LOD とセマンティック技術を用いたデータ連携の枠組みと課題の考察

A Framework and Issues on Data Association with LOD and Semantic Technology

細見 格 [1]

Itaru Hosomi[1]

[1]NEC セキュリティ研究所
[1]NEC Security Research Laboratories

Abstract: Linked Open Data is progressively popularizing in Japan via various efforts such as governmental open data promotion and the annual “LOD Challenge” contests. However those activities regarding LOD are independent for each other in most cases while one of the unique advantage of LOD is derived from semantic links of the data. This paper shows a framework consists of such activities and their issues on implementation. And the paper also mentions some issues and their solutions on data association at a test-bed project for public safety along with referring related activities in security domain.

1. はじめに

2011 年から続く Linked Open Data Challenge（LOD チャレンジ）や各地域で行なわれているハッカソン、国・自治体からのオープンデータ発信の取り組み、LinkData.org といったパブリックなデータ共有リポジトリの提供などにより、国内の Linked Open Data (LOD)が徐々に蓄積されてきている。また、LOD を用いたアプリケーションも LinkData.org などに蓄積され、自由に利用できるようになってきている。一方、LOD がその真価を発揮する鍵と期待されている、データ間の意味的リンクや、そのようなリンクを活用したアプリケーションの開発・展開は、これからといった状況である。データ間のリンクには、同一と見なせるリソース同士の owl:sameAs プロパティによるリンクがあるが、加えて、共通のクラス（上位概念）を持つインスタンス同士のリンクや他の意味的相互関係でもリンクすることにより、さらに幅広いデータ活用や新たなアプリケーションの創出に繋がると考えられる。しかし、データのオープン化に比べてデータ間のリンクやこれを活用した領域横断的なアプリケーションの開発には、新たなモチベーションやビジネスモデルが必要だろう。

これまで国内の LOD やそのアプリケーションが対象としてきた領域は、国や地域の情勢や文化、ライフサイエンス、図書/文献アーカイブに関するものが多くを占めていると見られる[1][2]。

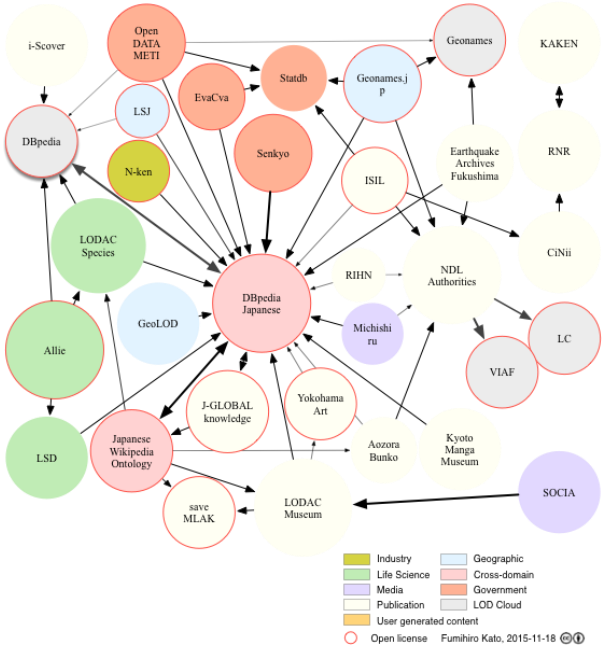

図 1　国内 LOD クラウド（JLDC 2015/11/18 版）

かたや、海外を含む産業界では、必ずしもオープンデータではないものの、RDF を用いたデータ形式の共通化により、リンクや SPARQL によるデータ連携/統合に積極的に取り組んでいる領域がある。その 1 つとして、本稿ではセキュリティやセーフティの領域におけるデータ連携の例を挙げ、LOD の世界におけるデータ連携促進のヒントを探る。

