オープンデータとクラウドソーシングの親和性
―タスク設計と品質管理に関する検討―

応般

1

大向 一輝（国立情報学研究所）

情報分野のクラウドソーシング応用

情報分野そのものを対象としたクラウドソーシングのアプリケーションとしては，ESP Game（Google Image Labeler）に代表される画像データに対するタギングのように情報の付加価値を高めるものや，reCAPTCHA のように紙の情報をテキスト化するプロジェクトが一定の成果を挙げている．これらはコンピュータによる自動化の難しいタスクを人間の力で解決するという構造が明確であり，それゆえに広く注目を集めた．

一方，種々の数値データについてはあらかじめ整理がなされており，そもそも何らかの処理を加える必要がないように思われるが，実際にはいくつかの理由から使い勝手に乏しく，過去に作られたデータが死蔵されたままになっているのが現状である．膨大なデータを利活用できるようにすることで，エビデンスに基づく議論・意思決定や新しいビジネスの創出などさまざまな効果が期待できるが，既存データの再加工を誰がどのような手段で取り組むべきかについては議論がなされておらず，そのためのコストの全容も明らかではない．本稿では，これらの課題に対して公共セクタにおけるオープンデータへの取り組みに着目し，クラウドソーシングの適用可能性について述べる．

オープンデータの技術的再利用性

オープンデータとは Web 上で公開された再利用性の高いデータ，あるいはそのようなデータを公開するための活動を指す．政府や地方自治体といった公共セクタにおける取り組みが代表例であり，組織としての透明性の確保や行政への市民参加を促進するいわゆるオープンガバメントの実現手段の 1 つとして位置づけられるとともに，ビッグデータの潮流の一端を担うものとしてイノベーションにも寄与するものと期待されている．オープンデータの概要については本誌 2013 年 12 月号の特集「オープンデータ活用」に詳しい [1)]．

オープンデータの根幹をなす再利用性の概念には，制度面と技術面の 2 つの観点がある．制度面では，データの利活用を行う際に知的財産権に関連する諸制約が可能な限り取り払われており，またその制約事項が利用者に対して明示されている状態をもって再利用性が担保されているといえる．たとえば米国では公共セクタが作成する情報は原則として著作権が存在しないパブリックドメインとして扱われるが，ほかの国においては公共の情報であっても著作権が発生する．これに対して，著作権の存在を認めつつ情報の自由な利活用を奨励する手段として，一定の条件下での 2 次利用や再配布をあらかじめ許諾するライセンスを付与することが望ましい．実際には，国際的に活動する非営利団体クリエイティブ・コモンズが策定した，商用・非商用の区別なく利活用が可能であり，利用者には原著作者のクレジットの明記のみを求める CC BY ライセンスならびにその互換ライセンスが広く用いられている．日本政府においても電子行政オープンデータ戦略の下，政府が公開するデータについては原則として CC BY のライセンスを付与することが定められている．

オープンデータの技術面での再利用性を定義するにあたっては，Web の発明者 Tim Berners-Lee が理想的なオープンデータの在り方を 5 段階のレベルで表現した「5 つ星オープンデータ」のスキームが

