

4-2-3 RDF を用いたデータ収集技法による太陽地球系科学セマンティック Web 構築の試み

4-2-3 An STP Semantic Web Technique and Service Using RSS 1.0 and RDF

村田健史 木村映善 山本和憲 亘 慎一 加藤久雄 森川靖大
是津耕司 木俣 豊 下條真司

MURATA Ken T., KIMURA Eizen, YAMAMOTO Kazunori, WATARI Shinichi,
KATO Hisao, MORIKAWA Yasuhiro, ZETTSU Koji, KIDAWARA Yutaka, and
SHIMOJO Shinji

要旨

本研究では、RSS 1.0 によるメタデータ自動収集システムを拡張し、セマンティック Web の構築には不可欠であるデータを RDF 形式でデータベース化するためのデータ収集技法を用いたシステム開発を行った。本システムでは、観測データファイルのメタデータだけでなく、情報通信研究機構 (NICT) が Web で公開している太陽地球環境予報、STARS を使って登録された研究者による各種現象のイベント情報、NICT が定義した太陽地球環境用語集、国立天文台が発行している理科年表情報など、太陽地球系科学分野に関連する様々な情報を統合する。そのために、HTML 形式から RDF 形式に変換するプログラムおよびリレーショナルデータベースのデータから RDF 形式にマッピングするプログラムを実装した。本研究で提案する RDF データで使用するメタデータ語彙に関しては、既存の標準・共通語彙のみでは太陽地球系科学分野独自の表現に対応できないため、本研究で独自の語彙群を定義した。この語彙の定義には W3C が定めたセマンティック Web 標準フォーマットである RDFS (Resource Description Framework Schema) と OWL (Web Ontology Language) を用いた。本研究の RDF データを利用する場合、提案した RDFS/OWL を共有することでメタデータ語彙を統一することが可能である。また、第三者の RDF データと本研究で提案する RDF データとを統合する場合、互いの RDFS/OWL において、語彙名と意味をマッピングすることにより、語彙の差異を吸収することが可能である。太陽地球系科学分野のセマンティック Web を地球科学分野全体のセマンティック Web に拡張する試行として、本研究では愛媛大学沿岸環境センターが構築している es-BANK (生物試料バンク) データ、インドシナ半島における水環境の化学汚染実態の解明と汚染除去技術の開発を目的とした RR2002 データ、理科年表 (暦部・天文部・気象部・物理/化学部・生物部・環境部) のデジタル情報を対象とし、必要なメタデータ語彙を定義した。これまではネットワーク上で独自公開されてきた地球観測データファイル、太陽地球環境予報、研究者イベント情報、地球環境情報などは、人間がそれらのデータ間にある時間・空間・概念・現象などの関係性を理解したうえで個別に処理するしかなかった。しかし、共通のメタデータ語彙を用いて RDF 化し RDF データベースに集約することで、計算機がデータ間の関係性を理解しながら処理可能となった。本研究で実装した RDF データベースには、RDF データ収集技法を用いて約 1 千万のステートメント (主語・述語・目的語のセット) を格納した。RDF データベースに対してクエリ文を用いて検索する方法と、ステートメントの関係性を一つずつ辿って検索する方法を実装し、それらの処理時間・検索結果の実用性と有用性を評価した。今後は、太陽地球系物理分野や地球環境分野以外の多くの地球科学分野の観測データを RDF 化することにより、分野横断的な全地球科学セマンティック Web の実現が期待される。そのためには、各分野のドメインエキスパートがメタデータ語彙の定義を的確に行い、さらにこれらを結合することにより、

本研究で構築した RDFS/OWL を整理・拡張する必要がある。また、これまでの RDF (データベース) の多くはテストベッドとして実験的に構築されてきた。今後は、学際的研究を可能とするセマンティック Web 環境の実現により、eGY の目的である、人類の生存に影響する複雑な地球システムを理解のための新たな知識発見や情報統合に貢献することが期待される。

In the Solar-Terrestrial Physics (STP), it is pointed out that circulation and utilization of observation data among researchers are insufficient. One of the reasons is that the data formats of STP observation data are not common. This is not only the issue of STP data but also of other natural science data. To archive interdisciplinary researches, we need to overcome this circulation and utilization problems.

Under such a background, the Solar-Terrestrial data Analysis and Reference System (STARS) has been designed and developed by the authors' group. The STARS has its own database that manages meta-data of satellite and ground-based observation data files. The STARS provides users with cross-over data file search services and download services over the Internet. It is noted that retrieving meta-data from the observation data and registering them to database have been carried out by hand so far in the STARS. It is hard to deal with a huge amount of observation data due to the lack of manpower.

We developed an automatic meta-data collection system for the observation data using the STARS RSS (RDF Site Summary) 1.0. The RSS 1.0 is one of the XML-based markup languages based on the RDF (Resource Description Framework), which is designed for syndicating news and content of news-like sites. Using the RSS 1.0 as a meta-data distribution method, the workflow from retrieving meta-data to registering them into the database is automated. This technique was applied for the DARTS (Data Archive and Transmission System), which is a science database managed by the PLAIN Center at ISAS/JAXA in Japan. We succeeded in generating and collecting the meta-data automatically.

Our final goal is to establish the STARS Semantic Web. The Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises, and communities. The most fundamental issue on the establishment is who manages meta-data in the Semantic Web. In the present study, we designed meta-data of the STARS along with the RSS 1.0 document. In order to describe the meta-data of the STARS beyond RSS 1.0 vocabulary, we defined original vocabularies for the STARS resources using RDF Schema.

Our system works as follows. The RSS 1.0 documents generated on data sites are automatically collected by a meta-data collection agent. The agent then extract meta-data to store them in an XML database. The XML database provides advanced retrieval processing that has considered property and relation. In the future, we develop a RDF database supported by inference engine, which leads to automatic processing or high level search for the data which are not only for observation data but for news and event information related to the STP.

【キーワード】

セマンティック Web, 宇宙天気, RSS 1.0, RDF, オントロジ
Semantic Web, Space weather, RSS 1.0, RDF, Ontology

1 まえがき

我々の住む地球にはメカニズムが判明していない自然現象が数多く存在し、人類だけでなく地球上に住む全ての生物や、地球資源、宇宙環境のためにも、太陽地球系科学観測の推進が求められている。太陽地球系科学の発展には、国際共同観測

データの自由な交換が不可欠である。1957～1958年の国際地球観測年 (IGY: International Geophysical Year) では多岐にわたる観測が世界各国の協力で実施され、IGY 実施前に比べて10倍以上のアナログデータが取得された。これらのデータは地磁気、電離層、気象学、地震学など様々な分野における地球科学の推進に役立ち、南

極条約の成立にもつながった[1]。1980年代までは郵送によるアナログデータの収集と交換が主流であったが、1990年代には観測データのデジタル化が積極的に進められ、最近ではWebベースでのデータ交換が一般的となった。

このような時代の潮流のもと、IGYから50年となる2007～2008年を国際デジタル地球年(eGY: electronic Geophysical Year)[2]とする計画が発足した。デジタル情報の利点であるネットワークによる高速データ交換や計算機による高速データ処理を活用し、人類の生存に影響する複雑な地球システムの理解と、情報統合による新たな知識発見環境の実現が期待されている。しかし、現在、Web上の多くのリソースは単純なメタ情報しか含んでおらず、その量も少ない。そのため、ネットワークを介したデータ交換においては、対象のデジタル情報の意味をコンピュータが自動的に理解する術がほとんど無く、情報の検索や活用が原始的かつ単純なレベルに留まっている。

これらの問題を解決する次世代Web技術の一つとして、セマンティックWebが提案されている[3]。セマンティックWebでは、データに意味を記述したメタ情報を付与することで、より複雑で精度の高い検索が可能になる。一方、セマンティックWebでは、誰がどのようにメタデータを記述するのかが問題の一つとして指摘されている。メタデータの標準記述フォーマットはRDF(Resource Description Framework)形式と定まっているものの、RDF形式はHTML形式ほどヒューマンリーダブルな記述形式ではなく、専門家以外にはRDF形式でデジタル情報を生成するのは容易ではない。

ネットワーク上に分散したデジタル情報の問題点を解決するための試みとして、天文学の分野で成果を上げているVO(Virtual Observatory)がある[4][5]。VOでは、世界中に分散する観測データを計算機内に実現した仮想宇宙空間に投影し、利用者にそれぞれの観測データの分散性や観測波長の違いを意識させない利便性の高いシステムを提供している。ただし、VOはOAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting)に準拠した形でメタデータの交換を行っているため、RDF形式のデータを標準とするセマンティックWebと協調は容易ではない。

本研究で対象とする太陽地球系科学分野においても、デジタル形式による観測データの取得と蓄積は進んでいるものの、それらのデータを利活用するための体制には不十分な点が多い。この問題を解決するため、情報通信研究機構電磁波計測研究センター宇宙環境計測グループでは、太陽地球系観測データ解析参照システム(STARS: Solar-Terrestrial data Analysis and Reference System)の開発を行ってきた[6][7]。STARSでは、太陽地球系物理観測データのメタ情報(ファイル名やデータ管理機関、観測日時など)をメタデータベースで集中管理することで、分散管理されている太陽地球系物理観測データの横断検索・取得・解析を実現している。さらに、RSS 1.0(RDF Site Summary)を用いることにより、メタデータベースに自動的にメタデータを登録している[8][9]。STARSの主機能は観測データファイルの一元的数据取得とデータプロットである。RSS 1.0によるメタデータ自動収集はSTARSの一部であり、メタデータ収集対象は観測データファイルに限定されている。本研究では、RSS 1.0によるメタデータ自動収集システムを拡張し、RDF形式でデータベース化するためのデータ収集技法を用いたシステム開発を行う。これにより、セマンティックWebの問題の一つであるメタ情報が付随されたRDF形式データの生成の問題の解決を試みる。デジタル情報にメタ情報を付随してデジタルアーカイブを構築した例として、東京大学大学院情報学環の文化資源統合アーカイブがある[10]。このデジタルアーカイブはセマンティックWebの技術であるRDFS(Resource Description Framework Schema)やOWL(Web Ontology Language)を用いてデジタル情報にメタ情報を付随し、歴史資料と文化資源に関するセマンティックWebの構築を試みている。本研究は、太陽地球系科学セマンティックWebの構築を目的とするが、RDFデータ収集対象を他分野のデジタル情報にも拡張することで、太陽地球系科学分野を越えた地球科学セマンティックWebの構築を提案する。

本論文の構成は以下になる。**2**では、太陽地球系科学分野におけるデジタル情報の現状と問題点について述べ、その問題点を解決するために我々がこれまで取り組んできた試みについて

紹介する。**3**では、RDF 文書にメタデータを記述する際に必要なメタデータ語彙の定義について述べる。**4**では、RDF を用いたデータ収集技法とそれにより収集した RDF データをデータベース化するための RDF データベースシステムの実装について述べる。**5**では、実装した RDF データ収集システムと RDF データベースの運用状況について述べ、実際に RDF データベースを利用した例とその処理時間、得られた結果について考察する。最後に、**6**でまとめと今後の課題について述べる。

