３．次世代電子政府での活用方法に関する調査研究

セマンティック Web 技術を次世代の電子政府に活用する視点から、電子政府関連の情報に対するセマンティック情報（メタデータ）の付与方法とそれを活用したセマンティック情報検索の方法について調査検討を行なった。

３.１ セマンティック情報（メタデータ）の付与に関する調査

各国の電子政府構築において、電子政府関連の情報に対してセマンティック情報を付与する事例としては、現状ではメタデータ付与の事例しかない。そこで、電子政府関連の情報に対するメタデータ付与に関して、下記の調査を行なった。

・国内の行政公開情報の調査
・海外政府のメタデータ付与に関する調査
・メタデータ階層モデル調査
・メタデータ定義調査
・利用者プロファイルのメタデータ定義に関する調査

３.１.１ 国内の行政公開情報の調査

国内の行政公開情報の調査を行ない、国内の行政公開情報について、どのようなメタデータの付与方式とメタデータの活用方法が相応しいかその方向性を検討した。

３.１.１.１ 経済産業省の Web サイトの調査

経済産業省を対象として、Web サイト上で掲載されている公開情報の調査を行なった。

対象の URL は経済産業省の情報公開の説明ページ[79]であるが、その配下には、次の 4 つの情報へのリンクがある。

(A) 行政文書
　1) 行政文書ファイル
　2) 特許庁行政文書ファイル
(B) 政策情報
　3) 政策 HP における掲載情報の紹介
　4) 報道発表資料

（A）行政文書

2) の特許庁行政文書ファイルについては、調査期間中アクセスできなかったため検討を行なっていない。また、1)の行政文書ファイル（管理簿）は、紙などで保存されている行政文書に対する情報（これ自体がメタデータ）であり、それに対する検索システムが稼動している。図 3-1に行政文書ファイル検索システムの条件入力画面を示す。図 3-2に示すように、行政文書ファイルの（メタデータの）内容としては、分類（大分類、中分類、小分類）、日付などの項目がある。現状の検索でも、詳細検索を用いれば、日付による検

[79] http://www.meti.go.jp/intro/consult/disclosure/a_main.html

索や、管理担当課室・係による分類などをメニュー的に選択しながらを検索を行なうことができる（図　3-3参照のこと）。行政文書ファイルの項目に関する考察については、３．１．４節を参照されたい。

（B）政策情報

3),4)は、ホームページ上に載っており、セマンティック Web 技術を活用しやすいのはこれらの情報である。3) はどちらかと言うとフリーフォーマットに近く、PDF 文書も多く存在する。4)は報道発表なので、概要、担当、発行日、資料内容へのリンクなどからなっている（図　3-4参照のこと）。

（C）その他

経済産業省では、上記の（A）（B）以外にも、経済産業省 Web サイトにおいて情報の提供を行なっている。上記以外には、

・What ’ s New
・入札
・公募
・イベント情報
・電子申請

などのページがある。

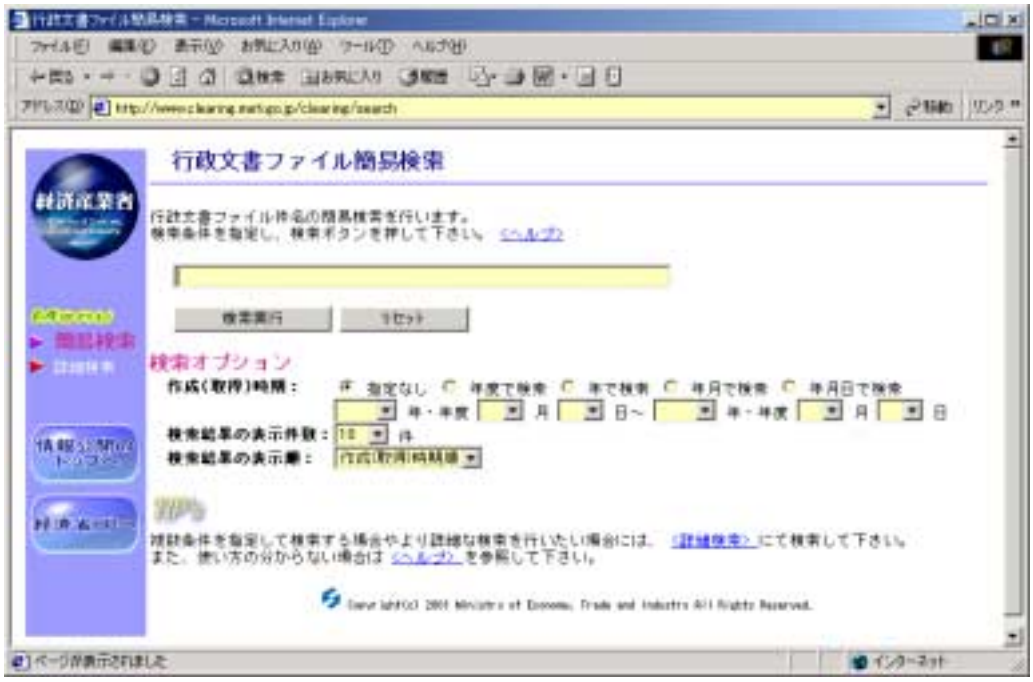


図　3-1　経済産業省の行政文書ファイルの検索画面

図　3-2　行政文書ファイルの検索結果

図　3-3　経済産業省の行政文書ファイルの詳細検索画面
分類項目をメニュー形式で選択することもできる

図　3-4　経済産業省の報道発表資料の例

（1）簡単な検索実験

経済産業省の Web サイトには、Web ページ上の情報を検索するホームページ検索（以下 HP 検索）機能もある。紙ベースを主とした行政文書を検索する行政文書管理簿の検索とは別であり、検索結果も異なる。表　3-1は幾つかの検索語を用いて両者の検索結果（件数）の比較を行なった結果である。

検索語	行政文書検索	HP 検索
JEITA	1	21
電子情報技術産業協会	0	95
JEIDA	0	36
日本電子工業振興協会	5	66
電子協	2	14

注）行政文書の 2 件は水産電子協会がヒット HP 検索は少なくとも JEIDA の意味のものがある。

表　3-1　METI 公開情報の検索実験

行政情報管理簿自体が紙情報のメタデータのようなものなので、キーワード検索では、上記のように検索結果の件数も少なくなっている。HP 検索では、文書内に電子協（JEIDA）などのように書かれるので検索できている。逆にあまり関係がない文書であっても、どこかに記述があれば検索結果に出てくる。

このような場合、JEIDA（日本電子工業振興協会）が JEITA（電子情報技術産業協会）に組織変更されたことや、日本電子工業振興協会が JEIDA と呼ばれているという知識（オントロジー）を活用し、検索を高度化する余地がある。

３.１.１.２ 国内の他の行政公開情報の調査

利用者側からみると、どの情報がどの省庁あるいはその他の施設等にあるかは判りにくく、省庁間で横断検索できるような機能は重要である。そのような横断検索を行なえるWebサイトとして、電子政府のポータルサイトである電子政府総合窓口がある。ここでは、電子政府総合窓口を中心に調査を行なった結果について述べる。

（１）電子政府総合窓口

電子政府総合窓口[80]では、各省庁のホームページの検索を行なうことができる。地方部局やその他施設なども検索できる[81]。（図 ３-5参照のこと。）

全省庁ホームページ検索に対し、前述と同様の検索実験を行なったところ、表 ３-2のようになった。

検索語	経済産業省		電子政府総合窓口				
	行政文書管理	HP 検索	産業省	関連省庁	関連団体	産業省以外	
JEITA	1	21	4	4	8	5	1
電子情報技術産業協会	0	95	28	39	48	56	17
JEIDA	0	36	22	22	35	35	13
日本電子工業振興協会	5	66	40	49	68	113	64
電子協	2	14	2	2	5	196	194

表 ３-2　電子政府総合窓口での検索実験

表 ３-2で電子政府総合窓口の「産業省」の欄は、検索対象に経済産業省を選択した場合、「関連省庁」の欄は、経済産業省に加え、資源エネルギー庁、特許庁、中小企業庁の関連庁を追加した場合、「関連団体」の欄は、さらに、地方部局や産業技術総合研究所などの関連機関を加えた場合、「産業省以外」の欄は、全本省庁で検索した結果及びそこから経済産業省関連機関を除いた結果の件数である。（図 ３-6参照のこと。）

経済産業省と電子政府総合窓口の検索結果が合わないことは、対象としているWebサイトの範囲が異なることや索引の作り方による検索漏れや同一情報の重複度合などの違いが原因と推定される。

意外に「産業省以外」の検索でも、経済産業省の関連団体である日本電子工業振興協会は引用されることが多い。経済産業省以外では、総務省が多かった[82]。

電子政府総合窓口では、行政文書ファイルに対する検索も省庁横断的に行なう事ができる。申請様式の横断検索も行なえるが、キーワード及び分類から選択していくようになっている。

---

[80] http://www.e-gov.go.jp/

[81] これに対して経済産業省のホームページ検索は http://www.meti.go.jp/ 以下のみを対象としていると思われる。

[82] 但し、経済産業省以外での「電子協」のヒットには、米国電子協会などでのヒットが多い。

図　3-5　電子政府のポータル「電子政府の総合窓口」

図　3-6　電子政府の総合窓口の詳細検索画面
本省庁のほか地方支分部局、その他施設等機関等も検索対象にできる

3.1.1.3 公開情報に対するセマンティック Web 技術活用の方向性

ここでは、上記公開情報を対象として、セマンティック Web 技術を活用し改良する一般的な方向性について述べる。

（1）一般的な機能拡張の方向性

前節で述べた公開情報のうち、（A）については、それ自体が紙媒体による情報に対する書誌情報であり、メタデータと呼べるが、これらのメタデータを追加・改良を行なうことを考えると、人手によりデータの追加・修正を行なうことになり、大きなコストがかかる。ここでは、主に（B）及び（C）、すなわち一般の Web を対象として、セマンティック Web の技術を用いてどのような改良が行なえるかの案を述べる。このうち幾つかは（A）の検索性能の向上にも寄与する。

セマンティック Web を用いた改良の方向性としては、以下が考えられる。

(1) リソース発見（検索機能）の高度化

- 検索条件の緩和：メタデータの情報を利用して、対象情報固有の類義語辞書を作成する。特に、メタデータを対象として検索を行なう場合、使用されている語が少ないため、条件を緩和しないとヒットしにくい。
- 知的検索機能の強化：メタデータやオントロジー情報を活用し、質問の意味内容に沿った検索を行なうことにより、検索の精度をあげることができる。
- 検索結果のグループ化：省庁を横断する検索システムによって検索対象となるドキュメントが豊富になる一方、検索結果として得られる項目も増大する。特に、類義語などを用いて検索条件を緩和する場合には、ヒットする項目が爆発的に増えることが予想され、利用者にとって必ずしも良い結果にはならない。したがって、検索結果のグループ分けや、不要な情報の隠蔽あるいは排除が重要になる。
- 他人の評価情報を活用する：他人がつけたメタデータやアノテーション、投票結果などを参考にする

⑵蓄積情報の構造化（リソース作成、管理）：メタデータやオントロジーで構造を記述することにより、一つのリソースを多角的に見せたり、構造から検索を行なったりすることが可能となる。

⑶提示情報の構造化（受け取り手の再利用）：情報を構造化することにより、利用者の望む方法での利用や再利用することができるようにする。

⑷上記を用いた自動化エージェント

（2）具体的改善案

経済産業省のホームページの内容（情報公開のページに限定せず）に則してより具体的には以下が考えられる。但し、これらもまだ、方向性を示すものであり、実現性の検証がなされていないものも含まれる。

⑴ 検索の高度化

1) 検索条件の緩和：

i) 電子政府総合窓口の検索では検索条件として「電子商取引」を指定した場合と「EC」を指定した場合では異なった結果が得られた。現状では、可能な全

ての同義語を利用者が順次指定して検索する必要があるが、意味情報を使うことでこの操作を自動化できる可能性がある。また、同義語間の変換を既に実現している検索エンジンにおいても、語彙の保守や拡張が容易にするために汎用のオントロジーを使うことは考えられる。

ii) 意味情報を使って包含関係にある概念にまで検索条件を緩和することも考えられる。例えば、検索条件に「自動車」を与えた時に、「自動車」に含まれる「乗用車」や「トラック」などについての検索も自動的に行なう場合で、このことで有効な情報が収集できる場合もあり得る。

iii) 行政文書管理簿の項で述べたが、各組織の正式名称に対する俗称や、MITI から METI、JEIDA から JEITA といった組織変更など、経済産業省関連の用語の関係性が、オントロジーとして記述されているとすると、それを用いて類語辞書を作れば、「JEIDA」で検索できなかったような漏れを防ぐことができる。

2) 知的検索機能の強化：

iv) 上記、ii)は、知的検索の点でも有効である。

v) 特定の語彙が、組織名であることが推定できれば、組織名である事を利用して検索及び絞り込みを行なうことができる。

vi) 時系列を考慮した資源内容変更の追跡

時系列情報のメタデータを活用した検索。具体的には組織の改編、消滅に伴う名称変更や申請プロセス自体の変更により、資源に対する問い合わせ先や資源の存在が不明になる場合がこれに相当する。

vii) オントロジーの公理の活用によりある程度の推論を行なうこともできるようになる。例えば、親の情報が明示されていなくても、子供の情報が記載されているサイトから、親の情報を推論するなど。

3) 検索結果のグループ化：

viii) HP 検索などのように検索結果が多いときに、概念階層の分類を用いて、大量に出現する解を分類表示するのも有効と考えられる。

例：**「JEIDA」で検索した結果、「2000 年問題」＊＊件、「リサイクル関連」＊＊件・・・**

また、検索の絞込み条件の候補として上記分類を提示することもできる。

ix) 意味的な関連を、検索後のドキュメント間の順位付けに利用することも可能かもしれない。例えば、検索条件に「乗用車」を与えた場合に「自動車」や「トラック」についても検索するが、表示順序は「乗用車」、「自動車」、「トラック」の順とする、などが考えられる。もし検索エンジンに手が入れられるならば、検索時のスコアリングを「乗用車」＞「自動車」＞＞「トラック」とすることで、よりエレガントな結果が得られるかもしれない。

x) 利用者プロファイルの活用：個人に関するメタデータである利用者プロファイルを活用して、個人の興味ある情報を優先して表示する。また、セキュリティーのために、相手に応じて表示する項目を変更することも考えられる。

4) 他人の評価情報の活用：

・ 他の利用者の意見やアノテーションを付加することにより、電子政府の

使い勝手を向上させることができる。

・ 電子投票やパブリックコメントにも関連すると思われる

(2) 蓄積情報の構造化（リソース作成、管理）

1) 概念階層を提示した検索・情報提示

・ 定義された概念階層を活用して、概念階層を利用した情報提示ページの自動作成（自動分類）が可能となる。

2) 部署や団体の変更(移動、名称変更)への対応：組織名称の一括変更など

・ 組織名称の一括変更や、組織と人との対応の更新など

3) 同一のリソースを、多様な観点で提供することが可能：(政策、部署、日付等々)：

・ ニュースリリースの検索

・ 各組織住所一覧などの管理の簡略化

・ 申請プロセスの生成、提示の枠組み

明文化されていない事実に関して、資源間の関連をユーザーの意向に沿って抽出し直すことにより、既存資源を異なった視点で閲覧することが可能となる。ある実施例のプロセスが明記されていない場合に、特定の要素に着目した処理手順の追跡やあるリポジトリに着目した場合に、存在する情報の抽象化・俯瞰的視点を提供する枠組みがこれに相当する。

4) リソース間の(半)自動リンクづけ：

・ 組織変更に伴うリソース移動の自動化

・ 関連組織からのリンクの更新の自動化

(3) 提示情報の構造化（受け取り手の再利用を助ける）

例えば、以下のようなことが考えられる。

1) ユニバーサルアクセス対応：

・ 住民からのアクセスのためには必要であり、障害者対応などを含め、省庁関連では重要である。情報に適切なメタデータを付加することにより、携帯電話や今後出現するモバイル端末への対応や、視覚障害者に対して音声で応答するシステムとの連携を、文字により良い情報が提供できるようにすることができる。

