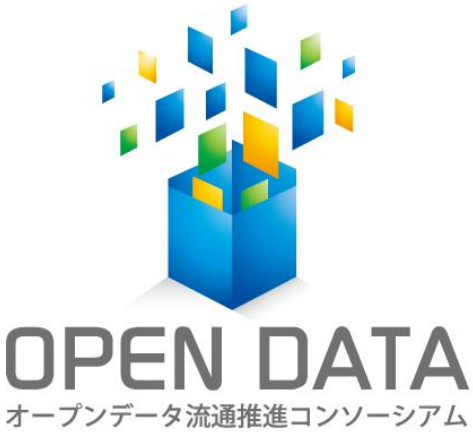
資料1-6



平成25年度第1回 技術委員会資料

オープンデータ流通推進コンソーシアム「オープンデータ化のための技術ガイド」作成案

2013.12.04

オープンデータ流通推進コンソーシアム 事務局

背景と目的

作成の背景

- 政府等によるオープンデータ化の推進
  - ✧電子行政オープンデータ推進のためのロードマップ（電子行政オープンデータ実務者会議）
  - ✧Open Data METI（経済産業省）
  - ✧情報通信白書のオープンデータ化（総務省）
  - ✧自治体によるオープンデータ化への取組（鯖江市・横浜市・流山市など）

- 政府・自治体職員がオープンデータ化を実施するうえでのガイドが必要
  - ✧当委員会において「オープンデータ化のためのデータ作成に関する技術ガイド」を作成
  - ✧電子行政オープンデータ実務者会議において「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」を作成
    - ◆本文書の内容は「実務者会議の議論の進展や関連技術の進展を踏まえ、随時改定する」とある（電子行政オープンデータ実務者会議・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）の概要）

作成の目的

- 昨年度のガイド・留意事項文書を精査し、政府・自治体職員がオープンデータ化を実施するうえで使いやすいガイドの作成を目指す。

「オープンデータ化のための技術ガイド」目次案

1. はじめに
   - 技術ガイドの位置づけや、記載概要を示す。
2. オープンデータ化の意義
   - オープンデータ化の背景と経緯を述べる。

データガバナンス委員会で検討中の、データガバナンスのガイドと整合性をとる。

3. オープンデータに関する技術背景・要求
   - オープンデータ化に際して参考になる技術や規格を列記し、それらを解説する。
     ✧識別子に関する規格についても、ここで解説する。
4. オープンデータ化のための技術的指針
   - 表形式／文書／地理データ／リアルタイムデータのそれぞれの形式ごとに、オープンデータ化を行う上での留意事項や推奨事項を解説する。
     ✧昨年度のガイドは、この部分のみが記載されていた。
   - メタデータを記述するための手法や留意事項、推奨事項を解説する。
     ✧Word/Excel/PDF等のプロパティ、Simple Data Format、データカタログの表現など。

昨年度版のガイドに追加する。

■ 上記構成を提案する理由

- 実務担当者がオープンデータ化を進める際には、オープンデータ化の技術背景を理解する必要がある。このため、これらの背景に関する解説を追記する。
- データ形式については、国内外で広く利用されている規格がすでにある。それらの調査結果を技術ガイドに反映させる。

参考とする識別子規格

- 調査の指標（← データを識別するうえでの要求事項）
  - URI表現可能性（RDFで利用できるか）
  - 唯一性保証の方法
  - 識別対象
  - 永続性
  - ID長（可変・固定）
  - 他の識別子体系の取り込み可能性
  - 運営主体（利用するための手続きなど）
  - 連番・分散管理の可能性　など
- 調査対象とする識別子案
  - ucode [ITU-T H.642.1]
  - EPC SGTIN/SSCC/SGLN
  - DoI (Digital Object Identifiers)　[ISO 26234]
  - UUID [ISO/IEC 11578]
  - ISBN [ISO 2108] / ISSN [ISO 3279]
  - 企業コード [ISO 6523]
  - 国名コード [ISO 3166-1] / 行政区画コード [ISO 3166-2]
  - OpenID
  - RFIDの固有ID [ISO/IEC 15963]　など
- 調査結果は次回報告する

参考とするファイル形式規格

- 表形式データ
  - Common Format and MIME Type for Comma-Separated Values (CSV) Files [RFC 4180]
  - Simple Data Format
  - Linked CSV
- 地理空間データ
  - GML
  - KML
  - shape
- リアルタイムデータ
  - Stream API
  - GTFS（General Transit Feed Spec）Realtime

表形式データに関する参考企画 (1/3: RFC 4180)

RFC4180(*1)の概要

- CSV（Comma-Separated Values）ファイルの書式と、それに関連づけられるMIMEタイプ（text/csv）を規定している。
  - ✧CSV形式の仕様と実装は多岐に渡っており、公式な仕様はない。RFC4180は、殆どの実装が解釈可能なCSV形式の書式を規定している。
- CSV形式のフォーマットだけでなく、ヘッダ行に関する規定もある。
  - ✧There maybe an optional header line appearing as the first line of the file with the same format as normal record lines. This header will contain names corresponding to the fields in the file and should contain the same number of fields as the records in the rest of the file (the presence or absence of the header line should be indicated via the optional "header" parameter of this MIME type).