2 太陽地球系科学分野のデジタル情報の現状

2.1 太陽地球系科学観測データのフォーマット

科学衛星により得られた観測データは、Level-0 データ、Level-1 データ、Level-2 データとデータ校正の度合いや公開レベルによってレベル分けされる。研究者が利用するデータは Level-2 データであり、観測班の管理のもとに配布される。Level-2 のデータフォーマットは観測班によって独自の形式で公開されるため他の研究者がそのデータを利用するには、専用の読み取りツールが必要となる。本研究で対象とする太陽地球系科学分野では、世界中に様々なフォーマットが散在しているため、複数の研究機関が独自に管理しているデータを総合的に利用することが困難となっている。

2.2 太陽地球系科学観測データのメタデータ

2.2.1 自己記述型データフォーマット

自己記述型データフォーマットとは、データ自身が持つ値に加えて、そのデータに関するメタデータが付随されたものである。太陽地球系物理観測データファイルのメタデータの例としては、ファイル内に使用されている数値の説明、単位、観測機器名やデータ管理者名、作成日時などが挙げられる。自然科学観測分野における代表的な自己記述型データフォーマットには、気象学分野では netCDF (network Common Data Format) 形式が一般的であり、天文学分野では FITS (Flexible Image Transport System) 形式が利用されている [11][12]。

2.2.2 CDF(Common Data Format)データ

CDF は、NASA/GSFC (National Aeronautics and Space Administration/Goddard Space Flight Center) の NSSDC (National Space Science Data Center) が提唱する汎用的データフォーマットである [13]。現在は、NASA のデータセンターで管理されている太陽地球系物理観測データの多くが CDF で公開されている。CDF では、データファイル全体のメタデータを格納する g 属性と、個々の変数のメタデータを格納する v 属性が用意されている点が特徴である。近年、我が国の太陽地球系物理観測データの多くでは CDF が導入されつつあり、2005 年 8 月打ち上げの REIMEI や 2007 年 9 月打ち上げの KAGUYA などの新しい衛星の Level-2 データでは CDF が採用されている。

2.3 太陽地球系科学分野のサービス

太陽地球系科学分野では各データ公開機関による観測データファイル公開の他に、以下のような情報の公開や配信サービスを行っている。

2.3.1 太陽地球環境予報

情報通信研究機構 (NICT: National Institute of Information and Communications Technology) の電磁波計測研究センター宇宙環境計測グループは、世界各地の観測データを基に、太陽フレア、太陽プロトン現象、磁気嵐等の状況を把握して、その発生を事前に予測した情報を電子メールや FAX 等で提供している [14]。太陽地球系科学分野の研究者や放送事業者や航空会社等は、研究や業務においてこの予報を参考にしている。

2.3.2 太陽地球環境 (STE: Solar-

Terrestrial Environment) 用語集

NICT の電磁波計測研究センター宇宙環境計測グループが提供している太陽地球環境の専門用語について解説した情報である。同センターの Web 上で一般に公開されている [15]。

2.3.3 3D Global MHD simulation Web

3D Global MHD simulation Web とは、NICT 電磁波計測研究センター宇宙環境計測グループのシミュレーションチームが公開している NICT リアルタイム MHD シミュレーションページのサービスの一つである [16]。NICT では、2003 年より数値宇宙天気シミュレーションを運用している。これは、実観測データを用いたリアルタイムでの

Global MHD シミュレーションにより、宇宙空間での災害予測を行うものである。最新のシミュレーション結果は図1のようにリアルタイムで3次元可視化され、ユーザはWeb上で視点移動、拡大・縮小、回転、平行移動等の幾何変換を用いて閲覧することができる。

2.3.4 太陽地球系物理イベント情報

学術論文やニュース等で紹介された太陽地球系物理観測分野に関係するイベント情報や、各研究者が発見した現象、研究した現象などが太陽地球系物理イベント情報として解析や研究の手がかりとなる場合が考えられる。STARS アプリケーションでは、イベント情報の参照機能と登録機能を有しており、研究者による各種現象情報を収集してデータベース化している。また、文部科学省科学研究費補助金(学術創成)のバーチャル観測所タスクフォースグループがイベントリストをWeb上で一般に公開している[17]。

2.4 理科年表

理科年表は、国立天文台が編纂する科学の全分野を網羅するデータブックである[18]。理科・科学データの原点として幅広く活用されている。理科

年表は、暦部、天文部、気象部、物理/化学部、地学部、生物部、環境部、附録で構成されており、とくに地学部では、太陽地球系物理に関連する地磁気および重力・電離圏の情報を取り扱っている。

2.5 太陽地球系科学分野の情報を総合利用する際の問題点

太陽地球系科学観測データ公開機関は各々が独立にデータを公開しており、公開データのデータフォーマットやデータ構造は統一されていない。このため、複数データの総合的利用の際には、情報の所在の分散とデータフォーマットの差異が問題となる。データ公開サイトが増えた場合には、アカウントを個別に管理しなければならないため、管理側・ユーザ側ともに負担となる。また、各データサイトの情報は定常的に更新されるため、ユーザは随時、更新情報を個別に確認することが求められる。また、データフォーマットの差異により、データのメタ情報はHTML形式の文中にある場合、CDF形式のヘッダ中に埋め込まれている場合、バイナリ形式で提供されている場合など様々である。このような場合に、ユーザは個々の読み込みツールを使い分ける必要がある。

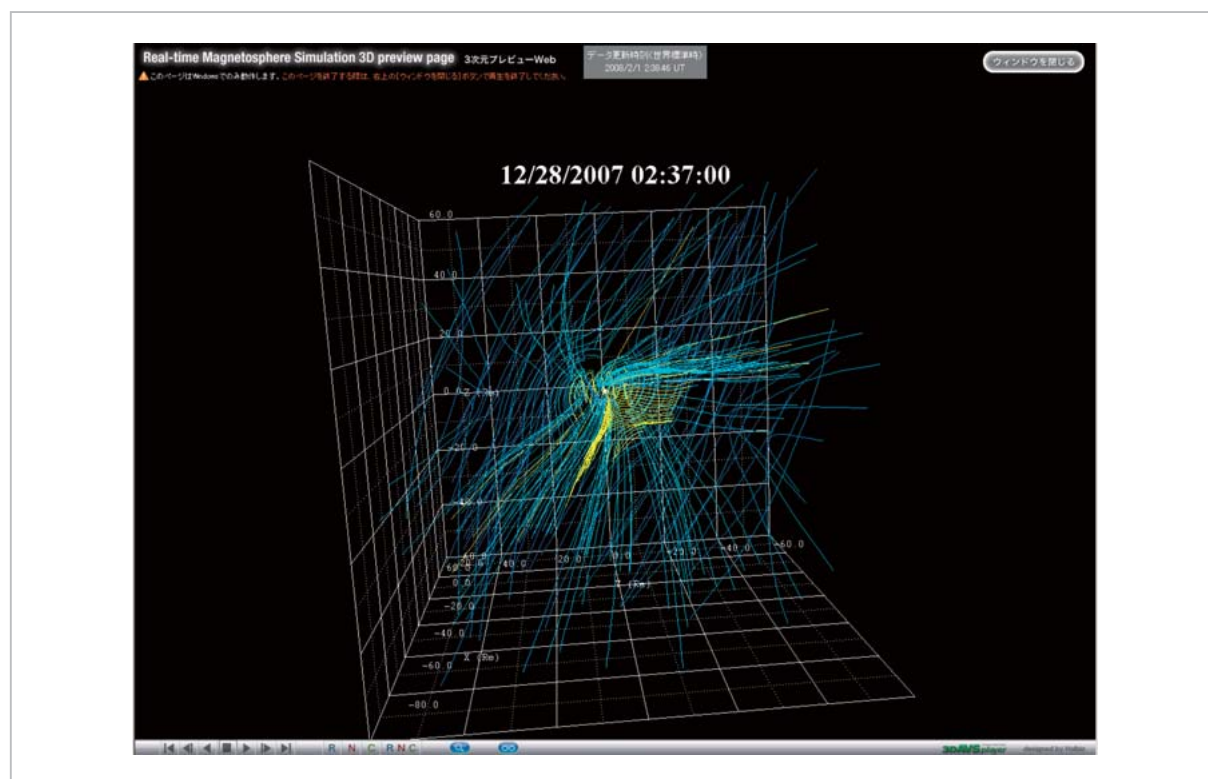


図1 3D Global MHD simulation Webによる3次元可視化の例(情報通信研究機構)

また、それぞれの情報は関連付けされていないため、異なるデータ間の関連情報を取得することは容易でない。例えば、ある研究者が NICT 太陽地球環境予報の太陽活動記事を参考にして解析を行う場合、次のようなステップが必要となる。(1) 研究者は太陽地球環境予報の太陽活動記事から太陽フレアの発生を記述している記事を発見する。(2) その記事の発行日時もしくは予報日時を取得する。(3) 各データ公開期間にアクセスし、その日時に観測された太陽活動に関するデータファイルの存在の有無を確かめる。(4) 必要なデータファイルをダウンロードし、各解析ツールを用いてプロットする。毎回このようなステップを踏むことは、研究者にとって負担となる。

2.6 太陽地球系観測データ解析参照システム (STARS)によるこれまでの取り組み

衛星観測データの利用において、2.5 で述べた問題点を解決するため、情報通信研究機構電磁波計測研究センター宇宙環境計測グループを中心に構築されてきた STARS (Solar-Terrestrial data Analysis and Reference System) では、各地に分散管理されたファイルの一元的利用と、ファイルフォーマットの差異を吸収したプロット機能を提供している。

2.6.1 STARS メタデータベース

2.1 で述べたように、観測データファイルのフォーマットは統一されておらず、データ構造が複雑なため、データベース管理システム (DBMS) により管理するのは困難である。そこで STARS では、観測データそのものではなく、観測データのメタデータをデータベースで管理する試みを行ってきた。STARS メタデータベースでは、太陽地球系物理分野の実際の管理形態に沿って、ユーザ管理を Mission-Team-Group-User の階層構造で、データ管理を Mission-Team-Data の階層構造で管理している。また、ユーザインタフェースとしてデータベースにアクセスするための Web サービスを提供しており、これらを利用した検索・ダウンロード・プロット機能を備えたアプリケーションとして STARS5 が配布されている[19]。表 1 に 2008 年 2 月現在の STARS メタデータベース登録数を示す。

表 1 STARS メタデータベース登録数 (2008 年 2 月現在)

テーブル名	レコード数
Mission	44
Team	114
Group	26
Data	300
File	2,649,783
User	154

2.6.2 SEDOC (Stars Event DOCment) クラス

STARS ではデータ読み込み処理において、衛星観測データのフォーマットの差異を吸収するために、SEDOC クラスを開発している。SEDOC クラスでは、個々のデータフォーマットの読み込みプログラムをカプセル化しており、ユーザは異なるデータフォーマットにおいても統一インタフェースを通じてデータを取得できる。