2) 電子申請の例：

・ 例えば、財団法人ニューメディア開発協会の提案している電子申請書類一式に対する文書のメタデータであるパッケージメタを活用すると、住民側で申請書類作成支援ソフトウェアを用いて、文書管理機能、添付書類の一覧表示機能、必須添付文書のチェック機能からその情報を利用できる。

3) 住所にメタデータをつける。それが住所だと分かれば、それを位置情報として変換して色々なサービスを行なうことが考えられる。

・ その位置を地図上に表示する。同じ地図上に複数の位置を表示することで相対位置関係がわかる。

・ その位置から最寄駅を調べる。「乗り換え案内」などと連動して、自分の乗車駅だけを入力すればそこに行くまでのルート、時間などが分かる。

・ その位置から自動車でのルートを調べる。アクセスするルートが分かれば、現在のそのルート交通情報などが分かる。

4) 日時にメタデータをつける。
・ カレンダー、年表ビューでの表示を行なう。
・ 時間軸線上にプロットし、他のイベントとの関係を見やすくする。

5) 数値にメタデータをつける。
・ グラフ化できる。自分の好きなグラフタイプで表示する。
・ 分析のデータとして利用する。

6) 会議などの時間場所（住所、会議室）などの情報にメタデータをつける。
・ PIM やスケジューラーとの連携を取る。
・ 3)のサービスなどとの連携を取る。

(4) 上記を用いた自動化エージェント

例えば以下のものが考えられる。

1) 新着情報のメールによる通知サービス[83]において、現在手動で行なっていると推定される作業を、ニュースリリースなどコンテンツの作成を行なうだけで、自動的に新着情報として配布できるようにする。この目的のために RSS のシンジケーションを用いることもできる。

2) 「What ’ s　New」 「入札」「公募」「イベント」などのページを定期的に見て利用者の興味ありそうな情報をメタデータにより判定し、利用者に通知するエージェント。

3) 例えば、各地方部局、関連施設等のスケジュール情報を定期的にチェックし、必要なスケジュールの一覧を作成するエージェント。

[83] 新着情報配信サービス（http://www.meti.go.jp/mailservices/index.html）

３.１.２ 海外政府のメタデータ付与に関する調査

海外政府におけるメタデータ分類方法について調査を行なった。

各国政府のメタデータエレメント標準は、表 ３-3の通りである[84]。表の中の◎は必須、○は任意のエレメントを意味する。

英国の e-GMF、EU の MIReG、及びオーストラリアの AGLS については、ダブリンコアの基本となる 15 エレメントに、いくつかの拡張エレメントを追加したものである。米国の GILS については、ダブリンコアを拡張したものではないが、e-GMF の文書「英国電子政府メタデータ標準」の中で、e-GMF との対応関係が記載されており、結局、ダブリンコアの 15 エレメントを GILS のエレメントにマッピングすることが可能である。このように、欧米諸国の政府機関においては、ダブリンコアがメタデータエレメントの事実上の国際標準（デファクト・スタンダード）として採用されている。

ダブリンコアの各エレメントの詳細については、３．１．４節で説明する。

---

[84]各メタデータ標準のリソースは、以下の通りである。ダブリンコアは http://dublincore.org/documents/dces/、e-GMF は「英国電子政府メタデータ標準」、MIReG は「MIReG メタデータ要素セット」、AGLS は http://www.naa.gov.au/recordkeeping/gov_online/agls/user_manual/agls_metadata_elements.html、GILS は http://www.gils.net/elements.html、Government of Canada は http://www.cio-dpi.gc.ca/clf-upe/6/6a_e.asp#6.3 である。

[85] 付録２を参照のこと。

エレメント名	ダブリンコア基本15エレメント（Simple DC）	e-GMF（英国）	MIReG（EU）	AGLS（オーストラリア：Australian Government Locator Service）	GILS（米国：Government Information Locator Service）	Government of Canada（カナダ）
Coverage	○	○	○	○	○（複数のエレメントから成る）	―
Description	○	○	○	○	○（Abstract）	○
Relation	○	○	○	○	○（Cross reference）	―
Source	○	○	○	○	○（Source of data）	○
Subject	○	◎	◎	◎※1	○（2つのエレメントから成る）	◎（Controlled Subject）
Title	○	◎	◎	◎	○（Folder title と Document title）	◎
Type	○	○	○	○	○	○
Creator	○	◎	◎	◎	○（Originator）	◎（Originator）
Contributor	○	○	○	○	○	○
Publisher	○	○	○	◎	○（Distributor）	―
Rights	○	○	○	○（Availability と Rights）	○（Availability と Access constraints）	―
Date	○	◎	◎	◎	○（Date of Publication 等3つのエレメントから成る）	◎
Format	○	○	○	○	○（Medium の詳細項目）	―
Identifier	○	○	○	◎※2	○（Schedule number）	―
Language	○	○	○	○	○（Language of resource）	◎（Language of resource）
Audience	―	○	○	○	―	―
Disposal	―	○	○	―	―	―
Location	―	○	○	―	―	―
Reservation	―	○	○	―	―	―
Function	―	―	―	◎※1	不明	―
Availavility	―	―	―	◎※2	不明	―
Mandate	―	―	―	○	不明	―
Keywords	―	―	―	―	不明	○
備考	―	ドラフト標準。各エレメントの下に詳細項目あり。各エレメントには、必須。任意、特定の状況で推奨、の区分がある	ドラフト標準。各エレメントの下に詳細項目あり。各エレメントには、必須。任意、特定の状況で推奨、の区分がある	※1「Subject」または「Function」のいずれかが必須 ※2「Identifier」または「Avilavility」のいずれかが必須	The UK e-Government Metadata Standard V2（英国電子政府メタデータ標準）において、e-GMF と GILS の対応関係が記載されている。	カナダ政府のサイトは上記の5つの◎のメタタグを採用しなければならない。 メタタグジェネレータ有り。 http://198.103.99.147/publications/metagen_e.html

表 3-3 海外政府のメタデータエレメント標準

３.１.３ メタデータ階層モデル調査

国内の行政公開情報にメタデータを付与するにあたって、どのようなメタデータの階層モデルが相応しいかを調査した。ここでは、メタデータ付与の方向性としてオントロジーの応用から、メタデータを目的に応じて階層的に構成する方式を提案する。この方式は、オントロジーのマッピングを用いないレベルでも有効である。

（１）メタデータとオントロジー

セマンティック Web 技術で議論されているメタデータは、２．１節で見たように、RDF の３つ組（リソース、プロパティ、値）で表現される。さらに、その上で規則の形での制約（公理と呼ばれることもある）が付加される。

３．１．２節で見たように、いくつかの海外政府においてはメタデータ標準が策定されている。このことは、RDF の 3 つ組の表現で言えば、政府が所有する情報（リソース）に対するプロパティが規定され、標準化されていることにほかならない。このような標準化とメタデータ付与によって、政府所有の大量で複数のリソースから、そのリソースの「日付」や「著者」といった何らかの意味のあるデータ（プロパティと値）を取り出すことが可能になる。

セマンティック Web 技術では、これに加え、上記 3 つ組の表現を用いて、オントロジーと呼ばれる語彙体系を定義することができる。

Uschold[86]によれば、さまざまな分野で用いられているオントロジーに対して、オントロジーの応用を分類すると大きくは以下の 3 つのカテゴリになるとしている。

１）ニュートラル・オーサリング
２）情報への共通アクセス
３）索引付け

１）は、共通に抽象的にオントロジーで記述したものを複数の環境で動作させるよう変換するものである。

２）は、（異種性を意識していない複数の）アプリケーションから異種のリソースに共通にアクセスを可能にするものである。

３）は、情報検索のための応用で、いわゆる知的検索とか概念検索とか呼ばれている応用であり、オントロジーの語彙体系を用いて検索の高機能化を図るものである。

Uschold はさらに２）を、

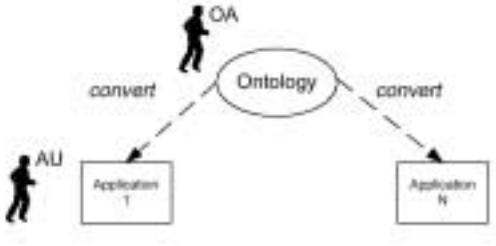
A）共有オントロジーを用いたデータアクセス、
B）オントロジーのマッピングによるデータアクセス、
C）共有サービス

のバリエーションに分類している。A）は共有されたオントロジーに基づきデータをアクセスするもの、B）はそれぞれオントロジーを持つデータに対してオントロジーをマッピングすることにより相互利用を行なえるようにするもの、C）は共有のオントロジーの仕

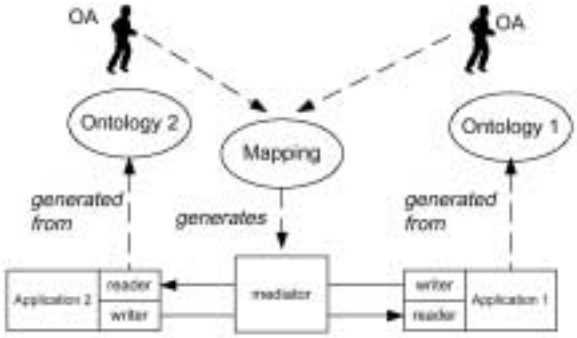
[86] Uschold, M., Jasper, R., A Framework for Understanding and Classifying Ontology Applications, *in Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5),* Stockholm, Sweden, August 2, 1999.

様に従い異なった環境でのサービスを実装するものである。

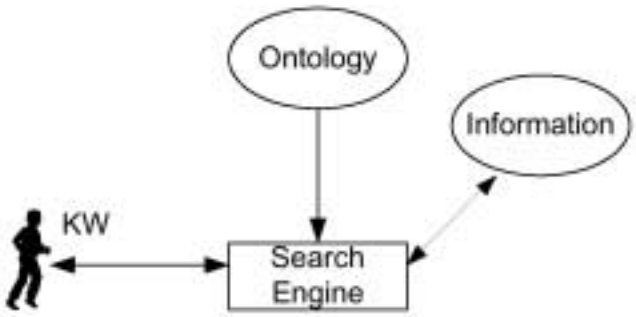
図 3-7は、Uschold[86]の分類の概略図である。オントロジーの一般的な適用方法の説明が目的でありさまざまなバリエーションも存在するので、セマンティック Web 技術の適用システム全体を表現するものではないが、この分類に従えば、セマンティック Web でのメタデータの応用では以下のように解釈されると思われる。1）は、メタデータ標準など典型的な応用のメタデータセットを作成して、それを基に多くの場所で活用する例と考えられる。あるいは、RDF スキーマの定義に基づき、個々のスキーマを記述するモデルと考えることができる。3）のリソース発見のための索引付けの利用でありる。2）は、この中では最も高度な活用法で、情報への共有アクセスのためのマッピングとしてオントロジーが活用される。前節まで議論されているメタデータ標準は、この2）の段階の A)のアプローチのひとつと言えよう。



1）ニュートラル・オーサリング



2）・オントロジーマッピングによるデータアクセス（情報への共通アクセスの B））



3）概念検索（索引付け）

図 3-7　オントロジーの適用パターン例（Uschold [86]より）

（2）メタデータの階層モデル

メタデータ付けの対象となるリソースにはさまざまなものが考えられる。一口に電子政府、あるいはその公開情報と言っても、定期刊行物やニュース、統計情報などその様式は多様であると考えられる。RDF の 3 つ組では、クラスを用いて、クラス毎に付加する情報（プロパティ）を定義することができる。

3．1．1．1節で述べたように、行政文書管理簿には、「管理担当課室・係」という属性がある。多くの場合、情報の管理責任部署が存在すると仮定するのは極めて自然である。管理部署による特有の表記方法なども存在する。その部署の管理にあった方法でメタデータを付与するのが自然である。メタデータ付与の課題となっていた、メタデータ付与者のメリットの問題も、管理責任組織が存在し、メタデータが、その管理目的に適合する場合は、問題なく、整合のとれたメタデータが保守管理されることが期待できる。逆に標準化として、使わないデータを大量に入力し・保守しなければならないとなると、普及は困難になる。

メタデータは、対象となるリソースに埋め込むことも、別のファイルにすることも、あるいは別のサイトに設置することもできる。このことは、管理部門ではないサイトのリソースや、複数のリソースサイトをまたがったリソースのメタデータを作成することもできることを意味している。

一方、利用者側からはしばしば様式や管理部署を問わず検索を行ないたいという要求がある。横断検索などを行なうためには、異種の管理体系のリソースをアクセスしてメタデータを作成・取得しなければならない。このためには、各データベースのアクセス方法や、プロパティの体系などを個別に知る必要があり、非常に煩雑な作業となる。この時、各リソースが既にあるメタデータを持っていたとするとそれを活用する事により、前述の B) のオントロジーのマッピングにより、その作業が軽減できる。この段階では、プロパティの規定に従い、プロパティどうしのマッピングを行なうレベルでもありうるが、プロパティやオブジェクト間の関係を記述した RDF によるメタデータが存在すれば、より柔軟な処理を行なうことが可能になる。

この結果、メタデータの階層が構成される。上位のメタデータは、下位層のメタデータを利用し必要な情報をそれに付加することで構成できる。目的に応じて、柔軟にメタデータを構成することができる。

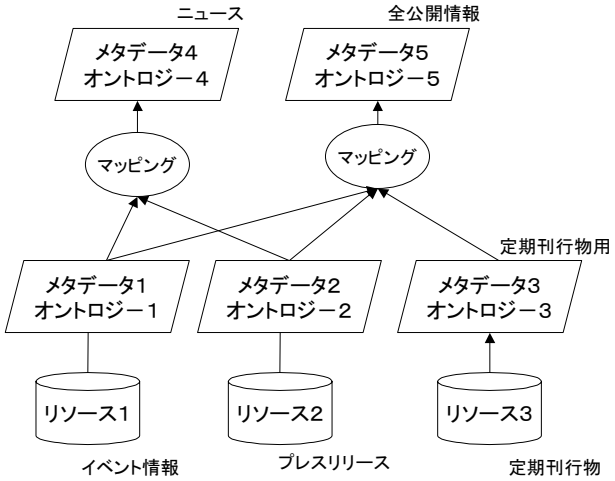


図　3-8　メタデータの階層モデル

このような構成を採ることで、もし各リソース管理部署の都合で、ある部署のオントロジーが変更になった場合や、上位のオントロジーを変更する場合などは、マッピングを変更するだけで対処することができる。

図　3-8では、マッピング機能を表現するために中間に置いている。マッピングは、上位と下位のメタデータ／オントロジーから自動生成されるのが理想的だが、最も簡易には、マッピングは上位のメタデータ／オントロジー（RDF 記述）に記述されていても良い。

例えば、メタデータにより、標準用語や標準語彙を定義し、領域や組織により異なる同義語や類義語を同等に取り扱う事ができるが､メタデータで定める標準用語や標準語彙は、永久的なものではありえず、時間と共に、その意味が変化することも起こり得る。このような時にメタデータの標準用語や標準語彙が変わった時、既存のメタデータを作りなおすことは、非効率である。この場合、より上位のメタデータ（RDF 記述）により、その新たな意味を定義することができれば、既存のメタデータに手を入れる必要がなくなる。

３.１.４ メタデータ定義調査

国内における行政機関公開情報の現状、及び行政文書に対するメタデータ付与の各国政府の現状を踏まえて、国内の行政機関公開情報に対するメタデータ定義について述べる。

意味情報検索の観点から情報資源をより有効に活用することが目的である。

（１）国内向け行政機関公開情報に対するメタデータ定義に関する一提案

既存の各国政府メタデータエレメント標準は、前述の通りダブリンコア（DC）に準拠している（３．１．２節参照）。ダブリンコアの基本エレメントは定義内容が明確で一般に普及し易いという利点がある一方で、精度の高い情報を記述する場合やコミュニティー独自の情報を記述するためにはメタデータエレメントの拡張や精密化が必要となる。

