武田英明　国立情報学研究所

日本におけるLinked Dataの現状と普及に向けた課題

本稿では日本における Linked Data の課題と現状についてまとめる．Linked Data には社会における情報循環の中で情報共有を促進するという社会的意義がある．情報循環とは利用・創造―公開―共有―収集―利用・創造―…という循環であり，これが社会での情報の価値を高めるプロセスになっている．我が国においても情報の価値を高めるにはこの情報循環を活発化する必要があり，その方法として Linked Data は期待できる．日本における Linked Data 化の推進には，公開の文化，コミュニティの成熟，中心的データの欠如，日本語の取り扱いという課題がある．現在は，国立国会図書館や国立情報学研究所等で自身の持つデータの Linked Data 化が行われ，公開を始めている．

私たちの Linked Data?

Linked Data はデータ共有の新しい方法として欧米で認知され，実践が進んでいる．日本においてはどうだろうか．セマンティック Web 自体の未普及もあって，まだ認知すらされているとはいえない状況である．日本においても Linked Data は可能であろうか．いやそれ以前にそもそも Linked Data は日本に必要なのだろうか．

本稿では日本における Linked Data 化活動を概観する．まず前提として，なぜ Linked Data が必要なのから考察を始める．これは本質的には社会における情報共有の問題である．したがって Linked Data だけにかかわる問題ではないのだが，Linked Data というのは情報共有の新しい世界である以上，避けて通れない．その上で，日本あるいは日本語固有の課題を挙げ，どのような解決法があるか考える．最後に具体的に大規模な Linked Data あるいは RDF を提供している活動を取り上げ，説明する．

Linked Data の社会的意義

冒頭で述べたように Web 技術の発展の先に Linked Data がある．その重要性は情報技術者や研究者にとっては比較的分かりやすいが，社会的意義をきちんと説明できないと，広く公開のデータを作ろうという Linking Open Data（LOD）活動は参加者や理解者を増やすことができない．そこで，まず Web の社会的位置付けから考えることで，なぜ LOD には社会的な意義があるかについて述べる．

●情報循環としての Web

Web の社会的意義とは，情報の社会的循環の大規模化・高速化こそが情報の価値を高めるということを実践的に知らしめた点である．

情報というのは単に作られただけでは価値がない．当然，他の人たちに伝達され，利用されてこそ価値

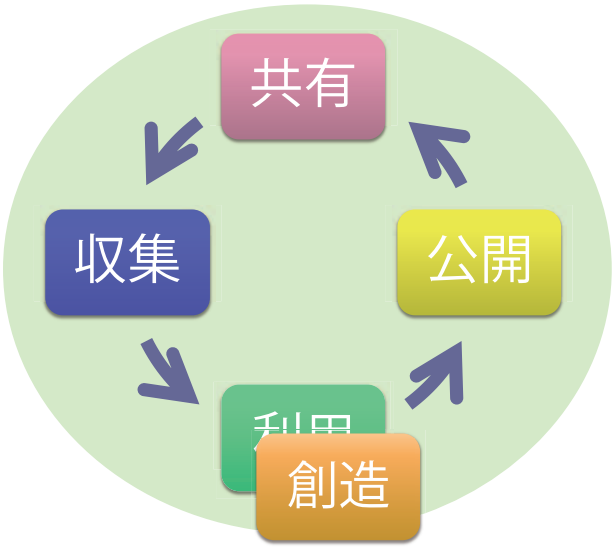

図 -1　情報循環

が生まれる．ある人によって他の人の情報に基づき新たな情報が作られ，それがまた他の人に使われて新しい情報が作られる．この循環こそが我々の社会での情報を豊かにさせてきた源泉である．個人的な情報伝達手段しかなかった時代には，利用・創造ー伝達ー利用・創造ー…という単純なものでしかなく，きわめて遅く小規模なものであった．

マスメディアの登場により公共的な情報循環が始まった．すなわち，利用・創造ー公開ー共有ー収集ー利用・創造ー…となった（図 -1）．個人的な情報伝達に比べ，格段に速く規模も大きくなった．この仕組みにより多くの職業的情報創造者（ジャーナリスト，作家，作詞者，作曲者等）が生まれた一方，情報を公開できる人間はそういった職業的情報創造者やメディア関係者に限られており，情報循環への関与という点では偏っていた．すなわち，情報を創造して公開できるのは一部の人々であり，多くの人々は単に利用者でしかないという偏りである．