2.6.3 RSS 1.0 によるメタデータ自動収集システム

RSS 1.0 (RDF Site Summary) とは、XML に準拠したメタデータ配信用のフォーマットである。STARS ではこれまで、メタデータの収集および登録作業がデータベース管理者の手動によって行われてきたことが問題であった。STARS では RSS 1.0 を用いたメタデータ自動収集システムを実装することにより、過去に蓄積された大量のデータファイルや毎日生成される最新のデータファイルのメタデータ収集が可能になった[9]。また、人為的ミスが無くなるとともに人的負担を減らすことができた。

2.6.4 STARS の問題点

STARS は、リレーショナルデータベースである STARS メタデータベースに依存している。STARS メタデータベースは衛星観測データの利用に特化しており、他のデータをそのままインポートすることができない。そのため、他のデータへの拡張を行う場合に別途テーブルまたはデータベースを作成する必要がある。また、分野を越えて横断的な検索を行いたい場合には、共通項を見つけ、正規化を行い、テーブル間の関連付けを行わなくてはならない。この関連付けはデータの意味を知っている人の手を介してレコード単位で

施されるため、拡張に比例して膨大な作業量を必要とする。これはリレーショナルデータベースに格納する値がリテラルであり、データの意味情報を持たないために機械的処理が行えないことが原因である。

3 RDF/RDFS/OWL を用いた太陽地球系科学メタデータ語彙の定義

3.1 RDFの概要

RDF は Web 上のリソースに関する情報を明瞭かつ論理的に表現するデータモデルであり、それを記述するための言語体系である。RDF の基本構造は単純であり、主語、述語、目的語の3つの要素により構成される。主語で表現されるのは URI (Uniform Resource Identifier) によって確実に識別できるリソースのみであり、URI によって一意に識別できないリテラル値は主語になることができない。述語は主語と目的語の間に存在する関係をメタデータ語彙によって表現する。このメタデータ語彙には、RDF の標準語彙のほか、主に書式や Web コンテンツの情報を表現する Dublin Core 語彙セット、人に関する情報を表現する FOAF (Friend of a Friend) 語彙セット、Web コンテンツの更新情報を表現する RSS 1.0 語彙セットなどの共通語彙がある [20] - [22]。これらの標準語彙・共通語彙を用いても主語と述語の関係性を表現できない場合、RDF は、独自に定義した XML 名前空間 (Namespace) とメタデータ語彙の使用を許している。主語、述語、目的語の3要素で表された1つの文は、ステートメントと称される。

3.2 RDFS/OWLの概要

RDFS/OWL は RDF で用いる語彙を定義する方法として W3C によって策定されたフォーマットである。RDFS と OWL はともに RDF 形式で記述されるため、データモデルは 3.1 で述べた RDF データモデルと同じである。RDFS は比較的シンプルな語彙定義を行う際に用いられ、RDFS では定義できない詳細な論理表現が必要な際に OWL が用いられる。また、RDFS/OWL は語彙の定義だけでなく、語彙同士の関係や、同じ性質のリソースをグループ化するクラス定義の手段も提供している。計算機が RDF データを処理

する際、RDFS/OWL を参照することで矛盾点の発見や推論処理、演繹を行うことが可能になると期待される。

3.3 RDF/RDFS/OWL による太陽地球系科学セマンティック Web 構築の試み

Web は HTML の普及によって爆発的に広がり、多くの Web 上のリソースが HTML によって記述されている。しかし、HTML は Web ページを記述するための言語として設計されたものであり、一般の文書やデータを記述するための要素が不足している。現状では HTML によって記述された Web ページは人間がその内容を確認してはじめて意味を成すものであり、計算機は HTML の記述に従ってその Web ページを表示するだけである。そのため、対象のリソースの意味をコンピュータが機械的に理解する術がほとんど無く、情報の検索や活用が原始的かつ単純なレベルに留まっている。これらの問題を解決する次世代 Web 技術の一つとして、セマンティック Web が提案されている。セマンティック Web ではリソースの意味を明示的に示すためのメタデータが必要である。メタデータを記述する標準フォーマットとして 3.1 で述べた RDF が W3C によって策定されている。

セマンティック Web の発展は、図 2 に示す階層の形で段階的に進められている [23]。最下層の URI、Unicode、XML、名前空間はセマンティック Web だけでなく、現在の Web の基盤となる

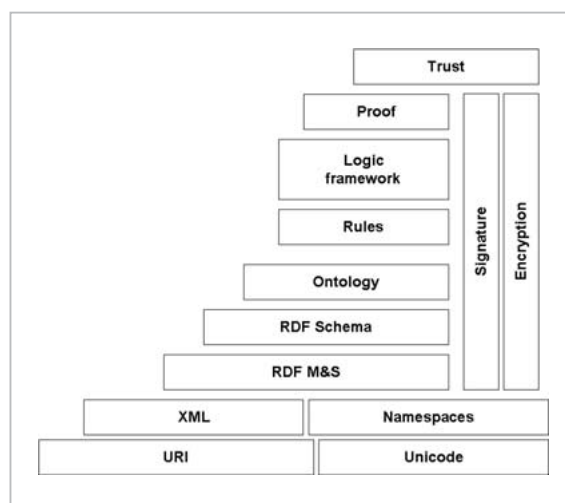


図2 セマンティック Web の技術群を示すレイヤーケーキ

技術である。これら既存技術の基盤の上に、セマンティック Web の要素技術が積み重ねられていく。本研究で作成している RDF のモデルの一部を図 3 に示す (RDF モデル図は一般的に、リソースは楕円、リテラルは四角形、述語は矢印で表現される)。また、図 3 に対応する RDF/XML 文書を図 4 に示す。図 3 と図 4 は GEOTAIL 衛星の磁場観測データを RDF 化したものの一部である。名前空間 `sw` は、本研究で独自に定義したメタデータ語彙であり、`<sw:mission>` はミッション名、`<sw:st>` は観測開始時刻のユリウス日、`<sw:et>` は観測終了時刻のユリウス日を表現する語彙である。RDF はこのような単純な定義を結合していくことで、複雑な構造を可能としている。

RDFS を用いて 2.3.1 で述べた NICT 太陽地球環境予報の語彙を定義した例を図 5 に示す。図 5 の左側部にある主語 (楕円) が、1998 年 4 月 17 日に発令された太陽地球環境予報リソースの URI を示している。主語から派生している述語 `<nict_stpfc:hasOverview>` は、太陽地球環境予報が概況記事を持っているという関係を示し、概況記事を識別する URI が目的語となる。述語 `<rdf:type>` は、主語リソースが目的語クラスのインスタンスであることを示す。RDFS は、述語の domain (定義域) と range (値域) を定義することができる。図 5 の右側部では、述語 `<nict_stpfc:hasOverview>` の定義域には太陽地球環境予報リソースをまとめたクラスである `<nict_stpfc:SpaceWeatherReport>` が示され、値域には概況記事リソースをまとめたクラスである `<nict_stpfc:Overview>` が示される。

次に、OWL を用いて太陽地球環境予報の語彙を定義した例を図 6 に示す。図 6 の左側部にある主語の楕円が、1998 年 4 月 17 日に発令された太陽地球環境予報の中の太陽活動に関する記事リソースの URI を示している。その主語から派生している述語 `<nict_stpfc:description>` は、記述内容が何であるかという関係を示し、目的語に実際の記事内容が示されている。OWL は RDFS の拡張言語であるため、RDFS の定義域や値域の定義を OWL でそのまま利用することができる。OWL 独自の語彙定義として、出現回数に関する制約やクラス間の関係を表現することができる。図 6 の右側部では、述語 `<nict_stpfc:description>` の最大出現回数が OWL 述語 `<owl:maxCardinality>` によ

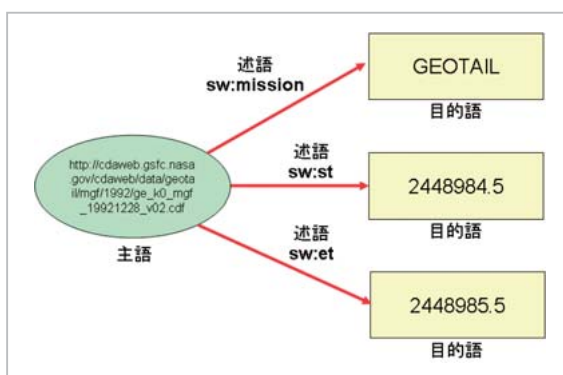


図3 RDF の 3 要素を用いた STARS メタデータのモデル例



図4 図 3 のモデルの RDF/XML 記述

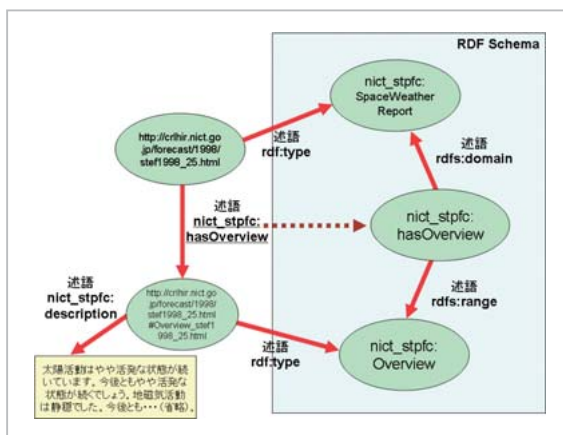


図5 RDFS を用いた語彙の定義域・値域定義 (太陽地球環境予報の例)

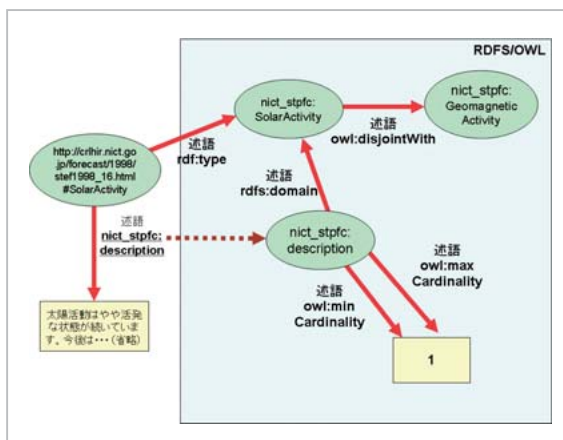


図6 RDFS/OWL を用いた語彙の定義域・値域、出現回数制約、クラス関係の定義 (太陽地球環境予報の例)

て1回であることが示され、最小出現回数もOWL述語〈owl:minCardinality〉によって1回であることが示されている。これは、太陽活動の記事は必ず一つの記述を持ち、2つ以上の記述を持たないことを定義している。OWL述語〈owl:disjointWith〉は主語クラスと目的語クラスは互いに素であることを示し、図6では太陽活動記事と地磁気活動記事は意味が異なるものであることを定義している。