こうした背景と各国政府の現状を踏まえ、以下にダブリンコアのエレメントを基礎として、国内の公開情報文書へメタデータを付与する場合に考慮すべき点を述べる。

○既存のメタデータに対する国内行政文書の親和性([]内はエレメント名)

[Title]タイトル、情報資源に与えられた名前

一般には作者もしくは公開者によって与えられる。 国内の行政文書における「タイトル」記述に対し直接的に写像できると思われるが、検索における有効性ならびに構造を定義したタイトル付けが必要である。

例えば、ある既存の行政文書情報ページでは表題項目が[大/中/小]の３分類に加え、タイトル項目が[1/2/3]行目といった詳細な分類が施されているが（環境省の行政文書情報トップ [87]> 書誌情報目次[88]を参照）タイトル列における意味構造は見られない。この行政文書情報ページの中の分類表を一部抜粋する（表 3-4参照のこと）。

大分類名	中分類名	小分類名	タイトル１行目	タイトル２行目	タイトル３行目
廃棄物・リサイクル	循環企画	リサイクル	第５、６回ごみ減量化国民大会	-	-
廃棄物・リサイクル	循環企画	リサイクル	分別収集促進計画、市町村分別収集計画	愛知県	-
廃棄物・リサイクル	循環企画	リサイクル	平成１２年度	保管施設一覧	岩手県、宮城県、秋田県、山形県、福島県
廃棄物・リサイクル	循環企画	広域臨海環境整備センター	フェニックス法　大綱・要綱等作成過程	大綱、要綱、五六予算、件名登録、関係資料	昭和五五年八月～五六年一月

表 3-4　行政文書の分類例

またこの形式では、正式名称とは異なる慣習的な名称が広く知られているプロジェクト等の文書を参照するのは困難と思われる。

既存の海外政府メタデータ標準の枠組みを直に利用するだけでも、こうした現状を大幅に改善することが可能である。[Title]エレメントには”別名”の記述が可能な詳細エレメントが用意されているため、これらを選択的に利用することで文書を効果的に同

[87] http://www.env.go.jp/guide/bunsho/index.html
[88] http://www.env.go.jp/guide/bunsho/syoshi_mokuzi.html

定できる。さらに年度や地域情報等の記述は後述の［Description][Date][Coverage]のエレメントによる別定義、また分類は[Subject]エレメントを用いて再定義することにより、可読性が高く、かつ機械的な検索にも有効なメタデータ定義が付与できる。