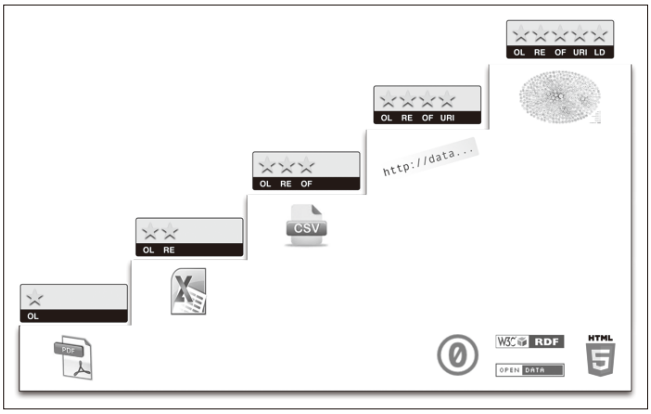


図 -1　5 つ星オープンデータスキーム（http://5stardata.info）

事業所規模別・企業規模別・業種別

区分		給与所得者数				
		3月末	6月末	9月末	12月末	年間月平均
		千人	千人	千人	千人	千人
平成20年分		54,672	55,269	55,000	54,739	55,124
21		56,232	54,909	54,035	53,884	54,967
22		55,817	54,887	54,367	54,153	54,792
23		55,169	54,688	54,459	54,273	54,647
24		54,432	54,466	53,948	54,221	54,267
25		55,982	55,935	55,673	55,354	55,736
		人	人	人	人	人
事業所規模別	10人未満	10,025,169	9,909,395	9,764,284	9,475,469	9,793,590
	10人以上	8,044,342	7,962,554	7,934,589	7,952,520	7,973,454
	30人以上 30人〃	9,029,604	8,866,174	8,826,926	8,645,401	8,842,012
	30人以上 100人〃	12,009,493	12,114,193	12,113,484	12,087,395	12,081,134
	30人以上 500人〃	4,033,197	4,068,114	4,044,791	4,057,085	4,050,800
	30人以上 1,000人〃	7,478,183	7,587,215	7,570,636	7,631,868	7,566,977
	30人以上 5,000人〃	5,361,658	5,426,879	5,418,072	5,504,484	5,427,772
	30人以上 計	37,912,135	38,062,575	37,973,909	37,926,233	37,968,695
	合計	55,981,646	55,934,524	55,672,782	55,354,222	55,735,739

図 -2　技術的再利用性の低いデータの例

知られている（**図 -1**）．

このスキームでは，1 つ星を獲得するためにはオープンライセンスの付与が必要である．これは前述の制度面での再利用性を担保するための必須条件である．次に，2 つ星を得るためには，データが機械的に取得可能であり自由に加工できるものでなければならない．これに反する例として，紙の書類をスキャンした画像ファイルが挙げられる．ファイル自体はコンピュータで取り扱うことができるが，記載された内容を転用するためには利用者が画像を見ながら再入力する必要があるため，再利用性が高いとはいえない．3 つ星を獲得するためには，公開されたデータのフォーマットが商用の独占的なものでないことが求められる．商用ソフトウェアの提供中止などによってデータへのアクセス手段が損なわれることを回避するための条項であり，CSV（カンマ区切りテキスト）や XML のような標準化されたフォーマットでの公開が推奨されている．4 つ星ならびに 5 つ星を獲得するためには，公開されたデータが RDF（Resource Description Framework）のようなセマンティクスを持つ形式で記述されている必要がある．

コンピュータを用いた文書作成が常態となっているいま，原資料となるファイル自体が 2 つ星や 3 つ星の要件を満たしており，そのまま公開することで一定程度のオープンデータ化は十分に達成できる．政府が公開する文書やデータにおいても PDF とともに編集可能なファイルが提供される例が増えており，オープンデータへの理解と実践が着実に進んでいることが分かる．

しかしながら，実際に得られるデータの中には 2 つ星や 3 つ星に該当していたとしても機械的に処理することがきわめて困難なものが多数存在している．ソーシャルメディア等ではこのようなデータを「ネ申（かみ） Excel」と呼び，その問題点や影響について議論が行われている．議論に興味のある方は上記キーワードで検索されたい．代表例として政府統計のポータルサイト e-Stat で提供されている民間給与実態統計（2013 年）の給与所得者数・給与額・税額に関するデータの一部を示す（**図 -2**）．技術的再利用性の観点からは，図の上半分と下半分とで異なる種類の情報が記載されていることは処理コストの大幅な増大につながる．また「平成 20 年分」の記載の下に 21・22 とあるように一部の表記の省略や，人数の単位が上下で異なることも個別の対応を要する．さらには「年間月平均」は各月の値から別途求めることができるため，並べて掲載することが不要あるいは有益でない場合がある．

こういったデータは，受け手にとって重要であろうと思われる情報を強調しつつ印刷時に定形のサイズに収めるための工夫の産物であるが，そのことが機械的な利活用の妨げとなっている．今後作成されるデータは公開の時点から再利用性の高い形式であ