文献[24]の付録Bの各表に、本研究で用いる太陽地球系物理観測データRDFで用いるメタデータ語彙一覧、STARS イベント情報RDFで用いるメタデータ語彙一覧、バーチャル観測所タスクフォースグループのイベント情報RDFで用いるメタデータ語彙一覧、太陽地球環境予報RDFで用いるメタデータ語彙一覧、理科年表地学部情報RDFで用いるメタデータ語彙一覧をまとめた。これらのメタデータ語彙を用いて太陽地球系科学分野のRDFデータを生成することにより、太陽地球系科学セマンティック Web を構築することを試みる。図7にRDF/RDFS/OWLによる太陽地球系科学セマンティック Web 構築例を示す。図7では、観測データRDF、STE用語集RDF、太陽地球環境予報RDFがメタデータ語彙を用いて各情報を展開している。観測データRDFは〈sw:mission〉語彙を用いてミッション名が“ACE”であることを示しており、STE用語集RDFは〈dc:title〉を用いて“ACE”用語について説明していることを示している。Web上で何も関連性を持たなかった2つの異質のデータが、セマン

ティック Web 上では“ACE”というリテラルを通して関連性を持つ。同様に観測データRDFと太陽地球環境予報RDFは、観測開始日時を表す〈sw:st〉、観測終了日時を表す〈sw:et〉、太陽地球環境予報が発行された日時を表す〈nict_stpf:Julian〉の語彙が持つ日時情報を通して関連性を持つ。このようにセマンティック Web では、これまで人が判断することではしか関連性を見出せなかったデジタル情報に語彙を用いてメタ情報を付与することで、計算機に情報の関連性を理解させて機械的処理を可能とする。

3.4 太陽地球系科学セマンティック Web を拡張した地球科学セマンティック Web の提案

3.3では太陽地球系科学分野に限定したセマンティック Web を構築することを提案したが、他分野のデジタル情報をRDF/RDFS/OWLによってセマンティック Web 標準のデータフォーマットで表現し、太陽地球系科学分野のデータと関係性を持たせることでセマンティック Web の拡張を行うことが可能である。図8に本研究で扱うデータの分類と実装対象としたデータ例の関係をまとめた全体像を示す。本研究では3.3で述べた太陽地球系科学分野のデジタル情報に加え、愛媛大学沿岸環境科学研究センターが構築しているes-BANK(生物試料バンク)データ[25]と、同センターのインドシナ半島における水環境の化学汚染実態の解明と汚染除去技術の開発を目的としたRR2002データ[26]、国立天文台が発行している理科年表の暦部・天文部・気象部・物理/化学部・生物部・環境部のデジタル情報を取り扱う。文献[24]の付録Bの各表に、本研究で定義したes-BANKのRDFで用いるメタデータ語彙一

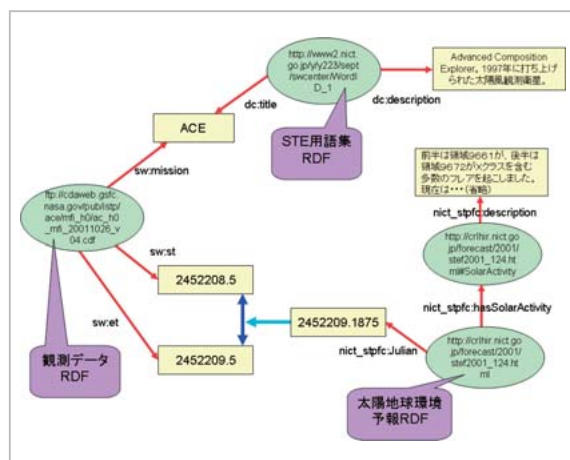


図7 RDF/RDFS/OWLによる太陽地球系科学セマンティック Web 構築例

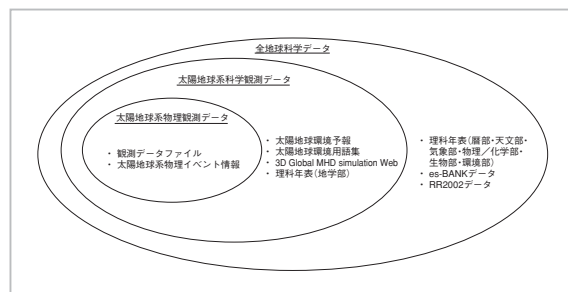


図8 本研究で扱うデータの分類と本研究の実装対象データ例

覧、RR2002 の RDF で用いるメタデータ語彙一覧、理科年表 RDF で用いるメタデータ語彙一覧が示されている。本研究で定義したメタデータ語彙を用いて太陽地球系科学分野だけでなく他分野の RDF データを生成することにより、太陽地球系科学セマンティック Web を拡張した地球科学セマンティック Web を構築することを提案する。

3.5 RDF データ収集の問題点

セマンティック Web を構築するためには、定義したメタデータ語彙を用いて既存のデータを RDF 形式に変換する必要や、今後生成される新規データを RDF 形式として生成する必要がある。しかし、RDF 形式は HTML 形式ほどヒューマンリーダブルな記述形式ではないため、専門家以外には RDF 形式でデジタル情報を生成するのは容易ではない。

1 で述べたようにセマンティック Web では、誰がどのように RDF データを生成するのが問題の一つとして指摘されている。本研究では上記の問題を解決する手段として、4 で RDF データを機械的に収集する技法について議論する。

4 RDF を用いたデータ収集と RDF データベースの実装

4.1 RDF を用いたデータ収集技法の提案

図 9 に、RDF を用いたデータ収集と 4.2 で述べる RDF データベースとを連携させたシステムの概要を示す。図 9 の左下部にある観測データ RSS File を生成する RSS Creator が RSS 1.0 によるメタデータ自動収集システムを示す。本研究では RSS 1.0 によるメタデータ自動収集システムを拡張し、太陽地球系科学分野のデータである STE 用語集、太陽地球環境予報、太陽地球系物理イベント情報を RDF 変換エンジンによって RDF 化して RDF データを収集する。他分野のデータである理科年表情報、es-BANK データ、RR2002 データも同様に RDF 変換エンジンによって RDF 化して RDF データを収集する。収集された RDF データは 4.2 で述べる RDF データベースに自動的に格納される。

本研究で提案する RDF を用いたデータ収集技法は、3 つに分類される。観測データファイルの情報から RDF データを収集する技法と、HTML 形

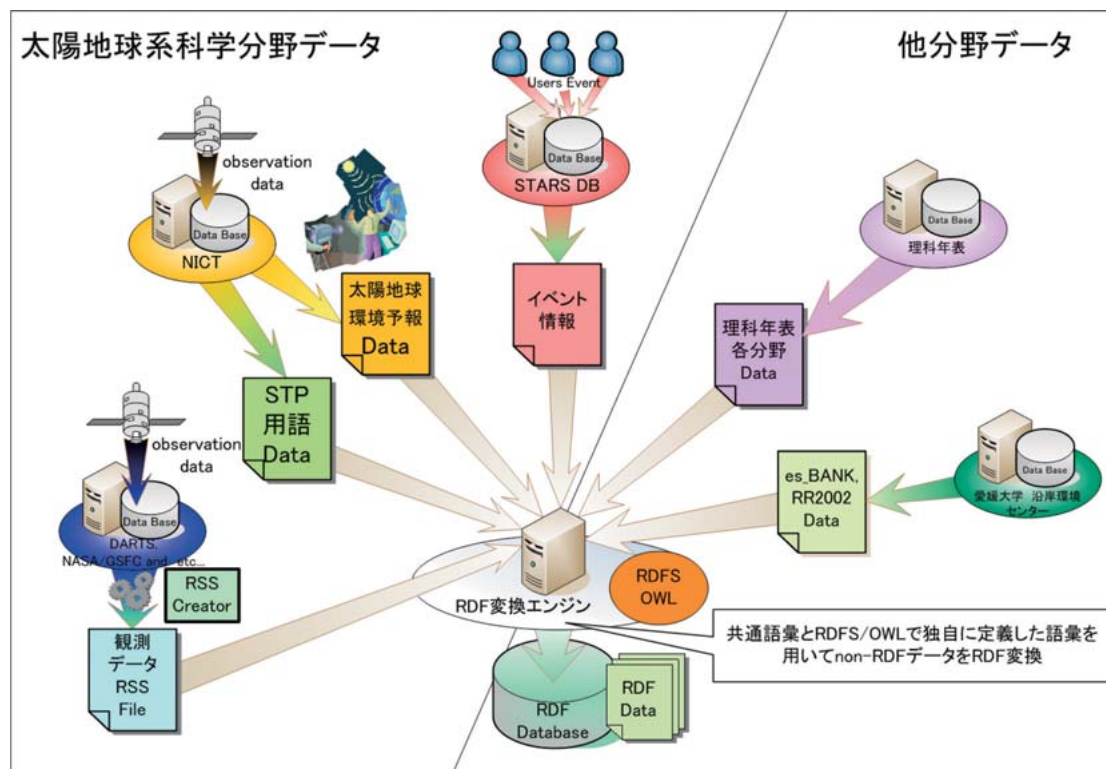


図9 RDF を用いたデータ収集と RDF データベースを連携させたシステムの概要

式の情報から RDF データを収集する技法、リレーショナルデータベース (RDB) の情報から RDF データを収集する技法である (図 10)。観測データファイルの情報から RDF データを収集する技法は、石倉らの RSS 1.0 によるメタデータ自動収集システムを利用する [9]。その他 2つの収集技法について、4.1.1、4.1.2 で詳しく述べる。

4.1.1 HTML 形式の RDF データ収集技法

Web 上の多くの情報は HTML 形式で公開されているが、HTML 形式と RDF 形式は非互換である。XSLT (XML Stylesheet Language Transformations) を用いて XHTML 形式を RDF 形式に変換する汎用ルールとして GRDDL (Gleaning Resource Descriptions from Dialects of Language) が W3C (World Wide Web Consortium) によって勧告されたが、HTML 形式を RDF 形式に変換する汎用ルールは見られない [27]。HTML の見出し要素が単純にフォントサイズの調整に用いられる場合や、テーブル要素が画面のレイアウトを整えるために用いられる場合があり、ページによって要素の使用目的が異なるため、RDF 形式への変換が困難である。

本研究では、構造が定まっている HTML 形式の文書を RDF 文書に変換する試みとして、太陽地球環境予報の HTML 文書を RDF 文書に変換するプログラムを作成した。図 11 に太陽地球環境予報 HTML ファイルを RDF 変換エンジンによって RDF 変換し、RDF データベースに格納するまでのシステム全体図を示す。NICT 電磁波計測研究センターの研究グループは、毎日の予報会議を経て太陽地球環境予報 (図 12) を作成し、

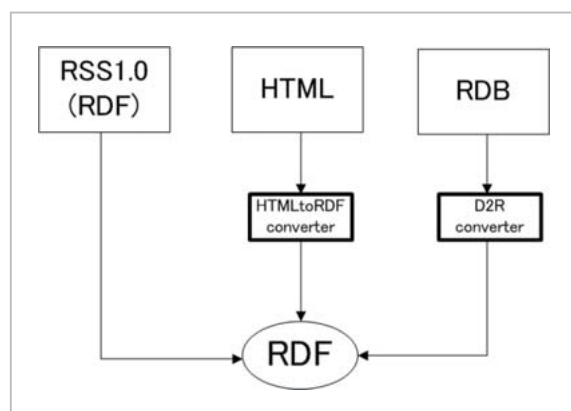


図10 本研究で提案する RDF を用いたデータ収集技法

HTML 形式で Web 上に公開している。本研究の RDF 変換エンジンは 1 日 1 回定期的に NICT のサーバにアクセスし、前回取得したファイルよりも更新日時が新しいファイルがあればそれらを取得する。取得した HTML ファイルを RDF 変換プログラムによって解析し、RDF 形式に変換する。最後に、変換した RDF データを、RDF データベース登録プログラムを用いて RDF データベースに格納する。