例： 付与部分例（例示目的のため、ダブリンコア以外のエレメントも混在して用いている。また記載の URL、e-mail アドレス等は全て架空のものである。）

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/RDF/RDF/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
     xmlns:dcq="http://purl.org/dc/terms/"
      <rdf:Description about="http://www.sample.go.jp/doc/sample.html">
  <dc:title>平成 n 年度 ～の実態に関する調査報告書</dc:title>
  <dc:creator>制作者; ○○課○○室 foo@sample.go.jp</dc:creator>
  <dc:creator>連絡先; ○○課○○室 foo@sample.go.jp</dc:creator>
  <dc:creator>責任者; ○○長 director@sample.go.jp</dc:creator>
  <dc:description>この文書は、△△計画における□□に関する動向調査の記録であり、××という現状から、～への対処が急務であるとの結論を得た。</dc:description>
   <dc:coverage.spatial.region>○○県,△△市</dc:coverage.spatial.region>
  <dc:coverage.temporal>2002-3-31/2003-3-31</dc:coverage.temporal>
  <dc:data>2002-02-21</dc:data>
  <dc:format>text/html</dc:format>
  <dc:language>ja</dc:language>
                :
                :
 </rdf:Description>
</rdf:RDF>
```

本エレメント記述を実際に行なう上で、考慮すべき点を以下に示す[89]。

- 情報資源の名前として一般的に知られているものを使用する。
- 本エレメントは必須とされているため、情報資源のタイトルが存在しない場合には新たに作成する必要がある。人目を引くためではなく、簡潔で意味のあるものにするべきである。
- 情報資源が e-mail の場合には、[Title]エレメントとして電子メールの件名である”Subject”を使用する。
- Alternative title（＝別称を定義するエレメント名）には、省略名や翻訳したタイトルを記載する場合もある。
- 情報資源の正式なタイトルが一般に不明確な場合、意味のある[Title]エレメントを Alternative title として付けるべきである。

[89] 以下、各エレメントにおける箇条書きの項目は、そのエレメントにおいて考慮すべき点を意味する。

- 何がタイトルか分からない資源に対しては、可能性のあるタイトルの[Title]エレメントを列挙しておく用法も認められる。
- 同じタイトルを持つ項目が複数ある場合、言語やバージョン番号、状態（例：草案、ドラフト）や日付などの値を[Title]エレメントに付加することで検索に役立てることができる。
- 情報資源が電子文書を含む電子フォルダの場合、フォルダ名を[Title]エレメントとして付けるべきである。
- [Title]エレメントは、センテンスとして書いた方がその内容が明確になる。

[Subject]主題及びキーワード

情報資源の主題あるいは内容を説明するキーワードや句を意味する。行政文書の参照方法として、大分類-中分類-小分類とカテゴリをドリルダウンする枠組みを動的に提供する場合に有効である。また、この[Subject]エレメントの定義として、全省庁を範囲とする統一的な分類カテゴリ、キーワード、体系コードに類するものを適切に策定できれば検索精度の向上が期待できる。このエレメントを用いた体系コード例(分類スキーマ)としては LCSH、IDC、MeSH、DDC、LCC、UDC 等が存在する。

- 本エレメントには、情報資源の最も重要なもの及びユニークな言葉を選ぶ。利用者が必要なものを正確に検索可能とするために必要十分な言葉を選択すべきである。
- 一般的な用語よりも特化された用語を使用することで、情報過多になるのを防ぐことができる。
- 本エレメントとして、同意語や言い換えた言葉、大文字と小文字を変えただけのものなどを列挙すべきではない。

[Creator]著者あるいは作者を表すラベル

情報資源の内容作成に関し、責任を持つ者や組織に関する記述。行政情報文書では作成者（課/室）がこれに相当する。責任者への連絡手段(e-mail)があることが望ましいが、作成元の組織改編・消滅に伴い文書追跡が困難になることも想定されるため、組織の完全な階層構造を記述しておくことが推奨される。

- 本エレメントに記載された人・組織の代表者は、法律上の責任・義務を負う。また、監査を行なう場合には人名を明記しておくことが必須となる。
- 役職や部署の省略名を用いることは、利用者にとって利用を困難にする恐れがあるため、正式な名前を使用するか、または（略称が記載されている）用語集へリンクを施すなどの配慮が必要である。

[Date]日付

情報資源が作成されたあるいは有効になった日付。この日付とは Coverage エレメントに記載されるものとは異なり、文書(情報資源)そのものが作成された日付である。文書内容が表す時間情報とは異なる。記述形式は YYYY(年の 4 桁表記)及び YYYY-MM-DD

（年-月-日）(例 2003-04-28) といったISO8601[90]準拠が強く推奨されている。元号表記はこの規格では対象外であるが、元号利用の手段としては(1)元号を正しく時間記述形式に写像できるような日時情報に関する(オントロジー変換としての)変換機構を備える (2)[Description]若しくは[Subject]エレメントに明記し文書情報の一部として扱うことが考えられる。

- [Date]エレメントは、世界中の人々が認識でき、且つ、ソフトウェアが解釈できる形式でなければならない。
- 時間が必要である場合、hh:mm を付加する。hh は時間を示す（24 時間を使用した場合）mm は分を示す。(例　13：23)

[Description]内容記述
情報資源の内容に関する説明記述。文書の場合の抄録、視覚的資料の場合の内容記述など。メタデータの付与対象（リソース）は文書に限らず動画・画像等もその範疇に入るため、今後公開情報として閲覧、検索が物理的に容易でないものも含む場合には(動画を含むメディア等)本エレメントの情報が検索に際し重要である。

- [Description]の記述内容として推奨される例を以下に示す。
  主題の補足説明
  情報資源の作成理由
  参照されるグループと組織
  適用されるイベント
  キーとなるフィールド（データベース）または章のリスト
  その他有用な情報
- [Description]エレメントの記述内容については、簡潔かつ正確であるべきである。
- 複数の[Description]エレメントを記述した場合、先頭のエレメントが利用者の目に入る可能性が高いため、一番大切な部分を同エレメント中で最初に記述する。
- 他のエレメント（[Title]エレメント, [Coverage]エレメント、[Subject]エレメントなど）内にある情報を[Description]エレメントと重複して記載すべきではない。

[Publisher] 情報資源を公開する人・組織
情報資源を現在の形態で利用可能にしたことに責任を持つ人や組織。情報資源の再発行や複製の許可を得る目的で参照される。行政文書の公開に関しては管理担当課室・係がこれに相当する。

- エレメントには、情報資源が存在する Web サイトを管理する組織を示す場合もある。
- 人名よりも職名を記述するのが適切な場合もある。
- 本エレメントは情報に関する法的な権利と責任を明記するという意味で、必須のエレメントである。

[Contributor] 貢献者

[90] http://www.w3.org/TR/NOTE-datetime.html

[Creator]エレメントには記載のない人あるいは組織で、情報資源を作り出すに当たって知的に重要な貢献をしたもの。編集者、翻訳者等にあたる。

- 本エレメントの記述内容は、Creator エレメントの記述との重複を避けるべきである。
- 人名よりも役職名・組織名を記述することが適切な場合もある。

[Type]情報資源の種類

例えば、ホームページ、ワーキングペーパー、テクニカルレポート、辞書・事典などの分類を記述する。行政機関文書では、カテゴリ別情報として提供されている分類ラベル(プレスリリース、FAQ、白書、報告書、調査研究結果、統計等)がこれに相当すると思われるが、相互利用性を保つためには、関係各省にてこのような用語の統一を図ることが必要である。

- 本エレメントを[Format]エレメントと混同して用いるべきではない。[Format]エレメントは例えばワード、エクセル、一太郎といったファイル形式を記述するために用いられるが、本エレメントは情報資源の内容そのもの指す。
- 公式の記録や文書は、ある大きなカテゴリの一部として存在することがある。本エレメントはそのような場合に、どの階層に属する情報資源であるかを明記するために用いることもできる（＝より詳細なエレメントである”Type.Aggregation Level”の利用による）。
- 複数の[Type]エレメントを記述することで、内容をより詳細に特定（記述）することができる。

[Format]情報資源のデータフォーマット

情報資源の表示・動作に必要なソフトウェアや、場合によってはハードウェアを識別するために利用できる情報を記述する。相互利用性を保つために関係省庁で統一見解が得られた用語リストの中から選ぶことが強く推奨される。

- 情報資源が別のフォーマットでも利用可能である場合には、[Relation]エレメントを併用してその旨を明記しておく。
- 本エレメントは情報資源が格納されている物理的な形式を記述する。例えば「本」、文書ファイルであれば「プレーンテキスト」、「一太郎」、「ワード」、データベースであれば「Access97 形式」等の記述となる。

[Identifier]識別子

情報資源を一意に識別するための文字列もしくは番号。国際標準図書番号(ISBN)や他の標準化された名前のように全世界的に一意に定まる識別子の定義を行なう。本エレメントとして URL を使用する場合には、（URL 変更に伴う）不整合が生じないよう注意すべき必要がある。

[Source]元情報資源

情報資源を作り出す元になった別の情報資源に関する情報。一般にエレメントには、当該情報資源に関する情報のみを記述することが推奨されているが、本エレメントには当該情報資源を見つけ出すために有用である別の情報資源に関する日付、作者、形式、識別子あるいは他のメタデータを書くことができる。実際の利用に際しては他文

書からの図、写真といった引用に近い参照の場合がこれに相当する。より抽象的な資源間の関連づけは[Relation]によって定義される。

[Language]言語
情報資源の内容を記述するために用いられている言語（コード）。実際のエレメント記述は RFC1766[91]に準拠した en、de、es、ja、th、zh 等の表記を用いる。日本語による文書は”ja”である。

[Relation]参照関係
別の情報資源との関係。例として、関連するプレスリリースの定義、同一のフォルダに属する文書の定義(Relation.IsPartOf)、参照先文書のバージョン情報(Relation.IsVersionOf)、参照先文書の識別子(Relation.References)等の定義を記述しておく。多様なカテゴリ分類が想定される行政文書においては、相互利用性を得るために情報資源間の関係を表す値は省庁間における統一が必要であり、少なくとも現在定義が進められている値のリスト中から選択して与えることが推奨される。なお関連先の指定は、参照先文書を一意に特定できるような識別子による指定が望ましい。

[Coverage]適用範囲
情報資源の内容に関する空間的(地理的)あるいは時間的特性。空間的範囲は物理的な範囲として座標(例えば経度と緯度)や、統一見解の得られた語彙リストの中から選ばれた地名（都道府県名・市町村名）を用いる。時間的範囲は当該情報資源が表している内容に関する時間的情報であり、情報資源の作成や公開に関する日付は[Date]エレメントによる。日付・時間の表記形式は[Date]と同様 ISO8601 に基づく[92]

[Rights]権利情報
権利管理に関する声明文、権利管理に関する声明文へのリンクを表す識別子、あるいは当該情報資源の権利管理に関する情報を提供するサービスへのリンクを表す識別子を記述しておく。

[Audience]利用者情報

情報資源の利用者として意図されているカテゴリ分類の記述。対象者カテゴリとしては例えば教職員、学生、技術者、○○市民、といったものが想定されている。本エレメントは情報資源の階層内における位置付けや、検索に際しある特定の人々を対象とした情報を抽出する際のフィルタリング目的で利用される。

[Disposal]保管・廃棄情報

情報資源の保管期間、廃棄時期に関する記述。情報資源を管理する上で必須の項目である。廃棄時期を迎えた場合の資源の処理方法も本エレメントの詳細項目である”Action”に記述できる。また、情報資源が Web ページである場合には、本エレメン

[91] http://ds.internic.net/rfc/rfc1766.txt
[92] http://www.w3.org/TR/NOTE-datetime.html

トの詳細項目の一つである”Review”に更新時期を記述しておくことで、更新されずに古いまま存在しているページを管理者が容易に把握するといった利用方法も可能である。

[Location]保存場所

情報資源の保存場所に関する記述。例えば、”執務室・棚番号 15”といった記述であり、電子的な情報であれば、DVD-ROM や DAT といった保存メディアの種別を付記しておくのも有効である。（例：地下倉庫・1-A 室・disc 番号 1832B[CD-ROM]）国内の行政公開文書には必須の項目と思われる。

[Preservation]永久保存データに関するコメント

公的文書を長期的に保存管理する上で必要な項目であり、管理者向けのエレメントである。技術の進歩に伴い新しい電子メディアが普及すると、旧来の電子媒体に蓄えられていた情報の読み出しが困難になることへの対処である。今後数 10 年、数 100 年といったスパンでの電子ファイルの読み出し・解釈に必要な情報を全て記載することを目標としているが、詳細は UK Public Records Office のメタデータ標準においても現在のところは検討中となっている。

[Function]業務内容

情報資源が属する業務内容に関する記述。オーストラリア政府ではこのエレメントと共に *Australian Governments' Interactive Functions Thesaurus* (AGIFT)を用いることで、業務内容（と対応する機関）を明確に分類している。エレメントの記述内容そのものは、比較的抽象度の高い分類であり、ある機関の Web ページを見た場合に、閲覧者にとって識別に役立つレベルの分類である。例えば「学校教育」「コミュニティーサービス」「チャイルドケアサービス」といった項目である。

[Availability]情報の入手・アクセスに関する情報

主として非電子化情報の入手方法やアクセス手段に関する記述。入手先情報の詳細項目として人名（組織名）、e-mail アドレス等を記載する点では Creator、Contributor、Publisher のエレメントと類似している。本エレメントではさらに（機関の）住所、郵便番号、電話、FAX、開庁時間・曜日、情報の入手に必要となる費用(有料/無料)等の記述がある。

[Mandate]関連条項・法律

当該情報資源の作成・公開の要求元となっている既存文書や条項、法律等についての記述。詳細エレメントとしては、法律・条令・事例（判例）等の種別がある。
(例：”http://www.austlii.edu.au/au/legis/cth/num_act/laa1989192/”
”England v Van Donk Matter No CA 40433/97”)

[Keywords]重要語の重み付け

情報資源中でキーワードとなる語彙に付与する。このメタデータを付与した語彙は、検索システムにとって特別の重み付けをもつ語彙として扱われることを想定する。直近の検索精度向上には有効と思われるが、”意味”構造を考慮していない点において、他のエレメントとは定義の方向性がやや異なる。

（2）海外政府メタデータ標準のエレメント名と国内行政文書との対応関係

以下ダブリンコアを軸に、海外政府で用いられているエレメント記述と、国内の行政文書の対応を表 3-5に示す。括弧書きの項目は現在の公開情報としては明確な記載が無いが、暗にその存在が仮定されるもの、記号「＊」は資源情報から定義可能なものを示す。重要度は意味情報検索の観点と実現可能性の観点から、既存の国内公開情報への付与が必須と思われる項目に「◎」記号を付けた。「参照」欄のメタデータ標準は次のものを表す。DC=ダブリンコア、MIReG=MIReG Metadata Element Set、AGLS=Australian Government Locator Service、Canada=Government On-Line(NR Can)である。

国内行政文書ファイルの情報項目	重要度	エレメント名	参照	エレメントの概要
（編集者・翻訳者）		Contributor	DC	(Creator に次いで)貢献した人・組織
＊	◎	Coverage	DC	情報資源内容の適用を受ける期間、場所
作成者	◎	Creator	DC	情報資源の作成にあたり最も責任を有するもの
作成(取得)時期	◎	Date	DC	情報資源そのものに関する日時情報
タイトル/(備考)	◎	Description	DC	情報資源に含まれる情報の説明
＊		Format	DC	情報資源の記述形式
＊	○	Identifier	DC	内容に対する一意の識別番号
(ja)		Language	DC	記述されている言語
管理担当課室・係	◎	Publisher	DC	情報資源を発行した責任者
＊	◎	Relation	DC	関連する情報資源への参照
＊		Rights	DC	情報資源全般の権利情報
＊		Source	DC	提示された情報資源への参照
大・中・小分類	◎	Subject	DC	主題及びキーワード
タイトル	◎	Title	DC	タイトル
(分類カテゴリ)	○	Type	DC	情報資源の種類
＊	○	Audience	MIReG	利用者のカテゴリ
保存期間・廃棄時期保存期間満了時の措置結果	◎	Disposal	MIReG	情報資源の保持と処理
保存場所	◎	Location	MIReG	情報資源の物理的な場所
(保存期間・保存場所)		Preservation	MIReG	永久保存資源の為の備考データ
＊	○	Function	AGLS	資源内容が関連する業務種別
＊	○	Availability	AGLS	(主としてオフライン資源に接触するための)資源作成者/管理者情報
＊		Mandate	AGLS	資源作成の為の法的根拠､条項の記述
＊		Keywords	Canada	資源のキーワード(検索システムに対する単語の重み付けに影響)

表 3-5 海外政府メタデータ標準のエレメント名と国内行政文書との対応関係

３.１.５ 利用者プロファイルのメタデータ定義に関する調査

利用者によって行政公開情報の利用目的は異なる。例えば、ある利用者は生活動態調査における消費指数を知りたいと思ったり、ある利用者は環境保護に関する法令情報を知りたいと思ったりする。これら利用者ごとの用途に柔軟に且つ効果的に対応できるようにするには、利用者個人に関するメタデータである利用者プロファイルの設定が必要である。本節では、利用者プロファイル用のメタデータ定義について調査を行なった。

３.１.５.１ P3P 調査

P3P1.0 の勧告案（The Platform for Privacy Preferences 1.0 Specification: W3C Proposed Recommendation 28 January 2002）[93]の調査を行なった。P3P とは、Platform for Privacy Preferences Project（プライバシー情報取扱いに対する個人の選好を支持する技術基盤）の略であり、W3C が開発中の、インターネット上のプライバシー保護を目的とした技術標準である。現状では、Web 上のプライバシーポリシー（個人情報の取扱い方針）は各社で記述形式がまちまちであり、また、その都度各社のプライバシーポリシーを読んで確認することは消費者には大きな負担となっている。P3P 標準は、プライバシーポリシーの掲載項目等を標準化し、かつマシンリーダブルな形式（XML 形式）で記述することによって、プライバシーポリシーを自動処理しようとするものである。

（１）カテゴリ

P3P1.0 勧告案では、利用者の個人情報に関するカテゴリとして、以下の 17 要素が挙げられている。

個々の個人データ（氏名、住所、性別、生年月日、年収、趣味等）をこれらのカテゴリ種別にしたがって分類することにより、Web サイトがそれらの個人データをどのように取り扱うかという規則を、個々の個人データごとにではなく、カテゴリごとに設定することができ、設定を容易化することができる。

<physical/>	|；実社会における連絡先情報
<online/>	|；オンライン連絡先情報
<uniqueid/>	|；ユニークな識別子
<purchase/>	|；購入情報
<financial/>	|；金融情報
<computer/>	|；コンピュータ情報
<navigation/>	|；ナビゲーションとクリックストリームのデータ
<interactive/>	|；インタラクティブデータ
<demographic/>	|；人口統計学的・社会経済学的データ
<content/>	|；文章の内容
<state/>	|；状態管理メカニズム
<political/>	|；市民情報

[93] http://www.w3.org/TR/P3P/

<health/>　　|; 健康情報
<preference/>　　|; プリファレンス（嗜好）データ
<location/>　　|; 位置データ
<government/>　　|; 政府発行の識別子
<other-category>　　|; その他

各カテゴリの説明は、以下の通りである。

<physical/>
実社会における連絡先情報：実社会において個人に問い合わせを行なったり、所在を突き止めたりすることを可能にするような情報。電話番号や住所など。
<online/>
オンライン連絡先情報：インターネット上で個人に問い合せを行なったり、所在を突き止めたりすることを可能にするような情報。電子メールアドレスなど。この情報は、ネットワークにアクセスするときに使用される特定のコンピュータには依存しないことが多い。
<uniqueid/>
ユニークな識別子：個人を整合的に特定したり、認識したりするために発行された識別子。金融機関の ID 番号を除く。また、政府発行の識別子を除く。Web サイトやサービスから発行された識別子を含む。
<purchase/>
購買情報：商品やサービスを購入することによって積極的に生成される情報。支払方法の情報を含む。
<financial/>
金融情報：口座、残高、支払い、借越し、購入、クレジットカード、デビットカードなどの個人の金融情報。 個人による個々の購買に関する情報は、それ単体では「金融情報」には属さない。
<computer/>
コンピュータ情報：個人がネットワークにアクセスするときに使用しているコンピュータシステムに関する情報。 IP アドレスやドメインネーム、ブラウザの種類、OS など。
<navigation/>
ナビゲーションとクリックストリームのデータ：Web サイトを閲覧することによって受動的に生じるデータ。訪問したページやページごとの滞在時間など。
<interactive/>
インタラクティブデータ：Web サイトを通じた、サービス提供者との明示的なやりとりから積極的に生じるデータ。また、そのようなやりとりを反映したデータ。検索エンジンでの検索事項やアカウント活動のログなど。
<demographic/>