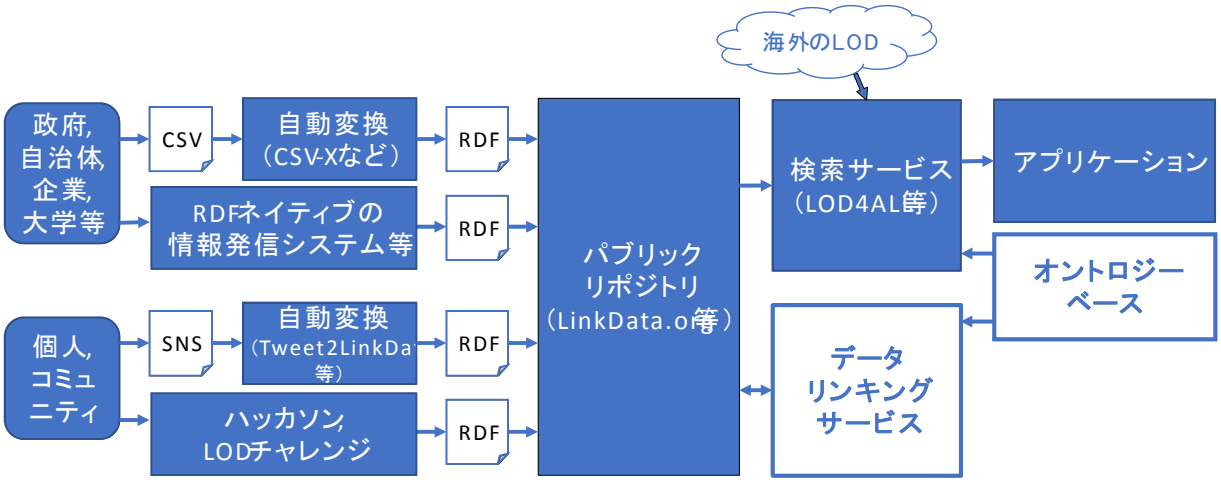

図2　国内LOD関連活動の全体像俯瞰図

2. 国内 LOD 関連活動の構造的俯瞰

2.1. LOD 関連活動の全体像

図 2 は、1 つの試みとして国内の LOD 関連活動の全体像をまとめたものである。ただし、図中の「オントロジーベース」および「データリンキングサービス」は、LOD を特徴づけるリンクの充実に必要と考えられるものの、現状は継続的に運用されている実体が無い。LOD の普及と活用を支える枠組みには、大きく(1)構築（construction）、(2)蓄積（storage）、(3)利用（consumption/exploitation）の 3 フェーズに分類される各種活動が必要となる。なお、公開や配信は、LOD の場合オープンな状態での蓄積と概ね同義と見なし、(2)のフェーズに含む。図 2 においては、政府や企業などの組織または個人による情報発信とその RDF（LOD）化が(1)の構築フェーズ、オープンなリポジトリへの蓄積とアクセス手段の提供が(2)の蓄積フェーズ、検索や LOD を用いたアプリケーションが(3)の利用フェーズに相当する。図 2 中央に例として記載した LinkData.org では、実際には(1)〜(3)の全フェーズをカバーするプラットフォームサービスが提供されている。

2.2. リンクされたオープンデータの意義

前節でまだ継続的運用実体が無いとした「オントロジーベース」および「データリンキングサービス」は、どのようなものになるか。オントロジー自体は、様々なものが公開され恒久的な URL を持つものもあるため、「オントロジーベース」はある程度提供されているとみなすこともできるが、要件に適したオントロジーを容易に見つける手段、特に日本語のリソースに対応したものは未整備である。（RDF ベースの）LOD は、リンクとオントロジーによってデータ同士を意味的に関係付けることができる。検索サービスや独自の検索機能を用いてデータを集め、そのデータ同士を目的に応じてシステム内で関係づけるアプリケーションは数多く存在する。そのように目的や文脈に依存したリンクは各アプリケーションで生成すべきかも知れないが、前述した owl:sameAs のような同一であることを表すリンクや状況に依らず成立する全体-部分関係などを表すリンクは、予め提供されていればそれ自体が検索（参照）条件としても有用だろう。そのような、意味的に関係付けられ任意の用途に利用できるデータであることが、“Linked” Open Data に最も期待したい特徴である。

