

## 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と太陽地球系科学セマンティック Web の試み

### Automatic Collection System of Metadata based on RDF and Development of Solar-Terrestrial Science Data Semantic We

# 村田 健史 [1]; 久保 卓也 [2]; 木村 映善 [3]; 山本 和憲 [4]; 石井 守 [5]

# Ken T. Murata[1]; Takuya Kubo[2]; Eizen Kimura[3]; Kazunori Yamamoto[4]; Mamoru Ishii[5]

[1] 愛大・メディアセンター; [2] 愛媛大・理工; [3] 愛媛大 CITE; [4] 愛媛大・理工; [5] 情報通信研究機構

[1] CITE, Ehime University; [2] Ehime Univ; [3] CITE, Ehime Univ.; [4] Ehime Univ.; [5] NICT

<http://www.infonet.cite.ehime-u.ac.jp>

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。国際地球観測年 (IGY) では多岐にわたる観測が世界各国の協力で実施された。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、2007 から 2008 年を国際デジタル地球年 (eGY) とする計画が発足された。ネットワークによる高速交換や計算機による高速処理が可能なデジタル情報の利点を活用し、人類の生存に影響する複雑な地球システムを理解し、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 上に存在する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。現在のネットワークを介したデータ交換の問題点として、対象のデジタル情報が何を意味するのかコンピュータが自動的に検知する術がなく、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とが挙げられる。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、情報を記述する際にデータに意味を記述したメタ情報を付与することで複雑で精度の高い検索が可能になり、特定の種類の情報を収集して活用することができる。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。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、それらのデータを提供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、愛媛大学総合情報メディアセンターでは、太陽地球系物理観測データ解析参照システム (STARS) の開発を行ってきた。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に集約することで、分散し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。また、RSS1.0 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するシステムを実現している。本研究では、RSS1.0 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 の構築に不可欠な RDF 形式で記述されたメタデータを自動的にデータベース化するため、RDF に基づいた汎用的メタデータ自動収集システムの研究開発を行った。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 上でデータベースに登録された研究者による各種の現象のイベント情報、NICT が定義した STE 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ために、HTML 形式から RDF 形式に変換するプログラム、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開発した。これにより、RSS によるメタデータ自動収集システムを拡張した RDF に基づいた汎用的メタデータ自動収集システムを構築した。

本研究で提案する RDF データ内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。ネットワーク上に公開されている観測データファイルと NICT 太陽地球環境予報、STP イベント情報などは、人間がそれらのデータ間にある時間・空間・概念・現象といった関係性を理解してそれぞれ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、それぞれのデータ間の関係性を計算機が理解しながら処理可能となった。実装した RDF データベースには、RDF に基づいた汎用的メタデータ自動収集システムによって約 1 千万のステートメント (主語・述語・目的語のセット) が格納された。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示し、それらの処理時間・検索結果の実用性と有用性について評価する。