人口統計学的・社会経済学的データ：個人の特徴に関する情報。性別や年齢、収入など。
<content/>

文章の内容：通信活動に含まれる言葉や表現。電子メールの文章や掲示板への掲示内容、またチャットルームでの通信内容など。

<state/>

状態管理メカニズム：利用者とのセッションを維持したり、また以前に特定サイトを訪問したことや特定コンテンツにアクセスしたことがある利用者を自動的に特定したりするメカニズム。クッキーなど。

<political/>

市民情報：宗教団体、労働組合、専門的な協会、政党などの会員、または所属。

<health/>

健康情報：個人の肉体的または精神的健康、性的志向、ヘルスケアサービスや製品の使用または調査、ヘルスケアサービスや製品の購入等に関する情報。

<preference/>

プリファレンス（嗜好）データ：個人の好みや嫌いなものに関するデータ。好きな色や音楽の好みなど。

<location/>

位置データ：個人の現在の物理的な位置情報と変更した場合にその位置の追跡のために使うことができる。GPS 位置データなど。

<government/>

政府発行の識別子：個人を識別するための政府が発行した識別子。

<other-category>

その他：上記の定義にあてはまらないその他のデータ。

上記のデータ要素のうち、利用者プロファイルとしての利用が可能なのは、<physical/>, <online/>, <uniqueid/>, <purchase/>, <financial/>, <computer/>, <demographic/>, <state/>, <political/>, <health/>, <preference/>, <government/>, <location/>である。ただし、金融情報<financial/>や健康情報<health/>、市民情報<political/>など、秘密性の高いデータも含まれるため、プロファイルに含まれる個々のデータについて、それを利用できる機関や、利用目的などに制限をかけることが望ましい。

その他の<navigation/>, <interactive/>, <content/>データ要素については、個人の行動履歴に関する情報であったり、一時的な情報であったりするため、利用者プロファイルの中に含めるのは適切ではない。

（2）データ要素

P3P1.0 勧告案では上記のカテゴリの他に、個々の個人データまたは個人の行動に関連したデータとして、いくつかのデータ要素が規定されている。基本データ要素としては、以下のものが挙げられている。

1. 日付

date	**カテゴリ**	**簡易表記名**

ymd.year	(可変カテゴリ)	年
ymd.month	(可変カテゴリ)	月
ymd.day	(可変カテゴリ)	日
hms.hour	(可変カテゴリ)	時
hms.minute	(可変カテゴリ)	分
hms.second	(可変カテゴリ)	秒
fractionsecond	(可変カテゴリ)	秒（小数点以下）
timezone	(カテゴリ)	タイムゾーン

2. 名前

personname	**カテゴリ**	**簡易表記名**
prefix	人口統計学的・社会経済学的データ	敬称
given	実社会における連絡先情報	名(Given Name)
family	実社会における連絡先情報	姓
middle	実社会における連絡先情報	ミドルネーム
suffix	人口統計学的・社会経済学的データ	名前の接尾語（Name Suffix）
nickname	人口統計学的・社会経済学的データ	愛称

3. ログイン

login	**カテゴリ**	**簡易表記名**
id	ユニークな識別子	ログイン ID
password	ユニークな識別子	ログインパスワード

4. 認証

certificate	**カテゴリ**	**簡易表記名**
key	ユニークな識別子	認証鍵
format	ユニークな識別子	認証フォーマット

5. 電話

telephonenum	**カテゴリ**	**簡易表記名**
intcode	実社会における連絡先情報	国番号
loccode	実社会における連絡先情報	局番
number	実社会における連絡先情報	電話番号
ext	実社会における連絡先情報	内線
comment	実社会における連絡先情報	注釈

6. 連絡先情報

contact	**カテゴリ**	**簡易表記名**
postal	実社会における連絡先情報, 人口統計学的・社会経済学的データ	郵便情報
telecom	実社会における連絡先情報	テレコミュニケーション情報
online	オンライン連絡先情報	オンラインアドレス

6.1 郵便

postal	**カテゴリ**	**簡易表記名**
name	実社会における連絡先情報, 人口統計学的・社会経済学的データ	氏名
street	実社会における連絡先情報	町・番地
city	人口統計学的・社会経済学的データ	市・区
stateprov	人口統計学的・社会経済学的データ	都道府県
postalcode	人口統計学的・社会経済学的データ	郵便番号
country	人口統計学的・社会経済学的データ	国
organization	人口統計学的・社会経済学的データ	団体

6.2 テレコミュニケーション

telecom	**カテゴリ**	**簡易表記名**
telephone	実社会における連絡先情報	電話番号
Fax	実社会における連絡先情報	ファックス番号
mobile	実社会における連絡先情報	携帯電話番号
pager	実社会における連絡先情報	ポケットベル番号

6.3 オンライン

online	**カテゴリ**	**簡易表記名**
email	オンライン連絡先情報	電子メールアドレス
uri	オンライン連絡先情報	ホームページアドレス

7. アクセスログとインターネットアドレス

7.1URI

Uri	**カテゴリ**	**簡易表記名**
authority	(可変カテゴリ)	URI 権限
stem	(可変カテゴリ)	URI ステム
querystring	(可変カテゴリ)	URI の照会列部分

7.2ipaddr

ipaddr	カテゴリ	簡易表記名
hostname	コンピュータ情報	完全なホストとドメイン名
partialhostname	人口統計学的・社会経済学的データ	ホスト名の一部
fullip	人口統計学的・社会経済学的データ	全 IP アドレス
partialip	人口統計学的	IP アドレスの一部

7.3 アクセスログ情報

loginfo	カテゴリ	簡易表記名
uri	ナビゲーションとクリックストリームデータ	要求されたリソースの URI
timestamp	ナビゲーションとクリックストリームデータ	要求とタイムスタンプ
clientip	コンピュータ情報、人口統計学的・社会経済学的データ	クライアントの IP アドレスまたはホスト名
other.httpmethod	ナビゲーションとクリックストリームデータ	HTTP 要求方式
other.bytes	ナビゲーションとクリックストリームデータ	レスポンスのデータバイト
other.statuscode	ナビゲーションとクリックストリームデータ	レスポンスステータスコード

7.4 その他 HTTP プロトコル情報

httpinfo	カテゴリ	簡易表記名
referer	ナビゲーションとクリックストリームデータ	ユーザーが要求した最後の URI
useragent	コンピュータ情報	ユーザーエージェント情報

上記のデータ要素のうち、利用者プロファイルとしての利用が可能なのは、「2.名前」、「3.ログイン」、「4.認証」、「5.電話」、「6.連絡先情報」、「6.1 郵便」、「6.2 テレコミュニケーション」、「6.3 オンライン」、「7.3 アクセスログ」の中の「clientip」、「7.4 その他 HTTP プロトコル情報」の中の「useragent」である。

その他のデータ要素については、個人の行動履歴に関する情報であったり、サーバ側の情報であったりするため、利用者プロファイルの中に含めるのは適切ではない。

3.1.5.2 OECD プライバシーステートメントジェネレーター調査

OECD のプライバシーステートメントジェネレーター[94]の調査を行なった。OECD プライバシーステートメントジェネレーターとは、Web サイト上に掲載された規定フォームの質問に答えていくことにより、自社サイトのプライバシーポリシーを作成、ダウンロードできるオンラインツールであり、2000 年 7 月頃に OECD が Web サイト上で公開した。OECD プライバシー・ガイドライン 8 原則をインターネット上で普及させることを狙いと

[94] http://cs3-hq.oecd.org/scripts/pwv3/pwhome.htm

するものである。

OECD プライバシーステートメントジェネレーターでは、Web サイトがサイト訪問者から収集する個人データとして、以下のような個人データの種別を規定している。

1. 主要な個人データ
   - 氏名
   - 性別
   - 住所
   - 電子メールアドレス
   - 電話番号／FAX 番号
   - その他

2. ビジネス情報
   - 勤務先／組織
   - 役職
   - 勤務先住所
   - 勤務先電子メールアドレス
   - 勤務先電話番号／FAX 番号
   - その他

3. 詳細な個人データとプロファイルデータ
   - 個人の詳細：ニックネーム、生年月日／年齢、出生地、国籍など.
   - 物理的な詳細：慎重、体重、その他の身体的特徴など
   - 家族構成：婚姻、パートナーシップ、扶養家族など
   - 教育とスキル：学歴、専門的興味など
   - ライフスタイルまたは嗜好：商品・サービスの購入情報、レジャー、スポーツ、個人または家族のふるまい、喫煙、飲酒、好きな色、好きな食べ物など
   - 金融資産：収入、財産など
   - その他

4. 識別子
   - オンライン識別子：Web サイトのパスワード、クッキー、訪問者の身分証明書（個人認証）など
   - 金融識別子：クレジットカード番号、銀行口座番号など
   - 公共機関に割り当てられた識別子：社会保障番号、個人識別番号など
   - バイオメトリック識別子：DNA、虹彩認識、指紋など
   - その他

5. 特殊なデータ
   - 民族または人種

- 政治的見解
- 宗教的信仰または哲学的心情
- 労働組合への加盟
- 医療健康情報
- 性生活
- 警察・司法データ
- その他

また、個人データ（個人を識別できるデータ）でないデータとして、以下のものを挙げている。

- IP アドレス、選好する言語、セッション（番号、キー）、Web ページへの滞在時間、閲覧したバナー広告、訪問した Web ページ
- ただし、上記データが特定個人と結び付けられている場合は、個人データとみなされる。

上記の個人データ種別のうち、利用者プロファイルとしての利用が可能なのは、「1.主要な個人データ」「2.ビジネス情報」「3.詳細な個人データとプロファイルデータ」「4.識別子」である。

「5.特殊なデータ」については、個人情報保護の観点から、利用者プロファイルの中に含めるのは適切ではない。

３.２ 情報のセマンティック検索に関する調査

電子政府関連の情報に関するセマンティック検索システムを実現するための要素技術として、下記の調査を行なった。

・メタデータ管理システム調査
・半自動メタデータ生成システム調査
・対話型メタデータ生成システム調査
・利用者プロファイル定義システム調査
・セマンティック検索システム調査

Web 情報にメタデータを誰が付けるのかという問題はセマンティック Web の実用化・普及に取って一番大きな課題であるとなると言われている。有効なメタデータが広く付けられることでセマンティック Web の有効性は加速度的に高まる。そこでメタデータを生成するツールが重要になってくる。自動的に生成できることが最も望ましいが、現在の Web ページはコンピュータが理解できるようには書かれていないので完全には自動生成できない。コンピュータが生成したメタデータを人手で手直しするとか、対話型メタデータ生成システムを用いて人の労力を削減するなどの半自動的なメタデータ生成ツールに期待が掛けられる。

３.２.１ メタデータ管理システム調査

メタデータ管理システムとしてどのようなものが現在存在しているか、また今後どのようなツールがあり得るかについて調査を行なった。

メタデータもデータの一種であることに変わりはないため、その管理に既存の DBMS（データベース管理システム）を利用することは十分可能である。実際、いくつかの研究では、インターフェイス・レイヤを介してリポジトリにアクセスすることで、リポジトリの実現によらないメタデータ管理システムを構築するアプローチがとられている。リレーショナルデータベース（RDB）は、スケーラビリティの点からリポジトリの有力な候補であるが、一般にはインターフェイスとの間の対応にギャップが存在する。メタデータの管理をオブジェクト指向データベースで行なうものなどもある。

またメタデータの記述は XML を用いるケースが増えているため、XML のデータ管理システムそのものがリポジトリとして有効である。XML データの管理システムとしては、XML データ中のテキストを RDB のデータ構造にマッピングしてデータベース化するものと、DOM のインターフェイスを用いてツリー形式の構造をそのままバイナリ化してデータベース化するものが存在する。大量のメタデータに複数のエージェントがアクセスする可能性があるセマンティック Web の世界では、高速性の観点から後者の XML データをオブジェクトとして管理する技術が、鍵を握る要素技術の一つとなると考えられる（既に商用製品は登場している）。

以下、既存の DBMS よりももう一段階上のレイヤーで行われているメタデータ管理の試みを紹介する。（１）ではメタデータ間のオントロジーの違いを吸収するアーキテクチャも検討されている管理システムとして、Multi-schema Metadata Format(MMF)を説明する。（２）ではセマンティック Web のメタデータとして、第一に挙げられる RDF に特

化したメタデータ管理の取り組みの事例を紹介する。（３）では、メタデータの設置や収集方法に焦点をあてて、その事例を紹介する。

（１）Multi-schema Metadata Format(MMF)

セマンティック Web が話題になる以前から、データベース毎にそれぞれ［生産県］、［生産地方」など、商品に関して異なるスキーマに基づくメタデータの差異をどのように吸収するかが問題になっている。これを解決するため、ディジタル・ビジョン・ラボラトリーズ(DVL)によって開発された、Multi-schema Metadata Format(MMF) [95]がある。商品特徴情報をメタデータ化し、WWW 上で提供するためのメタデータ形式である。MMF では、メタデータをメタデータインスタンス、スキーマ定義、スキーマオントロジー、標準属性辞書の４階層の構成とすることで、様々な人間が定義したメタデータ間のオントロジーの違いを吸収している。同時に、Metadata Mediator というエージェントを介して MMF の検索や、Metadata Mediator 間で交換するためのプロトコル Metadata Mediation Protocol (MMP) を規定し、MMF の運用を支援する。メタデータインスタンスは XML で記述され、Metadata Mediator 内のデータベースに保持されるが、どのような実装形式（DBMS も含め）でメタデータを管理するかは規定していない。MMF と MMP は、オントロジーの管理や運用までを考慮したメタデータ管理システム体系ということで、セマンティック Web に近いアプローチと言える。

（２）RDF の蓄積・検索に特化したデータベース

また、RDF に特化したデータベースの開発が研究レベルで進められている。

European Commission Joint Research Center (JRC)で開発されている RDFStore (Perl API for RDFStorage)[96]は、RDF の構造を解析、格納、管理する Perl のライブラリーである。このプロジェクトは 2001 年の 10 月に始まっている。RDF の 3 つ組（リソース、プロパティ、値）の検索を効率的に行なうことを目的としている。ここでは、RDQL/SQUISH と呼ばれる RDFStore で蓄積されたデータに対して SQL のような形式で検索するための表記法などについても開発している。

同様なシステムとして、RDFdb[97]がある。（RDFStore よりも先に開発されている模様）。こちらも Algae と呼ばれる SQL のような形式で RDF のステートメントを検索する言語が開発されている。

Sesame[98]は、EU 委員会 IST プログラムの１プロジェクトである On-To-Knowledge で Aidministrator Nederland によって開発された、RDF スキーマベースの Repository and Querying facility である。リポジトリには RDF データと RDF スキーマ情報を格納することができ、RQL と呼ばれる問合せ言語を用いて検索を行なう。リポジトリとして現在は Object-Relational DBMS である PostgreSQL が使われている（Repository Abstraction

---

[95] 坂田毅、多田浩之、大竹智久（1997)「メタデータの WWW 上への実装と電子商取引への応用」画像電子学会第 9 回メディア統合技術研究会

[96] http://rdfstore.sourceforge.net/

[97] http://www.w3org/2001/Talks/0505-perl-RDF-lib/

[98] http://sesame.aidministrator.nl

Layer を通じて様々なタイプのリポジトリを使用可能とされる[99]）。

RDF の取り扱いに特化したデータベースでは、RDF の 3 つ組を中心とした処理が行われると考えられる。RDF の 3 つ組によって記述されたリソース同士は柔軟に結合することができるが、それらを RDF データモデルのままで参照することが可能である。従来のデータベースは構造化が進み柔軟性を失っているのに対して、RDF データベースでは半構造的なデータ管理が可能であり、さまざまなアプリケーションからの要求にあわせてデータベースを変更するコストの削減などの効果も期待できる。

（3）セマンティック Web におけるメタデータの設置・収集・管理方法

ここまでは、データをシステムがハンドリングするための要素技術について述べたが、アプリケーションによる利用や、サービスという側面から考えると、メタデータをどこにおいて管理するかが問題となる。

（a）既存の Web サイトへのメタデータの設置

セマンティック Web では、Web の特性としてコンピュータネットワーク上のさまざまなユーザーがメタデータの生成や管理に携わることになる。個々のメタデータの置き場所が、各 Web サイトにあるというケースがまず考えられる。セマンティック Web のファーストステップとして、HTML 中に RDF を記述することが唱えられている。そして、RSS (RDF Site Summary) [100]はライトウェイトな RDF 準拠のメタデータフォーマットであるが、これも基本的には各 Web サイトがサイトの情報の要約などを提供するのに使われている。これらのケースは Web 全体が巨大なメタデータ管理システムという位置付けになる。

（b）メタデータの集中管理

RDF のリポジトリを構成する試みもある。W3C の Annotea プロジェクト[101]のように Annotea サーバをインターネット上に置いておくことによって、RDF で記述されたアノテーション情報を集中的に管理してユーザーに提供できる。ユーザーが Amaya などの Web ブラウザ上で表示している情報の URI に対応したアノテーション情報をサーバに問い合わせる形になる。

（c）メタデータのインデックス管理及び収集ツール

RDFWeb[102]のように、メタデータ自体は個々のユーザーが作成・設置するが、そのインデックスのようなものを管理する機構も存在する。DARPA による DAML プロジェクトにおいても DAML によって記述された Web 上のステートメントを収集する試みも 2001 年 5 月から DAML Crawler[103]によって開始されている。

また、RDF Crawler[104]もメタデータ収集ツールであり、インターネットから RDF フラグメントをダウンロードして知識ベースを構築する。RDF Crawler は Java API を提供す

---

99 http://www.On-To-Knowledge.org/downl/sesame.ppt

100 http://groups.yahoo.com/group/rss-dev/files/specification.html

101 http://www.w3.org/2001/Annotea/

102 http://rdfWeb.org/

103 http://www.daml.org/crawler/

104 http://ontobroker.semanticWeb.org/rdfcrawl/

るので、単独のアプリケーションとして利用する以外に、他のツールの中に埋め込むことも可能となっている。

SHOE[105]はWebのためのオントロジー言語であり、Webページの作者が機械処理可能な形式でドキュメントにアノテーションを付ける事ができる。SHOEでは、作者が付けた情報をWeb上に置くダイレクト・アクセスと、Web-crawlerを使って集めて知識ベースに格納するリポジトリ・ベースド・アクセスの２つのアプローチが考慮されている[106]。後者に関しては、SHOEデータの格納と検索エンジンへのアクセスのためのJava APIが提供されており、 Web-crawlerのExposéなど、このAPIを使うツールが各種提供されている。

[105] http://www.cs.umb.edu/projects/plus/SHOE/
[106] Jeff Heflin, Towards the Semantic Web: Knowledge Representation in a Dynamic, Distributed Environment, Ph.D. Thesis, University of Maryland

３.２.２ 半自動メタデータ生成システム調査

半自動メタデータ生成システムとしてどのようなものが現在存在しているか、また今後どのようなツールがあり得るかについて調査を行なった。半自動メタデータ生成システムの一部としてのメタデータ生成規則定義システムの調査（メタデータ生成規則の調査を含む）も行なった。

セマンティックWeb社会の実現においては、少なくともあるコミュニティー（電子政府や業界団体など）でメタデータの枠組み（属性、エレメントリスト）を決め、それに従ってメタデータを付与するという実際の作業が必要である。

特に、メタデータの付与については、Web時代でリソースが爆発的に増えている状況にあり、人手工数の削減は大きな課題となる。膨大で良質なメタデータを継続的に付与、更新していくことができなければ、セマンティックWeb社会は絵に描いた餅にすぎない。

こうした状況を背景に、コンピュータによる半自動メタデータ付与によって、膨大な人手工数の削減だけでなく、質のばらつきの少ないメタデータの作成が期待できる。本節では、半自動メタデータ生成について、関連する技術を調査した。

３.２.２.１ メタデータ自動付与における背景技術

メタデータ自動付与のためには、以下のような技術が必要である。

（1）HTMLパーザ

メタデータ付与対象がHTMLの場合、HTMLソース内の特定タグにメタデータとなるべき情報が含まれていることがある。その場合は、HTMLパーザによりソースのタグ構造を解析し、必要な部分を取出せば良い。例えば、以下のようなタグがある。

○METAタグ

METAタグはHTMLの規格としてフォーマットが決まっているだけで、エレメントは作者が自由に付けて良いことになっている。おおむね、以下の2つは多くのページで良く見られる。

- KEYWORD：キーワードを記入する（DC:Subject）
- DESCRIPTION: ページの説明文を記入（DC:Description）

○TITLEタグ（DC:Title）

（例）実際には、これらのメタデータは次のようにHTML内に埋めこまれる。

