政府 CIO 補佐官等ディスカッションペーパー
2021 年 8 月

メタデータルールと利用イメージの検討

2021 年８月

政府 CIO 補佐官
下山　紗代子、　関　治之、平本　健二、中村　弘太郎

内閣官房情報通信技術（IT）総合戦略室
長谷川　亮、　海老原　立弥

独立行政法人情報処理推進機構（IPA）
デジタルアーキテクチャデザインセンター（DADC）
竹本　和弘、　藤本　勝裕、　的場　大輔

要 旨

データ社会では様々なデジタルデータが活用、連携されるようになる。そのデジタルデータを見つけやすく、選定しやすくするために、データには見出し情報であるメタデータが付けられる。このメタデータは、各種コンテンツ、オープンデータ、ベース・レジストリなどで共通化されることで価値を増大させることができる。

本ペーパーではメタデータ導入や活用の考え方を示すとともに、その導入や活用のための方法を提案する。

目 次

１　はじめに

１．１　背景

行政機関は国内最大のデータホルダーであるが、そのデータは行政内部でも社会全体でも十分に使いこなせているわけではない。そもそも、各行政機関がどのようなデータを保有し、どこまで活用可能なのかということは、その組織外からわからないだけでなく、同じ組織内でもわかっていないことは多い。

世界各国では、行政機関の保有するデータを管理する必要性に早くから気がつき、W3C の定義するデータカタログのメタデータの標準である DCAT[1]を中心にデータ管理の取組を進めている。

また、従来のコンテンツの集積拠点としては図書館があり、既に体系的にコンテンツを管理している。他にも、教育、地理空間等の分野独自のメタデータを整備している分野もあり、イベントデータや施設等のデータも体系化が図られつつある。

これらのメタデータの基本構造は類似しており、相互運用性を確保することも可能と考えられている。

１．２　課題

現在、国内の行政機関でメタデータの標準が存在しないため、データホルダー毎に自由にメタデータを管理しており、また、メタデータ自体が存在しない場合もある。このようにメタデータの活用が統一されていない状況となるため、データ利用者、データを管理する行政職員の双方に負担がかかるとともに、データを効率的に活用することができていない。

利用者が感じる課題
- データを見つけられない
- データの存在があるかどうかもわからない
- 関連するデータに何があるのかわからない
- データを見つけてもライセンスなどの利用条件がわからない
- データの更新周期がわからない
- データの品質がわからない

行政職員が感じる課題
- 他の行政機関が持っている情報を活用して業務を高度化、効率化したい
- データに関する問い合わせへの対応に負荷がかかる

[1] https://www.w3.org/TR/vocab-dcat-2/

１．３　目的

本ペーパーは、行政の保有するデータを効率的に管理、活用できるようにし、コンテンツの利用や流通の促進を図るとともに重複の防止を図るためのメタデータの付与方式を検討することを目的とする。

コンテンツにメタデータを付けることにより、カタログサイトだけではなく、データの取引市場、検索サイト、個々のコンテンツの利用等、様々な場面でデータが使いやすくなると考えられる。

１．４　概要

メタデータとは、データを管理するために使われるデータであり、データに関するデータと言われている。

データは目的に応じた集合体で管理されることが多く、その集合体をカタログという。メタデータでは、そのカタログ情報も含め、そこにどのような情報が含まれるのかを階層的に管理するものであり、以下の構造で表される。