ることが望まれており，これを実現するためのガイドラインの整備等が進められている一方[2]，過去のデータについては何らかの変換が必須となる．変換にあたっては多大なコストを要することが予想されるが，誰がコストを負担すべきかを議論する以前にこのような変換が実際に可能であるかどうかに関する検討が必要であると思われる．

オープンデータとクラウドソーシング

これらの課題に対して，クラウドソーシングによるデータの質的向上や継続的な管理を目的とした研究が進められている．文献 3）ではクラウドソーシングが適用可能な領域を下記の 5 つに分類している．

- **同一性の明示化**

複数のデータにまたがって記載されている同一の概念（都道府県名や年度等）に対し共通のコード・ID を割り当てる．またはそれらの概念が同一であることを示すリンク関係を付与する．

- **補完・照合・修正**

データ項目の抜け漏れや記述ルールとの整合性，内容の正誤を確認し，修正する．

- **分類**

データ全体または記載された個々の情報を事前に与えられた体系に沿って分類する．

- **規則化・順序付け**

異なる形式のデータを統一的に扱うための変換処理や付加情報（ラベルやタイムスタンプ）の追加を行う．

- **翻訳**

複数の言語圏のデータに対応するためにデータ記述言語の統一化を行う．または利用者の可読性を高めるために各国語に翻訳する．

これらの対象領域によって，あるいはタスクとして与えられる個別のデータの内容によって，ワーカ（作業者）に求められるスキルの種類，レベルは大きく異なる．また得られた成果の品質をどのように担保するか，そして総コストの算出も重要であろう．

筆者らのプロジェクトではクラウドソーシングによるオープンデータ抽出・変換の試みを行っており，両者の親和性に関する検討を進めている[4]．以下ではその取り組みの一端を紹介する．

クラウドソーシングによるレガシーデータの抽出

これまでに述べてきたように，公共オープンデータにおいては編集不可能なファイルや複雑なレイアウトに起因する再利用性の低いデータ等が混在しており，単一の方法ですべてに対応することは難しい．ここでは初期的な検討として，クラウドソーシングとの親和性が高く，かつ再利用のニーズが大きいと思われる，白書に掲載されたグラフ画像を対象とした実証実験について述べる．

白書は国民に対する政策の周知を目的とした政府刊行物であり，現状分析や各種政策の効果の概要がまとめられている．ほかの行政文書・データと比較して読みやすい記述になっていることもあり注目度が高く，引用される頻度も高い．その白書に掲載されるデータは各機関が保有する膨大な情報の中でも代表的なデータであり，利活用の要求が強いものと予想される反面，これらのデータの多くはグラフの形で画像化されており再利用性が低い．そこでクラウドソーシングを用いた画像からのデータ抽出手法について検討を行った．

ここでの目的は，グラフ画像として与えられた情報からデータの項目と値を抽出することである．ワーカには漏れなくかつ正確にデータを抜き出してもらうことが期待される．これに対する単純なタスク設計としては，各ワーカにグラフ画像を提示し，データを CSV 形式で書き出してもらうことが考えられる．しかしながらこのようなタスクには以下の問題が内在している．

1）タスクの出力が数値データの羅列であり，ワーカ自身が入力の誤りに気がつきにくい．

2）CSV には本質的にデータ構造がなく，行・列の反転やヘッダの未入力等が容易に起こり得る．

本研究ではこれらの問題を踏まえて，少ないコ

ストで高い精度が得られるようなデータ抽出タスクの設計を行った．具体的には数値データの出力ではなく，与えられた画像とまったく同じグラフを表計算ソフトウェア上で再現するという課題とした．この方式ではワーカはグラフの再現タスクとして取り組み，リクエスタ（依頼者）はグラフの描画に用いられた数値データを得る．このようにタスク自体の目的とリクエスタの本来の目的を分けることによって，以下のメリットが考えられる．