```
<head>
<META NAME="AUTHOR" CONTENT="H. Tsuda">
<META NAME="KEYWORD" CONTENT="情報検索,テキストマイニング,Webマイニング,自然言語処理">
<title>テキストマイニング研究のページ</title>
</head>
```

（2） 情報抽出技術

情報抽出（Information Extraction）とは、文書から特定のイベント情報 （あるイベント5W1H情報）などを自動で抜き出す技術である。自然言語処理研究では、従来、新聞などのテキスト群を対象に、特定パターンの文字列から情報を抜き出すなど、多くの研究が行なわれてきた。Web文書を対象としての最近の動きとしては、Wrapper(ラッパー)と呼ばれる、タグを手掛りにした手法が多くなっている。ラッパーを使った応用としては、複数のショッピングサイトから商品の価格情報を取出して横断検索をできるようにするとかのように、Web情報をデータベース化するのが一般的である。

例えば、坂元論文[107]には、以下のようなHTML文書から国別の統計情報を抜き出しレコード化する例と、その自動化手法が解説されている。

(例)
元の HTML文書:
<HTML><TITLE></TITLE><BODY>
<B>Congo</B><I>242</I>
<B>Egypt</B><I>20</I>
<B>Belize</B><I>501</I>
<B>Spain</B><I>34</I>
</BODY></HTML>
抽出されるレコード情報:
((Congo,242),(Egypt,20),(Belize,501),(Spain,34))

情報抽出技術はメタデータの特定エレメント抽出にも利用可能である。例えば、DC:Rights（copyright表記を手掛り）、Location（住所表現を手掛り）などが考えられる。

（3）文書自動分類技術

文書自動分類技術は大別すると、以下の2種類に大別できる。

①クラスタリング (clutering)：文書群を与え、文書の類似度により複数のグループに分ける
②カテゴライゼーション (categorization)：文書群と既存クラスを与え、文書群をクラスに割り当てる。

メタデータ自動付与としては、カテゴライゼーションの方が関連が深い。というのも、例えばDC:Subject のようなエレメントは、統制されたシソーラス語彙を付与することが推奨されている。統制語によるクラスへの割当てと考えれば、カテゴライゼーションの問題に帰着する。

カテゴライゼーションは以下のような技術の組合せで実現される。

- 形態素解析：文を単語に分割する。日本語の場合は単語間にデリミタがないので、辞書やルールを用いた処理になる。英語の場合は、動名詞の活用の統一（stemmingと呼ばれる）などが必要。

[107] 坂元,有村「Webマイニング」人工知能学会論文誌, Vol.16, No.2,pp.233-238, 2001

- キーワード抽出: 単語の中から、文書やクラスに特徴的なキーワードを取出す。特徴的なキーワードとは、例えば、その文書に多く現われ、他にはあまり出現しないものというものが考えられる。TFIDFと呼ばれる有名な単語の重みづけ手法は、その一例である。
- 特徴素選択: 文書やクラスの特徴となる情報を選別する。たとえば、キーワードの列（ベクトル）とかが一般的。Web文書の場合には、URL特徴や、リンク元のアンカー文字列の組込みも有効とされる。
- 類似度マッチ: 文書やクラスの特徴素間の類似度を定義し、その近さを計算する。例えばキーワードベクトルに対しては、その方向の近さをベクトルの内積（二本のベクトルの作る角度のコサイン）で表すなどの手法がある。

（4）その他

例えば、Dublin Coreエレメントのうち、Date（日付）、Format（フォーマット）、Identifier（Webの場合はURL）などは、WebにおけるHTTPプロトコルの結果などから得ることができる。

3.2.2.2 メタデータ自動生成ツール[108]

（1）メタデータエディタ

メタデータエディタとは、メタデータの中身を入力するのを補助するツールやテンプレートのことである。メタデータの中身は基本的に作業者が付与し、出力フォーマット(RDF、HTML内のMETAタグ埋め込みなど)への変換はエディタが自動的に行なうことで、人手によるミスと労力を減らす。

○Dublin Core Metadata template
[開発元] Nordic Metadata Project
[URL] http://www.ub.lu.se/metadata/DC_creator.html
[対応メタデータ] ダブリンコア
[出力フォーマット] HTML (3,4) Metaタグ埋め込み
Webのフォームになっており、ダブリンコアの15エレメントのうち必要な情報を入れると、HTMLのMETAタグ埋め込み形式のダブリンコアメタデータを作成する。利用者は、その部分を自分のページにカット&ペーストする。出力形式はHTML(3,4)のみ。

○Reggie Metadata Editor
[開発元] Distributed Systems Technology Centre
[URL] http://metadata.net/dstc
[対応メタデータ] ダブリンコア, スキーマファイルを与えれば他のAGLSなどにも対応可
[出力フォーマット] HTML3.2(METAタグ), HTML4.0 (METAタグ), RDF, RDF Abbreviatied
ダブリンコアだけでなく、スキーマファイルを定義することで、独自のメタデータ形式にも対応できる。

---

[108] http://www.dstc.edu.au/Research/Projects/metaWeb/toolpost.html を参考にした。

○MetaWeb Generic Edit Tool
[開発] State Library of Tasmania
[URL]　http://www.dstc.edu.au/RDU/MetaWeb/generic_tool.html
[対応メタデータ] ダブリンコア, AGLS,ADMIN Core
[出力フォーマット] HTML (Metaタグ)
Tasmania Online serviceで用いられている。出力されるファイルをHTMLやテキストに埋め込む。

（2）メタデータジェネレータ
　エディタは作業者が全ての属性値を入力する必要があるのに対して、ジェネレータは、いくつかの属性値はシステムが自動で作成してくれるものである。メタデータのエレメントにより、自動化の精度はまちまちである(詳細は⑵参照)。したがって、ジェネレータは、実際はツールとしてはエディタ機能も持ち、誤って付与された値は簡単に修正できるようになっている。

○DC-dot
[開発] UKOLN, University of Bath
[URL]　http://www.ukoln.ac.uk/metadata/dcdot/
[対応メタデータ] ダブリンコア, 他の形式 (USMARC, SOIF, IAFA/ROADS, TEI headers, GILS, IMS)にも変換可能
[出力フォーマット] HTML (Metaタグ), RDF
URLを与えると、そのページ本文からダブリンコアメタデータエレメントを自動的に作成する。いくつかデータを与えた結果から類推するに、大体次のような動作をしているようである。

DC.Title： <title>タグを参照する。
DC.Subject： <meta>タグのKEYWORDを参照する。無い場合には、アンカー文字列から抜き出したキーワードを入れる場合もある模様。
DC.Description： <meta>タグのDescriptionを参照する。
DC.Date： 最終更新日を参照する。
DC.Format： htmlの場合は text/htmlとする。
Dc.Identifier： URLを参照する。

○HiSoftware TagGen
[開発] HiSoftware社
[URL]　http://www.hisoftware.com/taggen.htm
[対応メタデータ] GILS, WAGILS, ダブリンコア
HTML(XHTMLにも対応)ファイル群を選ぶと、ウィザード形式でメタデータを付与してくれるツール。
・キーワードは人手付与または本文からの自動抽出が可。スペルチェックや、シソーラスによる統制語への変換にも対応
・自動的にメタデータつきのページに変換し、サーバにアップする。(形式はHTML内METAタグで、RDFなどはない様子)

・検索エンジンコントロール用メタデータにも対応。例えば、ロボットがリンクを辿って良いかどうか、インデクシングして良いかどうかなど。

３.２.２.３ 実例１：オーストラリアの政府系メタデータ AGLS と半自動生成ツール

オーストラリアは非ヨーロッパ圏ではメタデータに関して最も進んでいる国と言われる。オーストラリアにおける政府系メタデータとして有名なものがAGLS（Australian Government Locator Service）である。AGLSは、オーストラリアのNational Archives of Australis (NAA)により、政府系公開情報リソース（文書、サービス）に対して付与されているメタデータエレメント標準である。NAAでは、メタデータを付与した政府系リソースをアーカイブとして管理している。

AGLSはダブリンコアを拡張した19のエレメントからなり、1997年から1998年にかけて策定された[109]。AGLSメタデータは基本的には、リソースの作者（Creator）または出版人（Publisher）が付与するものである。その作業において、メタデータ付与ソフトウェアが利用できる。

例えば、Subjectフィールドは、リソースの主題キーワードを記述する。シソーラスによる統制キーワードを付与することが推奨されている。例えば、

Library of Congress Subject Headingsの例
[LCSH] Biographies - Government - Australia

MeSH - Medical Subject Headingの例
[MeSH] Gene Expression Regulation, Bacterial

AGLSのメタデータ付与ツールには以下のようなものがある。

○Reg：前述のReggieの軽量版
メタデータエディタである。対応メタデータは、ダブリンコア, AGLS (ver. 1.0, 1.1, 1.2), EdNA, QLDGOV。他にもスキーマを定義すれば他メタデータにも拡張可能である。

○Klarity
[URL] www.klarity.com.au
URLを入れると、ページ内テキストを解析して、TITLE, KEYWORDS, DESCRIPTIONを入れて

---

[109] AGLS のエレメントは以下の 19 個である。ダブリンコアと同様に、qualifier も許される。（NAA のホームページ http://www.naa.gov.au/ よりアクセスできる）

Creator	Publisher	Contributor	Rights	Title
Subject	Description	Source	Language	Relation
Coverage	Function	Date	Type	Format
Identifier	Availability	Audience	Mandate	

AGLS はオーストラリアにおける各種メタデータの基本と言っても良く、実際 Australia and New Zealand Land Information Council（ANZLIC）以外の次のようなメタデータは、全てダブリンコアまたは AGLS 準拠している。

- Educational Network Australis (EdNA)
- Environmental Resoruces Information Network (ERIN)
- Business Entry Point (BEP)
- HealthInsite

くれる。文書自動分類を行なっていると思われる。メタデータ形式はダブリンコアである。出力形式はHTML, RDFに対応している。

３.２.２.４ 実例２：富士通研究所による Web ディレクトリ自動生成ツール

津田論文[110]では、Yahooのようなディレクトリを自動的に作成することを目的として、いくつかのページメタデータの自動付与を試みている。一部は商用に使われている。

Webディレクトリは、人気がありかつ情報の入口となるメニュー的なページがコンテンツとして望まれる。そのため、そのようなページを選別するためのメタデータを付与しているのが特徴的である。技術的には、Webのリンク関係の解析により、そうしたメタデータを付与しているのが特徴的である。ダブリンコアを拡張した、以下のエレメントの自動付与を行なう。

- Description：METAタグ内またはページ本文の最初
- Publisher：ホスト名
- Date：最終更新日またはcrawlerによる最終取得日
- Type：ページ(ナビゲーション)タイプ。
  メニュー、リンク集、コンテンツを、ページ間のリンク関係を元に判別している。例えば、メニューページはローカルリンク数や被リンク数が多い、リンク集はサーバ外へのリンク数が多いなどの特徴を使っている。
- Format：HTMLなど
- Identifier：URL。ミラーサイトは同一のものとして縮退。末尾のindex.htmlなどは適宜削除
- Source：リンク元URL集合
- Language：コード種別からヒューリスティックスで判定
- Relation：特定ドメインとの関連度
  リンクの近さで定義する。社内向けディレクトリであれば、社内ページからのリンクトポロジの近さにより、業務に関連した情報かそうでないかを判別する。
- Coverage (Location)：情報抽出により、ページ本文中の関連地域を取出す。
- Rank：ページ人気度。そのページがどのような他ページからリンクされているかを基準に算出し、時系列的動きの情報も加える。

このようにして、地域／ジャンル／時間と様々な観点を組み合わせた、多観点のディレクトリインターフェイスを実現している。図　３-9は、経済産業省に関係するカテゴリのページ画面である。

Rankの情報は、ダブリンコアにはないが、情報の新鮮度を与えるため、特にWebのように動きの激しい世界では重要となる。例えば、図　３-10に、厚生労働省の狂牛病関連のページの動きを示す。上に行くほど、そのページへのリンクが増え、世間から着目されていることがわかる（数字は、同ページの注目度順位を表し、小さいほど上）。世間で話題になった2001年10月中旬から、同ページへのリンクが増えているということがわかる。

---

[110] 津田,鵜飼,三末「Web ディレクトリのためのページメタデータの自動付与の試み」情報学シンポジウム 2002, pp.17-24, 2002

Webの場合、アクセスログがサーバにしか残らないこともあり、外部からページの人気度を測定するのは難しいが、リンク関係の解析によりある程度人気度を測定することが可能である。

Webリソースは雑多なので、著者がつけたメタデータだけでなく、外部からの評価指標も必要と考えられる。そのようなメタデータ付与には、ここで紹介したような技術が必要と考えられる。