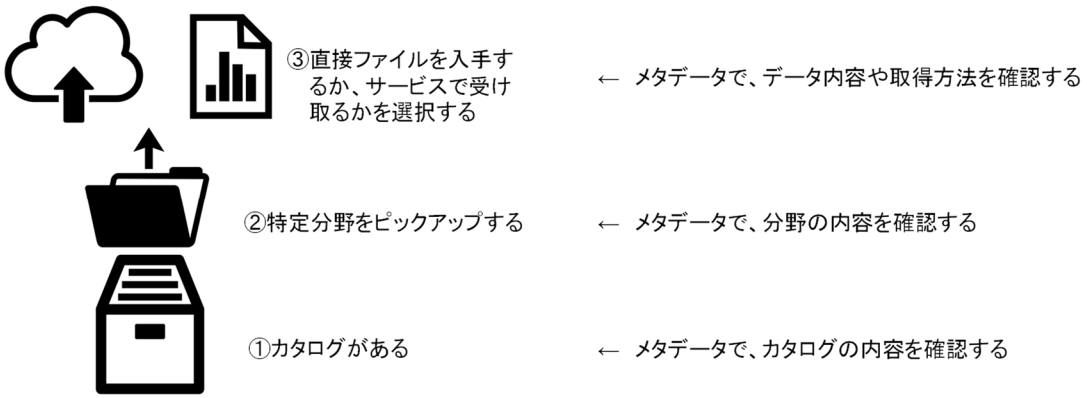

図 1　メタデータのイメージ

メタデータは、カタログとデータセットを中心に構成され、以下の構造を持つ。

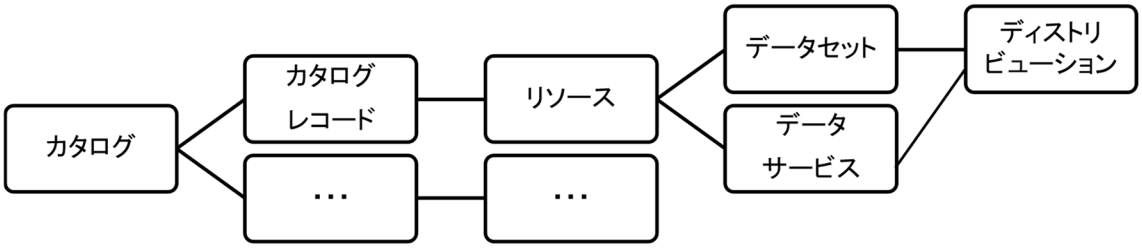

図 2　メタデータの構造イメージ

1)カタログ

特定目的に対するデータの集合体のことをカタログと呼ぶ。カタログは、より上位のカタログを持つことや、カタログ内にサブのカタログを持つこともある。

2)カタログレコード

カタログの中で分野を設定したいときに持つ。カタログ冊子の章のようなものである。

3)リソース

カタログの中のデータに関する項目である。データ項目は持たず、後述するデータセットやデータサービスにより構成される。

4)データセット

公開される情報の束であり、複数のデータの集合体であることもある。品質やライセンス等、利用するための情報を含む。

5)データサービス

サブスクリプションや API でデータを入手するための詳細情報である。サービス形式などの情報を含む。

6)ディストリビューション

個々のデータ単位の情報である。配信するための技術的内容などを含む。

DCAT[2]のクラスやプロパティは以下の構成である。

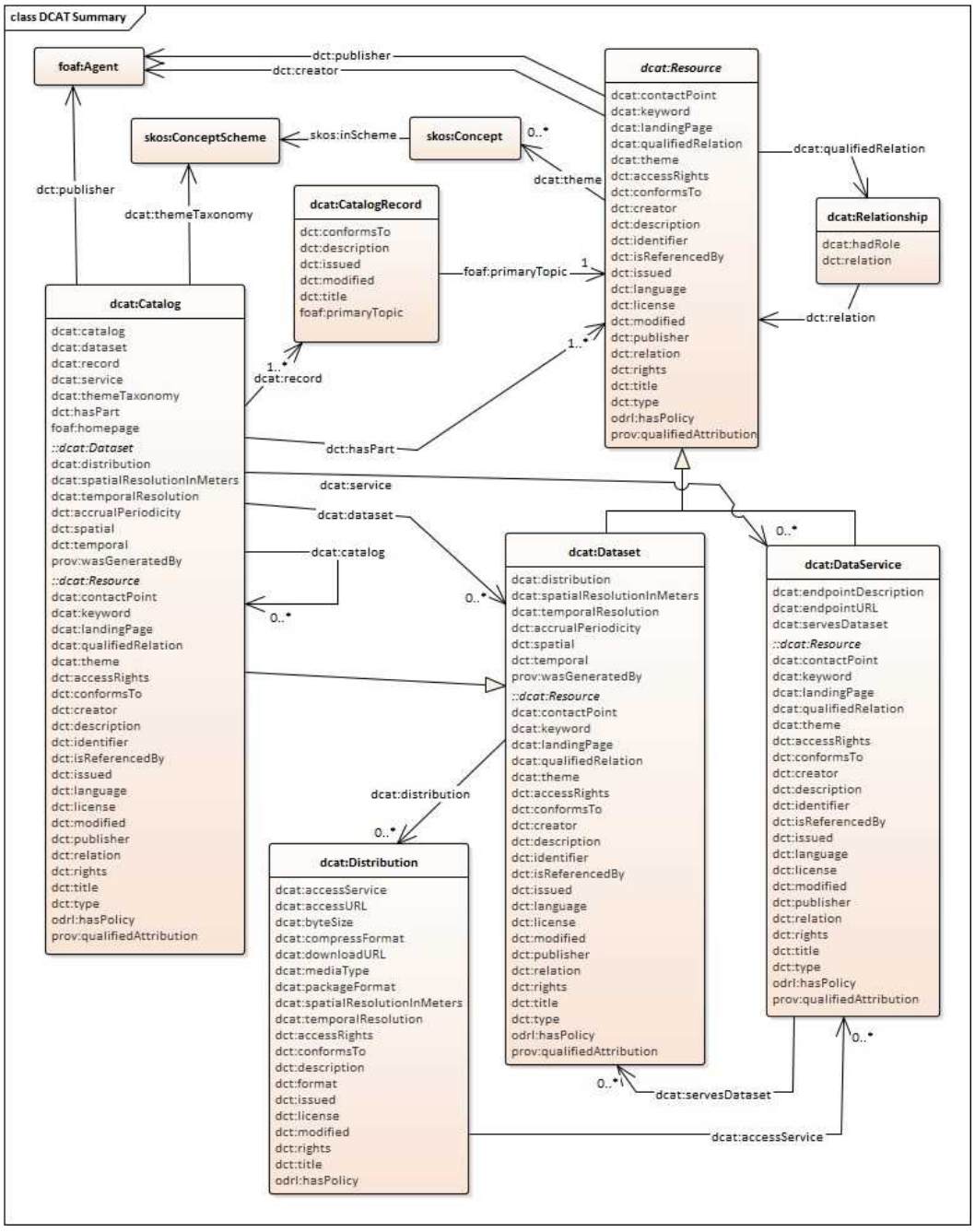

図 3　詳細メタデータ（DCAT）

[2] https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/

2　様々な分野での活用

広範な対象でメタデータが共通的に使えるように、様々な分野で検証を行った。

2.1　ベース・レジストリでの検討