RDF 変換前の太陽地球環境予報 HTML ソース例を図 12 に、RDF 変換プログラムを用いて RDF に変換した結果を図 13 に示す。RDF 変換プログラムは HTML ファイルから図 12 で示す記

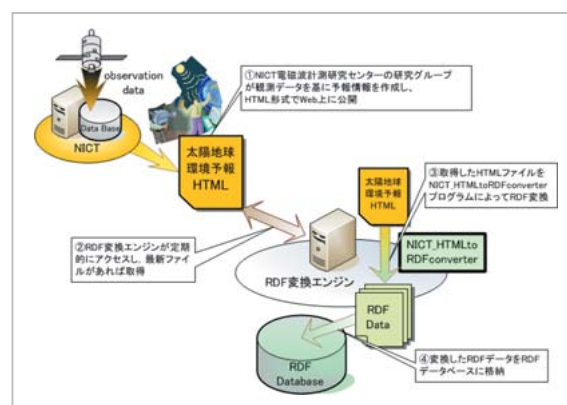


図11 太陽地球環境予報 HTML ファイルのメタデータ自動収集システム

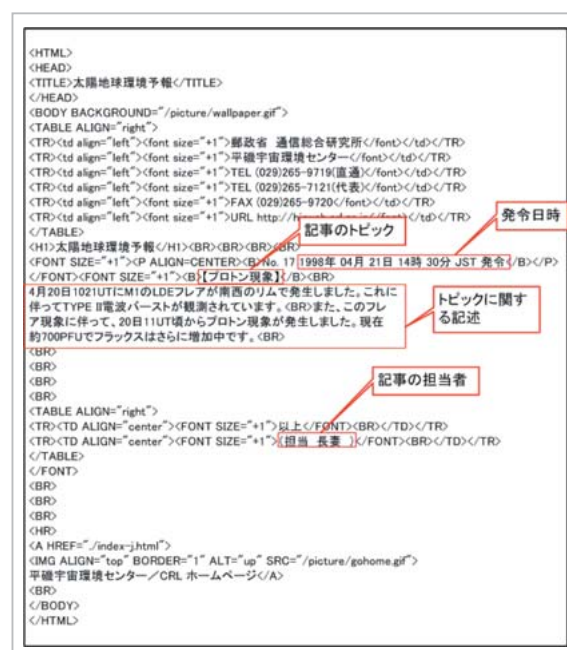


図12 RDF 変換前の太陽地球環境予報 HTML ファイルの一例

事のトピック、発令日時、トピックに関する記述、記事の担当者の情報を抜き出して RDF 化する。変換後の RDF データでは、ユリウス日表現の時刻情報を持つ観測データファイル RDF との親和性を考慮し、図 13 に示すユリウス日表現の時刻情報を追加する。太陽地球環境予報内で発令日時から過去 1 週間の観測データやフレア情報を記載するケースがあるため、発令日時の一週間前のユリウス日も記述している。

なお、RDF 変換の際には、3 で述べた RDFS/OWL で定義したメタデータ語彙を用いた。NICT の長妻氏の協力を得て、太陽地球環境予報

内で登場する記事のトピックをリストアップし、概念的に同じ意味と考えられる語彙の差異を吸収するために RDFS/OWL クラスを定義した。RDFS/OWL クラスと太陽地球環境予報内で登場する各トピックとの対応関係を表 2 に示す。太陽地球環境予報には毎週金曜日に発令される定期予報と、特筆すべき宇宙環境イベントが発生した場合に発令される臨時予報がある。定期予報では登場するトピックが定型化されているのに対し、臨時の発令時にはトピックが定型化されておらず予報担当者の裁量で記述されるため、表 2 のような表記の揺れが起こる。本研究では過去 10 年間の記事において登場したトピック名を網羅し、プログラム対応をしたが、今後新たなトピック名が登場した場合はその都度プログラム対応をしなければならない問題がある。

4.1.2 リレーショナルデータベースからの RDF データ収集技法

図 14 にリレーショナルデータベースの情報を RDF 変換するシステムの概要を示す。これまで STARS によってユーザから収集したイベント情報や、RR2002 データは既に関係データベースに格納されている。そこで、リレーショナルデータベースに対してクエリを発行し、その結果を用いて RDF データに変換する RDF データ変換プログラム D2R converter を作成した。この変換プログラムには、データベースのデータを

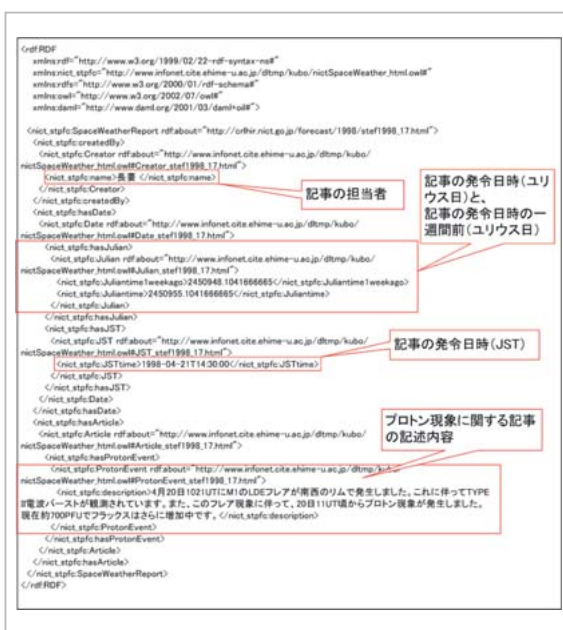


図 13 図 12 の HTML を RDF 変換プログラムを用いて RDF に変換した例

表 2 太陽地球環境予報で登場する記事のトピックと本研究で定義した RDFS/OWL クラスとの対応

RDFS/OWLクラス名	太陽地球環境予報で登場する記事のトピック
Overview	概況
SolarActivity	太陽、太陽活動、太陽活動概況、太陽活動概況(臨時)、太陽活動(概況解説)、太陽フレア、X クラスの太陽フレア、LDE フレア、太陽フレア(追加情報)
GeomagneticActivity	地磁気、地磁気活動、磁気嵐、磁気あらし、地磁気嵐、地磁気あらし、磁気じょう乱、地磁気嵐/訂正版、地磁気嵐終了、太陽風、衝撃波、SC、惑星間衝撃波、惑星間空間衝撃波、磁気雲、CME 現象、CME
ProtonEvent	プロトン現象、プロトン、プロトンイベント、プロトン現象/訂正版、プロトン現象継続、プロトン現象終了、プロトン現象(追加情報)、プロトン現象(訂正)、高エネルギー電子、高エネルギー、高エネルギー粒子
Ionosphere	電離層、電離層擾乱、スボラディックE 層、デリンジャー現象
RadiationBeltElectron	放射線帯電子、放射線帯粒子、高エネルギー粒子
LowLatitudeAurora	低緯度オーロラ、オーロラ予報
f10.7	f10.7

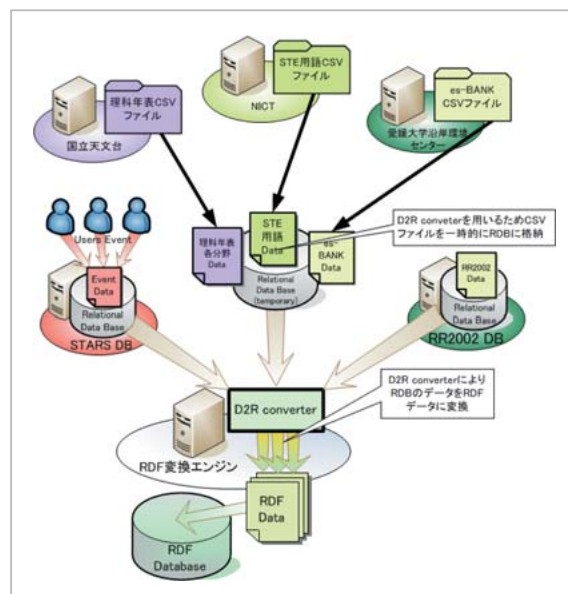


図 14 D2R を用いてリレーショナルデータベースの情報を RDF 変換するシステムの概要

RDF データにマッピングする機能を提供する D2R (Database to RDF Mapping Language and Processor) ライブラリを用いた [28]。

また、理科年表、es-BANK データ、STE 用語集データについては CSV 形式でデータを提供して頂いたため、一時的なりレーショナルデータベースを構築し、データベースに全データを格納してから D2R converter を用いて RDF に変換した。表 3 にレーショナルデータベースである STARS データベースの STARS イベント情報

表3 RDB に格納されている RDF 変換前の STARS イベント情報レコード例

CaseID(PK)	CaseName	CaseAttribute	UserID	iFlag	TimeStamp
1	Substorm	A huge energy transfer event ...	20002	1	2452974.5
2	AKR onset	A sudden enhancement of ...	20002	1	2452974.5
3	plasmoid entry	An event a satellite encounters a plasmoid.	20002	1	2452974.5
4	plasmopause skimming	An event that a satellite is located at the boundary ...	20002	1	2452974.5
5	Pi2 onset	An onset detected in Pi2 pulsation.	20002	1	2452974.5
6	magnetosphere	TBD	20002	1	2453612.9 4804398
7	magnetotail	TBD	20002	1	2453612.9 4804398
...(以下略)	...(以下略)	...(以下略)	...(以下略)	...(以下略)	...(以下略)



図15 表3の STARS イベント情報を RDF 変換するための D2R 設定ファイル(XML 形式)

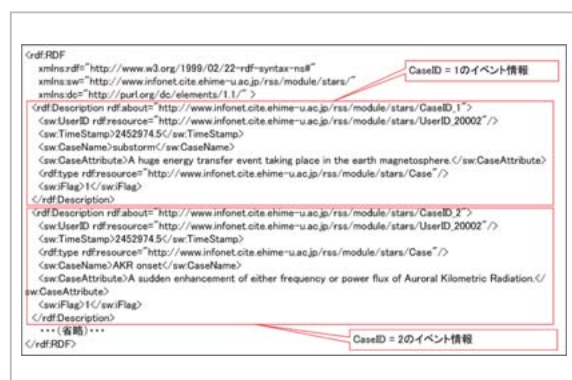


図16 表3を図15を用いて RDF 変換した結果の一部

テーブルのレコード例を示す。CaseID フィールドがイベントを識別する主キーであり、CaseName フィールドや CaseAttribute フィールドでイベントの名称や説明を管理している。

表3のレコード情報を D2R converter を用いて RDF に変換する際に必要な D2R 設定ファイルを図15に示す。D2R 設定ファイルには接続先のデータベース情報と実行する SQL クエリ文、クエリ取得結果を用いて RDF 形式にマッピングするルールを記述する。

