981～1985 年）　「5」＝第 5 期（1996～2000 年）
「3」＝第 3 期（1986～1990 年）　「6」＝第 6 期（2001～2005 年）

4 桁目 「X」ダミー記号。

5 桁目 「_」区切り記号。

6～10 桁目 各出版時期におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「教科書」レジスターのサンプル ID

  例：OT01_00002

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「T」教科書（TextBook）のサンプルであることを表す。

  3 桁目 「0～9」教科を表す。

  「0」=国語 「3」=社会 「6」=芸術 「9」=生活

  「1」=数学 「4」=外国語 「7」=保健体育

  「2」=理科 「5」=技術家庭 「8」=情報

  4 桁目 「1～3」学校を表す。

  「1」= 小学校 「2」= 中学校 「3」= 高校

  5 桁目 「_」区切り記号。

  6～10 桁目 各教科・学校におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「広報紙」レジスターのサンプル ID

  例： OP00_00001

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「P」広報紙（Public Relations）のサンプルであることを表す。

  3・4 桁目 「00～99」対象となった 100 自治体の通し番号を表す。

  5 桁目 「_」区切り記号。

  6～10 桁目 各自治体から取得したサンプルの取得順位を表す。

- 特定目的サブコーパス「ベストセラー」レジスターのサンプル ID

  例： OB0X_00001

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「B」ベストセラー（Best-seller）のサンプルであることを表す。

  3 桁目 「0～6」出版時期を表す。

  「0」＝第 0 期（1975 年以前） 「4」＝第 4 期（1991～1995 年）

  「1」＝第 1 期（1976～1980 年）「5」＝第 5 期（1996～2000 年）

  「2」＝第 2 期（1981～1985 年）「6」＝第 6 期（2001～2005 年）

  「3」＝第 3 期（1986～1990 年）

  4 桁目 「X」ダミー記号。

  5 桁目 「_」区切り記号。

  6～10 桁目 各出版時期におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「Yahoo!知恵袋」レジスターのサンプル ID

  例： OC01_00001

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「C」Yahoo!知恵袋（Chiebukuro）のサンプルであることを表す。

  3・4 桁目 「01～15」質問が投稿された大カテゴリ ID を表す。

  「01」＝ 「エンターテインメントと趣味」

  「02」＝ 「インターネット、PC と家電」

  「03」＝ 「ビジネス、経済とお金」

  「04」＝ 「職業とキャリア」

  「05」＝ 「ニュース、政治、国際情勢」

  「06」＝ 「スポーツ、アウトドア、車」

  「08」＝ 「暮らしと生活ガイド」

  「09」＝ 「健康、美容とファッション」

  「10」＝ 「子育てと学校」

  「11」＝ 「マナー、冠婚葬祭」

  「12」＝ 「教養と学問、サイエンス」

  「13」＝ 「地域、旅行、お出かけ」

  「14」＝ 「Yahoo! JAPAN」

  「15」＝ 「その他」

  5 桁目 「_」区切り記号。

  6～10 桁目 各大カテゴリにおけるサンプルの取得順位を表す。

- 特定目的サブコーパス「Yahoo!ブログ」レジスターのサンプル ID

  例： OY01_00005

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「Y」Yahoo!ブログ（Blog）のサンプルであることを表す。

  3・4 桁目 「01～15」記事が投稿された大カテゴリ ID を表す。

  「01」＝ 「ビジネスと経済」

  「02」＝ 「コンピュータとインターネット」

  「03」＝ 「生活と文化」

  「04」＝ 「エンターテインメント」

  「05」＝ 「家庭と住まい」

  「06」＝ 「政治」

  「07」＝ 「健康と医学」

  「08」＝ 「学校と教育」

  「09」＝ 「科学」

  「10」＝ 「出会い」

「11」＝「地域」
「12」＝「特集」
「13」＝「芸術と人文」
「14」＝「Yahoo!サービス」
「15」＝「趣味とスポーツ」

5 桁目 「_」区切り記号。

6～10 桁目 各大カテゴリにおけるサンプルの取得順位を表す。

- 特定目的サブコーパス「韻文」レジスターのサンプル ID

例：OV0X_00001

1 桁目 「O」特定目的サブコーパスに所属することを表す。

2 桁目 「V」韻文（Verse）のサンプルであることを表す。

3 桁目 「0～2」韻文の種類を表す。

「0」＝短歌 「1」＝俳句 「2」＝詩

4 桁目 「X」ダミー記号。

5 桁目 「_」区切り記号。

6～10 桁目 サンプルの取得順位を表す。

- 特定目的サブコーパス「法律」レジスターのサンプル ID

例： OL1X_00001

1 桁目 「O」特定目的サブコーパスに所属することを表す。

2 桁目 「L」法律（Law）のサンプルであることを表す。

3 桁目 「1～6」法律の公布時期を表す。

「1」＝第 1 期（1976～1980 年） 「2」＝第 2 期（1981～1985 年）
「3」＝第 3 期（1986～1990 年） 「4」＝第 4 期（1991～1995 年）
「5」＝第 5 期（1996～2000 年） 「6」＝第 6 期（2001～2005 年）

4 桁目 「X」ダミー記号。

5 桁目 「_」区切り記号。

6～10 桁目 各公布時期におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「国会会議録」レジスターのサンプル ID

例： OM11_00001

1 桁目 「O」特定目的サブコーパスに所属することを表す。

2 桁目 「M」国会会議録（Minutes of the Diet）のサンプルであることを表す。

3 桁目 「1～6」会議の開催時期を表す。

「1」＝第 1 期（1976～1980 年） 「4」＝第 4 期（1991～1995 年）
「2」＝第 2 期（1981～1985 年） 「5」＝第 5 期（1996～2000 年）
「3」＝第 3 期（1986～1990 年） 「6」＝第 6 期（2001～2005 年）

4 桁目 「1～8」会議の開催院・会議種別を表す。

「1」＝衆議院・常任委員会　「5」＝参議院・常任委員会
「2」＝衆議院・特別委員会　「6」＝参議院・特別委員会
「3」＝衆議院・本会議　「7」＝参議院・本会議
「4」＝衆議院・その他　「8」＝参議院・その他

