（別添２）

数値（表）、文章、地理空間情報のデータ作成に当たっての留意事項

目次

本書は、「二次利用の促進のための府省のデータ公開に関する基本的考え方（ガイドライン）」の「３　機械判読が容易なデータ形式による公開の拡大の考え方」のうち、数値（表）、文章、地理空間情報のデータ作成に当たっての留意事項を示すものである。

なお、本書でデータ構造の整形手順の説明のために示している表形式データは架空データサンプルであり、統計情報データベースを通じて提供される統計データ（「統計調査等業務の業務・システム最適化計画」に基づき統計情報データベースを通じた提供を推進している統計表管理システムの統計表を含む。）に本書を適用するということではない。

１．数値（表形式）データの作成に当たっての留意事項

（１）表形式データの定義

表形式データとは、行と列の縦横 2 次元状に配列されたデータである（図１）。

表形式データの架空データサンプル（その１）

集計項目	平成23年 (1,000円)	24年 (1,000円)	差分 注）
合計	55,000	127,768	a)232
あ　あ	1,000	1,100	110
い　い	2,000	2,200	110
う　う	3,000	3,300	110
え　え	4,000	4,400	110
お　お	5,000	3,300	66
か　か	6,000	2,200	37
き　き	7,000	1,100	16
く　く	8,000	5,500	69
け　け	9,000	9,900	110
こ　こ	10,000	10,000	100

注：平成23年から平成24年のうちの増減の割合を記載している。
a)脚注番号のサンプルを示している。

図1：表形式データの例

表形式データを構成する各要素の名称を、以下の通り定義する（図２）。

- キャプション（表題）：
  - 表形式データ全体を表す短い説明。
- カラム（Column）：
  - 表形式データの、縦方向の列。
- ロウ（Raw）：

- 表形式データの、横方向の行。

- セル（Cell）：
  - 表形式データの各項目。表計算ソフトでは、個々のマス目として表現される。
- データセル（Data Cell）：
  - 表形式データにおいて、数値データ本体が格納されるセル。
- タイトル（Title、題目）：
  - 表形式データの、各カラムの冒頭。カラムに含まれるデータセルの内容や単位を説明する。
- タイトル行:
  - タイトルが配置された行。
- テーブル（Table、表）：
  - 1 行以上からなるタイトル行、1 行以上のデータセル、0 行以上の脚注からなる、セルの集合。
- データセット（Dataset）：
  - テーブルを含む表形式データのまとまり。
- 脚注:
  - 表形式データに付与する、タイトルやデータセルに対する補助説明。
- 脚注番号:
  - タイトルやデータセルに付与する、脚注と結びつけるための番号。
- 単位:
  - 数値の基準となる、約束された一定量。例えば、 "m"（メートル）や"g"（グラム）に代表される物理単位や、「円」「ドル」に代表される貨幣単位等がある。
- 記数単位:
  - データセルの値の桁を示す数。たとえば、単位として「百万円」と書かれているカラムの記数単位は「1,000,000」である。実際の値は、データセルの値に記数単位を乗じたものである。

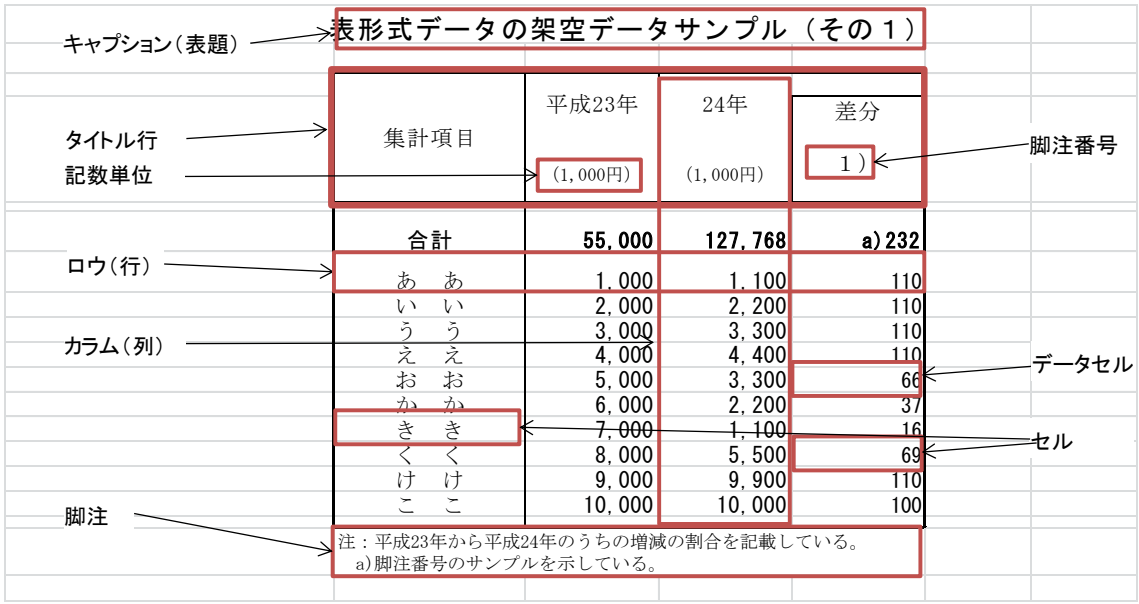


図2：表形式データの各要素の名称定義

（２）表形式データにおけるデータの構造の留意事項

表形式データを構造の整ったデータの構造にするための留意事項を以下に示す。留意事項に沿って構造を整えることで、機械判読に適したデータ形式に変換し利活用することが可能となる。

【留意事項１】
１つのデータセットには、１つのテーブルのみを含める。（複数個のテーブルを含めない）

【解説】

図３のデータセットには、複数の表を含んでいる。このようなデータセットをコンピュータが解読するためには、表の切れ目を扱う必要があり、解読手順が複雑になる。このため、１つのデータセットには、１つの表のみを持つべきである。複数の表が必要である場合は、その数だけ分割する（図４）。

１．架空データサンプル（その２）①

項　　目	α	β	γ	σ
ア　ア　ア　ア　ア　ア	1.012	1.014	1.041	1.041
イ　イ　イ　イ　イ　イ	1.035	1.019	1.081	1.000
ウ　ウ　ウ　ウ　ウ　ウ	1.040	1.028	1.059	1.022
エ　エ　エ　エ　エ　エ	1.011	1.009	1.007	1.012
オ　オ　オ　オ　オ　オ	1.039	1.027	1.030	1.030
合　　計	5.137	5.097	5.218	5.105

２．架空データサンプル（その２）②

項目	説明
α	あああ
β	いいい
γ	ううう
σ	えええ

３．架空データサンプル（その２）③

区　　分	X
A	1.032
B	1.062
C	1.024
D	1.055

図3：1つのデータセットに複数の表がある（留意事項1を満たさない）例

１．架空データサンプル（その２）①

項　　目	α	β	γ	σ
ア　ア　ア　ア　ア　ア	1.012	1.014	1.041	1.041
		019	1.081	1.000
		028	1.059	1.022
		009	1.007	1.012
		027	1.030	1.030
		097	5.218	5.105

２．架空データサンプル（その２）②

項目	説明
α	あああ

３．架空データサンプル（その２）③

区　　分	X
A	1.032
B	1.062
C	1.024
D	1.055

図4：図3の表を分割（留意事項1を満たす）

【留意事項２】
データセルに、整形や位取りのための文字（スペース、改行、カンマ等）を含めない。

【解説】

図５の集計項目カラムにある「ああ」「いい」等のデータセルは、整形のための空白を含んでいる。データセルに含まれる空白や改行に意味があるのか否かは、機械は判別できない。また、数値データには位取りのためのカンマが含まれている。カンマを除かなければ、機械はそのデータは正しい値として認識できない。従って、機械の解読に不要な空白や改行、カン