図16に表3のレコード情報を図15の D2R 設定ファイルを用いて実際に RDF 変換した結果の一部を示す。CaseID=1、CaseID=2 のイベント情報がそれぞれ RDF 形式で記述されているのが確認できる。

なお RDF 変換の際には、**3**で述べた RDFS/OWL で定義したメタデータ語彙(B.2)を用いた。変換した RDF データは RDF データベース登録プログラムによって RDF データベースに格納される。リレーショナルデータベースのテーブル構造が変更されない限り、定期的に D2R converter を実行することで対象のリレーショナルデータベースの情報を RDF データベースにレプリケーションすることが可能である。

4.2 RDF データベース

RDF データベースとは、RDF データの取り扱いに特化したデータベースである。RDF データベースは、格納対象が RDF フォーマットのデータであればテーブル構造やデータ型といった制約を意識せずに格納することができる。(データ構造やデータ型といったメタ情報は RDF データに含まれていると考える。)リレーショナルデータベースはテーブル構造やデータ型を厳密に定義するため、定義にそぐわないデータを格納することはできない。例えば、太陽地球系物理観測分野のデータを取り扱うのに特化した STARS メタデータベースでは海洋分野や天文分野などの他分野のデータを格納することができない。それに対し RDF データベースは、テーブル構造を自動的に管理するため、データごとにテーブルを設計管理する必要がある。また、それぞれの分野間を越えるためのメタ情報が含まれた RDF データを取り扱うため、分野間に存在するデータ構造の相違を

気にすることなく格納することが可能である。さらに、RDF データベースは RDF クエリ言語をサポートしており、RDF データの検索や操作を計算機上で実行することが可能である。

本研究では Aduna 社がオープンソースとして公開している RDF データベースフレームワークの Sesame 1.2.7 を用いて RDF データベースを構築した[29]。4.1 で述べた RDF データ収集システムと Sesame を用いた RDF データベースを連携させ、RDF データを RDF データベースに自動的に格納するシステムを構築した。本システムの技法を用いることで機械的に RDF データを収集することが可能となる。次章ではこの RDF データベースにより、3.5 で述べたセマンティック Web の問題の解決を試みる。

5 RDF を用いたデータ収集と RDF データベースの試験運用

5.1 RDF を用いたデータ収集の試験運用

5.1.1 RDF 収集対象データ

本研究では、表 4 に示すデータ一覧を RDF によるデータ収集対象とする。本研究でデータ収集の対象とした太陽地球系物理観測データ公開サイト一覧を表 5 に示す。本研究では国内外合わせて 8 つの研究機関が公開している合計 179 種類の観測データを RDF によって収集した。また、太陽地球系物理観測データを除いた各データのデータ公開機関や公開データの情報を表 6 に示す。本研究では国内の 5 つの機関が HTML、RDB、CSV 形式で公開・提供する情報を、RDF を用いて収集した。

表4 RDF を用いたデータ収集対象一覧と各データ収集方法と各データ形式

データ名	データ収集方法	データ形式
太陽地球系物理観測データファイル	RSS1.0によるデータ収集	CDF
NICT太陽地球環境予報	HTML形式からのデータ収集	HTML
NICT専門用語情報		HTML(CSV)
RR2002データ		RDB
理科年表データ		CSV
es-BANKデータ		CSV
学術創成イベント情報		HTML(CSV)
STARSイベント情報		RDB

表5 RDF を用いたデータ収集の対象とした太陽地球系物理観測データ公開サイト

研究機関名	ホスト名(略称)	収集対象データ種類数
宇宙航空研究開発機構(JAXA)	DARTS	32
情報通信研究機構(NICT)	3DWeb	5
名古屋大学	STELAB	20
金沢大学	CIE	1
アメリカ航空宇宙局(NASA)	GDWeb	87
アメリカ海洋大気庁(NOAA)	NOAA	7
ベルギー王位天文台	SIDC	1
カリフォルニア大学バークレー校	BERKELEY	26

5.1.2 RDF データ収集実績

文献[24]の付録 A の表に、本研究で実際に観測データファイルの RDF データ自動収集を行ったデータ公開サイトのホスト名、データ公開サイトが公開するデータ名、公開されるデータファイルが存在する期間が示されている。本システムによって RDF データに変換した NICT 太陽地球環境予報の実績を表 7 と表 8 に、RDF データベースに格納した RDF データのデータサイズとステートメント数を表 9 に示す。また、RDF データベースを運用する計算機環境を表 10 に示す。

表6 太陽地球系物理観測データファイルを除く RDF を用いたデータ収集の対象データ

公開データ	データ公開機関名	データ公開(提供)形式
太陽地球環境予報	情報通信研究機構(NICT)	HTML
STP 専門用語情報		HTML(CSV)
RR2002	愛媛大学沿岸環境センター	RDB
es-BANK		RDB(CSV)
学術創成イベント情報	バーチャル観測所タスクフォースグループ	HTML(CSV)
理科年表	国立天文台	CSV
STARSイベント情報	愛媛大学総合情報メディアセンター	RDB

表7 NICT 太陽地球環境予報の RDF データ変換実績(2007 年 2 月 12 日現在)

データ公開期間	1998年3月27日～2008年2月11日
変換ファイル数合計	1098

表8 NICT 太陽地球環境予報の年次別 RDF データ変換実績(2007 年 2 月 12 日現在)

発行年次	RDFファイル変換数
1998	58
1999	74
2000	111
2001	159
2002	131
2003	135
2004	102
2005	140
2006	103
2007	74
2008	11

表9 各情報の収集済み RDF データサイズと RDF データステートメント数

RDF変換対象	RDFファイルデータサイズ(単位:KB)	RDFステートメント数
太陽地球系物理観測データファイル	487,628	8,973,692
NICT太陽地球環境予報	11,736	83,866
NICT専門用語情報	55	720
RR2002データ	972	11,244
理科年表データ	3,020	50,214
es-BANKデータ	21,100	296,481
学術創成イベント情報	598	10,033
STARSイベント情報	100,691	974,741
合計	625,800	10,400,991

表10 RDF データベース運用計算機環境

OS	Microsoft Windows Server 2003 Enterprise Edition
CPU	Intel Xeon X3210 2.13GHz (Dual Core) × 4
RAM	8 GB
Disk	441 GB
DBMS	Sesame 1.2.7

5.2 RDF データベースの試験運用

5.2.1 SeRQL クエリを用いた検索

本研究で実装した Sesame による RDF データベースでは、RDF クエリ言語として SeRQL (Sesame RDF Query Language) と、RDQL (A Query Language for RDF) をサポートしている [30] [31]。RDF クエリ言語の構文は SQL クエリ言語の構文と似ており、SELECT 文、FROM 文、WHERE 文を用いてデータベースの検索や操作を行う。SeRQL は Sesame 独自のクエリ言語だが、RDQL の機能をほぼ包括しており、Sesame による RDF データベースと親和性が高いため、本研究では SeRQL を用いた。

SeRQL クエリを用いた検索例を図 17 と図 18 に示す。図 17 に示す SeRQL クエリは、太陽地球環境予報の太陽活動記事内で“X クラス”、“フレア”、“発生”のキーワード記述が含まれる時の発行日時を取得し、その発行日時を観測開始時間と観測終了時間で挟んでいる観測データファイルのファイル URL を取得するクエリである。図 17 の SeRQL クエリを実行した結果、表 11 に示す 7 つの太陽活動記事と、表 12 に示す 8 つのデータサ

イトが公開している 23 種類のミッションから計 606 個の観測データファイル URL を取得できた。

図 18 に示す SeRQL クエリは、RR2002 データのコンポーネント名とコンポーネントのデータタイプ、入力データ、観測日を取得し、そのデータが観測された地点の国情報を取得する。次に、理科年表地学部の“おもな国と属領の面積・人口・首都”情報より、データが主集された国の総人口数や人口密度、出生率、死亡率を取得するクエリである。図 18 のクエリを実行した結果、表 13 に示す 17 種類のコンポーネント情報とそれぞれのコンポーネントの入力データ (合計 510 個) と、表 14 に示す 3 つの国の人口・人口密度・出生率・死亡率の情報を取得できた。

5.2.2 Sesame を用いたステートメント関係の動的検索

RDF の基本構造は主語、述語、目的語の 3 つの要素により構成され、リソースはそれら 3 要素のいずれかに該当し、リテラルは必ず目的語に該当する。Sesame を用いた RDF データベースでは、SeRQL クエリによって得られた検索結果を

表 11 図 17 の SeRQL クエリ実行結果 (発行日時とキーワードを含む太陽活動記事)

発行日時	キーワードを含む太陽活動記事
1998-05-08T16:00:00	活動領域 #8214 は M/X クラスの大きなフレア… (省略)
1998-08-25T09:40:00	本日、午前 7 時 (JST) に X クラスの太陽フレアが発生… (省略)
1999-08-06T15:20:00	活動領域 #8645、#8647、#8649 が M クラスのフレアを… (省略)
2003-03-19T15:10:00	3月17日後半から活動領域 #314 が活発化しています… (省略)
2003-03-21T15:10:00	活動領域 #314 が 17 日後半から活動を開始しました… (省略)
2006-12-15T16:30:00	活動領域 930 における大規模フレアの発生により非常に… (省略)
2006-12-22T16:00:00	太陽活動は 17 日にやや活発な状態となった他は、概ね静穏な… (省略)

表 12 図 17 の SeRQL クエリ実行結果 (取得した観測データファイルのホスト名とミッション名、データファイル数)

ホスト名	ミッション名	観測データファイル数
DARTS	Geotail	26
	Cluster	71
	Reimei	4
3DWEB	NICT 3D Global MHD simulation	6
STELAB	OMTI	77
CIE	Akebono	7
CDAWeb	ACE	59
	Canopus	6
	DARN	11
	IMP8	8
	Interball	12
	SOHO	2
	FAST	5
	WIND	54
	Geotail	88
	LANL	26
	POLAR	39
	IMAGE	2
	OMNI2	15
	Cluster	10
NOAA	GOES	28
SIDC	Sun Spot Number	4
BERKELEY	THEMIS	46
合計: 8ヶ所	合計: 23 種類	合計: 606 個

```
SELECT DISTINCT SolarActivity, JSTtime, JulianTime, FileURL
FROM [stpf:root] nict:stpf:hasArticle [article],
[article] nict:stpf:hasSolarActivity [solar],
[solar] nict:stpf:description [SolarActivity],
[stpf:root] nict:stpf:hasDate [date], [date] nict:stpf:hasJulian [julian],
[julian] nict:stpf:JulianTime [JulianTime],
[date] nict:stpf:hasJST [JST], [JST] nict:stpf:JSTTime [JSTTime],
[stpf:root] sw:st [StartTime], [stpf:root] sw:et [EndTime],
[stpf:root] sw:mission [Mission], [stpf:root] rss:link [FileURL]
WHERE SolarActivity LIKE "*Xクラス*" AND SolarActivity LIKE "*フレア*"
AND SolarActivity LIKE "*発生*"
AND StartTime < JulianTime AND EndTime > JulianTime
```