5 桁目 「_」区切り記号。

6～10 桁目 各開催時期、開催院・会議種別におけるサンプルの取得順位を表す。

7.4.3 書誌 ID

書誌 ID（Bib_ID）列は、サンプルを取得した原本に対して一意に付された ID を表す。記載されている ID は、書誌情報データ（Bibliography.txt）の「書誌 ID（Bib_ID）」列に記載された ID に対応している（7.3.2 節参照）。

7.4.4 サンプル抽出基準点ページ

サンプル抽出基準点ページ（Sampling_page）列は、「サンプル抽出基準点」（7.4.5 参照）を含むページ番号を表す。

7.4.5 サンプル抽出基準点座標

サンプル抽出基準点座標（Sampling_point）列は、「サンプル抽出基準点」を同定する際、サンプル抽出基準点ページ内でランダムに指定されたある 1 点（交点）の座標を表す。

※ 横軸に0～9、縦軸にA～Jという目盛りを配置した10×10のマス目を準備し、それを印刷した透明なシートを実際のページに当て、ランダムに指定された交点（「3E」など）に最も近接している文字を「サンプル抽出基準点」として指定した。このサンプル抽出基準点をもとに、サンプルを取得した。

7.4.6 投稿日時

投稿日時（Timestamp）列は、「Yahoo!知恵袋」の質問、および「Yahoo!ブログ」の記事が投稿された日時を表す。

7.5 「人名録データ」（Directory.txt）

7.5.1 「人名録データ」の概要

人名録データ（Directory.txt）では、書誌データ（Bibliography.txt）の「責任表示（Bib_author)」列に記載されている人名や組織名（著者、編者、監修者など）や、各サンプルに含まれる記事を実際に執筆した著者名などの情報が、表 7-6 に示す 4 列によって表現されている。

表 7-6: 「人名録データ」の構成

1	**人名 ID**（Directory_ID）	人物や組織に対して一意に付された ID
2	**人名**（Name）	人物の氏名、または組織名
3	**性別**（Sex）	性別
4	**生年代**（BirthYear）	生年（10 年単位）

人名録データの例を、表 7-7 に示す。

表 7-7: 「人名録データ」の例

Directory_ID	Name	Sex	BirthYear
634	会田　雄次	男	1910
98948	アントニオ猪木	男	1940
153494	群　ようこ	女	1950
840303	厚生労働省労働基準局		
258003	講談社		
2502212	NHK「プロジェクト X」制作班		

7.5.2 人名 ID

人名 ID（Directory_ID）列は、人物の氏名または組織名に対して付された一意の ID を表す。

7.5.3 人名

人名（Name）列は、人物の氏名や組織名などを表す。

7.5.4 性別

性別（Sex）列は、人物の性別を表す。なお、組織の場合には記載していない。

7.5.5 生年代

生年代（BirthYear）列は、人物の生年を西暦の 10 年単位でまとめた年を表す。なお、性別・生年代については、原則として本人からの回答を記載しているが、国立国会図書館の典拠データなどから情報を補足しているものもある。また、組織の場合には記載していない。

7.6 記事情報データ（Article.txt）

7.6.1 「記事情報データ」の概要

記事情報データ（Article.txt）では、各サンプルに含まれる「記事」を対象として、「実著者」および「初出」に関する情報が、表 7-8 に示す 6 列によって表現されている。

表 7-8: 「記事情報データ」の構成

1	**サンプル** ID（Sample_ID）	各サンプルに対して一意に付された ID
2	**記事** ID（Article_ID）	各記事に対して一意に付された ID
3	**人名** ID（Directory_ID）	各記事を実際に執筆した著者に対して一意に付された ID
4	**役割**（Role）	著者の役割（実著者、原著者、翻訳者の別）
5	**初出情報**（First_appearance）	各記事の初出に関する情報
6	**初刊情報**（First_published）	各記事の初刊に関する情報

記事情報データの例を、表 7-9 に示す。

表 7-9: 「記事情報データ」の例

Sample_ID	Article_ID	Directory_ID	Role	First_ appearance	First_ published
LBa0_00002	LBa0_00002_V001	59986	実著者	1984	1986
LBq1_00026	LBq1_00026_F003	262756	実著者	2000-2001	2002
LBa1_00006	LBa1_00006_V001	459606	原著者	1986	
LBa1_00006	LBa1_00006_V001	108831	翻訳者	1986	
PB12_00059	PB12_00059_V001	189710	実著者	n.d.-n.d.	2001
PM11_00289	PM11_00289_F002	0	実著者	2001	
PN1a_00004	PN1a_00004_V003	256908	実著者	2001	

なお、記事情報データは、「書籍」「雑誌」「新聞」に対してのみ提供される。

7.6.2 サンプル ID

サンプル ID（Sample_ID）列は、サンプルに対して一意に付された ID を表す。記載されている ID は、サンプル情報データ（Sample.txt）の「サンプル ID（Sample_ID）」列に記載された ID に対応している。7.4.2 節を参照。

7.6.3 記事 ID

記事 ID（Article_ID）列は、サンプルに含まれる「記事」に対して一意に付された ID を表す。

例　PB15_00023_F001

LB29_00129_V001
PM11_00118_F002
PN1a_00013_F004

「記事」とは、「同一著者によって、同一のテーマについてまとまりをもって書かれた文章の範囲」を指す。ひとつのサンプル（可変長サンプル、固定長サンプルとも）は、ひとつの記事によって構成されている場合もあれば、複数の記事によって構成されている場合もある。記事 ID は、それが所属するサンプル ID の直後に「V001」「F002」などを続けて表される。「V001」は、そのサンプルに含まれる可変長（Variable_Length）サンプルの 1 番目の記事、「F002」は、そのサンプルに含まれる固定長（Fixed_Length）サンプルの 2 番目の記事であることを、それぞれ表す。

7.6.4 人名 ID

記事情報データにおける人名 ID（Directory_ID）列は、記事を実際に執筆した人物（著者）に対して付された ID を表す。記載されている ID は、人名録データ（Directory.txt）の「人名 ID（Directory_ID）」列に記載された ID に対応している。