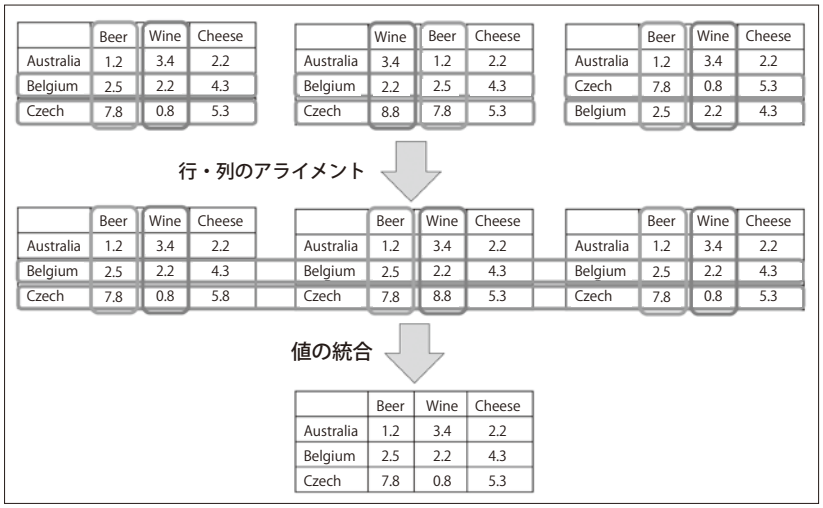


図 -3　表のアライメントと統合

1-1）与えられた画像と再現したグラフを比較することで，ワーカ自身がデータの入力誤りや漏れを検出しやすくなる．

1-2）同様にリクエスタも誤りを含むデータないしは明らかにタスクと無関係なデータを発見しやすい．

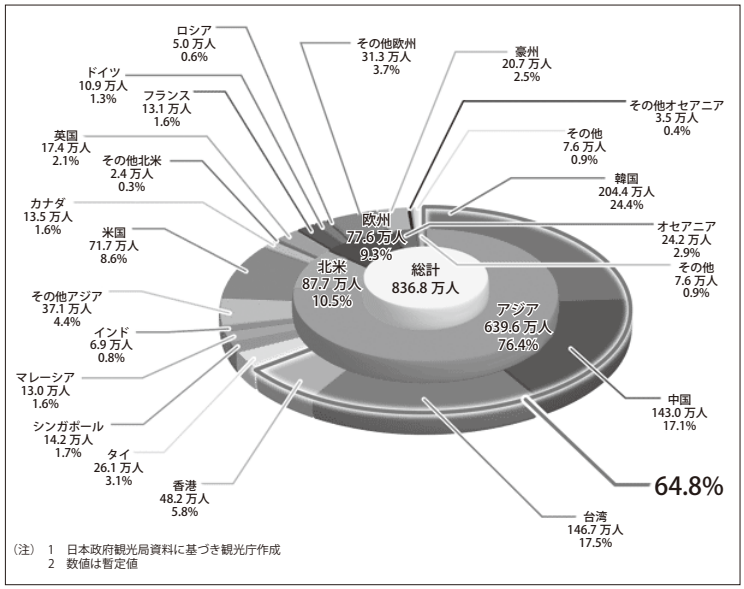
2）グラフを描画する過程でデータの構造そのものが抽出できる．

2）について，本研究で用いた Microsoft Excel ではグラフは複数の系列からなるオブジェクトの集合として扱われており，このオブジェクトのプロパティが表の行ラベルあるいは列ラベルに相当する．ワーカが 1 つの系列を行に沿って記述した場合でも列に沿って記述した場合でもグラフを描画する際に X 軸・Y 軸との対応関係を指定する必要があり，その際に系列オブジェクトとそのラベルが確定することになる．また Microsoft Excel では API を通じて外部のプログラムからこのオブジェクトや値にアクセスできるため，複数の表を横断した作業等が自動化できる利点もある．実証実験の目的の 1 つはこれらのメリットが実際に得られるかどうかの検証である．

一般的にクラウドソーシングにおいてはワーカの作業結果に誤りが混入することは避けられない．この誤りはワーカの知識不足や不注意，あるいは悪意によって生じる．このような前提において一定の品質を担保するために多重化や冗長化が用いられる．これは同じタスクを複数のワーカに依頼し，その結果を多数決などの方法によって統合することを意味する．また本研究が対象とする表形式のデータの統合にあたっては，ワーカごとに作成される表の行・列の順序が入れ替わっている恐れがある．このため，前処理として複数の表の行・列の並びを揃えるアライメントが必要になる（**図 -3**）．同じ系列について 2 人のワーカが抽出した値の傾向は類似していると仮定して，両者の表から任意の 1 行を選択し，その行に含まれる値の類似度を求める．これをすべての組合せに対して行い，最大の類似度となるような対応関係を得る．次に列同士についても同様の処理を行う．行・列ともに対応関係が確定した後は個々のセルの値について，ラベルなどの名目値の場合は多数決で，数値の場合は中央値によって確定させる．後者に中央値を用いる理由は，入力時の桁の誤りといった大幅な外れ値に対して頑健にするためである．