Web はこの偏りを直す仕組みを提供した．すなわち，だれでも自らの情報を公開し共有することができる．無料あるいはきわめて低料金で自らの情報を他者に利用可能な形で公開することができる．また公開された情報は一元的なコントロールなどを受けることなく自由に共有され，自由に利用することができる．この結果，情報循環はかつてないほど多数の参加者により大規模かつ高速に行われるようになった．

●情報循環としてのセマンティック Web

このように情報循環に新しい時代を作った Web であるが，さまざまな課題も生まれてきた．その中の 1 つがデータのコンピュータでの利用である．

Web の仕組みは当然のことながらコンピュータとコンピュータネットワークによって実現されている．しかし，情報循環には人間が関与することが前提になっていて，コンピュータにはあまり適切でない．顕著なのが HTML で，HTML による情報の構造は人間が理解するために使われており，これだけではコンピュータがそこに書かれている情報を適切に処理することができない．

その克服のための仕組みがセマンティック Web である．セマンティック Web は人間とコンピュータ双方が情報の内容をより多く理解し共有できるように，情報の意味を与える仕組みを用意している．それがセマンティック Web 言語である RDF（Resource Description Framework），RDFS（RDF Schema）や OWL（Web Ontology Language）である．セマンティック Web については本誌でも何度か取り上げている [1] ので詳しくはそちらを参照されたい．

●情報循環としての Linked Data

Linked Data はセマンティック Web のうち，個別の情報（インスタンス）を重視して情報公開・共有を行うというものである．セマンティック Web の構想はいくつかの階層からなる．図 -2 は最も初期のころのセマンティック Web の階層である．このうち，研究としては下位から上位へ，すなわち，RDF 記述のレベルからオントロジーへ，そしてさらに上位へと進んでいる．

しかし，言語が整備されたとしてもオントロジーを実際構築して共有していくのは大変なことである．オントロジーが広く共有されていれば，それに基づいた情報の共有は容易になる．しかし，それを待っていてはなかなか進捗しない．

Linked Data ではオントロジーの共有はひとまずおいておき，まずはデータの共有をしましょうというところに特徴がある．それが Tim Berners-Lee のいうところの Linked Data の 3 原則（本特集の第 1 編

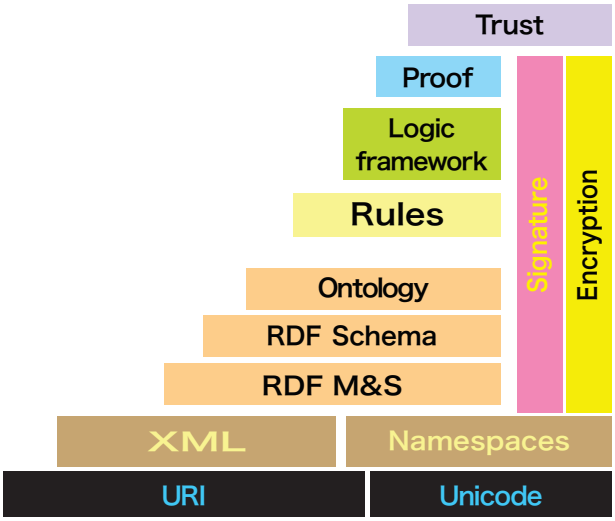

図 -2　セマンティック Web の階層

参照) である．概念レベルのオントロジーの共有は一朝一夕ではできないが，個別のデータの共有は比較的容易だということである．これが Linked Data の狙いであり，実際大きな勢いでデータが増えている．

●情報循環としての LOD

Linking Open Data（LOD)は Linked Data として情報を共有していこうという活動である．Linked Data の性質として相互につながってこそ意味があるので，そのつながりを集めて公開することでデータの利用を促進したり，より多くの参加者を集めようとしている．