図 3-9 経済産業省のカテゴリ

http://www.mhlw.go.jp/kinkyu/bse.html の

- [タイトル]「牛海綿状脳症(BSE)関係」ホームページ
- [説明] (なし)
- [キーワード] (なし)
- [ディレクトリ登録済] (厚生労働省) ()
- [収集日,最終更新日] 20020116,
- [ジャンル] 厚生労働省,
- [地域] ,,
- [順位] 21348
- [一次格納ディレクトリ] 20020112/185/18545#7 cache
- [内部ID] 128304472
- [ランク重み] 5594 , [参照数] 53(うちリモート 7), [被参照数] 57(うちリモート 57)

---

1. 人気度ランキングの変動履歴

10月末から
伸び

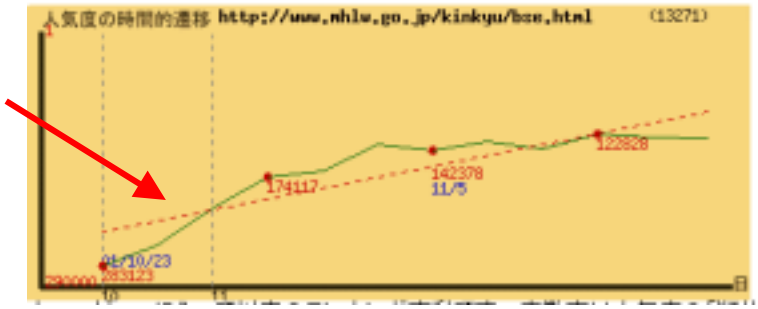


図　3-10　厚生労働省の狂牛病関連情報ページの動き

３.２.３ 対話型メタデータ生成システム調査

セマンティック Web 技術を利用するにあたって、ひとつの大きな課題はオントロジー（Ontology）と呼ばれる実世界の知識の記述を如何におこなうかということである。RDF による記述は、明確ではあるものの、定義された内容の相互関係を直感的に把握するのは困難である。複雑な知識を記述するためには、内容を図を用いて可視化したり、対話的に編集したりできる編集システムが必要である。このために開発されているいくつかオントロジーの作成・編集システムについて説明する。

（１）オントロジーの作成・編集システム

Amsterdam 大学 Social Science Informatics（SWI）では、WonderTools[111,112]というプロジェクトで、利用者が適切なツールを選択するのを支援するという目的で、オントロジーを作成するための以下のツールの評価をおこなっている。

- Ontolingua
- WebOnto
- ProtégéWin
- OntoSaurus
- ODE
- KADS22

この中からスタンフォード大学 KSL（Knowledge Systems Laboratory）の Ontolingua システム[113]を紹介する。

---

111 WonderTools, http://www.swi.psy.uva.nl/wondertools/

112 Duineveld, A. J., Stoter, R., Weiden, M. R., Kenepa, B., Benjamins,V. R. WonderTools? A comparative study of ontological engineering tools, In Proc. the 12th Internatinal Workshop on Knowledge Acquisition, Modeling and Management (KAW '99), 1999

113 Ontolingua, http://www-ksl-svc.stanford.edu/

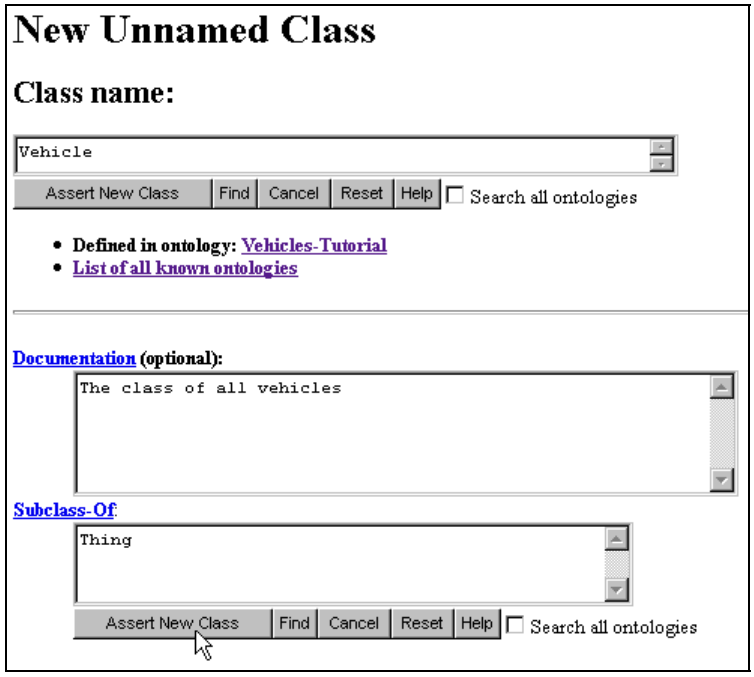


図　3-11　Ontolingua におけるクラスの定義画面（Vehicle クラスの作成）

Ontolingua システムは、サーバと記述言語からなり、Web ブラウザからサーバにアクセスしてオントロジーの作成、編集、挿入や統合を行なえる。また、複数の利用者が協同して設計できるという特長を持っている。図　3-11及び図　3-12に Ontolingua のオントロジー作成 Web 画面を示す。

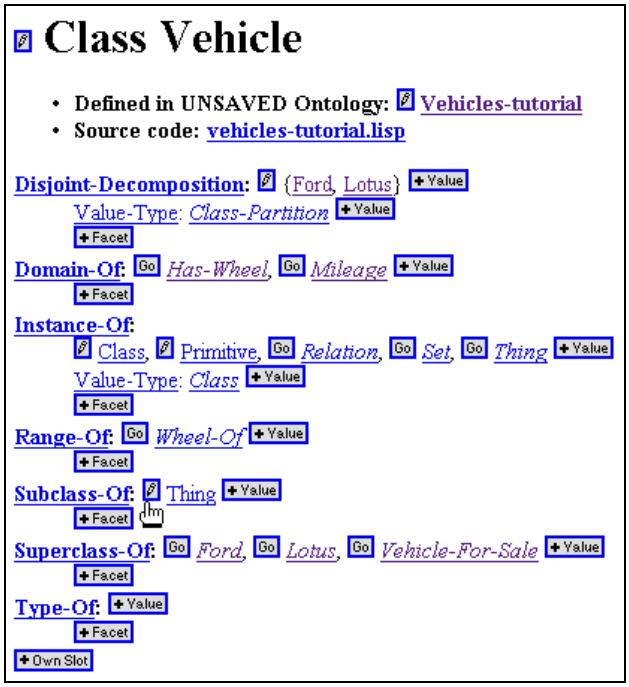


図　3-12 Ontolingua におけるクラスの定義画面（Vehicle クラスの属性を定義）

オントロジーは、一般的に概念とその関係性の記述として表現されるが、Ontolinguaでは、それぞれクラス及びスロットとして表現している。

図　3-11は、Vehicle-Tutorial という新しいオントロジーを定義したのち、Vehicle-Tutorial に「車」を表す新しいクラス Vehicle を定義している画面である。クラス Vehicle は、”the class of all vehicle”を記述したもので、既に定義されている「物」を表すクラス Thing のサブクラスとして定義している。図　3-12は、クラス Vehicle にいくつかのスロットを定義した後の画面である。

Ontolingua は、公理（Axiom）と呼ばれる規則を定義する機能も有している。例えば、「2 匹の動物が兄弟ならば、彼ら両方の母であるものが（1 匹）存在する」という規則（公理）によって、兄弟という概念と母子という概念の関係を記述している。

サーバは、スタンフォード大学のサイト[113]から使用できるようになっており、上述した機能のガイドツアーも用意されているので、詳細についてはそちらを参考にされたい。

上記 WonderTools によって挙げられたツール以外に、商用システムとして OntoEdit システムがある[114]。OntoEdit はカールスルーエ大学の AIFB が開発し、独 Ontoprise 社が商用化したもので、Ontolingua と同様にオントロジーの生成、処理等の機能を有し、さらに W3C 標準である RDF（Resource Description Framework）や、DAML+OIL、F-Logic の各形式に対する入出力機能を有している。OntoEdit のオントロジー作成画面を図　3-13 及び図　3-14 に示す。

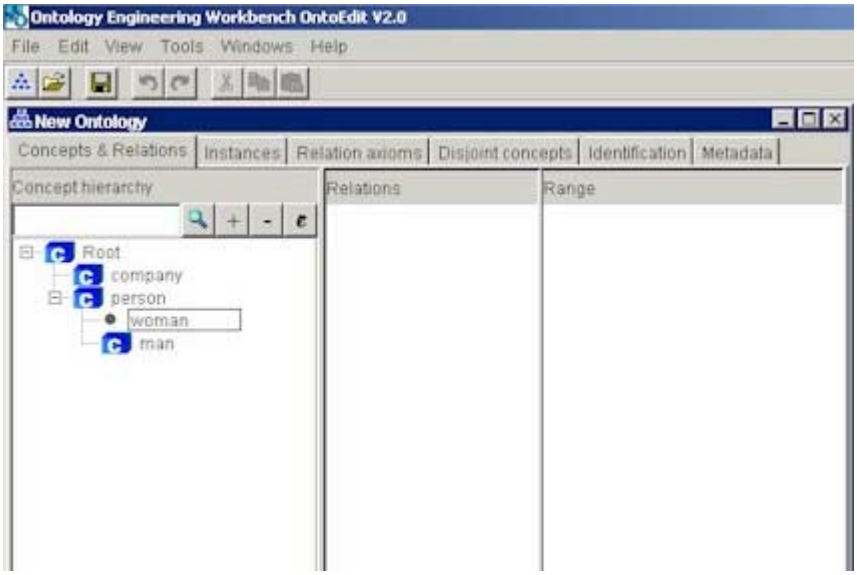


図　3-13　OntoEdit におけるコンセプトの定義画面

[114] OntoEdit, http://www.ontoprise.de/

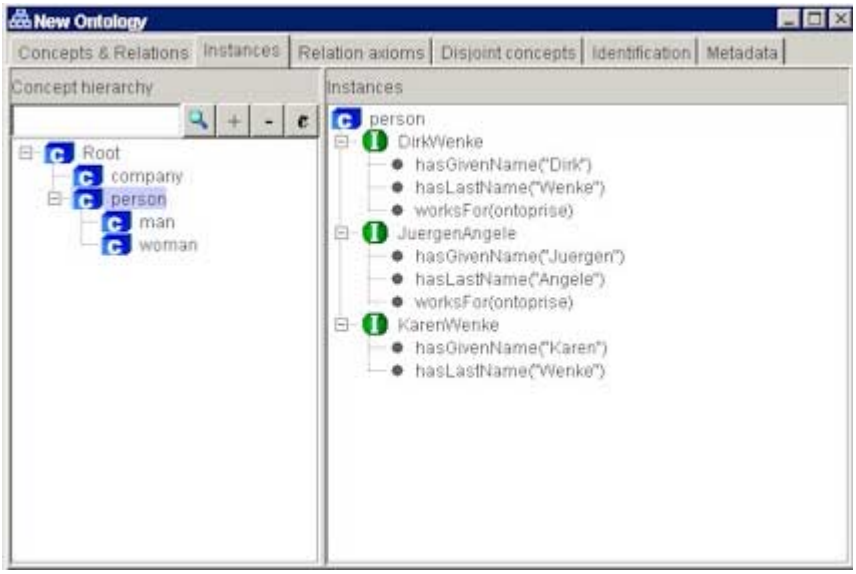


図　3-14　OntoEdit におけるインスタンスの作成画面

OntoEdit では、情報を階層的に表示できる。このため、概念階層を辿りながら必要な情報を編集できる。図　3-13は、人（person）という概念の下位概念として女性（woman）や男性（man）を定義している画面であり、図　3-14は、人（person）という概念の実体（インスタンス）として 3 名を定義している画面である。

また、OntoEdit では、公理を記述し、それを適用するために、F-Logic という述語論理の処理系を組み込んでおり、述語論理により規則（公理）を記述できる。しかし、F-Logic による公理の記述は、利用者には難解である場合が多いため、公理を設定するたに、「背反(disjoint)の概念[115]」や「対称的(symmetric)[116]」「反対(inverse)[117]」「[118]推移的(transitive)」な「関係」を指定できるようになっている。例えば「女性」と「男性」は背反の概念であるという公理を F-Logic で記述せずに、「女性」と「男性」のコンセプト間は「背反」であると設定することにより公理とすることができる。これにより実体が「女性」かつ「男性」を継承することはできなくなる（即ち、女性でありかつ男性であることはない）。

[115] 背反の概念：概念 A と概念 B の両方の下位概念になることはないこと。上記、女性・男性の例参照。
[116] 対称の関係：一つの関係 R に対して A と B とがある関係 R にあるとすると、B と A とも同じ関係 R にあること。例えば、A と B とが兄弟であれば、B と A とも兄弟であると言って良い。
[117] 反対の関係：二つの関係 R と S について、A と B とが関係 R にあるとき、B と A とは関係 S にあること。R が会員を表し、S が所属団体を表すとすると、A が B の会員であれば、B の所属団体に A があるというのは真である。
[118] 推移的な関係：一つの関係 R に対して、A と B が関係 R にあり、B と C とも関係 R にあるとき、A と C も関係 R であること。例えば、R が到達可能の関係を表すとき、A から B へ到達可能で、B から C へ到達可能であれば、A から C へ到達可能であることが成り立つ。

３．２．４ 利用者プロファイル定義システム調査

利用者プロファイルとは、サービスの提供者がその利用者各々に対して適切なサービスを提供するために参照するデータのことであり、いわゆる情報サービスのパーソナライゼイションやワン・トゥ・ワン・マーケティングに用いられる。電子政府においては、国民に対して年齢層や地域特性に応じた情報サービスを行なうために重要な役割を果たすものと考えられる。

利用者プロファイルの身近な例としては、ユーザーエージェント（Web ブラウザなど）で管理され Web サーバ側で参照される Cookie がある。Cookie の典型的な用途は、ある Web ページに対して各訪問者が何回訪れたかを記録しその回数をページ上に表示したり、会員登録した訪問者の登録内容（の一部）を記録して次回訪問時に自動的に認証を行なったりすることである。

（１）プライバシー情報取扱いに関連した利用者プロファイルの定義

利用者プロファイルは次のように分類できる。

(1) 利用者個人に関する情報（主に認証やサービスの個人向け最適化に利用）：氏名、性別、年齢、住所、電話番号、Ｅメールアドレス、公開鍵証明書など
(2) 利用者の嗜好情報（主にサービスの個人向け最適化に利用）：閲覧したコンテンツや利用したサービスの一覧（履歴）など
(3) 利用者の端末デバイス関連情報（主に端末向けのコンテンツ最適化に利用）：画面の解像度、カラー表示の可否、音声出力の可否、OS の種類など

これらは利用者個人のプライバシーに関わる情報であり、特に(1)や(2)に該当する情報をサービス提供者側が無断で収集したり、利用したりすると問題である。そこで、W3C では P3P[119]という技術標準を策定中であり、サービスを提供するサイトがプライバシーポリシーを利用者に明示したり、サイトとユーザーエージェントとの間でプライバシーポリシーに関する不一致を自動的に検出したりできるようにしている。また、サイトが収集する利用者の個人情報について、データの分類を行なっている（３．１．５．１節を参照のこと）。

P3P 1.0 勧告案ではプライバシーポリシーなどを XML で記述するようになっており、テキストエディタで直接記述できる。しかし、サイト管理者や Web ブラウザの利用者が全てのタグとその意味を理解して記述することは困難であるため、様々な設定支援ツールが開発されている。

Microsoft 社は現在最もポピュラーな Web ブラウザである Internet Explorer のバージョン 6 で P3P に対応し、Cookie のフィルタリング処理について利用者に簡易な設定手段を提供している[120]。また、財団法人ニューメディア開発協会では、P3P 準拠のプライバシーポリシーの作成を支援する各種ツールを開発、無償提供しており、サーバ（Web サイト）側とクライアント（ユーザーエージェント）側のそれぞれについて詳細な設定が可能になっている[121]。

---

[119] http://www.w3.org/P3P/
[120] Internet Explorer 6 のプライバシー機能：
http://www.microsoft.com/japan/developer/articles/dnpriv/html/ie6privacyfeature.asp
[121] プライバシー情報管理システム：http://www.nmda.or.jp/enc/privacy/

（２）Web サイト構築システムにおける利用者プロファイル定義

Web サイトから提供されるコンテンツやサービスを各利用者の個人情報や嗜好に応じて最適化するサーバ側のパーソナライゼイション機能は、多くの商用 Web サイト構築システムに搭載されている。IBM 社の WebSphere には WebSphere Personalization というオプション製品が用意されており、専用のルールエディタによって利用者プロファイルのルール定義とコンテンツ定義が可能になっている[122]。BEA Systems 社の WebLogic にも同様の機能を持つ WebLogic Personalization Server が用意されている[123]。

（３）利用者が用いる端末デバイスのプロファイル定義

一方、前述の “(3) 利用者の端末デバイス関連情報” を定義するための規格として W3C では Conposite Capabilities/Preference Profiles（CC/PP）の策定が進められている[6]。端末デバイスは各利用者がコンテンツや情報サービスにアクセスする手段であり、最近はデスクトップ PC やノート PC ばかりでなく PDA や携帯電話、セットトップボックスなど多様なデバイスを利用できる。端末デバイスの選択には利用者の嗜好や状況が反映されうることから、これも利用者プロファイルの一部と考えることができる。

CC/PP は RDF で記述するメタデータであり、将来的には P3P とも密接に関係してくるものとみられる。CC/PP Working Group のホームページ[124]にも記載されているように、HP Labs の DELI や W3C の Jigsaw など CC/PP に対応したサーバはいくつか発表されているが、定義ツールやクライアント端末側のサポートはまだこれからという状況である。

（４）電子政府における利用者プロファイル定義

現在、日本国内全ての都道府県庁公式サイトで、情報公開もしくは各種申請書用の文書フォーマットの１つとして Adobe Systems 社が開発した PDF が採用されている[125]。PDF 文書のオーサリングツールである Acrobat では、PDF 文書にその作成者のプロファイルを RDF 形式で自動的に埋め込むようになっている（図 3-15、図 3-16参照のこと）。

[122] IBM WebSphere Personalization：http://www-6.ibm.com/jp/software/Websphere/personalization/
[123] BEA WebLogic Personalization Server：
http://www.beasys.co.jp/products/Weblogic/commerce/index.html
[124] CC/PP Working Group：http://www.w3.org/Mobile/CCPP/
[125] アドビシステムズ（株）、全都道府県庁の公式サイトが Adobe PDF を活用と発表
http://www.adobe.co.jp/aboutadobe/pressroom/pressreleases/200202/20020206pdf.html

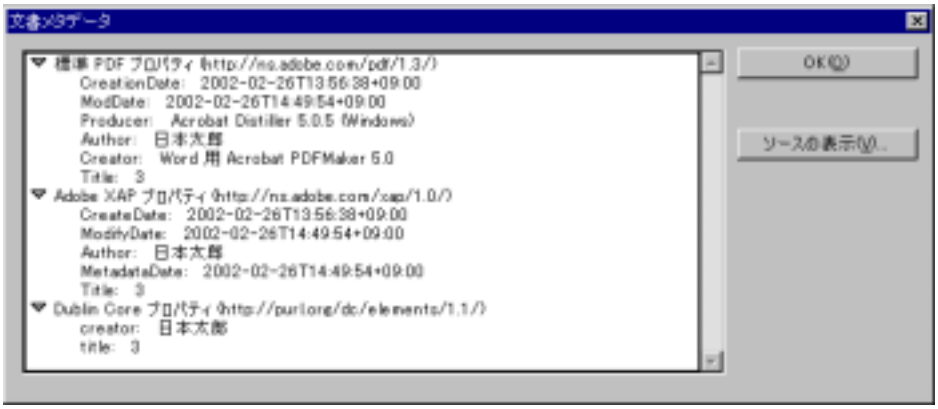


図　3-15　Acrobat で自動生成されたプロファイルの例（一覧表示）

