DSpaceを用いた超高層物理学のためのメタデータ・データベースの構築

河野　貴久† 小山　幸伸 堀　智昭 阿部　修司 吉田　大紀
林　寛生 新堀　淳樹 田中　良昌 鍵谷　将人 上野　悟
金田　直樹 田所　裕康

† 名古屋大学太陽地球環境研究所　〒 464–8601 愛知県名古屋市千種区不老町
E-mail: †kouno@stelab.nagoya-u.ac.jp

あらまし　超高層物理学の地上観測データは，風速，オーロラ，地磁気，太陽活動など多種多様であり，様々な機関によって分散管理されている．我々は，これら様々な観測データを組み合わせた総合的な研究を推進するために，超高層大気長期変動の全球地上ネットワーク観測・研究プロジェクト (IUGONET) において，分散管理された観測データに対するアクセシビリティを向上させるため，超高層物理学の観測データに付随したメタデータ・データベースを構築中である．我々は，このメタデータ・データベースの構築にリポジトリソフトウェアである DSpace を超高層物理学における観測データのメタデータ用に改良した．本論文では，超高層物理学のためのメタデータ・データベースの構築とそのメタデータ登録性能評価について報告する．

キーワード　メタデータ・データベース，観測データ，DSpace

Development of Metadata Database for Upper Atmosphere Observation by using DSpace

Takahisa KOUNO†, Yukinobu KOYAMA, Tomoaki HORI, Shuji ABE, Daiki YOSHIDA,
Hiroo HAYASHI, Atsuki SHINBORI, Yoshimasa TANAKA, Masato KAGITANI,
Satoru UENO, Naoki KANEDA, and Hiroyasu TADOKORO

† Solar–Terrestrial Environment Laboratory, Nagoya University
E-mail: †kouno@stelab.nagoya-u.ac.jp

1. はじめに

近年，地球温暖化の現象解明するための地球環境分野の研究が注目を集めている．そして，生物の多くが住む対流圏のみならず，地球大気圏の上層，すなわち超高層大気やそこでの様々な周期の長期変動と地球温暖化との関係が研究されつつある [1]．超高層大気中に見られるグローバルな諸現象は，多様なプロセスが複雑に絡み合った結果として観測される．そのため，超高層大気における長期変動のメカニズムを解明するためには，全地球規模の地上観測ネットワークにおける様々な観測データ，例えば，風速，オーロラ，地磁気，太陽活動等を組み合わせた総合的な解析が必要になる．超高層物理学の地上観測データは，観測を行った研究機関や観測プロジェクト毎にデータベース化され公開されているが，それらのデータベースを横断的に検索する手段が無く，個別の観測・研究に関係する特定分野の研究者のみに利用されることが殆どであった．このため，様々な観測データを組み合わせた総合的な解析を行うことは困難であった．

著者一同は，2009 年度より 6ヶ年計画として始まった「超高層大気長期変動の全球地上観測・研究 (IUGONET: Inter-university Upper atmosphere Global Observation NETwork)」プロジェクト [2] において，超高層大気長期変動に関する地上観測データのメタデータ・データベースを構築している．IUGONET プロジェクトは，国立極地研究所宙空圏研究グループ，東北大学大学院理学研究科地球物理学専攻太陽惑星空間物理学講座並びに東北大学惑星プラズマ・大気研究センター，名古屋大学太陽地球環境研究所，京都大学生存圏研究所，京都大学理学研究科附属地磁気世界資料解析センター，京都大学理学

研究科附属天文台および九州大学宙空環境研究センターの 5 機関 7 組織が参加している．これらの機関は，世界中の様々な地域・高度領域から各種装置を用いて超高層大気の観測を数十年にもわたり行っており，多種多様な観測データを蓄積している．IUGONET プロジェクトの目的の一つは，観測データの所在情報に代表される様々なメタデータをデータベース化することにより，複数の機関によって分散管理されている多種多様な観測データに対するアクセシビリティを向上させ，観測データの利用を促進することである．これにより，各研究機関が所有する各種観測データを有機的に組み合わせた分野横断的な研究の萌芽や，様々な現象が複雑に絡み合う超高層大気の長期変動のメカニズムを解明することで地球温暖化等の現象解明と予測に貢献することである．