LOD において公共セクタは重要な部分である．というのは元々公共セクタの情報は国民・市民に公開されている情報である．当然公開された情報は利用されることを期待されている．Web 以前は紙媒体や限定されたデータベースとして公開していたが，Web 以後は HTML や PDF で公開されるようになった．しかし，HTML や PDF で公開された情報はデータとしての利用は難しい．個別の処理をしないと，そこから必要なデータを抜き出すことができない．Linked Data の形式でデータを公開することで，こういった個別の処理なしでデータが利用可能になる．

また，公共セクタは社会において重要かつ大量のデータを抱えている．もちろん，プラバシーや国家機密にかかわることは，そもそも公開情報でないので除外するとしても，それ以外にも大量の情報を抱えている．この情報を Linked Data の情報循環に入れることは，情報循環が前提の社会としては必須なことといえよう．すなわち，公共セクタは LOD に情報提供をすることで情報循環のインフラを支えることが期待されている．

もちろん個別の企業や団体の情報も社会的な価値を多く持っている．その多くの情報の価値は社会における情報循環によって支えられている．とすれば公開可能な情報は，より利用されやすい形式で公開することがその価値を上げることになる．その仕組みとして LOD を使うのは企業的に見ても十分意義のあるものであると考える．

日本における Linked Data 化の課題

LOD の活動はヨーロッパおよびアメリカにおいて盛んであり，単に情報研究者の活動の域を超えて，個々の分野の専門家や政府などの組織を巻き込む活動になっている．

残念ながら日本ではさほど活動的であるとはいえない．それはなぜなのか，その解決はあるのかということをここで考えよう．ここで「日本」と呼んでいるのは，日本国内の活動 and/or 日本語での活動を指している．もちろん LOD は本質的にグローバルであり，こんな区分は本質的でないが，現状を把握するためにはあえて分けて考えてみる．

●情報公開・共有の文化

日本の社会，ことに組織においては情報公開・共有の重要性は十分に理解されているとはいえない．情報循環は情報の公共性を維持することであり，情報公開・共有はその情報循環を実現する要素として重要であるということが理解されていなければ，情報公開・共有はリスクだけが強調され，実際に自らの情報を公開・共有することができない．ことに公共セクタである組織のほうがより消極的なことが多いのは残念である．

これは日本の社会の文化的背景によるのか断言はできないが，いずれにしろ，この点から変えないと継続的・持続的な情報共有は実現できない．これは Linked Data 実現以前の問題であるが，特に Linked Data においては大規模なデータを持つ組織および

公共セクタの能動的な参画が重要であるので，Web 化よりこの点が効いてくる．

なお，政府系でもどこもが消極的というわけではない．国民への情報提供を主たる業務とするような組織は情報公開をより効果的にする手段として利用しつつある．本特集の第 5 編で触れた国土地理院は実質的に制限をつけずに情報の再利用を許しているし，後で述べる国立国会図書館や国立情報学研究所も新しい公開手段として利用しつつある．

政府系に関しては本特集の第 4 編で述べたようにオープンガバメントの動きが出ているので，より積極的に変わるチャンスがあると期待している．

●コミュニティの未成熟

Linked Data の実現には単に情報のネットワークだけではなく，人のネットワークも必要である．

Linked Data はその性質からして，異なる情報源からの情報が相互につながってこそ価値が出る．またデータそのものは各領域にあるので，単に情報研究者・技術者だけでなく領域の研究者・専門家の参画が必要である．Linked Data はまだ発展途上であり未解決な問題が多々あるので，このような人々が適宜インフォーマルにコミュニケーションをとって解決していかないといけない．

欧米を中心とするコミュニティではメーリングリストで小さい問題から大きな問題まで盛んに話し合われている．また分野ごとのコミュニティも形成しつつある．残念ながら国内ではこのようなコミュニティはまだ未形成である．これは筆者を含む本領域の研究者・専門家の宿題であるといえよう．

なお，後で触れるバイオサイエンス系はデータの性質上国内というよりは国際的な関係が重要であり，国際的コミュニティに加わることで Linked Data 化が推進されている．

国内ではまだ大きな動きとはいえないが，Google group には LinkedData.jp というコミュニティが作られ，少しずつ状況は変化している．後で説明する lod.ac プロジェクトでは美術館・博物館情報の Linked Data 化を進める中で，地域の NPO との連携も始まっている．