マ等を含めない（図６）。

表形式データの架空データサンプル（その１）

集計項目	平成23年 (1,000円)	24年 (1,000円)	差分 1）
合計	55,000	127,768	a)232
あ　あ	1,000	1,100	110
い　い	2,000	2,200	110
う　う	3,000	3,300	110
え　え	4,000	4,400	110
お　お	5,000	3,300	66
か　か	6,000	2,200	37
き　き	7,000	1,100	16
く　く	8,000	5,500	69
け　け	9,000	9,900	110
こ　こ	10,000	10,000	100

注：平成23年から平成24年のうちの増減の割合を記載している。
a)脚注番号のサンプルを示している。

図5：セルに整形のための空白、改行、カンマを含む（留意事項2を満たさない）例

表形式データの架空データサンプル（その１）

集計項目	平成23年 (1000円)	24年 (1000円)	差分 1）
合計	55000	127768	a)232
ああ	1000	1100	110
いい	2000	2200	110
うう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

注：平成23年から平成24年のうちの増減の割合を記載している。
a)脚注番号のサンプルを示している。

図6：整形のためのスペース、改行、カンマを除去（留意事項2を満たす）

【留意事項３】
年の値には、西暦表記とし、和暦を併記する。

【解説】

図７の年次カラムは和暦で書かれている。コンピュータのプログラムでは、年の値を数値の大小により比較することが多い。従って年の値は、年が経過するごとに値が単調増加する西暦とし（図８）、必要に応じて和暦を併記する。

また、内容によっては年度表記されていることもあるため、歴年と年度の判読が可能な記述をする必要がある。

表形式データの架空データサンプル（その３）

	年次		A (mg)	B (mg)	C (mg)
平成	5	年	0.01	0.01	0.00
	6		0.02	0.01	0.00
	7		0.01	0.01	0.00
	8		0.03	0.01	0.00
	9		0.20	0.01	0.00
	10		0.01	0.01	0.00
	11		0.02	0.01	0.00
	12		0.04	0.01	0.00
	13		0.01	0.01	0.00
	14		0.02	0.01	0.00
	15		0.03	0.01	0.00

図7：年が和暦で書かれている（留意事項3を満たさない）例

表形式データの架空データサンプル（その３）

	年次		年次 (西暦)	A (mg)	B (mg)	C (mg)
平成	5	年	1993	0.01	0.01	0.00
	6		1994	0.02	0.01	0.00
	7		1995	0.01	0.01	0.00
	8		1996	0.03	0.01	0.00
	9		1997	0.20	0.01	0.00
	10		1998	0.01	0.01	0.00
	11		1999	0.02	0.01	0.00
	12		2000	0.04	0.01	0.00
	13		2001	0.01	0.01	0.00
	14		2002	0.02	0.01	0.00
	15		2003	0.03	0.01	0.00

図８：西暦のカラムを追加（留意事項3を満たす）

【留意事項４】
数値等のデータの値やタイトル、単位以外の情報を、セルに含めない。

【解説】

図９の合計値は「a) 69」となっている。このセルには、値である「69」と注釈番号である「a)」の両方が含まれている。機械がこのセルを解読するには、事前に注釈番号「a)」を除かなければならない。このため、機械に解読させるべき数値やタイトル以外の情報を、セルには持たせない（図 10）。

表形式データの架空データサンプル（その１）

集計項目	平成23年（1000円）	24年（1000円）	差分１）
合計	**55000**	**127768**	**a) 232**
ああ	1000	1100	110
いい	2000	2200	110
ううう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

注：平成23年から平成24年のうちの増減の割合を記載している。
a) 脚注番号のサンプルを示している。

図9：セルにキャプション、注釈、注釈番号を含む（留意事項4を満たさない）例

集計項目	平成23年（1000円）	24年（1000円）	差分
合計	**55000**	**127768**	**232**
ああ	1000	1100	110
いい	2000	2200	110
ううう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

図 10：キャプション、脚注、脚注番号を除去（留意事項4を満たす）

【留意事項５】
すべてのセルは、他のセルと結合しない。

【解説】

図 11 のタイトルのセルが結合されている。通常、結合されたセルは、原則的にはすべて同じ値を持つ。これを機械が解読するためには、結合されたセルの値と、結合されている範囲を認識する必要がある。従って、セルは結合せず、同じ値を記載する（図 12)。

表形式データの架空データサンプル（その４）

年度	期	A (mg)	B (mg)	C (mg)
2005	上	0.01	0.01	0.00
	下	0.01	0.01	0.00
2006	上	0.02	0.01	0.00
	下	0.01	0.01	0.00
2007	上	0.01	0.01	0.00
	下	0.02	0.01	0.01
2008	上	0.03	0.01	0.00
	下	0.02	0.02	0.00
2009	上	0.02	0.01	0.00
	下	0.02	0.01	0.00
2010	上	0.01	0.01	0.00
	下	0.01	0.01	0.00

図 11：セルが結合されている（留意事項5を満たさない）例

表形式データの架空データサンプル（その４）

年度	期	A (mg)	B (mg)	C (mg)
2005	上	0.01	0.01	0.00
2005	下	0.01	0.01	0.00
2006	上	0.02	0.01	0.00
2006	下	0.01	0.01	0.00
2007	上	0.01	0.01	0.00
2007	下	0.02	0.01	0.01
2008	上	0.03	0.01	0.00
2008	下	0.02	0.02	0.00
2009	上	0.02	0.01	0.00
2009	下	0.02	0.01	0.00
2010	上	0.01	0.01	0.00
2010	下	0.01	0.01	0.00

図 12：年カラムのセル結合を解除（留意事項5を満たす）

【留意事項６】
値が存在しない場合を除き、データセルを空白にしない。（データ値を省略しない）

【解説】

図 13 の年次の平成 5 年以降の第 1 列及び第 3 列は、空白である。人間はこの部分のデータセルに「平成 6 年」が省略されていることがわかるが、機械には分からない。従って、このデータを機械判読に適した構造にするためには、値が存在しない場合を除き、データセルを空白にせず、値は省略しない（図 14）。

表形式データの架空データサンプル（その３）

年次		年次（西暦）	A（mg）	B（mg）	C（mg）
平成	5 年	1993	0.01	0.01	0.00
	6	1994	0.02	0.01	0.00
	7	1995	0.01	0.01	0.00
	8	1996	0.03	0.01	0.00
	9	1997	0.20	0.01	0.00
	10	1998	0.01	0.01	0.00
	11	1999	0.02	0.01	0.00
	12	2000	0.04	0.01	0.00
	13	2001	0.01	0.01	0.00
	14	2002	0.02	0.01	0.00
	15	2003	0.03	0.01	0.00

図 13：年のデータセル値が省略されている（留意事項6を満たさない）例

表形式データの架空データサンプル（その３）

年次		年次（西暦）	A（mg）	B（mg）	C（mg）
平成	5 年	1993	0.01	0.01	0.00
平成	6 年	1994	0.02	0.01	0.00
平成	7 年	1995	0.01	0.01	0.00
平成	8 年	1996	0.03	0.01	0.00
平成	9 年	1997	0.20	0.01	0.00
平成	10 年	1998	0.01	0.01	0.00
平成	11 年	1999	0.02	0.01	0.00
平成	12 年	2000	0.04	0.01	0.00
平成	13 年	2001	0.01	0.01	0.00
平成	14 年	2002	0.02	0.01	0.00
平成	15 年	2003	0.03	0.01	0.00

図 14：省略されている語句を補う（留意事項6を満たす）

【留意事項７】
データセルの内容を示すタイトルは、１行で構成する。

【解説】

図 15 のタイトルは構造化されており、2 行からなっている。4 列番目のカラムは、「差分（平成 23 年から平成 24 年の増減割合）」という意味であるが、これを機械は解読できない。タイトルの文言を工夫して、カラムのタイトルを 1 行で表現する（図 16）。

集計項目	平成23年（1000円）	24年（1000円）	差分 １）
合計	**55000**	**127768**	**232**
ああ	1000	1100	110
いい	2000	2200	110
うう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