平成24年度版技術ガイドとの関連

- RFC4180では、ヘッダ（表のタイトル部分）を最大1行にするように求めている。
- 一方、平成24年度版技術ガイド9「データセルの内容・単位・記数単位を示すタイトルが、それぞれ別の行に記載されている」を満たすと、RFC4180に準拠しなくなる。 → Simple Data Format

(*1) Y. Shafranovich. Common Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180, 2005. http://tools.ietf.org/html/rfc4180

表形式データに関する参考企画 (2/3: Simple Data Format)

- Simple Data Format(*2)の概要
  - Data Packages(*3)やJSON Table Schema(*4)等の規格を参照している。
  - 以下のようなフィールドを利用して、CSV形式データの（メタデータ）定義をJSON形式で行う。
  - name（データ名）
  - licenses（ライセンス）
  - datapackages_version（バージョン）
  - resources（CSVファイルの定義）
    - ✧url（データのURL）
    - ✧path（データのパス）
    - ✧schema（urlまたはpathが示すCSVデータの定義）
      - ◆ fields（CSVデータのカラム定義）
        name（カラム名）
        type（データ型／string, number, integer, date, time, datetime, boolean, binary, object, geopoint, geojson, array, any）
        description（カラムの説明）
  - フィールド名にボキャブラリを割り当てれば、RDFによるメタデータ表記にもなり得る。

(*2) http://dataprotocols.org/simple-data-format/
(*3) http://dataprotocols.org/data-packages/
(*4) http://dataprotocols.org/json-table-schema/

表形式データに関する参考企画 (2/3: Simple Data Format)

Simple Data Formatによる記述例

```
{
 "name": "my-dataset",
 "resources": [
  {
   "path": "data.csv",
   "schema": {
    "fields": [
     {
      "name": "var1",
      "type": "string"
     },
{
      "name": "var2",
      "type": "integer"
     },
     {
      "name": "var3",
      "type": "number"
     }
    ]
   }
  }
 ]
}
```

データセット名 “my-dataset”

データファイルのパス情報 “data.csv”

カラム定義
第1カラム: 「var1」という名前の文字列情報
第2カラム: 「var2」という名前の整数情報
第3カラム: 「var2」という名前の数値情報

表形式データに関する参考企画 (3/3: Linked CSV)

Linked CSV(*5)の概要

- RDF化を意識したCSVデータを記述フォーマットを規定しようとしている。
  - ✧ヘッダとデータ本体の間にメタ情報（type、see、langなど）を記述する。
  - ✧記述例（可読性を確保するためにCSVデータを表形式で示す）

	#	country	year	population
ヘッダ				
メタ情報	type	url	time	integer
	meta	index	url	/populations
	meta	license	url	http://creativecommons.org/publicdomain/mark/1.0/
データ本体		http://en.wikipedia.org/wiki/Afganistan	1960	9616353
		http://en.wikipedia.org/wiki/Afganistan	1961	9799379

RDF化

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>
@prefix rel: <http://www.iana.org/assignments/relation/>
@prefix : <http://example.org/af-population#>

<>
 rel:describedby
  <http://example.org/af-population#row:1>,
  <http://example.org/af-population#row:2> ;
 :index <populations> ;
 :license <http://creativecommons.org/publicdomain/mark/1.0/> ;
```

```
[ rel:describedby <http://example.org/af-population#row:3> ;
  :country <http://en.wikipedia.org/wiki/Afghanistan> ;
  :year "1960"^^xsd:gYear ;
  :population 9616353 ]

[ rel:describedby <http://example.org/af-population#row:4> ;
  :country <http://en.wikipedia.org/wiki/Afghanistan> ;
  :year "1961"^^xsd:gYear ;
  :population 9799379 ]
```

- ただし、厳密にはこの記法はRFC4180に準拠していない

(*5) Jeni Tennison. Linked CSV, 2013. http://jenit.github.io/linked-csv/

地理空間データに関する参考規格

- GML (Geography Markup Language)
  - Open Geospatial Consortium (OGC)によって開発された、地理的特徴を表現するXMLベースのマークアップ言語。
  - ISO 19136として標準化されている。
  - データ構造は、RDFに準拠している。
  - 平成20年4月から国土地理院が提供している基盤地図情報は、この形式で提供されている。
- KML (OGC KML)
  - Open Geospatial Consortium (OGC)が規格化する、地理的特徴を表現する XMLベースのマークアップ言語。
  - Google EarthやGoogle Maps、Google Mobileなどで利用されている。
- shape
  - 米国ESRI社の提唱する、ベクトル形式のGIS標準データフォーマット形式。
  - 国際標準化規格ではないが、業界標準フォーマットの1つになっている。

リアルタイムデータに関する参考規格

- Streams API(*6)
  - サーバ・クライアント間でのHTTPコネクションを継続し、値が更新されるごとにその結果を返す仕組み。
  - TwitterやTransport for Londonなどで利用されている。
- GTFS（General Transit Feed Spec）Realtime
  - GTFSは、公共交通機関の時刻表とその地理的情報に使用される共通形式。
  - GTFS Realtimeは、公共交通機関が運行車両に関するリアルタイムの最新情報をアプリケーション デベロッパーに提供できるようにするためのフィードの仕様。

(*6) Feras Moussa. Streams API. 2013. http://www.w3.org/TR/streams-api/

「オープンデータ化のための技術ガイド」: 昨年度部分の精査方針案

表形式データ

- 以下に代表される、各種規格との整合性・互換性を確認し、必要な修正を加える。
  - ✧RFC4180: Common Format and MIME Type for Comma-Separated Values (CSV) Files
  - ✧Simple Data Format

地理空間データ

- 既存のフォーマットや、それらの利用方法についての解説を追加する。
  - ✧GML
  - ✧KML
  - ✧shape　など

文書データ

- 文字列の抽出・検索に関する方式を調査し、必要な修正を加える。

リアルタイムデータ

- リアルタイムデータの記述・配信機構に対応した代表的な規格に関する解説を追加する。
  - ✧Stream API
  - ✧GTFS（General Transit Feed Spec） Realtime　など