●中心的データの欠如

本特集の第 1 編の LOD クラウドを見て明らかなのは DBpedia が LOD クラウドの中心になっているということである．LOD においてはさまざまな情報源同士が相互にリンクし合えるのであるが，そうはいってもデファクト的につなげることができるサイトがあれば，自身の情報を Linked Data 化するときの目標を定めやすい．いわば“参入障壁”を低くすることができる．それが DBpedia である．DBpedia はオンライン百科事典 Wikipedia を Linked Data 化したものであるので，きわめて広範な領域をカバーしている．たいていの分野で何らかの関係性を見いだすことができる．LOD において DBpedia はきわめて重要で，現在の LOD 活動はこの DBpedia の公開に始まると言っても過言ではない．

この DBpedia は日本語リソースとして使うには問題がある．DBpedia は英語版 Wikipedia を使っている．Wikipedia の各ページに相当する資源には Wikipedia の言語リンクを利用して多言語のラベルが付けられているので，日本語のラベルは存在する．しかし，Wikipedia は各国語版で大きく構成が異なるので，日本語の Linked Data には適切とはいえない．

これに関しては lod.ac プロジェクトでは，日本語のリソースを増やすために多様な種類の辞書・事典から用語を抜き出してリソース化した「ことはぶ」というものを開発している（後述）．

●日本語のリソースの記法

より技術的な課題としては，リソースの URI に日本語を使うかどうかという問題がある．クラス名やプロパティ名に日本語を使うか，あるいはそれに相当する英語名を使うかということである．URI の場合アスキー文字のみであるが，IRI（国際化 URI，Internationalized Resource Identifier）[RFC3987] に基づけば，unicode で書いた日本語文字列を含めることができるので，技術的には可能である．しかし，それだけで問題が解決するわけではない．

まずリソース名に日本語を混ぜることのメリットしては，

- 既存のデータ構造を流用できる
- 了解性(少なくとも日本人には)
- 同一性(翻訳による揺らぎがない)

ということが挙げられる．逆にデメリットは

- 関係システムが技術的に処理可能か不明(IRI に対応できていない)
- 日本人以外には意味不明
- 国際的なスキーマと合わせると英語・日本語が混交して不自然

ということが挙げられる．一方，元々日本語を使ったデータ構造を英語化して記述するとなると，メリットしてはこの反対であり，

- 技術的に安心(すべてのシステムが処理可能)
- 了解性
- 他の国際的なスキーマとスムーズに結合

ということになる．

一方のデメリットとしては

- 翻訳の必要，同一性の担保が難しい

ということがある．

Linked Data は国際的に流通するものであるという点においては，英語化したほうが適切だといえる．しかし，Linked Data 化されるものが常に国際的に流通を意図しているというのもおかしな話である．日本国内で流通することに意味があるものもある．そうであれば必ずしも英語化にこだわる必要はない．むしろ英語化が Linked Data 化の障害になるようならば，元々のデータで使われている日本語そのままの Linked Data 化で十分である．たとえばこの後に取り上げるもののうち，バイオサイエンスにおいては前者であり，日本語版 DBpedia では後者である．

中間的方式としては，英語化したリソースに日本語のラベルを張るという方法[☆1]や英語と日本語で2重にリソースを記述するなどの方法も考えられる．

現状では，データの性質を鑑み，方法を定めるということになろう．

なお，もう1つの日本語特有の問題は「読み」であるが，これは後述する．

☆1 DBpedia は label として各国語の表記が付加されている．これは元々の Wikipedia の言語間リンクを利用したものである．

現在の日本／日本語の Linked Data

ここでは日本において大規模に LOD あるいは RDF を公開している例をいくつか取り上げる．

●理化学研究所サイネス[☆2]

理化学研究所が運営しているバイオサイエンスデータの公開 DB サービスであるサイネス（SciNetS.org）においては，すべてのデータが OWL/RDF として利用可能である．バイオ系を中心に現在 100 個以上のデータベースが登録されている．全インスタンス数は約 900 万件，トリプル数で約 1 億，データサイズは約 11TB である．また，サイネスを使って国際的なデータ連携のプロジェクトが行われている（例：マウス表現型データの国際共有化／ InterPhenome[☆3]）．