しかし、データ間のリンクとそのリンクへの意味付け（プロパティの選択）には各データの理解が必要であり、そのための知識と多くの処理を要する。そのコストに見合う有用性とはどのようなものであり、どのような課題の解決が必要かを、産業界、特にセキュリティとセーフティの業界におけるアプリケーションを対象として考察する。

3. 安全と安心のためのデータ連携

3.1. 産業界におけるデータ連携

文献[3]では、産業用アプリケーションとして、出版、文化遺産管理、ヘルスケアそれぞれに関する LOD や Knowledge Graph の構築・運用事例が紹介されている。いずれも、非定型で多様な情報を整理し関連付けることが重要な領域である。これらの比較的オープンな情報を扱う事例については[3]などの

文献を参照頂くこととし、以下ではオープンデータを利用しつつも一般的にクローズドなセキュリティとセーフティ（防犯や災害対策）の業界における Linked Data や Knowledge Graph の利用例を紹介する。この業界では、様々な形式のデータを関連づけて分析する必然性があり、共通データモデルとしての RDF の利用およびオントロジーの構築に関する多くの取り組みがある。災害はいつ何によって生じるかを予め特定することが難しく、テロや犯罪の実行者は監視の網を潜り抜けるように行動する。そのため、あらゆるセンサーや情報共有手段を用いてデータを収集・統合し、複合的視点で事故や事件の予兆をいち早く検知することが重要と考えられている。

3.2 パブリックセーフティのプロジェクト

筆者らは、2013 年 4 月からの約 1 年間、様々なデータを収集・解析して街の治安維持に役立てる実験プロジェクトを行なった。データとしては、街頭のカメラやマイクで収集した映像や音に加え、インターネットで配信されている対象地域の交通情報や SNS などのオープンデータを利用した。プロジェクトは複数の企業で構成されたコンソーシアム単位で実施し、実験用システムも各メンバー企業（計 8 社）それぞれが提供したハードウェア、ソフトウェアを組み合わせて構築した。従って、扱うデータは種類や性質が異なる上に解析システムの開発企業が複数あり、データ連携に関する様々な課題が生じた[4]。

3.3 データ連携における課題

本プロジェクトで構築したシステムの大まかな構成を図 3 に示す。実際にはさらに幾つかのコンポーネントを含むが、ここでは省略している。

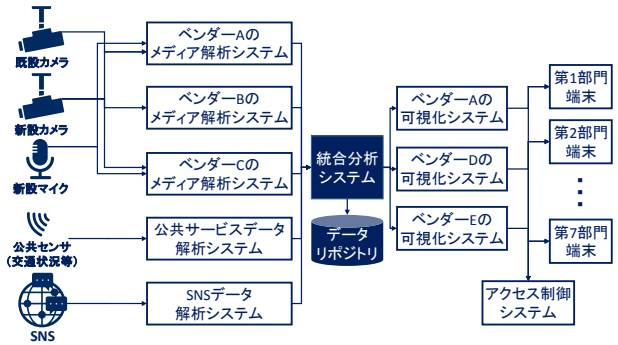

図 3　実験システムの構成

図 3 のように、各種センサーからのデータは各社のメディア解析システムや公共サービスデータの解析システム、SNS データ解析システムでそれぞれ解析し、それらで検知されたイベントは全て統合分析システムに集約される。統合分析システムでは、検知されたイベントのデータ形式を統一すると共に不要なイベントを除去し、関連するイベント同士をリンクさせて結果を各種の可視化システムに送る。その結果は、アクセス制御システムで参照部門ごとの閲覧可否を判定し、閲覧可能な情報のみが各部門の端末上で可視化される。このようなシステムの構築・運用においては、多様なデータを連携させる上で主に次のような課題を解決する必要があった。

