

既存のInDesignファイルから 二次利用データを抽出する

DTPデータの多くは、二次利用を考慮した作りになっていません。 当社が導入しているデータ抽出プログラムは、InDesignファイルから CSV形式など二次利用しやすいデータを効率的に抽出します。

■DTPデータ抽出の問題点

新聞、雑誌やカタログなどの紙メディアから、ニュースサイトやSNS、オンラインショップなどのインターネットメディアまで、現在の情報発信方法は多岐に渡っています。発信者としては、ひとつのコンテンツをなるべく多くのメディアに、効率良く展開することが理想ではないでしょうか。

マルチメディア展開を考えていくと、必ずたどり着くのがデータベース化です。データベース化を行うためには、コンテンツの内容を「タイトル」「見出し」「スペック」などの項目に細分化し、リスト化することが必要になります。

しかし、紙メディアの場合は紙面を完成させることが先に立ち、なかなかデータベースまで手が回らないのが現状ではないでしょうか。Webコンテンツであれば、HTML自体が構造化されたテキストなので比較的二次利用しやすいのですが、DTPは自由にデザイン、レイアウトができる反面、一旦紙面化されてしまうとそのデータは転用しづらくなってしまいます。

HTML

<h1>大見出し</h1>
<h2>中見出し</h2>
<h3>小見出し</h3>
<P>

InDesign

| 色味を記述られた
| 10,0,0,0,0-melim(0,0,0,0,0-melim(0,0,0,0,0
| colox大学-checktyfe(元大学)
| colox大学-checktyfe(元大学)
| colox大学-checktyfe(元大学)
| colox大学-checktyfe(元大学)
| colox大学-checktyfe(元大学)
| colox大学-checktyfe(元大学)
| colox checktyfe(元大学)
| colox checktyfe(元大学

HTMLの場合それぞれの要素がマークアップにより明確に区別されているが、InDesignは独自のタグがテキストに挿入されていてリスト化しづらい構造になっている。

■専用プログラムによるデータ抽出

当社ではDTPデータからのテキスト抽出をより効率的に行うため、データ抽出専用プログラムを導入しています。

このプログラムは、InDesignのデフォルトである索引抽出機能や目次抽出機能では補えない、「段落スタイル」「文字スタイル」「フォント名」「文字サイズ」「レイヤー」「正規表現」といった項目を検索条件に指定することができます。そのため、これまで手作業で行っていたInDesignファイルからの抽出作業を、ほぼ自動で行うことが可能です。

汎用性も高く、裏面で紹介しているように様々な商 品で活用できます。

そのほかにも以下のような特長があります。

①データベース構築

カタログ、ガイドブック、辞典など、データベース 化を前提として作られていないデータからでも、CSV 形式でテキストを抽出することで、データベースを容 易に構築できるようになりました。

②派生商品を効率的に制作

当社の自動組版技術を併用することで、オリジナルのコンテンツからこれまでより効率的に派生商品を制作することが可能です。学習教材やガイドブック、商品カタログ等に活用できます。

③画像データ抽出

画像リンク情報(ファイル名)も抽出データにひも付けできます。紙メディアから電子書籍やインターネットメディア、eラーニングコンテンツなどへ効率的に転用できます。

④インデックス (索引) 作成

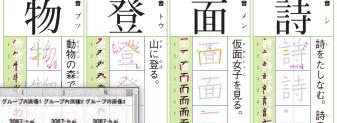
特定のスタイルの文字を、ノンブル情報と同時に抽出。これまで行っていたマーカー指定などが不要になり、索引を効率的に作成することができます。



抽出プログラムを使った事例紹介

●書籍からの抽出

「段落スタイル」と「文字スタイル」の情報を元にテキストを CSV形式で抽出します。画像はファイル名で抽出されています。



4	A	В	C	D	E	F	G	H	70	1	K		5 而		Ĕ	5	DV	なも
-	ベージ番号	漢字	読み仮名	用語例	意味	熟語	画数	部首名	グループ内医像1	グループ内画像2	グループ内面像3		6 m		を見る	6 計	- 7	_ `
	3	10	ý,	詩をたしなむ。詩の会。	自然や人事などから受ける感興・感動を、言語 形式で表現したもの。	詩句 詩句才 詩	13	こんべん	3067-nai	3067-h.ai	3067-bai		7 而 8 而 9 面	-	- 0	7 8 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	*	部の会
	3	面	x	仮面女子を見る。	物の表面	水面面面積	13	めん	3177-nai	3177-h.ai	3177-bai	11 # O	· (E)		いみ 物 の	10 計		式で
	3	登	トウ	山に登る。	のぼる	登 登 登 項 型	13	はつがしら	3142-n.ai	3142-hai	3142-bai		13=		表面	12 詩 13 調		表現したもの。
	3	物	ブツ	動物の森で遊ぶ	€ø	安異見金物	13	うしへん	3170-n.ai	3170-h.ai	3170-b.ai	登: 登: Us 禄: 山: in	王·	面。海。月接。面。	『水』。 『水』。 『面』。	シ ナチ	作流言語言	5. 誤

●カタログからの抽出

グループ化された商品単位で抽出されます。ひとつの商品の中に複数の画像がある場合は、「グループ内画像XX」という項目名で画像のファイル名が抽出されます。



	メーカー名		商品仕様			税抜価格	グループ内画像1	グループ内画像2	グループ内画像3
1	文具工業	マグネットクリップ	寸法/長さ50mm×幅45mm×厚さ25 mm	マグネット付きで この価格!	168	160	C:¥商品画像¥90586.eps		
1	文具工業		●水性インク●線幅/0.8mm●長さ/ 130mm●材質/再生プラスチック	定番のサインベン。 グリーン商品。	68.8	65.6	C.¥部品¥アイコン-オリジナルeps	C:¥商品画像¥90559.eps	C¥部品¥アイコン-値下げ_大eps

●表組みからの抽出

方眼罫ノートブック								
サイズ	罫内容	枚数	商品コード	税込価格	税別価格			
① B5サイズ	5mm方眼	30	az1303	¥315	¥300 *			
② A4サイズ	5mm方眼	40	az1304	¥420	¥400 *			
バインダー用ノート								
サイズ	罫内容	枚数	商品コード	税込価格	税別価格			
① B5サイズ	5mm方眼	30	az1305	¥525	¥500 *			
② A4サイズ	5mm方眼	40	az1306	+323	+500			

グループ化された表とその見出し単位で抽出されます。セルが 結合されている場合は、各行に共通の情報を付加して抽出します。

ı	ベージ番号	商品名	商品コード	梲込価格	税別価格
	1	方眼罫ノートブック	az1303	¥315	¥300
	1	方眼罫ノートブック	az1304	¥420	¥400
	1	バインダー用ノート	az1305	¥525	¥500
	1	バインダー用ノート	az1306	¥525	¥500
_					

●カセットからの抽出