図 15：タイトルが複数行からなる（留意事項7を満たさない）例

集計項目	平成23年（1000円）	平成24年（1000円）	平成23年から平成24年の増減割合
合計	**55000**	**127768**	**232**
ああ	1000	1100	110
いい	2000	2200	110
うう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

図 16：タイトルを 1 行にまとめる（留意事項7を満たす）

【留意事項８】
データの単位を明記する。

【解説】

データの単位（物理単位、貨幣単位）は、データ処理に必須である。このため、カラムにはデータの単位を明記する（図 17、18）。

なお、国際単位系に含まれる単位については国際単位系の利用を推奨する。日本独自の単位系を利用する場合は、国際単位系への換算値を併せて記載する。

集計項目	平成23年（1000円）	平成24年（1000円）	平成23年から平成24年の増減割合
合計	**55000**	**127768**	**232**
ああ	1000	1100	110
いい	2000	2200	110
うう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

図 17：タイトルに単位がない（留意事項8を満たさない）例

集計項目	平成23年（×1000円）	平成24年（×1000円）	平成23年から平成24年の増減割合（%）
合計	**55000**	**127768**	**232**
ああ	1000	1100	110
いい	2000	2200	110
うう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

図 18：タイトルの文言を修正し、単位を追記（留意事項8を満たす）

（３）表形式データにおけるデータ形式の留意事項

（２）に基づき、表形式データを構造の整ったデータの構造にし、更にデータ形式を機械判読に適したデータ形式にするための留意事項を以下に示す。留意事項に沿ってデータ形式を整えることで、機械判読に適したデータ形式にすることが可能となる。

【留意事項１】
項目ラベルの各値は、公開されているコードを積極的に活用する。

【解説】

項目ラベルの各値は、積極的に公開されているコード（地域コード、法人コード等）を活用することで、データ収集や各種コードによるデータの横断的利用が機械で自動的に容易となる。そのため、公開されているコードの所在を明らかにしつつ、そのコード値を活用した値にすることとする。

例えば、地域を表す情報（都道府県、市町村名等）に対応する地域コードの値を入れるデータセルを設けることで、地図情報との融合が容易に可能となる。

【留意事項２】
データセットは、オープンな標準データ形式で提供する。

【解説】

仕様が公開され、それが標準化されたフォーマット、すなわちオープンなフォーマットは、解読するツールが広く普及しており機械判読に適している。このため、文書形式データはオープンなフォーマットで公開する。

具体的には、CSV や XML を推奨するが、仕様が国際標準化されている Open Document Format（.ods）や Office Open XML 形式（.xlsx）でもよい。

【留意事項３】
保存するファイル名は、命名ルールに従う。

【解説】

公開ファイル名の命名の考え方として、より機械判読に適したものにす

るには、以下の要件求がめられる。

・1バイト文字列とする（日本語の全角文字等を含まない)。

・命名ルールが公表されていることが望ましい。

（ファイル命名の例）

ファイル名が「cas13it01_（任意の名称）.csv 」の場合

・最初の３ケタは作成機関　：　cas＝内閣官房

・４～５桁目は作成年　　：　13＝２０１３年

・６～７桁目はカテゴリ　　：　it＝IT 戦略分野

・８～９桁目は事業ＩＤ　　：　01＝白書情報

【留意事項４】

公開するデータは、URL リストの公開も行う。

【解説】

公開するデータについては、データの所在を明らかにするため、府省内における公開データに関する URL リストの公開も行う。

また、連続する表を公開する場合は、ファイル名を除く URL 表現の後、地域、年号、月等の表現を URL 表現として規定し、連続的に機械がデータを取得できることが望ましい。

【留意事項５】

公開前におけるファイル内容、プロパティには十分注意して公開する。

【解説】

各府省の Web コンテンツ作成ガイドラインにて規定されていることも多いと思われる注意事項であるが、オープンデータの取組においても同様の対応が求められる。以下にポイントを記す。

・ファイルの記載内容について

－ファイルのヘッダ部分に“【機密性 2 情報】”の記載が残っていたら、削除する（ホームページで公開される情報は“機密性 1 情報”」)。

－変更履歴、コメント等が非表示のまま残っていたら、変更履歴やコメント等は削除する。

－Excel でのデータ作成において、印刷範囲外のセルのメモや行や列を非表示にしたまま残っていたら、そのデータは削除する。

・ファイルのプロパティについて
ープロパティに他の団体名、個人名や資料名等が入ったまま残っていたら、他の団体名、個人名は削除し、資料名は公開する資料名に修正する。

＜参考：タイトルやデータ型の仕様記述方法＞

タイトルやデータ型は、利用者がデータの仕様を理解するよう公開する必要がある。その記述方法には、現在いくつかの技術コミュニティーで進められている取組みを含めて、記述箇所と記述形式の面で、以下で示すようないくつかの方法がある。更に、表形式データを取り扱う既存のツールとの親和性をふまえつつ、推奨する記述方法を今後検討することとする。基本的には、データの仕様が明確になり、データ処理を行なうプログラムが作成できることが重要であり、データの仕様が自明に理解できるデータについては、必ずしも明示的な仕様記述を行なう必要はない。

１．記述箇所について

記述箇所には、次の３つの方法が考えられる。
（ア）データの仕様を別ファイルに記述する方法
（イ）データカタログのメタデータに記述する方法
（ウ）データファイルの中に記述する方法

（ア）データの仕様を別ファイルに記述する方法

【解説】

データの仕様（データセットのタイトル名、データセットの作成者、データセットの公開日、データセットの基本言語、カラムの単位・記数、カラムのデータタイプ）をデータとは別のファイルにデータの仕様として作成し、データと合わせて公開する。このデータの仕様を公開することにより、利用者が機械で判読できるようソフトウェアを開発することが可能となる。

（イ）データカタログのメタデータに記述する方法

【解説】

データの仕様（データセットのタイトル名、データセットの作成者、データセットの公開日、データセットの基本言語、カラムの単位・記数、カラムのデータタイプ）について、データカタログのメタデータに記載する。現時点では、データカタログのメタデータ項目は決まっていないため、将

来的な実現方法の選択肢とする。

（ウ）データファイルの中に記述する方法

【解説】

単位やデータ型を、データファイル内に定型フォーマットで記述することにより、複数のデータセットを機械が同様に解読できる。

２．記述形式について

記述形式には、次の２つの方法がある。

（ア）@を利用する方法（csv ファイルヘッダ部分に記載する）

（イ）他で確立した同種の方法

（ア）@を利用する方法（csv ファイルヘッダ部分に記載する）

【解説】

表形式データのキャプション、タイトル、単位等のメタデータは、データセルの先頭に、表２に示すヘッダを利用して付与する。

ヘッダは”@”または”@@”から始める。”@”で始まるヘッダに対する値は、その行に記述する。”@@”で始めるヘッダに対する値は、次の行に記述する。

表1:本文書が規定するヘッダ

ヘッダ	意味
@Caption	データセットのキャプション
@Creator	データセットの作成者
@Date	データセットの公開日
@Language	データセットの基本言語
@@Title	タイトル行
@@Unit	カラムの単位
@@Baseval	カラムの記数単位
@@Datatype	カラムのデータタイプ

それぞれのヘッダについての詳細を、以下に記す。

① @Caption: データセットのキャプション

- @Caption は、データセットのキャプションを記述するヘッダである。@Caption、キャプション名、言語コードの 3 つのセルからな

る。

- 言語コードは省略可能であり、省略した場合は、@Language ヘッダが指定する言語コードが指定されたものとする。言語コードは ISO639-1 に基づく値である。

② @Creator: **データセットの作成者**

- @Creator は、データセットの作成者を記述するヘッダである。@Creator、作成者名、言語コードの 3 つのセルからなる。
- 言語コードは省略可能であり、省略した場合は、@Language ヘッダが指定する言語コードが指定されたものとする。言語コードは ISO639-1 に基づく値である。