国は今後ベース・レジストリを整備して行く予定である。日本のベース・レジストリは複雑な構造を持っており、法人の情報では、商業登記情報に法人名と本店所在地があるが、国税庁で法人番号を付番するとともに、漢字情報を一般のデジタル機器で使える範囲に代替するなどのクレンジング処理を行い公開している。また、その他のベース・レジストリ情報を統合して、経済産業省の gBizInfo[3]が情報の公開を行っている。このように関係性が分かりやすいものを明確にカタログ化して行くことが必要であり、そのためにはメタデータが重要である。また、メタデータでデータに関する情報をわかりやすく提供する際には、品質情報を含む利用を検討するために重要な情報も明確にして行く必要がある。

そのため、ベース・レジストリのカタログ化のためのメタデータが必要となり、欧州のベース・レジストリ用のメタデータである BregDCAT[4]をベースに検討を行った。

1)カタログやデータセットの単位に関する検討

例えば、医療機関のベース・レジストリを例として表現すると、厚生労働省が医療機関の一覧として各地方局単位で公開している「保険医療機関・保険薬局の指定等一覧及び保険医・保険薬剤師の新規登録一覧」から、以下のように表現される。

[3] https://info.gbiz.go.jp/
[4] https://joinup.ec.europa.eu/collection/access-base-registries/solution/abr-bregdcat-ap

図 4　医療機関の一覧

このページを、１つのデータセットととらえた場合は、以下のように、データを表現することができる。

カタログ　　：保険医療機関・保険薬局の指定等一覧及び保険医・保険薬剤師の新規登録一覧
データセット　　：関東甲信越厚生局管内
ディストリビューション　　：茨城県医科（PDF）
：茨城県医科（歯科併設）（PDF）
：茨城県歯科（PDF）
：茨城県歯科（医科併設）（PDF）
：茨城県薬局（PDF）
：栃木県医科（PDF）
：・・・