(1) どのようにしてデータ形式を統一するか

このプロジェクトでは、データが画像や音、テキストなど様々な上、画像は顔や不審な振舞い等の認識結果、音は声や物音、テキストはツイートや XML 形式の交通情報などであり、更には複数企業の解析システムを用いるなど、メディアの種類も設計も多様であった。しかも、その多くが完成品ではなくプロジェクト期間中に各企業で設計・開発されたため、最初にデータ連携のための標準形式を定義することができなかった。そこで、将来のセンサー追加への対応容易性も考慮し、イベントデータの記述に関する共通の上位オントロジーとメディアの種類毎のドメインオントロジーを構築し、ドメインオントロジーの追加・修正で順次提供される解析システムからの出力データに対応した。上位オントロジーには van Hage らが提案した Simple Event Model（SEM）を用いた[5]。また、例えば同じ画像認識結果でも図 4 のように解析システムによってデータの構造や単位、座標系が異なるため、データ形式の統一には解析結果の形式に対応した個別スキーマと統一スキーマ両方のオントロジーを定義し、前者から後者への変換を行なった。

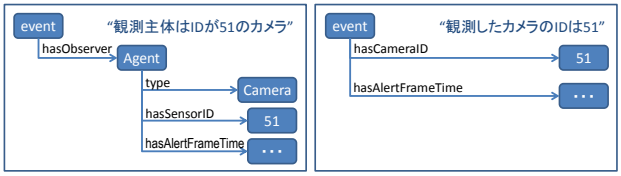

図 4　異なるデータ形式の例

(2) どこの時刻で同期をとるか

図 3 のように、実験システムではカメラ等のセンサーで得たデータから各種の解析システムで何らかの物体や状態に関するイベントの認識処理を行ない、その結果が統合分析システムに送られる。各処理はリアルタイムで実行されるが、現地のセンサーから実験用サーバー室の解析システムまでは有線または無線でデータが送られ、

また図示していないが映像や音声のデータはそれぞれ後で再生するためのビデオ/オーディオ管理システムを経由して解析システムに送られる。ここで、複数のカメラや異種のセンサーで同じイベントを捉えたかどうか判断するための同期用タイムスタンプをどこで打刻するかが課題となる。各センサーは必ずしもタイムサーバーで時計の調整をしておらず、後段の解析システムで受信した時刻は厳密にイベントの発生時刻とは言えない上、ディレイ時間も一定とは言い難い。そのため、両方のタイムスタンプを保持しつつ、統合分析システムで意味的矛盾（起こりえない時間順序等）をチェックするようにした。

(3) 関連付けられたデータを誰にどこまで見せるか

データ形式を統一しイベントの同期をとった後、統合分析システムでは関連イベントの紐づけ（リンク）やイベントの信頼度および重要度の評価を行なう。そして、信頼度と重要度が共に閾値以上のイベント群を各ユーザ端末上に可視化する。しかしながら、ユーザは複数の部門に分かれており、部門ごとに管轄範囲が異なるため、検知されたイベントの中には見て良いものとそうでないものがある。例えば、特定のビルやエリア内の監視カメラ映像は、そこの管理や警備を担当する部門の者にしか見せられない。そこで、ロール（役割）ベースのアクセス制御システムを導入し、検知したイベントとその根拠となる映像などの元データも参照できる場合、検知内容を表すメタデータのみを参照できる場合等、場所やセンサーによって見える情報を制御した[6]。なお、情報の見せ方にも二次元マップと主に屋内用の三次元マップそれぞれの可視化システムを備えた。

3.4 Situation Awareness

本プロジェクトでは、災害や事件に関わる各種のイベントを個別に自動検知し、共通の GUI 上に表示することに加え、個々のセンサーからのデータや一度のデータ解析のみでは検知し難い状況の認識（Situation Awareness、以下 SA）にも取り組んだ。SA には、その状況に名前や定義が可能なものとそうでないものがある。我々は、いずれも複数の断片的な情報（イベント）から判断されるメタなイベントと捉え、前者については複数のイベントの組合せによって導出される「上位イベント」を幾つか定義し、後者は「関連イベント」同士がリンクされた状態をそのまま可視化してユーザに提示することとした。ただし、実際にはいずれのタイプの SA も前段の解析システムで直接検知されたイベントを表すノードの間に統合分析システムで追加した「上位イベント」または「関連イベント」を表すノードを追加する形で、それぞれ図 4 のように表示した。