③ @Date: **データセットの公開日**

- @Date は、データセットの公開日を記述するヘッダである。@Date、公開日の 2 つのセルからなる。公開日は ISO 8610 に基づく値である。

④ @Language: **データセットの基本言語**

- @Date は、データセットの言語を記述するヘッダである。@Language、言語コードの 2 つのセルからなる。言語コードは ISO639-1 に基づく値である。

⑤ @@Title: **タイトル行**

- @@Title は、タイトル行を記述するヘッダであり、2 行で構成される。
- このヘッダの 1 行目は@@Title、言語コードの 2 つのセルからなる。
- 言語コードは省略可能であり、省略した場合は、@Language ヘッダが指定する言語コードが指定されたものとする。言語コードは ISO639-1 に基づく値である。
- このヘッダの 2 行目は、各タイトル名である。

⑥ @@Unit: **カラムの単位**

- @@Unit は、カラムの単位を記述するヘッダであり、2 行で構成される。
- このヘッダの 1 行目は@@Unit、言語コードの 2 つのセルからなる。言語コードは省略可能であり、省略した場合は、@Language ヘッ

ダが指定する言語コードが指定されたものとする。言語コードは ISO639-1 に基づく値である。

- このヘッダの 2 行目は、各カラムの単位である。単位に記数単位を含めてはならない。物理単位のべき乗数は、そのままテキストで記述する、たとえば加速度の単位「m/s2」は、「m/s2」と記述する。

⑦ @@Baseval：**カラムの記数単位**

- @@Unit は、カラムの記数単位を記述するヘッダであり、2 行で構成される。
- このヘッダの 1 行目は@@Baseval である。
- このヘッダの 2 行目は、各カラムの記数単位である。値を省略した場合、「1」が指定されたものと見なす。

⑧ @@Datatype：**カラムのデータタイプ**

- @@Unit は、カラムのデータタイプを記述するヘッダであり、2 行で構成される。
- このヘッダの 1 行目は@@Datatype である。
- このヘッダの 2 行目は、XML Schema に基づくデータタイプ値である。

（イ）他で確立した同種の方法

【解説】

データの仕様を記述する同種の取組として、Simple Data Format（SDF）[1]、Google DataSet Publishing Language (DSPL)[2]、Linked CSV[3]等が存在する。

- SDF については､表形式のデータを表す CSV をデータに利用した場合、JSON 形式の別ファイルにデータの定義を行うものである。
- Google DataSet Publishing Language (DSPL) については、表形式のデータを表す CSV をデータに利用した場合、XML 形式の別ファイルにデータの定義を行うものである。
- Linked CSV は、将来の LOD 化に向け、RDF として解釈されるべき CSV ファイルのデータ定義を CSV ファイル内で行なう方法である。

これらは、データの仕様を記述する取組であり、今後の普及動向や対応す

[1] http://www.dataprotocols.org/en/latest/simple-data-format.html
[2] https://developers.google.com/public-data/faq#how_do_i_decide
[3] http://jenit.github.io/linked-csv/

るツールの整備状況をみて判断することが適当と考えられる。

（4）ケーススタディ（データ構造の整形）

図 21 を例に、表形式データのデータ構造を整形する手順を示す。

表形式データの架空データサンプル（その１）

集計項目	平成23年 (1,000円)	24年 (1,000円)	差分 注)
合計	**55,000**	**127,768**	**232**
あ　あ	1,000	1,100	110
い　い	2,000	2,200	110
う　う	3,000	3,300	110
え　え	4,000	4,400	110
お　お	5,000	3,300	66
か　か	6,000	2,200	37
き　き	7,000	1,100	16
く　く	8,000	5,500	69
け　け	9,000	9,900	110
こ　こ	10,000	10,000	100

注：平成23年から平成24年のうちの増減の割合を記載している。
a)脚注番号のサンプルを示している。

図 19:整形前のオリジナルデータ

まず、表形式データが満たすべき条件のうち、図 19 が満たしていない箇所を列記する。その結果は表 3 の通りである。

表2:図 21 の条件確認結果

項目	留意事項	評価
(1)	1 つのデータセットに、1 種類の表形式データ（1 つのテーブル）が掲載されている。	○
(2)	整形のためのスペース、改行、位取りのカンマを含まない。	×
(3)	年の値を西暦で表記している。	×
(4)	数値やタイトル以外の情報（ラベル、注釈等）が、テーブルに含まれない。	×
(5)	すべてのデータセルが、他のデータセルと結合されていない。	○
(6)	値がない場合を除き、データセルの値が空白でない。	○
(7)	データの単位が明記されている。	×
(8)	カラムのタイトルに、単位や記数単位が含まれない。	×

それぞれの項目について、条件を満たしていない箇所を Microsoft Excel を

利用して整形する手法を記す。

○手順１：複数のテーブルに分割

新しいシートをテーブルの個数分作成し、それぞれのシートにテーブルを移動させる。これにより、1 つのデータセットに 1 つのテーブルを掲載することができる。

○手順２：脚注、脚注番号、キャプションを削除

セルの値として脚注、脚注番号、キャプションが記載されている場合は、それを取り除く。

脚注番号がセルの書式設定として付与されている場合は、セルの書式設定メニューを利用して除去する。Microsoft Excel 2007 以降であれば、「ホーム」メニューの「セル」タブにある「書式」メニュー（図 20）を利用する。Microsoft Excel 2003 以前であれば、「書式」→「セル」メニューを利用する。「セルの書式設定」ウィンドウの「分類」項目が「ユーザ定義」になっているので、これを「数値」に変更すれば、脚注番号を除去できる。

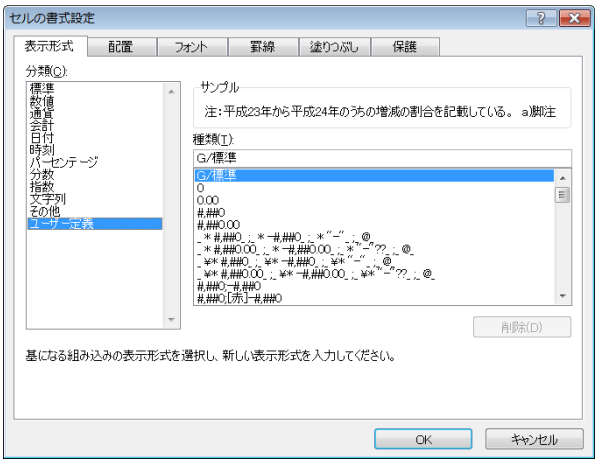


図 20: セルの書式設定ウィンドウ（Microsoft Excel 2007）

○手順３：不必要なスペース、改行、カンマの除去

不必要なスペース、改行、カンマがカラム全体、行全体、または複数のセルにまたがっている場合は、セルの置換機能を利用して、スペースと改行を除去できる。

Microsoft Excel 2007 以降であれば、除去対象のカラム全体、行全体、または複数のセルを選択し、ホームメニューの「編集」タブにある「検索と選択」というメニュー（図 21）を選択する。Microsoft Excel 2003 以前で

あれば、「編集」→「置換」メニューを選択する。検索する文字列欄に空白を入力し、置換する文字列欄を空にして「置換」ボタンを押すと、スペースを除去できる。

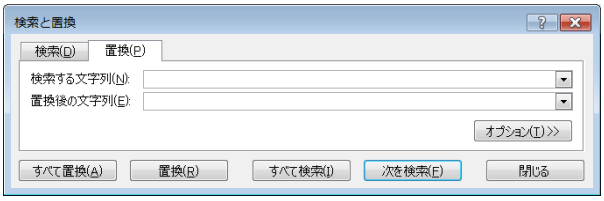


図 21：検索と置換ウィンドウ（Microsoft Excel 2007）

位取りのためのカンマは、セルの書式設定メニューを利用して除去する。「セルの書式設定」ウィンドウの「分類」項目から「数値」を選択し、右側にある「桁区切りを使用する」チェックボックスを外せば、位取りのためのカンマを除去できる（図 22）。