しかし、このようにするとディストリビューションが多過ぎて検索性が低く

なってしまう。また全国の都道府県単位で活用したいこともあることから、都道府県単位に以下のように表現することもできる。

カタログ	：保険医療機関・保険薬局の指定等一覧及び保険医・保険薬剤師の新規登録一覧
カタログレコード	：関東甲信越厚生局管内
データセット	：茨城県
ディストリビューション	：医科（PDF）
	：医科（歯科併設）（PDF）
	：歯科（PDF）
	：歯科（医科併設）（PDF）
	：薬局（PDF）

また、医科、歯科、薬局のようにカテゴリで表現することも可能である。

カタログ	：保険医療機関・保険薬局の指定等一覧及び保険医・保険薬剤師の新規登録一覧
カタログレコード	：関東甲信越厚生局管内
データセット	：医科
ディストリビューション	：茨城県医科（PDF）
	：茨城県医科（歯科併設）（PDF）
	：栃木県医科（PDF）
	：栃木県医科（歯科併設）（PDF）
	：・・・

カタログレコードを使わずに、以下のようにカタログを親子形式でネスト化[5]して表現することも可能である。

親カタログ	：保険医療機関・保険薬局の指定等一覧及び保険医・保険薬剤師の新規登録一覧
カタログ	：関東甲信越厚生局管内
データセット	：茨城県
ディストリビューション	：医科（PDF）
	：医科（歯科併設）（PDF）

[5] あるカタログがある際に、カタログを束ねた上位のカタログがある場合や、サブ分類による下位のカタログがある場合がある。

：歯科（PDF）
：歯科（医科併設）（PDF）
：薬局（PDF）

このように、カタログは目的に応じて編集が可能で、様々な対象の記述のために自由度を持ち、かつ、検索の利便性も維持した形式にする必要がある。

2)データ項目のコード化、詳細化

DCAT をベースに項目の検討をして行くと、項目の記述方法に困ることがある。

例えば、公開者（Publisher）という項目は、Agent というデータ項目で表現することとなっている。つまり、厚生労働省が Publisher の場合、Agent としての詳細情報として「名称」「法人番号」「所在地」等の情報を持つことができるが、どこまでデータ項目を持つべきかと行った問題がある。また、DCAT の品質情報は自由記述になっている。しかし、自由記述ではデータ利用に関する品質評価を客観的に行うことが困難であり、さらなる詳細化が求められている。

このようなデータ項目に関して実装のための記述方法の検討が必要である。

２．２　教育分野での検討

教育分野では、教材のメタデータである IEEE LOM[6]やコースウェアのメタデータである IMS の CC（Thin Common Cartridge）[7]が存在する。このように、既存のメタデータがある場合には、その整合性が重要となる。

1)既存メタデータとの比較

既存のメタデータである IEEE LOM と DCAT、及び、以前文部科学省が運用を行っていた教育コンテンツデータベースである NICER のメタデータ[8]との比較を試みた。以下は IEEE LOM と DCAT の比較の簡易整理である。

[6] https://www.ieeeltsc.org/working-groups/wg12LOM/lomDescription/
[7] https://www.imsglobal.org/activity/common-cartridge#:~:text=Thin%20Common%20Cartridge%20%28Thin%20CC%29%20is%20a%20profile,CCs%20only%20contain%3A%20Learning%20Tools%20Interoperability%20%C2%AE%20Links
[8] https://nier.jp/papers/6-8.pdf

大項目	内容	互換性	
		LOM	DCAT
1. 基本情報	学習内容についての基本的な情報。タイトル、説明、キーワードなど。	✓	✓
2. 教育情報	学習内容の教育面に関する情報。学習タイプ（講義/ワークショップ/テスト）、対象者（学習者/教師/保護者）など。	✓	
3. 分類情報	学習内容の分類情報。学習指導要領コードなど。	✓	
4. 関連情報	学習内容に関連する資料等の情報。教科書の該当する章、ページ等。	✓	✓
5. 技術情報	学習内容の技術面に関する情報。コンテンツの形式、ファイルサイズ、URLなど。	✓	✓
6. 権利情報	学習内容の権利に関する情報。ライセンス、利用条件など。	✓	✓
7. 状態に関する情報	学習内容の公開状態等の情報。バージョン、作成者、公開日など。	✓	✓
8. リスト管理情報	学習内容情報単体に関する情報ではなく、学習内容情報をまとめた教材リスト等の一覧データ自体についての情報。	✓	✓
9. 注釈情報	学習内容に対する注釈（コメントや申し送り事項等）に関する情報。	✓	✓