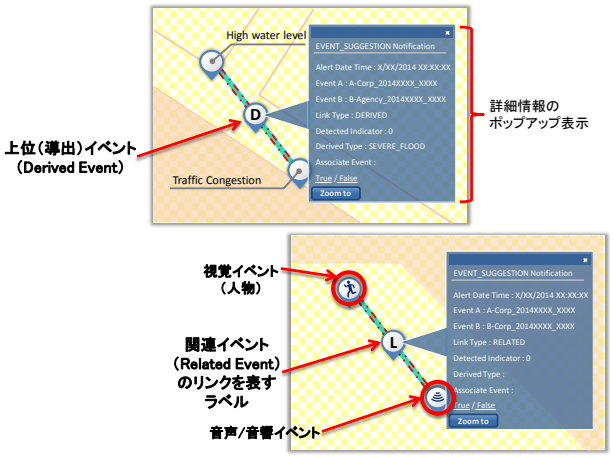

図 4　上位イベントと関連イベントのイメージ

このような機能を実現するための統合分析システムの概要構成を図 5 に示す。統合分析システムは、主にリアルタイムの複合イベント処理（CEP）と、SPARQL を用いた SA 処理を行なう。各種の解析システムで検知されたイベントデータは、CEP で個別に重複イベントの除去などをリアルタイムで行なうと共に、前述のオントロジーに基づく RDF 形式に変換し、ローカルな位置情報を緯度経度情報に変換した共通の位置情報などのメタデータを付与してデータリポジトリに保持する。また、主要な SA 処理として、新たに取得した複数のイベントから「上位イベント」を、さらに保持していた過去のイベントも参照して「関連イベント」を、それぞれ検出する。

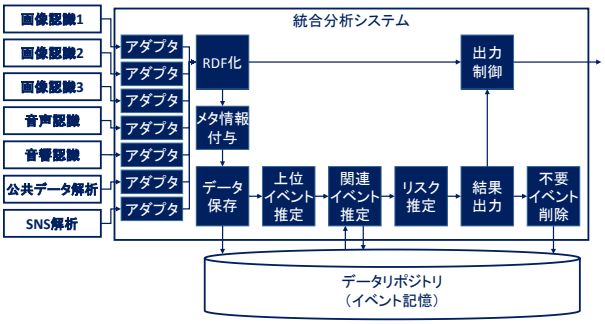

図 5　統合分析システムの構成

統合分析システムでは、SPARQL を用いて次のような複数の機能を実現している。

① イベントの信頼性評価：ASK または SELECT クエリで時刻や場所、値の範囲などの制約条

件を満たす正しいイベントか否かを判定

② 「上位イベント」の生成：CONSTRUCT クエリで複数のイベントから上位イベントの RDF データを生成

```
CONSTRUCT {
 er:Res1 a sem:Event .
 er:Res1 er:hasEventCategory 'DerivedEvent' .
 er:Res1 sem:eventType er:SevereFlood .
 er:Res1 er:hasObserver er:Res2 .
 er:Res2 er:hasCameraID ?cameraID .
 er:Res1 er:hasPublishTime ?time2 .
 er:Res1 er:hasRating ?score .
 er:Res1 sem:hasSubEvent ?event1 .
 er:Res1 sem:hasSubEvent ?event2 .
 ?event1 sem:eventType xx:Flooding .
 ?event2 sem:eventType yy:HeavyTraffic .
 }
 WHERE {
  ?event1 xx:hasTypeDescription xx:Flooding .
  ?event1 xx:hasCameraID ?cameraID .
  ?event1 xx:hasAlarmDefinition ?score .
  ?event1 xx:hasAlarmTime ?time1 .
  ?event2 yy:hasType yy:HeavyTraffic .
  ?event2 yy:hasCreateDate ?time2 .
  FILTER ( xsd:dateTime(?time2) > xsd:dateTime(?time1)
   && xsd:dateTime(?time2) - xsd:dateTime(?time1)
     <= 'PT1H'^^xsd:duration )
 }
}
```