なお、データセルに直接カンマを入力している（「セルの書式設定」の数値分類の桁区切りによる桁区切り表示を行っていない）場合、カンマは削除する。

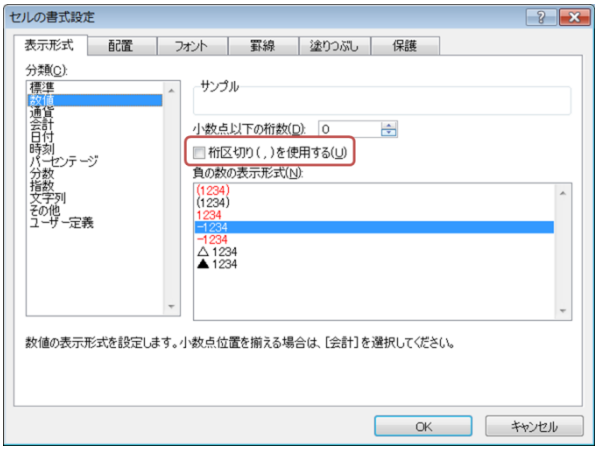


図 22：セルの書式設定ウィンドウ（Microsoft Excel 2007）

この作業が完了した時点で、データセットは図 23 のようになる。

表形式データの架空データサンプル（その１）

集計項目	平成23年（1000円）	24年（1000円）	差分 １）
合計	**55000**	**127768**	**a)232**
ああ	1000	1100	110
いい	2000	2200	110
ううう	3000	3300	110
ええ	4000	4400	110
おお	5000	3300	66
かか	6000	2200	37
きき	7000	1100	16
くく	8000	5500	69
けけ	9000	9900	110
ここ	10000	10000	100

注：平成23年から平成24年のうちの増減の割合を記載している。
a)脚注番号のサンプルを示している。

図 23:不必要なスペース、改行、カンマを除いたデータセット

○手順４：年の値を西暦で記載

西暦の年を記載するためには、2 つの方法がある。

- 和暦を記載しているセルを書き換える。
- 和暦を記載しているカラムの隣に、西暦を記載するカラムを追加する。

今回の例では、前者の方法をとる。

○手順５：セルの結合を解除

セルの結合を解除する。解除した結果生じた空白セルには、解除前に記載されていた値をコピーする。

○手順６：省略されたセルをコピー

前行と同じ値であるため記載が省略されているセルには、前行の値をコピーする。

○手順７：タイトルを１行にまとめる

図 25 のタイトルは構造を持っており、セルの結合を利用してその階層を表現している。これを 1 行で表現するために、タイトルの文言を変更する。たとえば、左から 4 番目のセルのタイトルを「2011 年から 2012 年の増減割合（%）」とする。

これまで整形を行った結果は図 24 の通りである。これは、表形式データの留意事項を満たしている。

集計項目	2011年（×1000円）	2012年（×1000円）	2011年から2012年の増減割合（%）
合計	**55000**	**127768**	**69**
ああ	1000	1100	-105
いい	2000	2200	-52
うう	3000	3300	0
ええ	4000	4400	26
おお	5000	3300	27
かか	6000	2200	-40
きき	7000	1100	-109
くく	8000	5500	-33
けけ	9000	9900	52
ここ	10000	10000	50

図 24:整形完了後のデータシート

（５）ケーススタディ（データ形式の整形）

表形式データのデータ形式を整形する手順を示す。

○手順１：地域コード等の設定

地域を表す情報（都道府県、市町村名等）に対応する地域コードの値を入れる列を設ける。（※これにより、地図情報との融合が可能となる。）

○手順２：特定アプリケーションに依存しない形式で保存

Microsoft Excel で作成した表を、拡張子「.csv」のファイルとして保存する。

```
集計項目,2011 年のコスト（×1000 円）,"2012 年のコスト（×1000 円）",2011 年から 2012 年の増減割合（%）
合計,55000 ,127768 ,232
ああ,1000 ,1100 ,110
いい,2000 ,2200 ,110
うう,3000 ,3300 ,110
ええ,4000 ,4400 ,110
おお,5000 ,3300 ,66
かか,6000 ,2200 ,37
きき,7000 ,1100 ,16
くく,8000 ,5500 ,69
けけ,9000 ,9900 ,110
ここ,10000 ,10000 ,100
```

図 25:図 24 を CSV 形式で出力

○手順３：特定アプリケーションに依存しない形式で保存

Microsoft Excel で作成した表を、拡張子「.csv」のファイルとして保存する。保存の際は、複数のシートをまとめて保存できないため、シート個別毎に、CSV のファイルを作成していることが望ましい。

なお、保存にあたって、ファイル名は、公開時の URL 表記のルール（既に設けられている場合は、そのルールに従う）に従って付与する。

○手順４：プロパティ情報の編集

ファイルのプロパティ情報に不適切な記載が居ないかチェックし、適切な記載を記入する。

２．文書形式データの作成に当たっての留意事項

（１）文書形式データの定義

文書形式データとは、文字を主な構成要素とし、一部図表を含んだデータである。

文書形式データに関する主な用語について、以下に解説する。

- プレインテキスト:
  - コンピュータ上で文章を扱うための一般的なファイルフォーマット、または文字列の形式の１つ。文字情報以外の情報、たとえば文字の色や大きさ、形状、文章に含まれる図表等の情報を含まない。
- 見出し:
  - 文章において内容の要点を非常に短い言葉にまとめ、本文より大きな字で章や節の最初に置かれる言葉。大きい方から編（部）、章、節、項、目といった名称が付けられる慣習がある。
- タグ:
  - 文章に対する構造（章、節、図表等）や見栄え（色、大きさ、形状等）に関する指定。
- マークアップ言語:
  - 文章の構造や見栄えに関する指定を、文章とともにテキストファイルに記述するための言語。

（２）文書形式データにおけるデータの構造の留意事項

文書は、基本的には人間が読む事を主目的としたデータである。文書形式データを構造の整ったデータの構造にするための留意事項を以下に示す。留意事項に沿って構造を整えることで、機械判読に適したデータ形式に変換し利活用することが可能となる。

【留意事項１】
文章に存在する部、章、節、図表等の構造が、コンピュータが明快に認識できる形で記述する。

【解説】

文章は、部、章、節、段落、図表等の構造を持っている。たとえば機械が文章の第１章を抽出したいとするならば、第１章が文章のどの部分にあるのか分からなければならない。このためには、タグやマークアップ言語を利用して、部、章、節、段落、図表等の見出しを追加し、タイトルを区別する（図

25)。

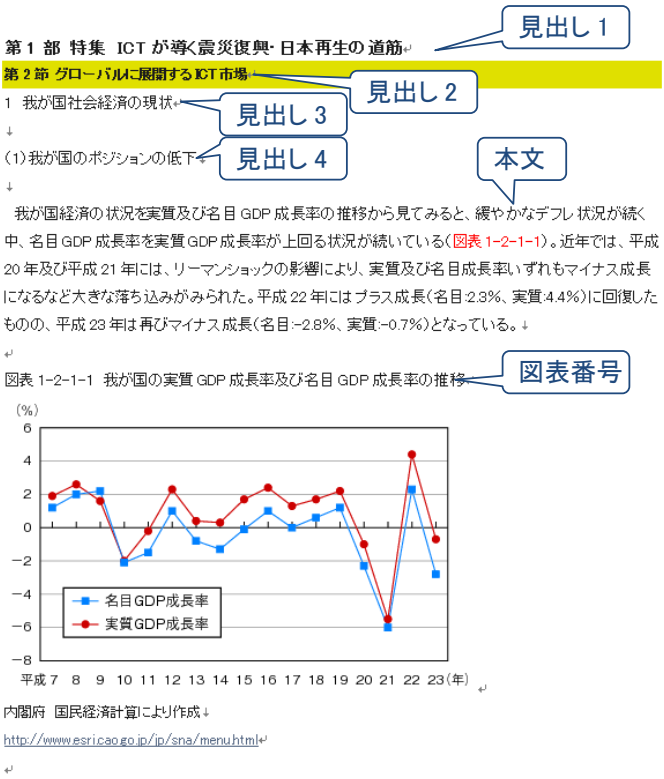


図 25:見出しを利用して文章を執筆[4]