```
<rdf:RDF xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
 xmlns:iX='http://ns.adobe.com/iX/1.0/'>

 <rdf:Description about=''
  xmlns='http://ns.adobe.com/pdf/1.3/'
  xmlns:pdf='http://ns.adobe.com/pdf/1.3/'>
  <pdf:CreationDate>2002-02-26T13:56:38+09:00</pdf:CreationDate>
  <pdf:ModDate>2002-02-26T14:49:54+09:00</pdf:ModDate>
  <pdf:Producer>Acrobat Distiller 5.0.5 (Windows)</pdf:Producer>
  <pdf:Author>日本太郎</pdf:Author>
  <pdf:Creator>Word 用 Acrobat PDFMaker 5.0</pdf:Creator>
  <pdf:Title>3</pdf:Title>
 </rdf:Description>
```

図　3-16　上記プロファイルのソース（一部）

Acrobat では、このようなメタデータに加え、PDF 文書上に設定されたフォームに必要事項を入力し、電子署名を付加した電子申請書としてオンラインで提出することが可能となっている。（図　3-17参照のこと。）

図　3-17　電子署名（右上部分）を付加した PDF 文書の例

このほか、特に電子政府においては、利用者プロファイルに登録すべき情報が手書きやプリントされた紙の帳票の形で提出、管理されている場合も多いと考えられる。既に紙の文書として蓄積された情報は、人手によるキーボード入力とスキャン画像を組み合わせて電子化されるほか、OCR（光学的文字認識）を用いて自動的にテキスト情報を入力し、全文検索が可能な製品もある。

さらには、定型的な紙面を OCR で認識し、プロファイルとして適切な情報を抽出する技術も開発されている。実用化している典型例としては名刺管理システムがある。専用のスキャナにより OCR を通じて氏名、会社名、住所、電話番号などを自動的に認識し、データベースへ登録することができる[126]。今後は PDF や Web ページを用いた電子的帳票入力と OCR を用いた紙帳票の電子化を統合したプロファイル入力・管理システムが、各電子政府・電子自治体に導入されてくるものと予想される。

[126] 名刺管理ソフトの例：http://www.mediadrive.co.jp/products/package/win/econ/econ40/

３.２.５ セマンティック検索システム調査

セマンティック Web の電子政府への応用として、政府や自治体が収集・発行・管理する情報に対する、より人間に近い知的なセマンティック検索が期待されている。本節では、セマンティック検索を必要とする応用事例と、そこで使用あるいは要求される技術について述べ、セマンティック検索の可能性・将来性を示す。

３.２.５.１ 応用事例

（1） ホームページ検索

政府には多くの省庁や関連機関が存在し、それぞれが独自のＨＰを立ち上げて情報を提供しているが、一般利用者にとっては、求める情報の管轄省庁がわからないとなかなか情報にたどり着けない。このため、電子政府全体を統合的に検索する仕組みが必要である。現状では、例えば、総務省のＨＰ[127]では、「電子政府の総合窓口」として政府系の全ＨＰに対する検索がかけられるようになっているが、まだキーワード検索の域を出ないため、必ずしも使い勝手の良いものではない。（図３-18参照のこと。）

図 ３-18 総務省「電子政府の総合窓口」

[127] http://www.e-gov.go.jp/

また、地方自治体による情報提供においても同様の問題が存在する。欲しい情報はわかっていても、それがどこの場所（自治体）にあるか知らない場合には、情報にたどり着くのは容易ではない。このため、財団法人地方自治情報センター（総務省関連機関）では、NIPPON-Net[128]と呼ばれる、地方自治体のＨＰ検索のためのポータルサイトを運営しているが、キーワード入力の仕組みに工夫は見られるものの、まだ使い勝手は十分ではない。（図　3-19参照のこと。）

図　3-19　地方自治情報センターの「NIPPON-Net」

（2）　デジタル・ミュージアム

日本には国や地方自治体が運営する美術館・博物館等が多く、電子政府の一環として、文化財の保護と国民への情報公開サービスの両方の観点からデジタル・ミュージアムが構築されている。代表的なものとして、デジタル・ミュージアム推進協議会（総務省関連組織）が運営する、全国のミュージアムが保有している美術品を横断的に検索することのできるデジタル・ミュージアム[129]がある。（図　3-20参照のこと。）

128　http://www.nippon-net.ne.jp/
129　http://www.digital-museum.gr.jp/

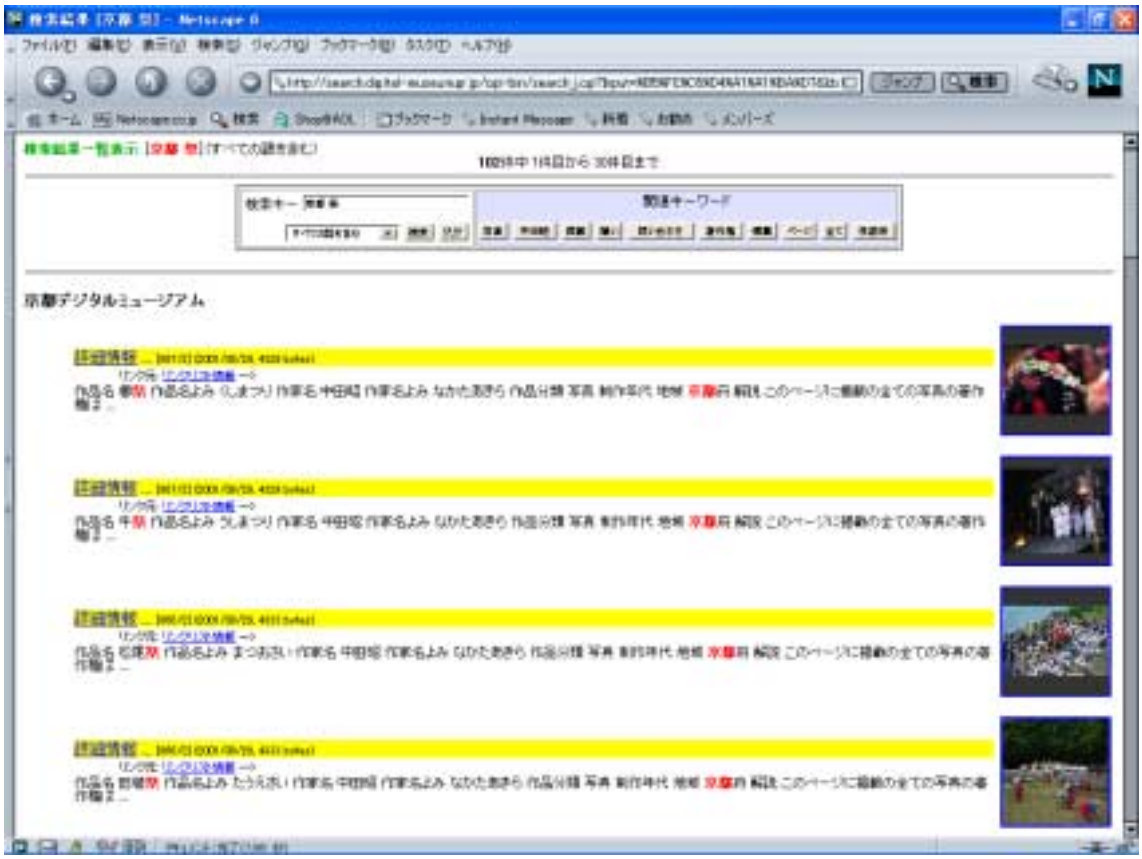


図　3-20　デジタル・ミュージアム推進協議会の「デジタル・ミュージアム」

同様の取組みはヨーロッパにおいても数多くのプロジェクトが進められている。例えば、英国では、大英博物館[130]の所蔵品のオンラインデータベース（図　3-21参照のこと）、National Trust[131]の管理する歴史的資産のデータベース（図　3-22参照のこと）、Wildscreen Trust による絶滅危機生物のデータベース構築を目的とした ARKive プロジェクト[132]（図　3-23参照のこと）などがある。

130 http://www.thebritishmuseum.ac.uk/
131 http://www.nationaltrust.org.uk/main/
132 http://www.arkive.org.uk/default.asp

図　3-21　大英博物館の所蔵品のオンラインデータベース

図　3-22　　National Trust の管理する歴史的資産のデータベース

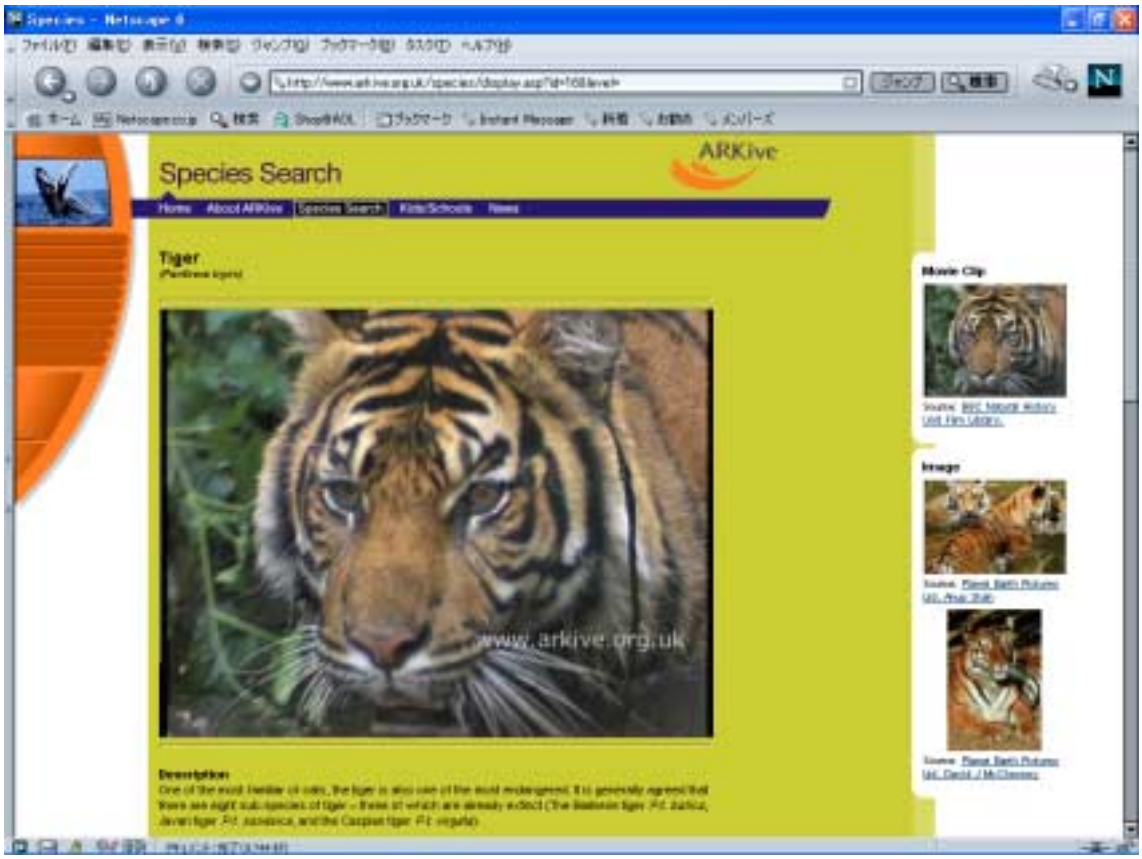


図　3-23　　Wildscreen Trust による ARKive プロジェクト

３.２.５.２ 必要な技術

（1）知的検索

上記の応用事例のホームページ検索やデジタル・ミュージアムの多くは、インターネット上に分散した情報源を統合的に検索するものである。それぞれの情報源は独立して設計・構築されるものであるため、横断的に検索することは容易ではない。この問題を解決するため必要な要素技術としては次のようなものがある。

○オントロジー変換

分散型のデジタル・ミュージアムでは、それぞれの情報源ではシソーラスを使用しているかもしれないが、横断的に見た場合、情報アイテムの属性の名前ですら統一されている保障はない。例えば、美術品の製作された年は、“製作年”“年代”“時代”などの表現のゆれが見られる。このため、表現の同義性を判断する必要があり、あらかじめ一つの統合的なオントロジーにマッピングしておくか、オントロジー（表現）間の同義関係を表す変換規則を用意しておき検索時に変換するなどの手法が要求される。

○自然言語表現からの情報抽出（専門用語抽出、テキストマイニング）

検索の対象となるホームページやデータベース・エントリは自然言語で表現されているため、検

索要求に適合しているかを判断するには、自然言語表現からの専門用語抽出、さらには、意味解析に基づくテキストマイニングによる情報抽出技術が必要となる。

○検索要求の自由記述入力

現状の検索要求入力はキーワード入力が基本で、AND/OR の組合せは可能なものの、ユーザーの意図が十分に表現できないという問題がある。このため、検索要求の自然言語による自由記述入力が必要である。検索システム側は、入力された検索要求を解析し、検索に必要な用語と意図を抽出し、検索式に変換する機能が求められる。

○意味的マッチング

現状の検索アルゴリズムは、入力されたキーワードに対する全文検索が基本である。しかし、単純なキーワードマッチングでは、表現のゆれに対応できないため、オントロジーによってサポートされる表現間の意味的な関係に基づいて、検索対象を展開あるいは絞込みする機能が求められる。

（2）知的問合せ（クエリー・サポート）

データベースに格納されている情報に対して、知的な問い合わせを行なう手法について、セマンティック Web では研究が進んでいる。近年、コンピュータネットワーク上でやり取りされるさまざまな情報を XML で取り扱う動きが盛んだが、セマンティック Web でも知的な問い合わせの対象となる情報を XML で記述し、相互運用性を向上させようとしている。しかし、XML は任意のデータ構造を表現できるが、構造が何を意味するかについては規定していない。そこでセマンティック Web では、XML をベースに情報の意味を記述するためのデータモデルを規定した RDF をメタデータの記述に活用しようとしている。RDF のデータモデルに従って、既存のデータベースのデータを変換したり、新たに登録する情報やデータに意味を付与したりしておけば、全文検索に頼らずに、よりユーザーの質問の意図に沿った回答を返すことができる。

例えば、官公庁のホームページを利用して「ある種の申請書の提出をどの窓口が今日受け付けているか」調べたいとする。官公庁のホームページ内を全文検索しても、たまたま自分が知りたいことがすべて載っているホームページが見つからない限り、検索結果を基にいくつものホームページを人間が読んで答えを探す必要がある。申請を受け付けない日があることを書いてあるページや、各窓口の取り扱い申請書類一覧の表などを見つけてやっと目的の知識を得ることがきるかもしれない。

ここで、各「窓口」のメタデータとして、RDF による「取扱申請」、「受付日時」、「場所」に関する記述がされていると仮定する。また、「取扱」と「受け付け」は似た意味であることが記述されたメタデータ（オントロジー）があったとする。そのようなデータベースが整った条件下で、知的な問い合わせを行なうと、ユーザーが申請したいと考えているものが、まずどの窓口で取り扱われているかがすぐにわかる。そして、「今日」という言葉は時間の一種で、現在の日にちをシステムで調べて、その日にちが「受付日時」の範囲内に収まっているかを調べることができるかもしれない。そしてユーザーに対しては、ユーザーの住所などの利用者プロファイルを参照して、「あなたの現在いる場所から一番近い窓口はどこにあって、本日何時までなら申請書の提出が可能です」と表示されるかもしれない。（さらに、実際に窓口にいかなくて済む電子的な申請方法もあわせて勧めるかもしれない。）

上記のような世界を実現するシステムにおいて最もコアとなる要素技術の１つは、RDF 専用のデータの取得・変更・挿入・削除などを行なうクエリー（Query）言語の開発とそれを利用した検索システムであると考えられる。

オランダの AIdministarator 社で開発された Sesame[133]と呼ばれる RDF スキーマベースのリポジトリとその検索機能を備えたツールがある（これは将来的には DAML+OIL ベースの検索もサポートする予定がある）。EU 委員会の IST プログラムにおいて活動している On-To-Knowledge プロジェクトの主要な発表成果の１つである。検索言語に OQL ベースのギリシャの ICS-FORTH で開発された RQL (RDF Schema Query Language)を用いている。

この他にも 3.2.1 で述べたように、RDFStore の RDQL/SQUISH、RDFdb の Algae、さらに HP の RDQL[134]など、SQL のような形式で RDF のステートメントを検索する言語が開発されている。

今後は、クエリー機能の向上と、このクエリーを操作するための扱い易いユーザーインターフェイスの開発が必要である。

現在世の中に提供されているセマンティック検索デモシステムの多くは、ある程度閉じたサイト内に蓄積された RDF データや、検索できる RDF データのボキャブラリに対して制約を課している。今後は、オントロジーの相互変換や統合などを複数のサイト間などで動的に行なうフレームワークの開発が必要になると考えられる。

DAML プロジェクトの ITTALKS[135]は、IT 分野の講演の Web ページにメタデータを付与し、さらに個人のプロファイルやスケジュール、住所などの情報と組み合わせて、エージェントが推論を行ない、参加が推奨される講演をユーザーに提供するポータルサービスである。ITTALKS はいくつかのエージェントを持ち、その中の MapQuest エージェントと呼ばれるものは、別のサービスに対するオントロジーの翻訳機能を提供し、ユーザーの興味のある講演を探すエージェントと協調して動作する。ここでは、ACM のトピックと UMBC (University of Maryland Baltimore Country) のトピック間でオントロジー交換を実現している。このような技術が任意のオントロジーの相互変換フレームワークの開発の進展につながると考えられる。

さまざまなオントロジーが存在し、それらを相互に参照して動作するシステムの場合は、従来の AI システムでは許されなかった矛盾などについても処理できる仕組みが必要になる。現在の Web の検索結果に誤り（ユーザーの意図と結果が異なっているという意味で）があっても、検索エンジンがそれなりにユーザーの役に立っているように、まずは柔軟なルール運用による実用性の検証が求められている。また、オントロジーの信頼性に関してもデジタル署名や複数ユーザーのレイティングなどによってフィルタリングするアイデアがあるが、実用性の検証が求められている。

---

133 http://sesame.aidministrator.nl/doc/sesame-dissemination.ppt
134 http://www.hpl.hp.com/semWeb/rdql.html
135 http://www.semanticWeb.org/SWWS/program/full/paper41.pdf

セマンティック Web 技術と次世代電子政府での
活用方法に関する調査研究

調査報告書

平成１４年３月

発行　　財団法人ニューメディア開発協会
〒108-0073　東京都港区三田１－４－２８
TEL　　０３－３４５７－０６７２