各記事を実際に執筆した人物を「実著者」と判定し、その人名および人名 ID を記録した。翻訳書については、実著者に替えて、「原著者」と「翻訳者」の人名と人名 ID の組を記録した。実著者、原著者、翻訳者は、各サンプルの印刷紙面や、原本の目次、奥付に表示されている人名、著作権処理の過程で判明した実著者の情報などをもとに判定した。新聞については、実著者の記名が新聞記者と思われる場合は、その人名に替えて「朝日新聞社」などの新聞社名を記録した。なお、当該の文章を執筆した人名が確定できない場合は「実著者不明」として「0」という人名 ID を与えた。

7.6.5 役割

役割（Role）列は、「人名 ID（Directory_ID）」列に記載された ID に対応する人物の役割を表す。

例　「実著者」
　　「原著者」
　　「翻訳者」

7.6.6 初出情報

初出情報（First_appearance）列は、当該の記事に含まれる文章が、雑誌や新聞などで初めて発表・出版された年を表す。

7.6.7 初刊情報

初刊情報（First_published）列は、当該の記事に含まれる文章が、初めて書籍として刊行された年を表す。

なお、「初出情報」「初刊情報」は、次のような問題意識および方法によって、情報を取得した。ある書籍に含まれる文章は、その書籍の発行時において初めて世に発表されたものと、そうでないものとに分かれる。このうち前者は、一般的には「書き下ろし」と呼ばれる。一方、後者には、雑誌や新聞に連載されていた小説が単行本として出版される場合、単行本が改版して出版される場合、単行本が文庫として出版される場合などがある。中には、100 年以上前に出版された本が 2005 年に文庫として出版されている例もある。

そこで、取得した文章がそれ以前に出版されたことがあるかどうかについて、可能な限り調査した。原本の奥付や目次の周辺、後書きなどを確認し、初出・初刊に関する情報を取得した。同時に、『文芸雑誌小説初出総覧』（日外アソシエーツ）、『日本近代文学大事典』（講談社）も参照した。さらに、一部については、国立国会図書館（NDL-OPAC）を使って調査を行なった。情報が取得できた場合、その年を初出情報・初刊情報として記録した。なお、初出が確認できなかった場合は、初出情報は空欄とした。また、初刊が確認できなかった場合は、出版年を初刊情報として記録した。

なお、当該の書籍が一連のシリーズとして刊行された場合や、複数年にわたる連載記事として出版されていた場合は、「1965-1971」のように、年号をハイフンでつないで表示した。また、雑誌や新聞などに連載された原稿が書籍になった旨が明記してあるものの、その年が確定できない場合には、「n.d.（no date）」と記録した。

上記で述べた書籍の場合と同様に、雑誌・新聞についても初出情報を調査した。初出情報が確認できなかった場合は、出版年を初出の年とした。また、書籍として刊行された年を表す初刊情報は、雑誌・新聞には付与されていない。

付録 7-A: 書誌情報データ「ジャンル」情報の詳細

7.A.1 「書籍」のジャンル情報の詳細

- ジャンル(1)

「書籍」の「ジャンル(1)」列には、国立国会図書館で付与された「NDC（日本十進分類法）第 9 版」の第 1 次区分（類目）を表す数値と、その分類名が記載されている。

例　「0 総記」、「1 哲学」、「2 歴史」、「3 社会科学」、「4 自然科学」、
「5 技術・工学」、「6 産業」、「7 芸術・美術」、「8 言語」、「9 文学」、
「分類なし[1]」

- ジャンル(2)

「書籍」の「ジャンル(2)」列には、国立国会図書館で付与された「NDC（日本十進分類法）第 9 版」の第 3 次区分（要目）を表す数値が 3 桁で記載されている。詳細については、『日本十進分類法新訂 9 版』（日本図書館協会）などを参照。

- ジャンル(3)

「書籍」の「ジャンル(3)」列には、「C コード（図書分類コード）」が記載されている。「C コード」は日本図書コードの一部で、4 桁の数値で構成される。左から 1 桁目は「販売対象コード」で、対象読者を表す。2 桁目は「発行形態コード」で、発行形態を表す。3・4 桁目は「内容コード」で、書籍の内容を表す。詳細は、『ISBN コード／日本図書コード／書籍 JAN コード利用の手引き』（日本図書コード管理センター）などを参照。

※ 「C コード」の 1 桁目 「販売対象コード」の分類を、以下に示す。
「0」＝ 一般、「1」＝ 教養、「2」＝ 実用、「3」＝ 専門、「4」＝（欠番） 、
「5」＝ 婦人、「6」＝ 学参 I（小中） 、「7」＝ 学参 II（高校）、「8」＝ 児童、「9」＝ 雑誌扱い

※ 「C コード」の 2 桁目 「発行形態コード」の分類を、以下に示す。
「0」＝ 単行本、「1」＝ 文庫、「2」＝ 新書、「3」＝ 全集・双書、
「4」＝ ムック・その他、「5」＝ 事・辞典、「6」＝ 図鑑、「7」＝ 絵本、
「8」＝ 磁性媒体など、「9」＝ コミック