【留意事項２】
文章内に、整形のための符号や文字（空白、改行等）を含めない。

【解説】

文章に含まれる空白、改行が有意であるか否かを、機械は判断できない。文書の解析や読み上げを行う際に、これらの空白、改行が支障となる。このため、機械の解読に必要のない空白や改行は、事前に除く。

（３）文書形式データにおけるにおけるデータ形式の留意事項

文書形式データを構造の整ったデータの構造にし、更にデータ形式を機械判読に適したデータ形式にするための留意事項を以下に示す。留意事項に沿ってデータ形式を整えることで、機械判読に適したデータ形式にすることが可能と

4 図中の文章は，総務省「 平成 24 年版　情報通信白書」より引用.
http://www.soumu.go.jp/johotsusintokei/whitepaper/

なる。

【留意事項１】
文書データ、オープンな標準データ形式で提供する。

【解説】
仕様が公開され、それが標準化されたフォーマット、すなわちオープンなフォーマットは、解読するツールが広く普及しており機械判読に適している。このため、文書形式データはオープンなフォーマットで公開する。
具体的には、プレインテキストにタグを挿入した XML 形式や HTML 形式のようなマークアップ形式を推奨するが、仕様が国際標準化されている Open Document Format（.odt）や Office Open XML 形式（.docx）もよい。 また、文字列のみである場合、テキスト形式（.txt）でもよい。

【留意事項２】
文書形式データが図表を含む場合、それらを構成する表形式データが添付されているべきである。

【解説】
図表やグラフを多く含む文書の、それら図表やグラフを形成した元になる表形式データが、機械判読に適したフォーマットで取得できるならば、それらのデータを利用したマッシュアップが容易になる。

【留意事項３】
公開前におけるファイル内容、プロパティには十分注意して公開する。

【解説】
「１．（３）表形式データにおけるデータ形式の留意事項」の【留意事項５】と同様である。

（４）ケーススタディ（データ構造の整形）

文書形式データのデータ構造を整形する手順を示す。Microsoft Word を利用して文書データを成型する例を示す。

○手順１：スタイルの設定

部、章、節等の構造と、見出しレベルとを対応づける。

たとえば、部は「見出し 1」、節は「見出し 3」、小節は「見出し 3」、小々節は「見出し 4」、図表タイトルは「図表番号」に対応づける（図 26）。

○手順２：スタイルの活用

対応づけた規則に従って文章を執筆する。その際、整形のために空白や改行を挿入しないように留意する。

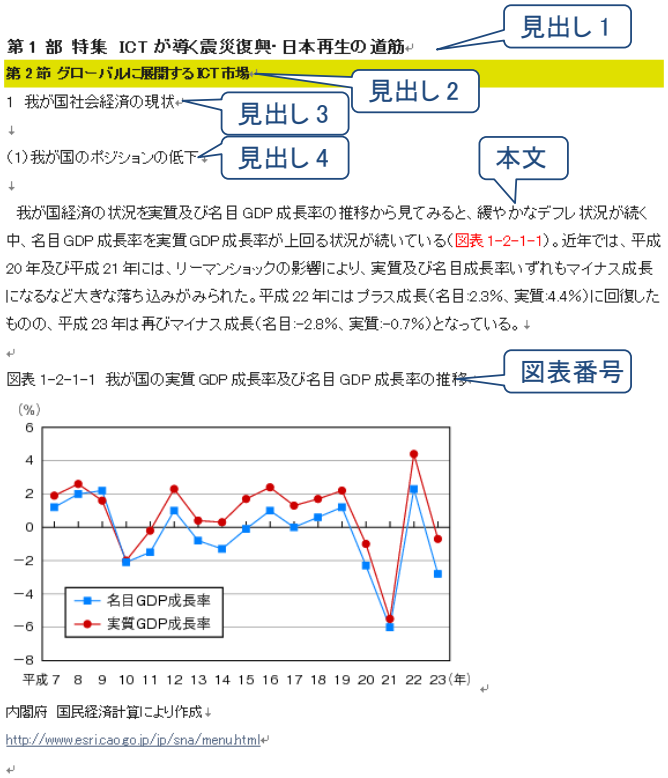


図 26: 文章の構造と見出しを対応付ける例[5]

[5] 総務省「平成 24 年版　情報通信白書」による.
http://www.soumu.go.jp/johotsusintokei/whitepaper/index.html

（５）ケーススタディ（データ形式の整形）

文書形式データのデータ形式を整形する手順を示す。Microsoft Word を利用して文書データを成型する例を示す。

○手順１：特定アプリケーションに依存しない形式で保存

編集した文書を、Open Document 規格準拠の XML 形式で書き出す。Microsoft Word であれば「ファイル」→「名前をつけて保存」の順に選択し、「ファイルの種類」を「OpenDocument テキスト（.odt）」に指定し、OpenDocument 規格準拠の XML 形式で書き出す。

※.odt ファイルは zip 形式で圧縮されている。ファイルの拡張子を.zip に変更して展開してみると、複数の XML ファイルと画像データから構成されていることが分かる。

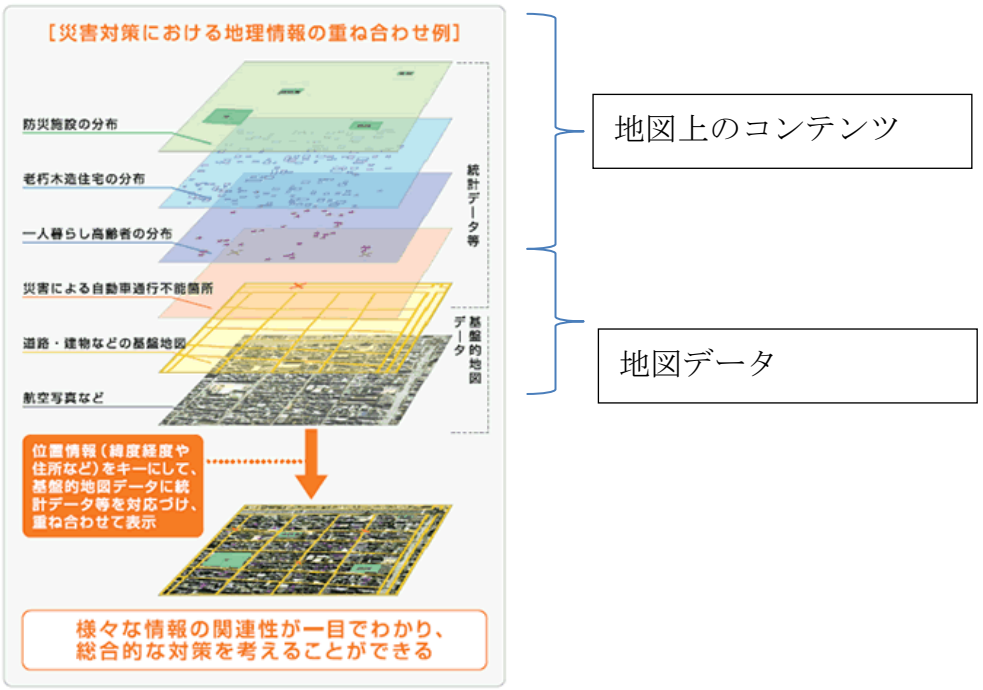
○手順２：リンクを追加

生成したファイルに、図表を構成する表形式データのリンクを追加する。

３．地理空間情報の作成に当たっての留意事項

（１）地理空間情報の定義

地理空間情報とは、空間上の特定の地点又は区域の位置を示す情報（当該情報に係る時点に関する情報を含む。）及び当該情報に関連付けられた情報（地理空間情報活用推進基本法第 2 条第 1 項）を指す。このデータは、地図データと地図上のコンテンツに分類される（図 27）。