図 6 「上位イベント」生成クエリの例（部分）

③ 「関連イベント」のリンク生成：CONSTRUCT クエリで必須属性（イベントタイプや観測主体）が共通の同一イベント（sameAs）または時刻と場所に加えて対象に共通性のある関連（related）イベントのリンクを生成

④ 追跡用「関連イベント」のリンク生成：顔認識で同一人物と判定されたイベント間に CONSTRUCT クエリでリンクを生成

以上のように、本プロジェクトの実験では SPARQL でヒューリスティックに条件を設定し、関連するイベントデータに対して各種のリンクや上位イベントを生成した。望ましくは、それぞれの判定条件を過去の事例から最適化、または機械学習を用いた判定モデルも適用すべきだが、ここでは各センサーや解析システムの多くが初めて導入されるもので過去にデータの蓄積が無く、加えて災害や事件など発生頻度の低いイベントでは機械学習に必要な量のデータを集め難い。このような場合、SPARQL のような標準クエリ言語で様々なタイプのデータ連携条件を仮で設定し、実データから判定の精度と網羅性を評価して条件を修正しつつ、機械学習も利用できるようにデータを蓄積していくアプローチが現実的と考える。

3.5 関連動向

前述のプロジェクト実施期間とほぼ同時期より、主に米国の軍や諜報・情報機関の主導で Open Source Intellingene（OSINT）や Imagery Intelligence（IMINT）、SIGINT（Signals intelligence）など様々なインテリジェンス（検知情報）を統合分析するマルチ・インテリジェンス・システム（Multi-INT System）の研究開発が盛んになり、MarkLogic 社などのベンダーがデータモデルに RDF を用いたソリューションを提供し始めている。また、関連する学会としては Semantic Technology for Intelligence, Defense, and Security（STIDS）があり、毎年オントロジーの提案や活用の議論が活発に行なわれている。

地理データに関する標準化機構 OGC（Open Geospatial Consortium）では、パブリックセーフティ等における地理空間データ統合のために SPARQL を拡張した GeoSPARQL を策定しており[7]、地理空間オントロジーの構築にも取り組んでいる[8]。また、トロント大ではスマートシティ実現のために環境や教育、エネルギーなど様々なドメインオントロジーを構築しており、その 1 つに都市評価指標の国際標準 ISO 37120 に基づくパブリックセーフティ・オントロジーがある[9]。

国際テロ対策としての前述したマルチ・インテリジェンス・システムでも、こうしたオントロジーの構築と利用が進められているが、同領域ではその殆どが非公開またはベンダー独自のものと見られる。一方、テロ対策の中でもサイバーテロ対策は一般のサイバーセキュリティと多くの点で共通し、こちらに関しては標準語彙やオントロジーが積極的に公開されている[10][11]。近年、国家レベルのサイバー攻撃や世界規模でのランサムウェア被害などが深刻化し、サイバーセキュリティは世界的な重要課題となっている。日々膨大な数の新たな攻撃手段が生み出されており、攻撃用のツールやサービスが充実してきている[1]。こうした背景から、脅威情報と呼ばれる攻撃の事例を体系的に記述したデータの作成と共有が重要視され、センサーで検知された断片的な情報

[1] 最近は闇市場で個人情報やマルウェアを安価に購入でき、クラウド型のランサムウェア作成サービスもある。

からどのようなタイプのサイバー攻撃を受けた可能性があるかを素早く判断できるようにするためのデータ統合・照合手段として STIX[12]などの標準語彙やオントロジーの整備が推進されている。

4. おわりに

本稿では、LOD のリンクによるデータ連携促進のヒントを探る試みとして、多くの異なるデータをメディアの違いや提供元の違いを超えて連携させる一つの事例を紹介し、それによって可能となる状況認識（Situation Awareness）のセマンティック技術による実現例を述べた。また、セキュリティ＆セーフティ業界における１つの課題認識とその解決のためのオントロジー構築に関する動向を紹介した。