※ 「C コード」の 3・4 桁目 「内容コード」の分類を、表 7-10 に示す。

[1] 2005 年 10 月時点において国立国会図書館で NDC が付与されていなかった場合に相当する。

表 7-10: 「C コード」の 3・4 桁目 「内容コード」の分類

「00」= 総記
「01」= 百科事典
「02」= 年鑑・雑誌
「04」= 情報科学
「10」= 哲学
「11」= 心理（学）
「12」= 倫理（学）
「14」= 宗教
「15」= 仏教
「16」= キリスト教
「20」= 歴史総記
「21」= 日本歴史
「22」= 外国歴史
「23」= 伝記・系譜
「25」= 地理
「26」= 旅行
「30」= 社会科学総記
「31」= 政治（国防・軍事含む）
「32」= 法律
「33」= 経済、財政、統計
「34」= 経営
「36」= 社会
「37」= 教育
「39」= 民俗・風習
「40」= 自然科学総記
「41」= 数学
「42」= 物理学
「43」= 化学
「44」= 天文・地学
「45」= 生物学
「47」= 医学・歯学・薬学
「50」= 工学・工学総記
「51」= 土木
「52」= 建築
「53」= 機械
「54」= 電気
「55」= 電子通信
「56」= 海事・兵器
「57」= 採鉱・冶金
「58」= その他の工業
「60」= 産業総記
「61」= 農林業
「62」= 水産業
「63」= 商業
「65」= 交通・通信
「70」= 芸術総記
「71」= 絵画・彫刻
「72」= 写真・工芸
「73」= 音楽・舞踊
「74」= 演劇・映画
「75」= 体育・スポーツ
「76」= 諸芸・娯楽
「77」= 家事
「78」= 生活
「79」= コミックス・劇画
「80」= 語学総記
「81」= 日本語
「82」= 英（米）語
「84」= ドイツ語
「85」= フランス語
「87」= 各国語
「90」= 文学総記
「91」= 日本文学総記
「92」= 日本文学詩歌
「93」= 日本文学小説・物語
「95」= 日本文学評論・随筆・その他
「97」= 外国文学小説
「98」= 外国文学その他

7.A.2 「雑誌」のジャンル情報の詳細

- ジャンル(1)

「雑誌」の「ジャンル(1)」列には、表 7-11 に示す 6 種類の「大ジャンル」の情報が、雑誌タイトルごとに記載されている。

例　「1 総合」、「2 教育・学芸」、「3 政治・経済・商業」、「4 産業」、
「5 工業」、「6 厚生・医療」

- ジャンル(2)

「雑誌」の「ジャンル(2)」列には、表 7-11 に示す 27 種類の「中ジャンル」の情報が、雑誌タイトルごとに記載されている。

- ジャンル(3)

「雑誌」の「ジャンル(3)」列には、表 7-11 に示す 71 種類の「小ジャンル」の情報が、雑誌タイトルごとに記載されている。

- ジャンル(4)

「雑誌」の「ジャンル(4)」列には、雑誌タイトルの「刊行形態」が記載されている。

例　「月刊」「週刊」「月 2 回刊」「隔月刊」「隔週刊」「季刊」「年刊」
「年 2 回刊」「年 4 回刊」「年 5～6 回刊」「年 3 回刊」

なお、大ジャンル、中ジャンル、小ジャンル、および刊行形態の分類は、『雑誌新聞総かたろぐ』（メディア・リサーチ・センター）での記載に基づく。

表 7-11: 雑誌の大ジャンル・中ジャンル・小ジャンルの一覧

大ジャンル	中ジャンル	小ジャンル
総合	総記／マスコミ	総記
		マスコミ（新聞・放送）
		出版・読書・図書館
		出版情報・書評
	一般	一般週刊誌
		総合誌
		女性週刊誌
		婦人誌
		読み物
		東京都／タウン・地域誌
		関東地方／タウン・地域誌
		近畿地方／タウン・地域誌
	家庭／生活	生活情報
		ファッション
		料理・栄養
		住居・インテリア
		育児・家庭教育
	児童	少年
		少女
	娯楽／芸能	ヤング
		テレビ・ラジオ・芸能・映画
	レジャー／趣味	レジャー
		旅行・観光
		趣味の乗り物
		釣り・狩猟
		写真・カメラ
		家庭園芸
		ホビー・クラフト・日曜大工
		模型・無線・コンピュータゲーム
		音楽・オーディオ
		囲碁・将棋
		ペット

表 7-11: 雑誌の大ジャンル・中ジャンル・小ジャンルの一覧（続き）

大ジャンル	中ジャンル	小ジャンル
総合（続き）	スポーツ	スポーツ一般・陸上競技
		アウトドア・海／山
		球技
		ゴルフ
		武道・格闘技
教育・学芸	教育	教育技術
	学習／語学	小・中学生
		高校・大学生
	文学／芸術	文学文芸総合
		大衆文芸
		俳句
		短歌
		芸術・美術
	人文科学	宗教
	社会科学	歴史一般
	自然科学	自然科学一般
		地球宇宙科学
政治・経済・商業	政治／外交	国会行政
		海外情勢外交
	経済／経営	経営／経済
	金融／財政	金融財政
	商業／消費者	広告宣伝・ＰＲ
	国勢／民力	国勢／民力
		所得・物価・消費
産業	農林水産	農業経営
	食料／食品	醸造業
	運輸／通信	海事・海運・港湾
工業	工業一般	公害・環境保全
	建設／土木	建設一般
	機械	機械一般
		自動車・オートバイ・自転車
	電気機／電子	家電・弱電・照明
		エレクトロニクス
		コンピュータ／情報処理
		電波・電気通信
厚生・医療	厚生	福祉
	医学	医学総合
		家庭医学・健康

7.A.3 「新聞」のジャンル情報の詳細

- ジャンル(1)

「新聞」の「ジャンル(1)」列には、その新聞タイトルが配達される範囲の違いによって、以下の分類が記載されている。

例　　「全国紙」「ブロック紙」「地方紙」

なお、「新聞」の書誌 ID の 4～5 桁目で表される ID（01～31）は、新聞の母集団に含まれる 16 タイトル、および朝夕刊の別に対して独自に付与した ID である。各タイトルに対応づけられたジャンル（配達エリア）との対応関係を、表 7-12 に示す（著作権の都合で採録対象から外した 2 タイトルは表示していない)。

表 7-12: 新聞の書誌 ID（4～5 桁目）の内訳

ID	タイトル	朝夕刊	配達エリア	ID	タイトル	朝夕刊	配達エリア
01	朝日新聞	朝刊	全国紙	17	河北新報	朝刊	地方紙
02	朝日新聞	夕刊	全国紙	18	河北新報	夕刊	地方紙
03	毎日新聞	朝刊	全国紙	19	新潟日報	朝刊	地方紙
04	毎日新聞	夕刊	全国紙	20	新潟日報	夕刊	地方紙
05	読売新聞	朝刊	全国紙	21	京都新聞	朝刊	地方紙
06	読売新聞	夕刊	全国紙	22	京都新聞	夕刊	地方紙
09	産経新聞	朝刊	全国紙	23	神戸新聞	朝刊	地方紙
10	産経新聞	夕刊	全国紙	24	神戸新聞	夕刊	地方紙
11	北海道新聞	朝刊	ブロック紙	25	中国新聞	朝刊	地方紙
12	北海道新聞	夕刊	ブロック紙	26	中国新聞	夕刊	地方紙
13	中日新聞	朝刊	ブロック紙	27	高知新聞	朝刊	地方紙
14	中日新聞	夕刊	ブロック紙	28	高知新聞	夕刊	地方紙
15	西日本新聞	朝刊	ブロック紙	30	琉球新報	朝刊	地方紙
16	西日本新聞	夕刊	ブロック紙	31	琉球新報	夕刊	地方紙