実証実験とその意義

本提案手法の有効性を観光白書平成 25 年版 [5] に記載されている 61 個のグラフ画像を対象とした実証実験によって確認する．これらのグラフには値が明記されているため，実際には値を読みとるだけで正解が得られる比較的容易なタスクである．これら



（注） 1 日本政府観光局資料に基づき観光庁作成
2 数値は暫定値

図 -4 提示されたグラフ画像の例（観光白書平成 25 年版より）

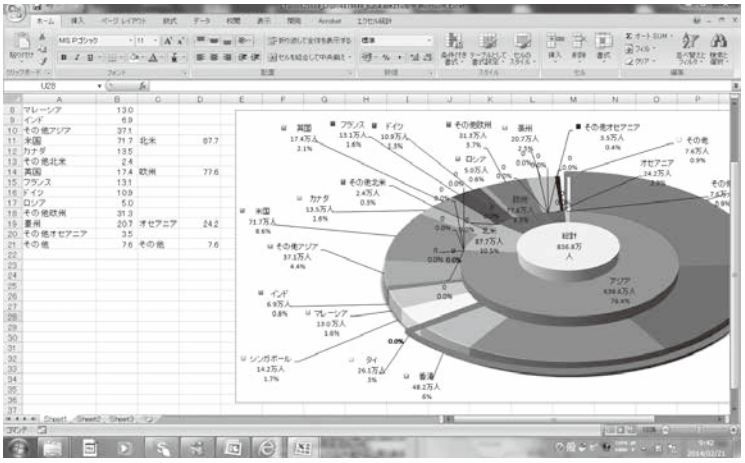


図 -5 グラフ再現タスクの結果の例

の画像に対して，グラフの再現を求めるタスク（以下グラフ再現タスクと呼ぶ）を 1 画像あたり 3 件，数値の書き起こしを求めるタスク（以下数値タスクと呼ぶ）を 1 画像あたり 2 件募集し，得られたデータの精度等を評価した．タスクの単価は 200 円である．グラフ再現タスクに参加したワーカの総数は 20 名，1 画像あたり平均 2.7 件の回答があった．また数値タスクに参加したワーカの総数は 23 名，1 画像あたり平均 1.9 件の回答があった．

グラフ再現タスクで提示した画像（**図 -4**）とその結果（**図 -5**）の例を示す．折れ線グラフや棒グラフ以外の複雑なグラフでも忠実に再現されており，ワーカのソフトウェア操作スキルはかなり高いことが分かる．一方で正しくデータが得られなかった例としては，提示された画像をそのまま提出するといったタスク内容の誤解に基づくもの，ほかのタスクの結果をアップロードした事例が少数存在した．以下ではこれらを除いた，正常に実行されたタスクにおいて抽出されたデータの精度を分析する．

ラベルの抽出精度は数値タスクが 96.8%，グラフ再現タスクが 92.2%，数値の抽出精度は数値タスクが 92.1%，グラフ再現タスクが 94.4% であり，ラベルについては数値タスクが，値についてはグラフ再現タスクが優れた結果となっている．エラーの中には「オーストラリア」を「オーストリア」と入力するといった明らかな誤りのほかに，「1 月」を「1」とだけ入力するような情報の不足，当該の項目が未入力であるといったパターンが見られる．**表 -1** にタスク種別とエラーの分類を示す．明らかな誤りに分類されたエラーに注目すると，数値タスクよりもグラフ再現タスクの方が誤り率が低下している．これはグラフの再現によってワーカ自身が入力内容を確認できるために低下したものと推測される．実際に数値データの平均二乗誤差は数値タスクで 28.4，グラフ再現タスクで 0.55 と大きく差があり，桁の間違いの影響が出ている．一方で，未入力についてはグラフ再現タスクの方がエラー率が高い．グラフ画像の再現に関係のないデータが欠損している例や，凡例やラベルを表ではなく描画されたグラフに直接テキストボックスを使用して記述している例が多い．