サイネスでは，バイオ研究者が求める検索を実現するために通常の SPARQL エンジンではなく，統計処理機能を拡張した独自開発の検索エンジン（GRASE）を採用している．また，RDF のままでは Web ブラウザや JavaScript が直接処理しにくいという欠点を補うために，簡易な方式でも同じデータにアクセスできるよう Semantic-JSON というインタフェースを提供している[☆4]．Semantic-JSON API ではすべての情報に ID が付けられ，データ取得の指示(命令)と，この ID を含んだ URI をサーバに投げることでデータを取得する．この API は各種言語（Ruby, Perl 他）のライブラリとして用意されており，さらにはこのサイト上でスクリプトを書いて実行する環境も用意している．

●ライフサイエンス統合データベースプロジェクト[☆5]

大学共同利用機関法人 情報・システム研究機構ライフサイエンス統合データベースセンター（DBCLS）ではさまざまなアプローチでバイオサイエンスデータのセマンティック Web 化を進めている．たとえば各種ライフサイエンス系の Web サー

☆2 サイネスについては豊田哲郎氏（理化学研究所）にご教授いただきました．感謝いたします．
☆3 http://www.interphenome.org/
☆4 https://database.riken.jp/sw/wiki/ja/cria160s1ria160s2i/ この URI が Semantic-JSON の例になっている．
☆5 DBCLS での取り組みについては，中尾光輝氏（DBCLS）にご教授いただきました．感謝いたします．

ビスの標準的な方法でアクセス可能にする TogoWS ☆6 では出力を RDF として得られるようにしている．DDBJ-PDBj-KEGG RDF 化プロジェクトではタンパク質データベース PDBj の RDF 化などを行っている．ほかにも小規模用データベースシステム TogoDB では RDF 出力をサポートする予定である．

また，DBCLS ではバイオ系におけるプログラミング技術の向上と知識共有のために，合宿形式で行う DBCLS BioHackathon を主催している．そこではバイオ系のデータに対するセマンティック Web 技術を適用したプログラミングも行われている．

●国立国会図書館の NDLSH

図書館の世界では，いま世界的に急速に Linked Data 化が進んでいる．LOD クラウドの右上に publication 関係が集まっているが，その中でも図書館に関係する LOD は LCSH を中心にまとまっている．LCSH はアメリカ議会図書館の件名標目表（subject heading）のことである．件名標目とは図書を分類するときの統制語彙で，多くは階層的な構造を持っている．各国の中央図書館は自らの管理する件名標目や著者名典拠や書誌を Linked Data 化して公開を始めている．

日本では国立国会図書館自らが管理する国立国会図書館件名標目表（NDLSH）を Linked Data 化して公開を始めている☆7．規模としては約 130 万トリプルである．また SPARQL エンドポイントも用意されており，恐らく日本で最初の実用的な SPARQL エンドポイントである．図 -3 に SPARQL でのクエリ例を示す．

データ構造は単純で基本は Dcterms と SKOS と呼ばれる語彙を使ったものである．SKOS は元々図書館系の情報構造に基づいているので相性はいい．対応する LCSH がある場合は rdfs:seeAlso でつなげている．

日本語特有の問題としては「読み」がある．読みというのは他の言語には存在しない．しかし日本語のデータにおいては重要な要素である．NDLSH においては独自の transcription というタグを定義してそれをタイトルの下部構造として埋め込んでいる．これはタイトルに限らず他のリソースでも読みが存在し得るので，