7.A.4 「白書」のジャンル情報の詳細

「白書」の「ジャンル(1)」列には、白書のタイトルおよび内容によって分類した 9 種類のジャンル名（「安全」「外交」「科学技術」「環境」「教育」「経済」「国土交通」「農林水産」「福祉」）が記載されている。各ジャンルと白書のタイトルは、表 7-13 のように対応している。

表 7-13: 「白書」のジャンル情報

ジャンル	白書タイトル
安全	警察白書
	原子力安全白書
	原子力白書
	交通安全白書
	公害紛争処理白書
	消防白書
	犯罪白書
	防衛白書 / 日本の防衛
	防災白書
外交	外交青書 / わが外交の近況
	政府開発援助（ODA）白書 / 我が国の政府開発援助
科学技術	科学技術白書
	情報通信白書 / 通信白書
環境	環境白書
	循環型社会白書
教育	文部科学白書 / 我が国の文教施策
経済	エネルギー白書
	ものづくり白書 / 製造基盤白書
	経済財政白書 / 経済白書
	公益法人白書
	地方財政白書
	中小企業白書
	通商白書
	独占白書 / 独占禁止白書
	労働経済白書 / 労働白書
国土交通	観光白書
	国土交通白書 / 運輸白書 / 建設白書
	首都圏白書
	土地白書 / 国土利用白書
農林水産	食料・農業・農村白書 / 農業白書
	森林・林業白書 / 林業白書
	水産白書 / 漁業白書
福祉	厚生労働白書 / 厚生白書
	高齢社会白書
	国民生活白書
	少子化社会白書
	障害者白書
	人権教育・啓発白書
	青少年白書
	男女共同参画白書

※「防衛白書 / 日本の防衛」のように、「/」で区切られている白書タイトルは、1976 年から 2005 年までの間にタイトルの変更があったことを表す。

7.A.5 「教科書」のジャンル情報

- ジャンル(1)

「教科書」の「ジャンル(1)」列には、教科の別が記載されている。

例　「国語」「数学」「理科」「社会」「外国語」「技術家庭」「芸術」「保健体育」「情報」「生活」（ただし、「外国語」は中学校と高等学校のみ、「情報」は高等学校のみ、「生活」は 小学校のみとなる）

- ジャンル(2)

「教科書」の「ジャンル(2)」列には、学校の別が記載されている。

例　「小学校」「中学校」「高校」

- ジャンル(3)

「教科書」の「ジャンル(3)」列には、学年の別が記載されている。

例　「1」「2」「3」「4」「5」「6」「(空文字)」

※　「学年」の情報は、小学校・中学校の場合にのみ記載される。高校の場合は空文字になる。

7.A.6 「広報紙」のジャンル情報

- ジャンル(1)

「広報紙」の「ジャンル(1)」列には、当該の自治体の地域が記載されている。

例　「北海道地方」「東北地方」「関東地方」「中部地方」「近畿地方」「中国地方」「四国地方」「九州・沖縄地方」

- ジャンル(2)

「広報紙」の「ジャンル(2)」列には、当該の自治体の都道府県名が記載されている。

例　「北海道」「青森県」「秋田県」「沖縄県」など

なお、「広報紙」の書誌 ID の 4～8 桁目で表される ID は、「全国地方公共団体コード」の上 5 桁と一致しており、広報紙を発行している自治体に対応する。ID と自治体の対応関係を、表 7-14 に示す。

表 7-14: 「全国地方公共団体コード」と自治体名の対応

01103 北海道札幌市東区
01109 北海道札幌市手稲区
01213 北海道苫小牧市
01230 北海道登別市
01631 北海道十勝支庁音更町
02202 青森県弘前市
03208 岩手県遠野市
04101 宮城県仙台市青葉区
04361 宮城県亘理郡亘理町
05201 秋田県秋田市
06207 山形県上山市
07203 福島県郡山市
07447 福島県大沼郡会津美里町
08203 茨城県土浦市
08217 茨城県取手市
08235 茨城県つくばみらい市
09202 栃木県足利市
09213 栃木県那須塩原市
09361 栃木県下都賀郡壬生町
10201 群馬県前橋市
10205 群馬県太田市
10208 群馬県渋川市
11107 埼玉県さいたま市浦和区
11208 埼玉県所沢市
11219 埼玉県上尾市
11461 埼玉県北葛飾郡栗橋町
12104 千葉県千葉市若葉区
12203 千葉県市川市
12206 千葉県木更津市
12229 千葉県袖ケ浦市
13104 東京都新宿区
13108 東京都江東区
13112 東京都世田谷区
13116 東京都豊島区
13120 東京都練馬区
13203 東京都武蔵野市
13209 東京都町田市
13221 東京都清瀬市
14101 神奈川県横浜市鶴見区
14107 神奈川県横浜市磯子区
14114 神奈川県横浜市瀬谷区
14133 神奈川県川崎市中原区
14204 神奈川県鎌倉市
14208 神奈川県逗子市
14212 神奈川県厚木市
15106 新潟県新潟市南区
15210 新潟県十日町市
16202 富山県高岡市
17204 石川県輪島市
18201 福井県福井市
19201 山梨県甲府市
19208 山梨県南アルプス市
20203 長野県上田市
20385 長野県上伊那郡南箕輪村
21204 岐阜県多治見市
21217 岐阜県飛騨市
22103 静岡県静岡市清水区
22136 静岡県浜松市浜北区
22213 静岡県掛川市
22222 静岡県伊豆市
23113 愛知県名古屋市守山区
23211 愛知県豊田市
23302 愛知県愛知郡東郷町
24203 三重県伊勢市
24210 三重県亀山市
25206 滋賀県草津市
25209 滋賀県甲賀市
26108 京都府京都市右京区
26204 京都府宇治市
26407 京都府船井郡京丹波町
27109 大阪府大阪市天王寺区
27123 大阪府大阪市淀川区
27141 大阪府堺市堺区
27210 大阪府枚方市
27220 大阪府箕面市
28102 兵庫県神戸市灘区
28201 兵庫県姫路市
29204 奈良県天理市
29340 奈良県生駒市
30201 和歌山県和歌山市
30206 和歌山県田辺市
31202 鳥取県米子市
32201 島根県松江市
33461 岡山県小田郡矢掛町
34108 広島県広島市佐伯区
34205 広島県尾道市
35210 山口県光市
36341 徳島県名西郡石井町
37201 香川県高松市
38202 愛媛県今治市
39205 高知県土佐市
40203 福岡県久留米市
40305 福岡県筑紫郡那珂川町
41401 佐賀県西松浦郡有田町
42201 長崎県長崎市
43215 熊本県天草市
44202 大分県別府市
45206 宮崎県日向市
46218 鹿児島県霧島市
47209 沖縄県名護市