図 5　メタデータ比較の例

保有するコンテンツに DCAT のメタデータがついていたとしても、LOM をベースとした教育システムでは必要な情報が不足することがある。また、連絡先などのデータ項目の構成や表記が LOM と DCAT の実装で異なる場合もある。その場合には、データのコンバータやタグ追加の仕組みなどが必要となってくる。

2)共通的なタグの必要性

教材には学習指導要領コードの付与が推進されているが、一般のコンテンツに学習指導要領コードは付与されていない。そこで汎用的なコードとの連携が重要になる。

国内で広く普及しているコードが望ましく、その点からは、統計を中心に使用される日本標準産業分類と図書館の分類に使用される日本十進分類法である NDC[9]の汎用性が高いと考えられる。また、グローバルなデータ流通を考えても、このような汎用性の高いコード体系を参照しておくことが重要になる。

[9] https://www.ndl.go.jp/jp/data/NDC10code201708.pdf

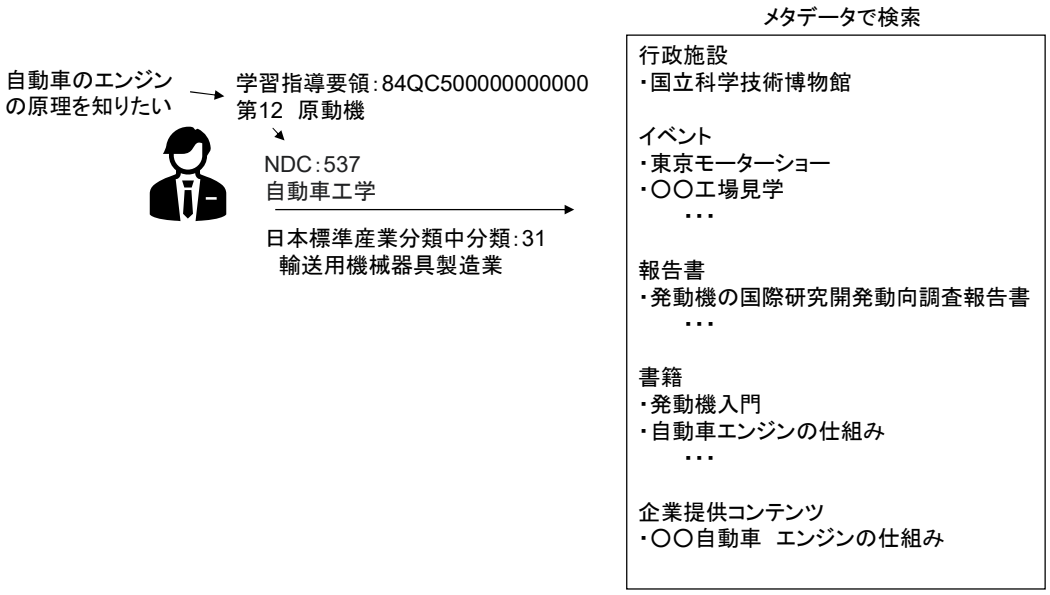

図 6　共通的なタグ活用のイメージ

２.３　会議資料での検討

行政機関の会議資料は、国内外の動向をまとめているなど価値のあるデータであるが、報告書のみしか参照されないことが多い。また、インターネットで検索すると会議資料が検索結果として提示されるが、そこにメタデータは付与されていない。特に、「資料１」と記載されているが、その会議体が何かわからないということが多く、メタデータの付与が求められている（昨今の行政サイトはディレクトリ構造ではなくパーマリンク[10]が使われることが多いため、関連情報を探すことが難しくなっている）。

会議資料のメタデータは、以下のように表現できる。

カタログ	: 〇〇会議
カタログレコード	: △△タスクフォース
データセット	: 第一回〇〇会議△△タスクフォース
ディストリビューション	: 資料１
	: 資料２
	: 参考資料１
	: データ