図 17 太陽地球環境予報の太陽活動記事内の特定キーワードから観測データファイルのファイル URL を取得する SeRQL クエリ

```
SELECT DISTINCT ComponentName, CountryName, Data Type, Unit, MaxValue, MinValue,
Note, InputData, ObsDate, Population, PopulationDensity, BirthRate, DeadRate
FROM [ComponentID] ns:year:ComponentName [ComponentName],
[ComponentID] ns:year:Data Type [Data Type],
[ComponentID] ns:year:Unit [Unit], [ComponentID] ns:year:Max [MaxValue],
[ComponentID] ns:year:Min [MinValue], [ComponentID] ns:year:Note [Note],
[DataID] ns:year:ComponentID [ComponentID], [DataID] ns:year:InputData [InputData],
[DataID] ns:year:ObsDate [ObsDate], [DataID] ns:year:LocationID [LocationID],
[LocationID] ns:year:CountryID [CountryID],
[CountryID] ns:year:CountryName [CountryName], [CountryID] ns:year:Population [Population],
[CountryID] ns:year:PopulationDensity [PopulationDensity],
[CountryID] ns:year:BirthRate [BirthRate], [CountryID] ns:year:DeadRate [DeadRate]
```

図 18 RR2002 データのコンポーネント情報から国情報を取得し、理科年表からその国の人口に関する情報を取得する SeRQL クエリ

表13 図18のSeRQL クエリ実行結果 (RR2002のコンポーネント情報)

コンポーネント名	データタイプ	入力データ数
TBT resistance 200	bacteria	20
Estron(sediment)	endocr disrupt	14
Hopanes	POPs	45
Nonylphenol(water)	endocr disrupt	10
coprostanol	markers	47
TBT resistance 10	bacteria	41
MP/P ratio	markers	46
TBT resistance 100 pct	bacteria	21
TBT resistance 100	bacteria	22
Liner Alkyl Benzenes markers 54	bacteria	54
Polycyclic Aromatic Hydrocarbons	POPs	46
TBT resistance 10 pct	bacteria	36
Nonylphenol	endocr disrupt	26
TBT resistance 0	bacteria	42
TBT resistance 200 pct	bacteria	20
Estron(water)	endocr disrupt	5
Bisphenol A	endocr disrupt	10
Bisphenol A(water)	endocr disrupt	5
合計:17種類	合計:4種類	合計:510個

表14 図18のSeRQL クエリ実行結果(理科年表地学部のおもな国と属領の面積・人口・首都情報の一部)

国名	人口(万人)	人口密度	出生率(‰)	死亡率(‰)
ベトナム	7,655	231	28.9	7.9
カンボジア	1,052	58	38.2	14.1
ラオス	519	22	45.2	15.2

利用して、ステートメント関係の動的検索を行うことが可能である。SeRQL クエリを実行するとリソースとリテラルで表現された検索結果が得られる。ステートメント関係の動的検索では、得られたリソースもしくはリテラルのいずれか一つを選択すると、そのリソースを主語とするステートメント関係一覧、そのリソースを述語とするステートメント関係一覧、そのリソースを目的語とするステートメント関係一覧を動的に取得することが可能である。

図19～図24にステートメント関係の動的検索を行った時の画面遷移の様子を示す。

図19ではまず、画面上半分のSeRQL クエリ入力部分に5.2.1の図18で示したSeRQL クエリを入力しており、クエリ結果を画面下半分のクエリ結果出力部分に示している。ここで、クエリ結果の中から“カンボジア”リテラルを選択すると、図20に遷移する。図20では、“カンボジア”リテラルを主語とするステートメント関係一覧、“カンボジア”リテラルを述語とするステートメント関係一覧、“カンボジア”リテラルを目的語とするステートメント関係一覧を出力している。リテラルは目的語にしかかなりえないため図20のような出

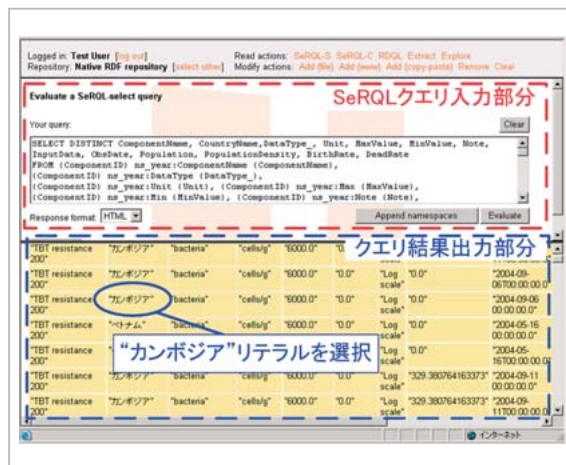


図19 ステートメント関係の動的検索画面1 (SeRQL クエリ入力部分とクエリ結果出力部分)



図20 ステートメント関係の動的検索画面2 (国を識別するための国IDリソースを選択)

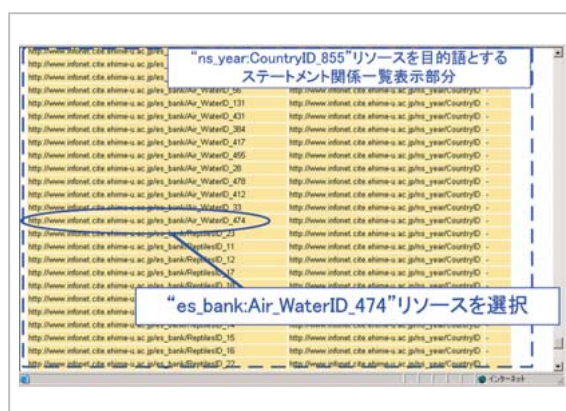


図21 ステートメント関係の動的検索画面3 (es_BANK データの大気・水データを識別するためのIDリソースを選択)

力結果となっている。ここで、国を識別するための国IDである“ns_year:CountryID_855”リソースを選択すると、図21に遷移する。図21では、

“ns_year:CountryID_855” リソースを目的語とするステートメント関係一覧を示している。ここで、es_BANK データの大気・水データを識別するための ID である“es_BANK:Air_WaterID_474” リソースを選択すると図 22 に遷移する。図 22 では、“es_BANK:Air_WaterID_474” リソースを主語とするステートメント関係一覧を表示している。ここで、採取された日時を表現する“es_BANK:SamplingDateTime” 述語によって示される“2004-

09-10T00:00:00” リテラルを選択すると図 23 に遷移する。図 23 では、“2004-09-10T00:00:00” リテラルを目的語とするステートメント関係一覧を示している。ここで、観測データファイルを識別する URI リソースを選択すると図 24 に遷移する。図 24 では、観測データファイルを識別する URI リソースを主語とするステートメント関係一覧を出力している。

このように、Sesame を用いた RDF データベースでは、出力されたリソースやリテラルを次々と選択していくことでステートメント関係を動的に辿っていくことが可能である。

5.3 考察

表 15 に、図 17 と図 18 に示した SeRQL クエリの平均処理時間(対象クエリを 10 回試行したときの平均処理時間)と検索の対象になったステートメント数を示す。

図 17 のクエリと同様の結果を、RDF データベースを用いずに手動で取得する場合について考える。この場合、まず、Web 上で公開されている 1,000 以上の太陽地球環境予報の HTML ファイルの中から特定のキーワードを含む記事を検索する。さらに特定のキーワードを含んだ記事の発行日時を確認し、各データサイトにアクセスをして、その発行日時を観測期間とする観測データファイルの所在を検索する。表 12 に示した 8 つのデータサイトが管理する 23 種類のミッションの観測データファイルの所在を手動で取得することは容易ではない。

本研究では、RDF データを自動収集し RDF データベースを構築するシステムを提案した。本システムにより、上述の作業を 1 度のクエリを実行するだけで機械的に処理することが可能となった。大量に存在する観測データファイルのステートメントを全て検索しているため約 2 分半の処理時間を要したが、手作業による検索処理と比べると高速されたとと言える。

図 18 のクエリと同様の結果を、RDF データ

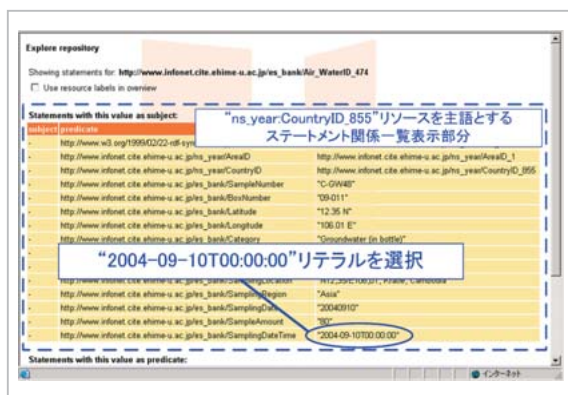


図22 ステートメント関係の動的検索画面 4 (大気・水データを採取した日時を示すリテラルを選択)

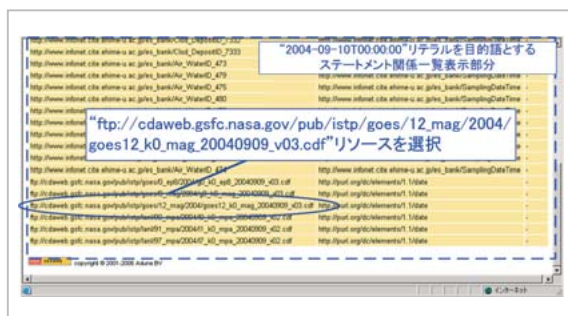


図23 ステートメント関係の動的検索画面 5 (観測データファイルを識別するためのファイル URL リソースを選択)

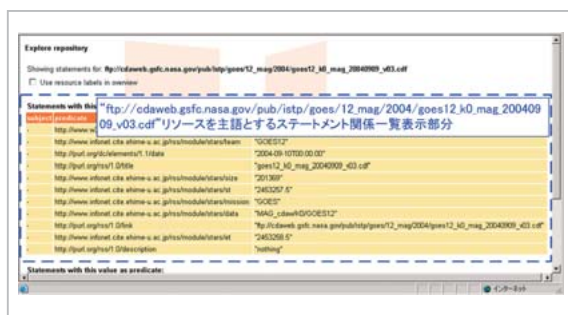


図24 ステートメント関係の動的検索画面 6 (選択した観測データファイルリソースを主語とするステートメント関係一覧表示)

表 15 SeRQL クエリの平均処理時間と検索対象ステートメント数

クエリ例	平均処理時間(単位: ms)	検索対象ステートメント数
図17の場合	155,390.6	9,057,558
図18の場合	252.5	61,458

ベースを用いずに手動で取得する場合、異なる分野の情報である理科年表地学部情報と RR2002 データの関係性を理解しなければならない。RR2002 データではコンポーネントに対してそれぞれ入力データがあり、どこの国の入力データであるかを示す国情報を有している。理科年表地学部情報では、“おもな国と属領の面積・人口・首都”情報があり、これが RR2002 データの国情報と関連を持つと考えられる。このようなステップで異なる分野間の情報の関連を理解し、それぞれの情報公開元からデータを取得する作業は容易ではない。しかし、本研究の技法を用いて異なる分野の情報を収集し、RDF データベースを構築することによって、2つの情報の関係性を機械的に結び合わせる事が可能である。水環境の化学汚染実態を示す RR2002 データに理科年表の人口に関する情報を結び合わせることで、これまで気付かなかった水質情報と出生率や死亡率との因果関係を発見できる可能性がある。