複数人が作成した表をアライメントを通じて統合することの効果については，個別のデータ抽出精度が 90% 前後を推移している一方で，3 名の表の統合や 5 名の統合によって 95% 以上の精度を得ることに成功している．

以上の結果より，グラフ画像からのデータ抽出を

タスク種別	セルの種類	明らかな誤り	情報の不足	未入力
数値タスク	ラベル	3.4%	2.4%	3.9%
	数値	17.4%	2.3%	2.3%
グラフ再現タスク	ラベル	2.5%	5.0%	6.4%
	数値	10.9%	1.6%	8.2%

表 -1　タスク種別とエラーの分類

クラウドソーシングで実現することは原理的に可能であるが，精度を高めるには適切なインストラクションが必要であることが明らかになった．グラフの再現は数値データの誤りを軽減することに貢献するが，ラベル等を網羅的に抽出するためには直接的な指示が必要である．また抽出されたデータ構造に基づいて複数のワーカの作業結果を統合することで精度が向上することも確認できた．本研究は初期的な検討であり，今後の課題としては数値が記述されていないグラフへの対応がある．このためにはグラフの自動読みとりソフトウェアを活用するとともに，目視によるクラウドソーシングによって大きなエラーが生じないようにするなど，複合的な対応が必要であると思われる．

オープンデータの取り組みから見た本研究の意義としては，過去のデータを変換するためのコスト算出に寄与できた点が挙げられる．変換作業の単価が定められたことで，オープンデータ化のための投資額と得られるリターンの関係が明確になる．また，より再利用性の高い 4 つ星・5 つ星データの生成については，現在セマンティック Web の分野で検討されている RDF Data Cube[6] への対応が考えられる．RDF Data Cube では次元・属性・測度の 3 つの要素からなるグラフ構造として捉えることでセマンティクスを保持することが可能である．クラウドソーシングによってこれらの要素を抽出し，そこから機械的な変換によってグラフ構造を入手することで高い互換性を持つデータを半自動的に得ることが可能になると思われる．

クラウドソーシング研究としての本提案の特徴は，本来の目的とワーカに教示する目的が直接関連しないようなタスク設計によって得られる結果の品質を高めた点にある．グラフの再現というタスクはワーカにとって自己完結的であり，結果の評価も容易である．あらゆる問題に対してこのようなインストラクションとフィードバック機構を与えられるとは限らないが，今後も多様なデータの変換タスクに取り組むことでオープンデータとクラウドソーシングの親和性に関する検討を深めていく所存である．

参考文献

1）庄司昌彦 : オープンデータ活用，編集にあたって，情報処理，Vol.54, No.12, pp.1202-1203 (Dec. 2012).
2）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）別添 2，https://www.kantei.go.jp/jp/singi/it2/densi/kettei/data/gl26_betten2.pdf
3）Simperl, E., Norton, B. and Vrandečić, D. : Crowdsourcing Tasks in Linked Data Management, 2nd Workshop on Consuming Linked Data（COLD 2011）Co-located with the 10th International Semantic Web Conference（ISWC 2011）(2011).
4）小山　聡，馬場雪乃，大向一輝，堂腰裕明，鹿島久嗣 : クラウドソーシングを用いたレガシーオープンデータの機械可読化，信学技報，Vol.114, No.181, AI2014-11, pp.1-6 (2014).
5）観光白書平成 25 年版, http://www.mlit.go.jp/common/001013847.pdf
6）The RDF Data Cube Vocabulary, http://www.w3.org/TR/vocab-data-cube/

（2015 年 6 月 25 日受付）

大向一輝（正会員）■ i2k@nii.ac.jp

国立情報学研究所准教授．博士（情報学）．セマンティック Web やソーシャルメディア，オープンデータの研究とともに，学術情報サービス CiNii の開発に携わる．著書に『Web がわかる本』（岩波書店），『Web らしさを考える本』（丸善出版）がある．