我々は，超高層物理学のためのメタデータ・データベースを構築するにあたり，全てを独自に開発するのではなく，既存のソフトウェアを組み合わせ，一部改良することでシステムを構築することにした．本論文では，超高層物理学の観測データのメタデータ・データベースを構築する上で行った調査と改良した点について報告する．

2. IUGONET メタデータ・データベースシステム

2.1 メタデータ・フォーマット

IUGONET プロジェクトでは，超高層物理学の地上観測データのメタデータのフォーマットに Spase Physics Archive Search and Extract(SPASE) メタデータ・フォーマット [3] [4] を採用した．SPASE メタデータ・フォーマットは，NASA の Virtual Magnetospheric Observation や Virtual Heliospheric Observation などアメリカを中心とした多くのプロジェクトで採用されているメタデータ・フォーマットである [5]．我々は，この SPASE メタデータ・フォーマットを超高層物理学の観測データ向けに拡張した．

SPASE メタデータ・フォーマットは，観測データのメタデータを観測機関，観測装置，研究代表者，観測データセット，観測データファイルなどの 12 種類のリソースタイプに分けて，XML で記述する．各メタデータはユニークにつけられた ResourceID を持ち，異なるリソースタイプであっても，関連するメタデータ同士が相互に参照できる [6]．代表的なメタデータ要素を挙げると，観測データセットにおける観測開始時刻や終了時刻，観測データファイルにおける観測データの所在情報 (URL)，観測所における装置が置かれている緯度・経度等である．

2.2 メタデータ・データベース

IUGONET プロジェクトでは，メタデータ・データベースのシステム構築にあたり，機能面や運用面を検討した結果，学術機関リポジトリとして多く採用されている [7] DSpace [8] を基盤となるデータベースソフトウェアとして用いることにした．しかしながら，DSpace は，デジタルコンテンツに加え，例えば，著者名，タイトル，登録時刻などの書誌的な Dublin Core [9] 形式のメタデータを扱うリポジトリソフトウェアであるため，標準では超高層物理学における観測データのメタデータを扱うことは困難である．しかしながら，DSpace はオープンソースソフトウェアとして公開されていて改良し利用することができることと，国立情報学研究所で行われている学術機関リポジトリ構築連携支援事業 [10] において多くの大学の機関データベースとして多く採用され構築・運用に関する情報は豊富にあることから，DSpace を超高層物理学の観測データ向けに改良して用いることにした．

3. DSpace の改良と評価

IUGONET メタデータ・データベースが求める主な機能として，1. 時間・地理空間の範囲検索，2. 外部提供インターフェース，3. 登録システムの開発，4. メタデータの追加登録速度の評価，の 4 点について改良・調査した結果を報告する．

3.1 時間・地理空間の範囲検索

観測データのメタデータは観測開始時刻と観測終了時刻を含む．この場合，観測データはこれら二つの時刻の間に存在する．超高層物理の研究者が求める検索では，指定した二つの時刻の間に存在する観測データに紐付いたメタデータを検索結果として抽出することが求められる．DSpace は，オープンなデジタルリポジトリを構築するための機関リポジトリとして開発されており，Dublin Core メタデータフォーマットを標準でサポートしている [11]．Dublin Core メタデータフォーマットは作成時刻や登録時刻を記述出来る．そのため，DSpace は”時点”に関する検索は可能であるが，”二つの時点に挟まれた時間帯”の検索はできない．この時間の範囲検索を実現するために，我々は DSpace の検索クエリーをカスタマイズした．