出典：国土交通省 HP「GIS とは」に一部追記。
http://www.mlit.go.jp/kokudoseisaku/gis/guidance/guidance_1.html

図 27：地図データと地図上のコンテンツの整理

地理空間情報に関する主な用語について、以下に解説する。

- ラスタ形式:
  - 画像を点（ドット）の羅列によって表現したもの。PNG、JPEG、GIF、BMP、TIFF、PICT 等がある。
- ベクタ形式:
  - 2 次元コンピュータグラフィックスをコンピュータ内部で表現するデータ形式。GML[6]、KML[7]、SVG[8]等がある。

---

[6] GML（Geography Markup Language）：XML ベースのマークアップ言語であり、JIS X 7136（地理マーク付け言語）が策定されている。

[7] KML（Keyhole Markup Language）：XML ベースのマークアップ言語であり、地理情報システムのオープンソース化を目指す団体の規格に OGC KML として取り入れられている。

（２）地理空間情報におけるデータの構造の留意事項

地理空間情報を構造の整ったデータの構造にするための留意事項を以下に示す。留意事項に沿って構造を整えることで、機械判読に適したデータ形式に変換し利活用することが可能となる。

①地図データ

データの構造については、既存の取組み（基盤地図情報の提供等）で整備されており、特に留意する事項はない。

②地図上のコンテンツ

地図上のコンテンツとは、点・線・面を示す地理空間情報をいう。東日本大震災において、公開されている避難所情報に緯度経度座標が入っていないため、避難所の場所を地図に可視化しようとした際に時間を要した。位置情報によって、可視化等の利活用が進む地理空間情報においては、その公開にあたり、その位置情報を付与することは重要である。

「１ 数値（表形式）データの作成に当たっての留意事項」に加え、以下の留意事項がある。

【留意事項１】
地理空間情報のうち、位置情報に関するデータを付与する場合は、緯度経度座標を付与する。付与する際、準拠している座標参照系（世界測地系等）を明記する。

【解説】

座標の付与方法としては、以下が考えられる。

ア）地図から座標を取得する。

- 国土地理院の公開する数値地図情報[9]や、基盤地図情報の座標を利用する。
- 国土地理院で公開の電子国土 Web[10]の地図上から座標を取得する。
- 民間サービス（Yahoo!ジオコーダ API、GoogleGeo コーディング API 等）の座標変換サービスを利用する。

---

8 SVG（Scalable Vector Graphics）：SVG は、JIS X 7197（SVG に基づく地図の表現及びサービス）、並びに JIS X 4197（変倍ベクタグラフィックス）が策定されている。

9 http://www.gsi.go.jp/MAP/CD-ROM/cdrom.htm

10 http://portal.cyberjapan.jp/index.html

イ）住所から座標取得する。

・国土交通省「街区レベル位置参照情報アドレスマッチングツール[11]」「位置参照情報ダウンロードサービス[12]」等のサービスを利用する。

・民間サービス（Yahoo!ジオコーダ API、GoogleGeo コーディング API 等）で提供されるアドレスマッチングで得た座標を利用する。

表3 避難場所の一覧データ（留意事項1を満たさない例）

種別	避難場所名	住所
広域避難場所	日比谷公園	東京都千代田区日比谷公園１
避難所	日比谷高校	東京都千代田区永田町２丁目１６－１

表4 避難場所の一覧データに緯度経度座標を追記（留意事項1を満たす例）

種別	避難場所名	住所	緯度（※）	経度（※）
広域避難場所	日比谷公園	東京都千代田区日比谷公園１	35.675652	139.754426
避難所	日比谷高校	東京都千代田区永田町２丁目１６－１	35.674994	139.740512

※：世界測地系を使用

なお、平成 13 年の測量法改正以降、同法第 11 条 2 項に基づき、測量の基準としては、世界測地系が使用されている。もし、法改正前の日本測地系での測量成果を使用して新たにデータを作成する場合は、国土地理院が提供する「緯度・経度を世界測地系に変換するためのソフトウェア[13]」により、日本測地系に基づく測量成果を世界測地系に基づく測量成果に変換することが望ましい。

（３）地理空間情報におけるデータ形式の留意事項

地理空間情報を構造の整ったデータの構造にし、更にデータ形式を機械判読に適したデータ形式にするための留意事項を以下に示す。留意事項に沿ってデータ形式を整えることで、機械判読に適したデータ形式にすることが可能となる。

11 http://portal.cyberjapan.jp/

12 http://nlftp.mlit.go.jp/cgi-bin/isj/dls/_choose_method.cgi

13 http://vldb.gsi.go.jp/sokuchi/tky2jgd/about.html

①地図データ

【留意事項１】
データの提供に当たっては、機械判読が可能なベクタ形式に依るものとする。ベクタ形式のデータの作成にあたっては、最新の ISO 規格及び JIS 規格に基づいた地理空間情報標準プロファイル（JPGIS）、地理空間情報のメタデータの共通仕様を規定する日本版メタデータプロファイル（JMP）を用いる。

【解説】

仕様が公開され、それが標準化されたフォーマット、すなわちオープンなフォーマットは、解読するツールが広く普及しており機械判読に適している。このため、地図データはオープンなフォーマットで公開する。

府省において、地図データは主に、ラスタ形式、ベクタ形式が用いられている。また、地理情報システム(以下、GIS) 等で利用する空間データや位置情報を含む各種のコンテンツを記述するための XML ベースのマークアップ言語である GML も利用されている。

公開においては、ラスタ形式と比較して、同一の情報を表すのに必要な容量の小さくなるベクタ形式や、GML 形式が望ましい。また、公開にあたり、準拠している座標参照系（世界測地系等）を表記することで、データ利用の際の座標変換が容易になる。

JPGIS は、ISO の地理情報に関する専門委員会（ISO/TC 211）で策定された国際規格を基にした国内実用標準であり、異なるシステム間で地理空間情報データを相互利用する際の互換性の確保を主な目的に、データの設計、品質、記述方法等のルールを定めたもので、GIS 関係省庁連絡会議では政府の技術的標準と位置づけられている。JPGIS 及び JMP に基づいて地理空間データ及びメタデータを整備・提供することで、データを相互利用しやすい環境が整備され、異なる整備主体で整備されたデータの共用、システム依存性の低下、重複投資の排除等の効果を期待することができる。

また、利活用の更なる拡大を図るためには、SVG 形式 に変換して公開することが望ましい。

前記、ベクタ形式、GML 形式の場合、それに対応した GIS を用意する必要があり、当該ソフトウェアの知識や経験がある者の利用に限られるという課題があるが、SVG は、XML 形式の 2D ベクタ画像記述言語であり、HTML5 に組み込まれている（HTML5 対応の Web ブラウザ（Firefox 3.5

以降や Internet Explorer 9 以降等）があれば利用できる）。Web 標準の知識や経験がある者は多く、様々な活用が期待できる。なお、SVG 形式に変換するツールキット等も公開されている。