☆6 http://togows.dbcls.jp/
☆7 http://id.ndl.go.jp/auth/docs/sparql

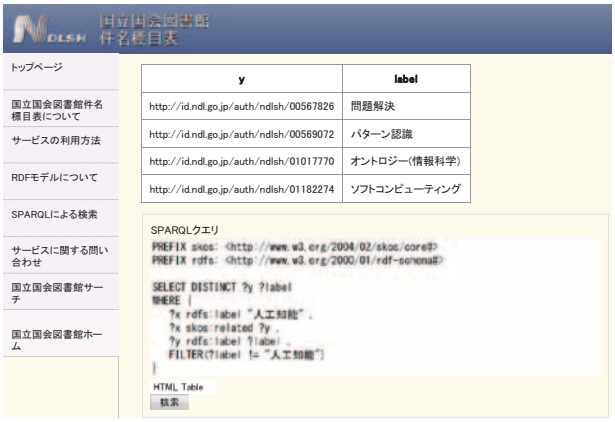

図 -3　国会図書館件名標目表の SPARQL クエリ

統一的構造としては分かりやすい．反面，ブランクノードを含む構造になり，利用側では注意が必要である．

●国立情報学研究所の CiNii および Kaken

国立情報学研究所が提供するデータベースサービスでは通常の HTML によるデータ提供に加えて，RDF によるデータ提供も始めている．

CiNii ☆8 は国内論文の書誌および本文検索サービスであり，現在 1,300 万件以上のデータを提供しており，月間 6 億以上のアクセスのあるサイトである．CiNii における主要な情報オブジェクトは書誌情報と著者情報であるが，主に書誌情報を RDF として提供している（著者情報の RDF は簡易版）．その例を図 - 4 に示す．HTML の URL ＋ “.rdf” の URL としてアクセスできる．基本的には Dcterms と PRISM（The Publishing Requirements for Industry Standard Metadata），Foaf を組み合わせて表現している．日本語と英語の混在については言語タグ（en と jp）を付けて，別のリソースとして扱っている．

Kaken は文部科学省科学研究費補助金の報告書のデータベースである．主な情報オブジェクトは報告書と研究者で，件数にして 100 万件程度の報告書および 18 万人程度の研究者がデータベース化されている．メタデータとしてはタイトルなどに Dcterms，人物情報に Foaf を使うもののほかは独自のタグを定義して使っている．RDF へのアクセスは http の content negotiation を使ってできるようになっている．実験的に SPARQL エンドポイントを構築している．

☆8 http://ci.nii.ac.jp/

```
−<rdf:RDF>
 −<rdf:Description rdf:about="http://ci.nii.ac.jp/naid/110007338463#article">
    <foaf:isPrimaryTopicOf rdf:resource="http://ci.nii.ac.jp/naid/110007338463.rdf"/>
    <dc:title>Webにおけるアイデンティティとセマンティックスの表現と利用(<特集>WebアイデンティティとAI)</dc:title>
    <dc:creator>武田 英明</dc:creator>
    <dc:publisher>社団法人人工知能学会</dc:publisher>
    <prism:publicationName>人工知能学会誌</prism:publicationName>
    <prism:issn>09128085</prism:issn>
    <prism:volume>24</prism:volume>
    <prism:number>4</prism:number>
    <prism:startingPage>512</prism:startingPage>
    <prism:endingPage>518</prism:endingPage>
    <prism:publicationDate>2009-07-01</prism:publicationDate>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/web" dc:title="web"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/semantic_web" dc:title="semantic_web"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/HTTP" dc:title="HTTP"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/RDF" dc:title="RDF"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/RDFS" dc:title="RDFS"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/Linked_Data" dc:title="Linked_Data"/>
    <dc:date>2009-07-01</dc:date>
  </rdf:Description>
 −<rdf:Description rdf:about="http://ci.nii.ac.jp/naid/110007338463#article" xml:lang="en">
  −<dc:title>
      Representation and Use of Web Identity and Semantics(<Special Issue>Web Identity and Artificial Intelligence)
    </dc:title>
    <dc:creator>Takeda Hideaki</dc:creator>
    <dc:publisher>The Japanese Society for Artificial Intelligence</dc:publisher>
  −<prism:publicationName>
      Journal of Japanese Society for Artificial Intelligence
    </prism:publicationName>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/web" dc:title="web"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/semantic_web" dc:title="semantic_web"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/HTTP" dc:title="HTTP"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/RDF" dc:title="RDF"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/RDFS" dc:title="RDFS"/>
    <foaf:topic rdf:resource="http://ci.nii.ac.jp/keyword/Linked_Data" dc:title="Linked_Data"/>
  </rdf:Description>
 −<rdf:Description rdf:about="http://ci.nii.ac.jp/naid/110007338463#article">
  −<foaf:maker>
   −<foaf:Person>
      <foaf:name>武田 英明</foaf:name>
      <foaf:name xml:lang="en">Takeda Hideaki</foaf:name>
     −<con:organization>
      −<foaf:Organization rdf:about="http://ci.nii.ac.jp/org/%E5%9B%BD%E7%AB%8B%E6%83%85%E5%A0%B1%E5%AD%A6%E7%A0%94%E7%A9%B6%E6%89%80%253A%E6%9D%B1%E4%BA%AC%E5%A4%A7%E5%AD%A6%E4%BA%BA%E5%B7%A5%E7%89%A9%E5%B7%A5%E5%AD%A6%E7%A0%94%E7%A9%B6%E3%82%BB%E3%83%B3%E3%82%BF%E3%83%BC">
          <foaf:name>国立情報学研究所</foaf:name>
          <foaf:name>東京大学人工物工学研究センター</foaf:name>
          <foaf:name xml:lang="en">National Institute of Informatics</foaf:name>
          <foaf:name xml:lang="en">RACE, The University of Tokyo</foaf:name>
        </foaf:Organization>
      </con:organization>
    </foaf:Person>
  </foaf:maker>
 </rdf:Description>
</rdf:RDF>
```