[10] 資料の URL を、乱数等を使って一意に特定する仕組みである。URL がサイトの構造と独立しているため、サイトの更新があってもコンテンツのリンク切れを防ぐことができる。

：議事録

また、カタログをネスト構造にして会議資料のメタデータを以下のように表現することもできる。

親カタログ	：〇〇会議
カタログ	：△△タスクフォース
データセット	：第一回〇〇会議△△タスクフォース
ディストリビューション	：資料１
	：資料２
	：参考資料１
	：データ
	：議事録

どちらの表記も可能であるが、利用者の利便性を考えると、会議資料に関して統一的なルールを作ることが望ましい。

２．４　イベントや施設情報との共通化

メタデータはコンテンツに付けられることが多いが、利用者の立場から考えると、「図　6　共通的なタグ活用のイメージ」のように施設やイベント情報と共通的に検索することができることでさらに利便性が向上する。タイトルや概要、タグ等のデータ定義を共通化しておくことで、このような検索も実現可能となる。

２．５　CKAN[11]への実装の検討

データカタログサイトを作る際に、データカタログのオープンソースである CKAN を使用することが多い。しかし、CKAN は完全に DCAT に準拠しているわけではない[12]。実装にあたっては、DCAT エクステンションを使用するが、個々のデータ定義やコード設計含め、メタデータの設計は他のサービスとの連携も含め精査して行く必要がある。

[11] https://github.com/ckan
[12] https://github.com/ckan/ckanext-dcat#rdf-dcat-to-ckan-dataset-mapping

3　メタデータ付与による効果

メタデータ付与の検討を行うと、現場の負荷が増えるため協力することができないという反対を受けることが多い。その作成する瞬間だけを考えると多少作業は増えるが、１時間もかからない簡単な作業である。

一方、メタデータがないと利用者はそのデータを検索するために長い時間を浪費することになる。それでも見つけることができればよい方であり、必要なデータにたどり着けず、そのデータを使用した効果を得られないことも多い。通常は、メタデータを付ける機会は一度きりであるが、利用する機会は複数回に及ぶことが多い。その全体の効果を作成者も考える必要がある。

メタデータ作成者にとっても効果がある。効果を以下に示す。

・過去の情報検索が容易になる
・人事異動の際に引継ぎが容易になる
・問い合わせに対して迅速に対応することができる
・データの利用が増え、政策的な効果が上がる
・データカタログなどを用いて広く広報することができる

それでも、メタデータを付けるのは負荷であると考える部門もある。メタデータは、納品物の一部として納品してもらう仕組みにして、そのデータをアップロードするだけにするといった工夫も可能である。

自分にとってのメリット、利用者にとってのメリット、データの価値最大化の視点から、メタデータの付与を推進して行くことが重要である。

4　まとめ

これまで、データベースやデータカタログを整備する際に個々にメタデータの整備を行ってきた。しかし、分野横断でのデータ活用の可能性が広がってきており、メタデータの共通化やコード共通化の重要性が増してきている。各分野の独自データ項目やコードと共存しながら検索しやすい環境を検討して行く必要がある。

また、カタログは親子関係でネスト化するのか、カタログレコードで分類するのかといったことも様々な事例を見ながら検証を進めて行く必要がある。

５　解説

５．１　データ標準

1)W3C　DCAT 2.0

インターネットに関する標準化団体の W3C が 2020 年４月に公開したデータカタログ用メタデータであり、現在、バージョン３の検討が進められている。

https://www.w3.org/TR/vocab-dcat-2/

2)EC　DCAT-AP v2.0.1

DCAT 2.0 をベースに EC 諸国で導入するために実務的に編集した実装モデルである。

https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/201-0

3)EC　BregDCAT-AP v2.00

欧州委員会が、DCAT をベースにベース・レジストリのカタログ用に作成したメタデータである。

https://joinup.ec.europa.eu/collection/access-base-registries/solution/abr-bregdcat-ap

4)Schema.org

Web 検索のデファクトスタンダードのデータモデルを策定している Schema.org が提供するデータセットのメタデータモデルである。

https://schema.org/Dataset

5)Google 検索　データセット

Google の検索で使用されるデータセットのメタデータである。

https://developers.google.com/search/docs/data-types/dataset?hl=ja