以上のような課題に加え、日本市場には特有の課題がある。日本では IT 投資の優先度が多くの場合

ハードウェア＞ソフトウェア＞運用・保守

となっている。これは、昔の製造業中心だった時代の国内産業の考え方が IT 市場にも根強く踏襲されているためと考えられる。しかし、今ではハードウェアはコモディティ化が進み、ソフトウェアはオープン化・サービス化し、世界中から安価で入手可能となった。一方、運用・保守は事情が異なる。現場で対象領域の知識を持った者が行なわなければならず、クラウドサービスも普及してきたが、それぞれの現場を知っていなければできないことも多い。

翻って、セマンティック技術、即ち意味やコンテキストを考慮して問題を解決する知的作業は、領域に関わらず共通の部分もあるが、現場やトレンドに合わせる必要がある部分も多い。欧米では、現場に応じたオントロジーを構築し更新していくための教育プログラムが整備され、専門のコンサルタント業もある。日本人の美徳と言われる「おもてなし」の考え方は、市場では無償や低価格のサービスとして期待される面もあり、常に更新し最新の状況に合わせていく必要のあるデータ中心の産業やセマンティック技術にとっては１つの重要な課題と思われる。

こうした課題に対してオープンデータ推進の観点からできうる提案としては、共通・類似の環境を持つコミュニティ同士で互いに意味の分かる、説明のつく情報を共有し、問題にできるだけ早く低コストで対応していく仕組みの構築がある。サイバーセキュリティ業界における世界的な脅威情報の発信と共有は、ある側面でこれを実践している例と言える。日本の商習慣を容易に変えられないならば、共通の利益や安心・安全に関わる情報は共有し、且つどのシステムにもすぐに反映できる標準的でマシンリーダブルな形で共有することは１つの策だろう。官学民で改めて検討する価値のあるテーマと考えるが、民間企業では、ビジネスとしての成立性とも併せて検討する必要がある。

参考文献

[１] 加藤: 日本語 Linked Data Cloud 図 2015-11-18 版, http://linkedopendata.jp/?p=616

[２] LinkData データセット一覧, http://linkdata.org/work

[３] J. Z. Pan, et al: Industrial Applications and Successful Stories, In: Exploiting Linked Data and Knowledge Graphs in Large Organiations, pp.213-236, Springer, (2017)

[４] P. Wang, K. W. Woo, S. K. Koh: シンガポールにおけるより安全な都市「セーファー・シティ」の構築, NEC 技報, Vol. 67, No. 1, pp.73-77, (2014)

[５] W. R. van Hage, et al: Design and use of the Simple Event Model (SEM), Journal on Web Semantics 9(2), pp.128-136, (2011)

[６] P. Wang, et al.: 組織間の安全な情報共有を実現する「MAG1C」情報ガバナンス・ソリューション, NEC 技報, Vol.67, No.1, pp.64-67, (2014)

[７] OGC, GeoSPARQL - A Geographic Query Language for RDF Data, (2011), http://www.opengeospatial.org/standards/geosparql

[８] I. Simonis and S. Fellah: OGC Testbed 10 Cross Community Interoperability (CCI) Ontology Engineering Report, (2014), http://www.opengeospatial.net/doc/PER/testbed10/cci-ontology

[９] K. Khazei and M. S. Fox: A Public Safety Ontology for Global City Indicators (ISO37120), EIL Working Paper, (2017), http://eil.mie.utoronto.ca/wp-content/uploads/2015/06/GCI-PublicSafety-Ontology-28apr2017.pdf

[１０] J. A. Wang and M. Guo: OVM: An Ontology for Vulnerability Management, Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies. CSIIRW '09, pp.34:1-34:4, (2009)

[１１] M. B. Salem and C. Wacek: Enabling New Technologies for Cyber Security Defense with the ICAS Cyber Security Ontology, Proceedings of The Tenth International Conference on Semantic Technologies for Intelligence, Defense, and Security, pp.42-49, (2015)

[１２] Structured Threat Information eXpression (STIX) 1.x Archive Website, http://stixproject.github.io/