図 -4　CiNii における RDF 記述

● lod.ac プロジェクト

このプロジェクトは情報・システム研究機構新領域融合研究センターのプロジェクトの一環として「学術リソースのためのオープン・ソーシャル・セマンティック Web 基盤の構築」というタイトルで実施しているものである．日本における学術に関するデータを Linked Data の方式で公開・共有するということを実践的に実施して，実践を通じてのプラットフォーム作りと構築知識の獲得を目的としてる．

《美術館・博物館情報》

その最初の対象は，分散かつ未統合のデータのテストケースとして美術館・博物館情報の統合とした．日本における美術館・博物館の情報は，各館が独自に所蔵品情報を公開する程度で情報の統合が行われていない．そこで本プロジェクトで日本全国の美術館・博物館情報を Linked Data として共有して統合できる仕組みを作ることにした．このような試みはヨーロッパでは EU のプロジェクトとして Europeana というものが行われている☆9．Europeana では EU27 カ国の博物館の収蔵情報を統合して扱えるサービスを構築している．Europeana においても一部の情報を Linked Data 化して提示する実験システムを公開している．

LOD Museum☆10 では美術シソーラス[2]，作品データベース，個別美術館・博物館といった異なる情報源からの情報を統合して構築される．このようなそれぞれが自身の情報のオーソリティで複数の情報源を統合するときには，どのようにデータを統合するかという統合ポリシーが必要である．今回はオーソリティ統合に関して次のような原則を用意した．

- 自分がオーソリティを持つ情報オブジェクトは自ら ID を付与して管理する
- 他の情報源がオーソリティを持つ情報オブジェクトはその ID を流用した独自の情報オブジェクトとして記述する
- 自分がオーソリティを持つ情報オブジェクトから他の情報源がオーソリティを持つ情報オブジェクトとは参照関係(owl:isPrimaryTopicOf または他のプロパティ)で結ぶ

このような構造にしたのは，オーソリティの異なるデータをその違いを残して管理するためである．データの追加や更新においてこの違いを保持しておくことは重要である．

LOD Museum では作品，作者，所蔵館が基本の

☆9 http://www.europeana.eu/portal/
☆10 http://lod.ac/

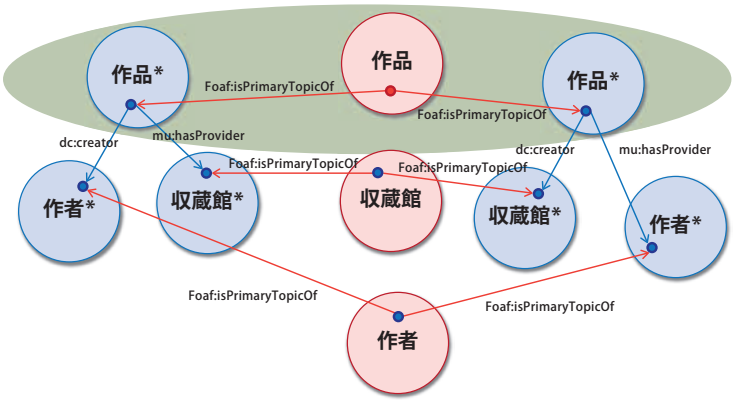

図 -5　異なる情報源のデータ統合

情報オブジェクトであり，それぞれを一元的に ID を付けて管理する．しかし，LOD Museum が生成した情報オブジェクトは ID と最小限の記述しか持たず，これらに関して外部の情報源から取り込んだ情報はそれぞれ別の情報オブジェクトとして記述される（図 - 5）．たとえば，ある作品に関する情報は 2 個以上の owl:isPrimaryTopicOf でつながった情報オブジェクトの和として表現される．