②地図上のコンテンツ

「1 数値（表形式）データの作成に当たっての留意事項」と同様である。

（4）ケーススタディ（データ構造の整形）

①地図データ

既存の方法を活用するため、本書では記載しない。

②地図上のコンテンツ

位置情報に関するデータは、留意事項 1 に示されているサービスを活用し、「住所」から「緯度」「経度」を取得し、列に追記する。

（5）ケーススタディ（データ形式の整形）

①地図データ

既存の方法を活用するため、本書では記載しない。

②地図上のコンテンツ

「1 数値（表形式）データの作成に当たっての留意事項」と同様である。

４．用語定義

本書が使用する用語の定義を表５に示す。

表5:用語の定義

用語	意味
二次利用	公開されているデータを引用、転載、複製、改変等を行うことにより利用すること
機械判読に適した	コンピュータプログラムに代表される機械が、データを自動的に解読し、技術的に二次利用できること
表形式データ	行と列の、縦横二次元状に配列されたデータ
文書形式データ	一次元状に配列された文字を主な構成要素とし、一部、図表を含み、人間がそれを読むことにより、人間に何らかの作用を与えることを目的としたデータ。
メタデータ	あるデータ自身について記述した、抽象度の高い付加データ
オープンなフォーマット	仕様が公開されており、それが国際標準化団体によって標準化されているファイルのフォーマット・形式特定のアプリケーションに依存しないデータ形式)
表計算ソフト	数値データの集計、分析に用いられるアプリケーションソフトウェア。画面上に格子状のマス目を表示し、そのマス目にデータを入れることにより表を作成する機能を有する。
キャプション（Caption、表題）	表形式データ全体を表す短い説明
カラム（Column）	表形式データの、縦方向の列
ロウ（Row）	表形式データの、横方向の行
セル（Cell）	表形式データの各項目。表計算ソフトでは、個々のマス目として表現される
データセル	表形式データにおいて、データ本体の値が格納されるべきセル
タイトル（Title、題目）	表形式データの、各カラムの冒頭、カラムに含まれるデータセルの内容や単位を説明する
タイトル行	タイトルが配置された行
データセット（Dataset）	機械がセルを取得する対象となる、表形式データの基本単位。表計算ソフトでは、1 シートにあたる。CSV 形式ファイルでは、1 ファイルにあたる。
テーブル（Table、表）	一行以上からなるタイトル行、一行以上のデータセル、0 行以上の脚注からなる、セルの集合

用語	意味
脚注	表形式データに付与する、タイトルやデータセルに対する補助説明。
脚注番号	タイトルやデータセルに付与する、脚注と結びつけるための番号。
単位	数値の基準となる、約束された一定量。「m」「g」に代表される物理単位、または「円」「ドル」に代表される貨幣単位がある。
記数単位	データセルの値の桁を示す数。たとえば，単位が「百万円」である場合，記数単位は「1,000,000」である。実際の値は、データセルの値に記数単位を乗じたものである。
データ型	機械が扱うデータの形式。文字列型、整数型、実数型、日付型等を指す。
CSV（Comma Separated Values）	表形式数値データの行を改行で区切り、セルを半角のカンマ「,」で区切る、テキストデータの表現形式。RFC 4180 により標準化されている。
XML（Extensible Markup Language）	個別の目的に応じたマークアップ言語作成のため、汎用的に使うことができる仕様、および仕様により策定される言語の名称。
RDF（Resource Description Framework）	主語、述語、目的語の 3 つ組で物事を表現するモデル。Web 技術の標準化団体 World Wide Web Consortium (W3C) が標準化している。
見出し	文章において内容の要点を非常に短い言葉にまとめ、本文より大きな字で章や節の最初に置かれる言葉。大きい方から編（部）、章、節、項、目といった名称が付けられる慣習がある。
プレインテキスト	コンピュータ上で文章を扱うための一般的なファイルフォーマット、または文字列の形式の 1 つ。文字以外の情報、たとえば文字の色や大きさ、形状、文章に含まれる図表等の情報を含まない。
タグ	文章に対する構造（章、節、図表等）や見栄え（色、大きさ、形状等）に関する指定、またはその指定方法。
マークアップ言語	文章の構造や見栄えに関する指定を、文章とともにテキストファイルに記述するための言語。
ワープロソフト	コンピュータ上で動作する、文章の入力、編集、印刷機能を実現したソフトウェア。

【補足情報】データの改ざんに対する技術対策

機械判読が容易な形式でデータが公開されることにより、データの改ざんに対する懸念が生じることがある。以下、改ざんへの技術的な対処方法について述べる。

（１）改ざんの定義

ここでは、データの改ざんとは、「オリジナルデータを改変し、それをオリジナルデータだと偽る」と定義する。

（２）技術的な対処方法

基本的にデータの改ざんを完全に防止するためのソフトウェア上の仕組みはない。実際にとりうる技術的な手法は、データの改ざんの検知及びデータの改ざん者を特定できる仕組みを用意することである。それによって利用者が改ざんされていないデータの入手を容易にし、またデータの改ざんを抑止する。

なお、技術的な対処方法は、データ利用の容易性を損うことや暗号処理などの計算負荷が大きいため、データの内容により、その必要があるものについて行うことが適当であり、基本的にはルールやリテラシーにより対応することが望ましい。

①改ざん検知技術

元データと改ざんされたデータとの間で、改ざんの有無を検知する技術として、チェックサム、電子署名、タイムスタンプといった方法がある。

表6 改ざん検知技術

改ざん検知技術	改ざん検知方法	検知できる内容
チェックサム（CRC/SHA-256）	データ保有者は、公開するデータに対して誤り検出関数（ある一定のルール）によって数値を算出し、公開データと合わせて誤り検出関数、数値を公開する。利用者（データ保有者自身含む）は、誤り検出関数、	・元データの改ざん有無

改ざん検知技術	改ざん検知方法	検知できる内容
	数値を用いて、公開データが改ざんされていないことを確認する。[14]	
電子署名	データ保有者は、公開するデータに対して電子署名をつけ、自身の公開鍵と合わせて公開する。利用者（データ保有者自身含む）は、公開鍵を用いて、データについている電子署名を検証して改ざんされていないことを確認する。[15]	・元データの作成者・作成機関 ・元データの改ざん有無（ただし、電子署名付与者による改ざんは検知不能）
タイムスタンプ	データ保有者は、公開するデータに対し、通常保存する際に記録されるタイムスタンプとは別に、専門機関からタイムスタンプを取得し、公開する。利用者（データ保有者自身含む）は、専門機関にタイムスタンプが正しいことを確認することで、改ざんされていないことを確認する。[16]	・元データの最終更新時刻 ・元データの改ざん有無（電子署名と併用する際、電子署名付与者とタイムスタンプ刻印者を別とすることで、電子署名付与者による改ざんを検知可能）

以下、３つの改ざん検知技術のうち、セキュリティ性及びコストが中である電子署名（暗号技術を利用した技術）について、ア）～イ）に具体的な手法と活用できる仕組みを記載する。

ア）暗号技術を利用した改ざん検知手法

データの改ざんを検知するためには、暗号技術を活用した、電子署名やデータのハッシュ値を付与することが有効である。特に公開鍵暗号系の技術によって付された電子署名については、その安全性の管理をきちんと行なうことができることが知られている。

具体的には、オリジナルデータには、ハッシュ値や電子署名を付した形で公開すればよい（ハッシュや電子署名の利用に際しては、「電子政府推奨暗号リスト」に掲載の暗号技術を利用する。また、ハッシュ値は Web サイト等の改ざんが困難な環境にて公開し、電子署名の利用に際しては、政府認証基盤（GPKI）を活用する。）。それによって、改ざんされたデータのハ

14 参考 URL：http://www.atmarkit.co.jp/fsecurity/rensai/inci03/inci01.html
15 参考 URL：http://www.jipdec.or.jp/esac/intro/shikumi.html
16 参考 URL：http://www.dekyo.or.jp/tb/system/system_7.html

ッシュ値や電子署名はオリジナルデータのハッシュ値は電子署名と異なるものとなるので、容易に発見できる。

なお、正しいハッシュ値や電子署名を計算して偽造することは極めて困難であることが知られている。

イ）アプリケーションソフトウェアの備えられた仕組みの利用

現在、様々なデータフォーマットにおいて、電子署名をつけることができるように整備されているものがある。例えば、以下のデータ形式には、そうした仕組みが備わっている。

docx、xlsx、pptx:　　Microsoft Office 形式

ods:　OpenDocument の SpreadSheet 形式

こうしたデータを主に扱うアプリケーションソフトウェア側にも、この仕組みを処理できるようにしており、改ざんされたデータをアプリケーションソフトウェア側で検知する機能を備えている。従ってこれらのアプリケーションを活用することで、比較的簡単に電子署名などのメカニズムを利用することができるようになっている。

（注）本留意事項は、機械判読に適したデータ形式でのデータの作成手順を記載する趣旨から、PDF 形式は例示していませんが、人が読む観点からの PDF 形式での公開やそれへの電子署名付与を否定するものではありません。