一方、本システムの問題点としては、時刻フォーマットの相違により、本来ならば同じ時間軸上である情報を結び合わせる事ができない点がある。観測データファイルとイベント情報はユリウス日フォーマットで時刻を表現しているのに対し、理科年表や RR2002 データ、es_BANK データは xsd:dateTime フォーマットもしくは独自のフォーマットで時刻を表現しており、これらの情報は時刻によって結び合わせる事ができない。なお、4.1.1 で述べた太陽地球環境予報 HTML ファイルの RDF 変換エンジンでは、独自のフォーマットで示された発行日時情報をユリウス日フォーマット、xsd:dateTime フォーマットの両方で表現して RDF 化している。そのため、太陽地球環境予報に関しては上述したどの情報も時刻によって結び合わせる事が可能である。このように複数の時刻フォーマットに対応した時刻情報を含めておくことで、メタ情報は冗長になるが、時間軸に沿った検索処理に対応できる。

6 結論

セマンティック Web の構築では、RDF 形式で記述されたメタデータのデータベース化が不可欠である。本研究では、愛媛大学総合情報メディア

センターで研究開発が進められてきた RSS 1.0 による太陽地球系物理観測メタデータ自動収集システム [9] を拡張し、RDF に基づいた汎用的メタデータ自動収集システムの構築を行った。

本システムでは、観測データファイルのメタデータだけでなく、NICT が Web で公開している太陽地球環境予報、STARS を使って登録された研究者による各種現象のイベント情報、NICT が定義した太陽地球環境用語集、国立天文台が発行している理科年表情報など、太陽地球系科学分野に関連する様々な情報を統合した。そのために、HTML 形式から RDF 形式に変換するプログラムおよびリレーショナルデータベースのデータから RDF 形式にマッピングするプログラムを実装した。

本研究で提案する RDF データで使用するメタデータ語彙に関しては、既存の標準・共通語彙のみでは太陽地球系科学分野独自の表現に対応できないため、本研究で独自の語彙群を定義した。この語彙の定義には、W3C が定めたセマンティック Web 標準フォーマットである RDFS と OWL を用いた。本研究の RDF データを利用する場合、提案した RDFS/OWL を共有することでメタデータ語彙を統一することが可能である。また、第三者の RDF データと本研究で提案する RDF データとを統合する場合、互いの RDFS/OWL において、語彙名と意味をマッピングすることにより、語彙の差異を吸収することが可能である。

太陽地球系科学分野のセマンティック Web を地球科学分野全体のセマンティック Web に拡張する試行として、本研究では愛媛大学沿岸環境センターが構築している es-BANK (生物試料バンク) データ、インドシナ半島における水環境の化学汚染実態の解明と汚染除去技術の開発を目的とした RR2002 データ、理科年表(暦部・天文部・気象部・物理/化学部・生物部・環境部)のデジタル情報を対象とし、必要なメタデータ語彙を定義した。これまではネットワーク上で独自公開されてきた地球観測データファイル、太陽地球環境予報、研究者イベント情報、地球環境情報などは、人間がそれらのデータ間にある時間・空間・概念・現象などの関係性を理解したうえで個別に処理するしかなかった。しかし、共通のメタデータ語彙を用いて RDF 化し RDF データベースに集

約することで、計算機がデータ間の関係性を理解しながら処理可能となった。

本研究で実装した RDF データベースには、RDF データ収集技法を用いて約 1 千万のステートメント(主語・述語・目的語のセット)を格納した。RDF データベースに対してクエリ文を用いて検索する方法と、ステートメントの関係性を一つずつ辿って検索する方法を実装し、それらの処理時間・検索結果の実用性と有用性を評価した。

今後は、太陽地球系科学分野や地球環境分野以外の多くの地球科学分野の観測データを RDF 化することにより、分野横断的な全地球科学セマンティック Web の実現が期待される。そのためには、各分野のドメインエキスパートがメタデータ語彙の定義を的確に行い、さらにこれらを結合す

ることにより、本研究で構築した RDFS/OWL を整理・拡張する必要がある。また、これまでの RDF (データベース) の多くはテストベッドとして実験的に構築されてきた。今後は、学際的研究を可能とするセマンティック Web 環境の実現により、eGY の目的である、人類の生存に影響する複雑な地球システムの理解のための新たな知識発見や情報統合に貢献することが期待される。

謝辞

本論文は、愛媛大学理工学研究科電子情報システムコース 2007 年度久保卓也君の修士論文をもとに作成したものです。同君に感謝します。

参考文献

- 1 南極条約 (Antarctic Treaty), 外務省, http://www.mofa.go.jp/mofaj/gaiko/kankyo/jyoyaku/s_pole.html, 2007.
- 2 EGY: The Electronic Geophysical Year, <http://www.egy.org/>, 2005.
- 3 W3C, The SemanticWeb, <http://www.w3.org/2001/sw/>, 2001.
- 4 大石雅寿, 白崎裕治, 他, バーチャル天文台を実現する分散データベース・計算資源の国際連携機構, <http://www.ieice.org/iss/de/DEWS/DEWS2005/procs/papers/4A-o3.pdf>, DEWS2005, 2005.
- 5 本田敏志, 白崎裕治, 他, 天文学連携データベースシステム (ヴァーチャル天文台) の開発, <http://www.ieice.org/~de/DEWS/DEWS2005/procs/papers/1D-d2.pdf>, DEWS2005, 2005.
- 6 村田健史, “国際太陽地球系物理観測の広域分散データベースシステム”, 電子情報通信学会論文誌(B), Vol.J86-B, No.7, pp.1331-1343, Jul. 2003.
- 7 村田健史, 岡田雅樹, 阿部文雄, 荒木徹, 松本紘, “太陽地球系物理観測の分散メタデータベースの設計と評価”, 電子情報通信学会論文誌(B), Vol.J87-B, No.2, pp.309-313, Dec. 2004.
- 8 木村映善, 村田健史, “RSS/RDF を利用した太陽地球系物理観測データのメタデータ配信の検討”, 情報処理学会論文誌, Vol.47, No.4, pp.1051-1062, Apr. 2005.
- 9 RSS 1.0 を利用した科学衛星・地上観測データのメタデータ自動収集, 石倉諭, 村田健史, 久保卓也, 木村映善, 山本和憲, 篠原育, 電子情報通信学会論文誌(B), Vol.J91-B, No.4, pp.499-509, Apr. 2008.
- 10 研谷紀夫, 馬場章, “文化資源オントロジの構築とその活用”, 情報知識学会誌, Vol.17, No.2, pp.129-134, Feb. 2007.
- 11 NetCDF, <http://www.unidata.ucar.edu/software/netcdf/>, 2006.
- 12 The FITS Support Office, <http://fits.gsfc.nasa.gov/>, 2006.
- 13 The Common Data Format (CDF), <http://cdf.gsfc.nasa.gov/>, 2006.
- 14 太陽地球環境予報, http://crlhir.nict.go.jp/forecast/stef_latest.html, 2008.
- 15 STE 現象用語集, <http://www2.nict.go.jp/y/y223/sept/swcenter/yogo.html>, 2006.
- 16 NICT リアルタイム MHD シミュレーション, <http://www2.nict.go.jp/y/y223/simulation/realtime/>

- homej.html, 2007.
- 17 Nobeyama Radio Observatory: Solar Group, <http://solar.nro.nao.ac.jp/>, 2008.
 - 18 理科年表オフィシャルサイト, 国立天文台, <http://www.rikanenpyo.jp/>, 2008.
 - 19 STARS Web Page, 情報通信研究機構電磁波計測研究センター宇宙環境計測グループ, <http://www.seg.nict.go.jp>, 2009.
 - 20 DCMI: Dublin Core Metadata initiative, Dublin Core Metadata Element Set Version 1.1, <http://dublincore.org/documents/dces/>, 2006.
 - 21 FOAF: The 'friend of a friend' project, <http://rdfweb.org/foaf/>, 2005.
 - 22 RDF Site Summary (RSS) 1.0, <http://web.resource.org/rss/1.0/>, 2000.
 - 23 W3C, Semantic Web Layer Cake, <http://www.w3.org/2004/Talks/0412-RDF-functions/slide4-0.html>, 2004.
 - 24 久保卓也, RDF を用いたデータ収集技法による太陽地球系科学セマンティック Web 構築の試み, 愛媛大学理工学研究科修士論文, 2007.
 - 25 es-BANK, 愛媛大学沿岸環境科学研究センター生物環境試料バンク, <http://www.ehime-u.ac.jp/~cmes/esbank/esbank.htm>, 2003.
 - 26 Research Revolution 2002 (RR2002), 新世紀重点研究創世プラン, <http://kyousei.aesto.or.jp/>
 - 27 W3C, GRDDL: Gleaning Resource Descriptions from Dialects of Languages, <http://www.w3.org/TR/grddl/>, 2007.
 - 28 D2R (Database to RDF Mapping Language and Processor), <http://sites.wiwiw.fu-berlin.de/suhl/bizer/d2rmap/D2Rmap.htm>, 2005.
 - 29 Open source RDF Database framework Sesame, <http://www.openrdf.org/index.jsp>, 2007.
 - 30 SeRQL: Sesame RDF Query Language, <http://www.openrdf.org/doc/sesame/users/ch06.html>, 2003.
 - 31 RDQL: A Query Language for RDQL, <http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/>, 2004.



むら た け し
村田健史

電磁波計測研究センター宇宙環境計測
グループグループリーダー 博士(工
学) 宇宙情報工学、福祉情報工学

き むら えい ぜん
木村映善

愛媛大学大学院医学系研究科准教授
博士(医学)
医療情報学、医療経済学



やま もと かず のり
山本和憲

愛媛大学工学部研究生
Virtual observatory、セマンティッ
クWeb



むら いち
村田 一

電磁波計測研究センター宇宙環境計測
グループ研究マネージャー 博士(理学)
太陽地球系物理学、宇宙天気

かとうひさお
加藤久雄

電磁波計測研究センター宇宙環境計測
グループ主任研究員



もりかわしゅう
森川靖大

電磁波計測研究センター宇宙環境計測
グループ有期技術員
惑星大気モデリング、地球流体力学、
グリッドコンピューティング



さいつこうし
是津耕司

知識創成コミュニケーション研究セン
ター知識処理グループ主任研究員
博士(情報学)
情報検索、データベース、Web情報
システム



きくち たか
木俣 豊

知識創成コミュニケーション研究セン
ター知識処理グループグループリー
ダー 博士(工学)
情報検索、データベース、ユビキタス
コンピューティング



しもしょうしん
下條真司

上席研究員 工学博士
マルチメディア応用システム、Peer-
to-Peer、インターネット、ユビキタ
スネットワーク、グリッド技術