それぞれのメタデータは，Dcterms，Foaf，NDLSH，CIDOC CRM といったメタデータから必要な項目を抜き出したタグを集めて構成した．このメタデータでは作品の詳細なデータを記述するのではなく，共通性のある属性を列挙している．なお，美術関係においては作者名義は作者とは別に重要である．LOD Museum では，作品には作者名義と作者を（もし違えば）別のプロパティで表現し，作者情報においては，foaf:nick で作者名義を記述するようにしている．

日本語に関しては，作品名や作者名等は基本的に言語タグ（@ja-hani, @ja-hrkt 等）を用いて同一プロパティに多重に値を与えて表現する．

現在，日本美術シソーラス，国指定文化財等データベース☆11，14 館の所蔵データベースに基づいて構築している．より具体的な事項については文献 3）を参照されたい．

《ことば，事典情報》

先に述べたように，DBpedia の汎用的なリソースがあると参照先として使えるので LOD 化を進めやすい．そのために，まず日本でのことば，用語を集めてリソースとして参照できるサイト「ことはぶ」☆12 を用意した．「ことはぶ」は各種辞書・事典（Wikipedia, はてなキーワード，ニコニコ大百科（仮），Yahoo! 百科事典等）の掲載語を集め，集約して RDF によって記述したものである．NICT による日本語化された WordNet も含まれている．集約の結果，約 225 万語あった．個別のリソースごとの RDF あるいは SPARQL エンドポイントとしてアクセスできる．

また Wikipedia の Infobox を利用した LOD 化は東京大学の中山浩太郎氏と共同で日本語版 DBpedia として開設する予定である．

未来に向けて

本稿では日本における Linked Data にかかわる活動を紹介した．まだ個別の取り組みにとどまっており，大きな動きになっているとはいえない．しかし，国内においてもオープンガバメントの動きが出てきたように☆13，海外の動きに合わせて大きく変化することも考えられる．そのときに備えて国内においてもコミュニティを作り技術や情報の共有を進めるべきであろう☆14．

参考文献
1）萩野達也：セマンティック Web，0. 編集にあたって，情報処理，Vol.43, No.7, pp.707-708（July 2002）．
2）福田博同，五十殿利治：美術シソーラスデータベース形成の諸問題，情報管理，Vol.40, No.9, pp.790-809（Sep. 1997）．
3）嘉村哲郎，加藤文彦，大向一輝，武田英明，高橋　徹，上田　洋：Linked Open Data による多様なミュージアム情報の統合，人文科学とコンピュータシンポジウム論文集，情報処理学会シンポジウムシリーズ，Vol.2010, No.15, pp.77-84, 情報処理学会（2010）．

（平成 22 年 10 月 31 日受付）

武田 英明（正会員）takeda@nii.ac.jp
大学共同利用機関法人情報・システム研究機構国立情報学研究所情報学プリンシプル研究系・教授，同学術コンテンツサービス研究開発センター長．人工知能，Web 情報学などの研究に従事．人工知能学会，電子情報通信学会，AAAI，Design Society 各会員．

☆11 http://www.bunka.go.jp/bsys/

☆12 http://wordnet.jp/kotohub/

☆13 オープンガバメントラボ http://openlabs.go.jp/

☆14 本稿をまとめるに当たって，lod.ac プロジェクトでの議論が大変参考になりました．特に大向一輝氏※，加藤文彦氏※，嘉村哲郎氏（総合研究大学院大学／東京芸術大学），濱崎雅弘氏（産総研），Tran Duy Hoang 氏※には感謝いたします．※印：NII（国立情報学研究所）