7.A.7 「Yahoo!知恵袋」のジャンル情報の詳細

「Yahoo!知恵袋」の「ジャンル(1)～(3)」列には、14 種類の「大カテゴリ」、59 種類の「中カテゴリ」、および 130 種類の「小カテゴリ」という 3 階層のカテゴリがそれぞれ記載されている。大カテゴリ・中カテゴリ・小カテゴリの一覧を表 7-15 に示す。小カテゴリの違いは / で区切られている。

表 7-15: 「Yahoo!知恵袋」のジャンル情報

大カテゴリ	中カテゴリ	小カテゴリ
エンターテインメントと趣味	ゲーム	ゲーム / オンラインゲーム / トレーディングカード
	テレビ、ラジオ	テレビ、ラジオ / CM / ラジオ
	映画	映画
	音楽	音楽 / 楽器 / 邦楽 / 洋楽
	芸能人、タレント	芸能人、タレント / あの人は今 / 話題の人物
	占い、超常現象	占い、懸賞
	本、雑誌、コミック	本、雑誌、コミック / コミック / 雑誌
インターネット、PC と家電	インターネット	インターネット
	パソコン、周辺機器	パソコン、周辺機器
	家電、AV 機器	家電、AV 機器 / オーディオ
	携帯電話、モバイル	携帯電話、モバイル
ビジネス､経済とお金	家計、貯金	家計、貯金 / ローン / 家計、節約 / 貯金
	株と経済	株と経済 / 株式 / 経済、景気
	企業と経営	企業と経営 / 会計、経理、財務 / 会社情報、業界動向 / 企業法務、知的財産 / 起業
	保険、税金、年金	保険、税金、年金 / 税金 / 年金 / 保険
職業とキャリア	資格、習い事	資格、習い事 / 資格 / 専門学校、職業訓練
	就職、転職	就職、転職 / 就職活動 / 退職、入社手続き
	派遣、アルバイト、パート	派遣、アルバイト、パート / アルバイト、フリーター / パート / 派遣
	労働問題、働き方	労働問題、働き方 / 失業、リストラ / 労働条件、給与、残業 / 労働問題
ニュース、政治、国際情勢	ニュース、事件	ニュース、事件 / 事件、事故、流行 / 話題のことば
	政治、社会問題	政治、社会問題
スポーツ､アウトドア、車	アウトドア	アウトドア / キャンプ / 釣り
	スポーツ	スポーツ / オリンピック / サッカー / ダイビング、サーフィン / 格闘技、武術 / 野球
	バイク	バイク
	自動車	自動車 / 新車 / 中古車
暮らしと生活ガイド	ショッピング	ショッピング / これ、探してます
	ボランティア、環境問題	ボランティア、環境問題

	家事、住宅	家事、住宅 / 家事 / 不動産、引越し
	公共施設、役所	公共施設、役所 / 美術館、博物館、図書館 / 役所、手続き
	福祉、介護	福祉、介護
	法律、消費者問題	法律、消費者問題 / 消費者問題 / 法律相談
	料理、グルメ、レシピ	お酒、ドリンク / レシピ、調理法 / 飲食店、デパ地下 / 料理、食材 / 料理、グルメ、レシピ
健康、美容とファッション	コスメ、美容	コスメ、美容 / エステ、マッサージ / コスメ、化粧品
	ファッション	ファッション
	メンタルヘルス	カウンセリング、治療 / ストレス / 心の悩み、相談
	健康、病気、ダイエット	健康、病気、ダイエット / ダイエット / 病気、症状、ヘルスケア
	恋愛相談、人間関係の悩み	恋愛相談、人間関係の悩み
子育てと学校	子育て、出産	子育て、出産 / 子どもの病気とトラブル / 子育ての悩み / 妊娠、出産
	受験、進学	受験、進学
	小・中学校、高校	小・中学校、高校
	大学、留学	大学、留学 / 大学 / 留学
	幼児教育、幼稚園、保育園	幼児教育、幼稚園、保育園
マナー、冠婚葬祭	マナー	マナー / あいさつ、てがみ、文例
	冠婚葬祭	冠婚葬祭 / 結婚 / 葬儀
	祭りと年中行事	祭りと年中行事
教養と学問、サイエンス	一般教養	一般教養
	芸術、文学、歴史	芸術、文学、歴史
	言葉、語学	言葉、語学
	数学、サイエンス	数学、サイエンス
	天気、天文、宇宙	天気、天文、宇宙
	動物、植物、ペット	動物、植物、ペット
地域、旅行、お出かけ	海外	海外
	交通、地図	交通、地図
	国内	国内 / 花火大会
Yahoo!JAPAN	Yahoo!オークション	Yahoo!オークション
	Yahoo!サービス	Yahoo!サービス
	Yahoo!知恵袋	Yahoo!知恵袋
その他	アダルト	アダルト
	ギャンブル	ギャンブル

7.A.8 「Yahoo!ブログ」のジャンル情報の詳細

「Yahoo!ブログ」の「ジャンル(1)～(3)」列には、15 種類の「大カテゴリ」、54 種類の「中カテゴリ」、および 316 種類の「小カテゴリ」という 3 階層のカテゴリがそれぞれ記載されている。大カテゴリ・中カテゴリ・小カテゴリの一覧を表 7-16 に示す。小カテゴリの違いは / で区切られている。

表 7-16: 「Yahoo!ブログ」のジャンル情報

大カテゴリ	中カテゴリ	小カテゴリ
ビジネスと経済	金融と投資	通貨、為替 / 株式 / 保険 / 貯蓄、預金 / 銀行 / 不動産 / その他金融と投資
	雇用	就職 / 転職 / アルバイト / 人材派遣 / 失業、無職 / その他雇用
	ビジネス	会社経営 / 起業 / その他ビジネス
	職種	事務職 / 営業職 / 技術職 / 企画職 / 専門職 / 公務員 / その他職種
	経済	景気 / 国際経済 / その他経済
コンピュータとインターネット	インターネット	ホームページ / ネットサービス / その他インターネット
	コンピュータ	ソフトウェア / パソコン / 周辺機器 / Windows / Macintosh / その他コンピュータ / UNIX
生活と文化	祝日、記念日、年中行事	クリスマス / 正月 / 誕生日 / バレンタインデー / 花火 / ホワイトデー / 花見 / エイプリルフール / その他祝日、記念日、年中行事
	グルメ、ドリンク	レシピ / 飲食店 / 食べ物 / 飲み物 / 菓子、デザート
	環境問題	その他環境問題 / 省エネ / 自然保護 / リサイクル / ごみ問題 / 地球温暖化
	事件・事故	事件 / 事故 / 防犯
	災害	火災 / 地震 / 台風 / 火山活動 / その他災害
	文化活動	宗教 / ボランティア活動 / 祭りと伝統 / その他文化活動
	季節	冬 / 秋 / 夏 / 春
エンターテインメント	映画	俳優、女優 / その他映画 / 映画祭 / 映画レビュー / 映画監督
	テレビ	アナウンサー / コマーシャル / その他テレビ / ドラマ番組 / バラエティ番組
	音楽	その他音楽 / 音楽祭 / 洋楽 / 邦楽 / 音楽レビュー / ミュージシャン
	占い	心理テスト、性格診断 / タロット占い / 星占い / 血液型占い / 風水 / その他占い
	芸能人、タレント	男性 / 女性 / グループ
	超常現象	幽霊、心霊 / 都市伝説 / UFO / 超能力 / その他超常現象

	テーマパーク	ディズニーリゾート / ユニバーサル・スタジオ・ジャパン / 遊園地 / その他テーマパーク
家庭と住まい	住まい	ガーデニング / 修理とリフォーム / 住居 / インテリア
	ペット、動物	昆虫 / 観賞魚、水草 / 鳥 / ウサギ / ハムスター / 犬 / 猫 / その他ペット
	家庭電化製品	オーディオ / 季節家電 / 映像機器 / 調理器具 / その他家電
	家庭	家計 / 育児 / 家族 / 家庭環境
政治	政界と政治活動	政党､団体 / 選挙 / 政界 / 地方自治 / 軍事 / 国会 / その他政界と政治活動
	国際情勢	中東情勢 / アジア情勢 / アフリカ情勢 / アメリカ情勢 / ヨーロッパ情勢 / オセアニア情勢 / その他国際情勢
健康と医学	美容と健康	フィットネス / スキンケア / ボディケア / ネイルケア / ダイエット / その他美容と健康
	病気、症状	子どもの病気 / メンタルヘルス / 生活習慣病 / アレルギー / その他の病気 / 花粉症
学校と教育	学校	小学校 / 中学校 / 高校 / 専門学校 / 大学 / その他学校 / 受験
	教育	習いごと / 幼児教育 / 社会教育 / その他教育
科学	社会科学	人類学と考古学 / 経済学 / 心理学 / 政治学 / 法学 / その他社会学
	自然科学	化学 / 工学 / 物理学 / 天文学 / 気象学 / 生物学 / その他自然科学
出会い	恋愛	失恋 / 遠距離 / アドバイス / 片思い / 初恋 / その他恋愛
	結婚	離婚 / 結婚式 / 見合い / 再婚 / その他結婚 / 婚約、結納
地域	日本	北海道 / 青森県 / 岩手県 / 宮城県 / 秋田県 / 山形県 / 福島県 / 東京都 / 神奈川県 / 埼玉県 / 千葉県 / 茨城県 / 栃木県 / 群馬県 / 山梨県 / 新潟県 / 長野県 / 富山県 / 石川県 / 福井県 / 愛知県 / 岐阜県 / 静岡県 / 三重県 / 大阪府 / 兵庫県 / 京都府 / 滋賀県 / 奈良県 / 和歌山県 / 島根県 / 岡山県 / 広島県 / 山口県 / 徳島県 / 香川県 / 愛媛県 / 高知県 / 福岡県 / 佐賀県 / 長崎県 / 熊本県 / 大分県 / 宮崎県 / 鹿児島県 / 沖縄県
	世界の地方	アジア / アフリカ / オセアニア / 北アメリカ / 中東 / ヨーロッパ / ラテンアメリカ
特集	趣味とスポーツ	CLUBKEIBA
芸術と人文	芸術、アート	イラストレーション / 絵画 / 写真 / 工芸 / 書道 / その他芸術、アート

	文学	ノンフィクション、エッセイ / 小説 / 詩 / 俳句、川柳 / 短歌 / その他文学 / 伝記、自伝
	デザイン	ファッション / 工業デザイン / 建築デザイン / その他デザイン
	舞台、演劇	観劇 / 伝統芸能 / その他舞台、演劇
	人文科学	倫理学 / 哲学 / 歴史 / その他人文科学
Yahoo!サービス	Yahoo!ブログ	練習用
	Yahoo!オークション	出品 / 落札 / ウォッチリスト / Yahoo!オークションストア
	Yahoo!ゲーム	その他 Yahoo!ゲーム
	Yahoo!アバター	アバター作成
	Yahoo!スポーツ	ファンタジーサッカー
	Yahoo!ショッピング	Yahoo!ショッピングストア
趣味とスポーツ	スポーツ	野球 / サッカー / ゴルフ / テニス / 格闘技 / モータースポーツ / スキー / スノーボード / マリンスポーツ / その他スポーツ / 陸上競技 / バスケットボール / オリンピック / バレーボール / ラグビー / 卓球
	レジャー	旅行 / 釣り / 登山 / 散歩 / キャンプ / その他レジャー
	趣味	読書 / 漫画、コミック / アニメーション / ゲーム / おもちゃ / カラオケ / 携帯電話 / その他趣味
	乗り物	鉄道、列車 / 自動車 / オートバイ / その他乗り物 / 飛行機 / 自転車
	ギャンブル	パチンコ、パチスロ / 競馬 / 宝くじ / その他ギャンブル

7.A.9 「韻文」のジャンル情報の詳細

ジャンル(1)

「韻文」の「ジャンル(1)」列には、韻文の種別が記載されている。

例 「短歌」「俳句」「詩」

7.A.10 「法律」のジャンル情報の詳細

ジャンル(1)

「法律」の「ジャンル(1)」列には、データの取得元である「法令データ提供システム」で採用されている、法務省『日本現行法規』に基づく法律のジャンルが記載されている。一覧を表 7-17 に示す。

表 7-17: 「法律」のジャンル情報

「01 憲法」	「19 災害対策」	「35 金融・保険」
「02 国会」	「20 建築・住宅」	「37 陸運」
「03 行政組織」	「21 財務通則」	「38 海運」
「04 国家公務員」	「23 国税」	「39 航空」
「05 行政手続」	「24 専売・事業」	「40 貨物運送」
「07 地方自治」	「25 国債」	「42 郵務」
「08 地方財政」	「26 教育」	「43 電気通信」
「09 司法」	「27 文化」	「44 労働」
「10 民事」	「28 産業通則」	「45 環境保全」
「11 刑事」	「29 農業」	「46 厚生」
「12 警察」	「30 林業」	「47 社会福祉」
「14 国土開発」	「31 水産業」	「49 防衛」
「15 土地」	「32 鉱業」	「50 外事」
「16 都市計画」	「33 工業」	
「17 道路」	「34 商業」	

7.A.11 「国会会議録」のジャンル情報の詳細

ジャンル(1)

「国会会議録」の「ジャンル(1)」列には、開催院の別が記載されている。

例　「衆議院」「参議院」

ジャンル(2)～(3)

「国会会議録」の「ジャンル(2)～(3)」列には、4 種類の会議種別（「常任委員会」「特別委員会」「本会議」「その他」）と会議名称が、それぞれ記載されている。会議種別と会議名称は、表 7-18 のように対応している。

表 7-18: 「国会会議録」のジャンル情報

会議種別	会議名称
本会議	本会議
常任委員会	安全保障委員会、運輸委員会、科学技術委員会、外交防衛委員会、外務委員会、環境委員会、議院運営委員会、経済産業委員会、決算委員会、決算行政監視委員会、建設委員会、厚生委員会、厚生労働委員会、行政監視委員会、国土・環境委員会、国土交通委員会、財政・金融委員会、財政金融委員会、社会労働委員会、商工委員会、総務委員会、大蔵委員会、地方行政委員会、逓信委員会、内閣委員会、農林水産委員会、文教委員会、法務委員会、予算委員会
特別委員会	ロッキード問題に関する調査特別委員会、安全保障特別委員会、沖縄及び北方問題に関する特別委員会、科学技術振興対策特別委員会、個人情報の保護に関する特別委員会、交通安全対策特別委員会、公害対策及び環境保全特別委員会、国会等の移転に関する特別委員会、国旗及び国歌に関する特別委員会、国際平和協力等に関する特別委員会、災害対策特別委員会、世界貿易機関設立協定等に関する特別委員会、政治倫理の確立及び公職選挙法改正に関する特別委員会、青少年問題に関する特別委員会、物価等対策特別委員会、物価問題等に関する特別委員会
その他	議院運営委員会庶務小委員会、憲法調査会、国民生活・経済に関する調査会、国民生活・経済に関する調査特別委員会高齢化社会検討小委員会、産業・資源エネルギーに関する調査会、少子高齢社会に関する調査会、文教委員会入試問題に関する小委員会、予算委員会公聴会、予算委員会第三分科会、予算委員会第四分科会、予算委員会第五分科会、予算委員会第六分科会、予算委員会第八分科会

付録 7-B: サンプル ID ベース書誌情報データの構成

サンプル ID ベース書誌情報データ（Joined_info.txt）は、表 7-1（Bibliography.txt）、表 7-4（Sample_info.txt）、表 7-6（Directory.txt）、表 7-8（Article.txt）の情報を結合し、サンプル ID を単位として生成したものであり、『中納言』での書誌情報表示に用いられているデータである。ここに記録された「人名 ID」「人名」「生年代」「性別」の情報は、各サンプルを実際に執筆した人物に関する情報を表している。同一サンプルに対して複数のレコードが存在する場合は、重複する情報をスラッシュで区切って並べている。

表 7-19: サンプル ID ベース書誌情報データの構成

フィールド名称	結合元のファイル
サンプル ID	---
書誌 ID	Bibliography.txt
タイトル	Bibliography.txt
副題	Bibliography.txt
巻号	Bibliography.txt
責任表示	Bibliography.txt
出版者	Bibliography.txt
出版年	Bibliography.txt
ISBN	Bibliography.txt
サンプル抽出基準点ページ	Sample_info.txt
ジャンル 1	Bibliography.txt
ジャンル 2	Bibliography.txt
ジャンル 3	Bibliography.txt
ジャンル 4	Bibliography.txt
責任表示 ID	Bibliography.txt
人名 ID	Article.txt
人名	Directory.txt
生年代	Directory.txt
性別	Directory.txt
corpusName	サブコーパスの略称（新規追加）