観測開始時刻を start_time，終了時刻を end_time とし，時間検索クエリーの二つの時刻を from_time と to_time とした場合，時間の範囲検索のクエリーは

```
( start_time:[from_time TO to_time] OR
  end_time:[from_time TO to_time] )
OR
( start_time:[00000101000000 TO from_time] AND
  end_time:[to_time TO 99991231235959] )
```

で記述される．時刻は文字列としてデータベースに登録しており，00000101000000 と 99991231235959 は時刻の最小限度と最大限度を表す．ここで，範囲検索は DSpace の検索エンジンである Apache Lucene [12] の Range 検索 (e.g. field:[A TO Z]) を用いた．

DSpace は地図検索機能が無いため，時間と同様に検索クエリーをカスタマイズすることで地理空間の範囲検索を実装した．IUGONET で扱う超高層物理学のメタデータは，観測装置の設置場所や測定を行っている範囲 (矩形) の情報を含む．観測装置の設置場所の検索は地点の検索である．この検索は，緯度と経度をそれぞれ Range 検索し AND を取ることで実現した．測定を行っている範囲に関しては，時間の範囲検索と同様である．緯度の範囲と経度の範囲をそれぞれ検索し AND を取ることで地図上の矩形範囲の中で観測されたデータを検索可能にした．

3.2 外部提供インターフェース

IUGONET プロジェクトでは，メタデータ・データベースの構築に加えて，可視化・解析ソフトウェアを開発している．IUGONET メタデータ・データベースを利用する方法として，WEB ブラウザを用いた検索方法に加え，外部提供インターフェースを用意することで可視化・解析ソフトウェアからの利用も想定している．

具体的には，可視化・解析する対象である観測データファイルの所在情報などをメタデータ・データベースから検索・取得できるようにすることで，可視化・解析ソフトウェア側から観測データへのアクセシビリティを良くすることを目指している．図 1 にメタデータ・データベースと可視化・解析ソフトウェアとの連携を示す．メタデータ・データベースにより取得した観測データの所在アドレスをもとに実データをダウンロードして，可視化・解析を行う連携例である．

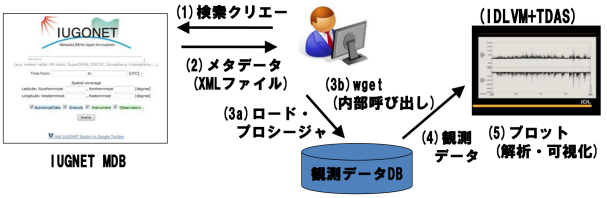

図 1　解析ソフトウェアからメタデータ・データベースの利用例．

3.3 登録システムの開発

IUGONET メタデータ・データベースはプロジェクト終了後も観測データのメタデータの登録が行われ利用されていく予定である．データベースには複数の機関や多くのプロジェクトによって観測されているデータのメタデータを登録する必要がある．超高層物理学における典型的な観測データは 1 日 1 ファイル生成され，観測データ生成後速やかにメタデータ・データベースに登録されることが望まれる．さらに，IUGONET プロジェクトでは過去に観測されたデータの登録も行うため，1 度に大量の件数を登録することもある．このようなことから，メタデータ・データベースに登録するまでの作業と時間を短縮するためにデータベースへのメタデータ登録を自動化するシステムを開発する必要がある．

我々はバージョン管理システムである Git を用いたメタデータの登録システムを開発している．Git は主にソースコードのバージョン管理に使われるソフトウェアであるが，このテキストファイルを対象にした履歴管理機能を，メタデータファイル (XML ファイル) の管理に応用することにした．データ提供機関が作成したメタデータファイルは，一度 Git で管理される．そこで，メタデータの妥当性が検証された後，DSpace にメタデータが自動登録される．開発した登録システムを使うことで以下の利点がある．

- ユーザは，DSpace の登録コマンドを使わずに，通常のディレクトリ・ファイルシステムに近い形でメタデータを登録することができる．
- 更新履歴情報を管理することでメタデータの信頼性を高めることができる．信頼性とは，メタデータを誰がいつ作成し，間違ったメタデータがいつ修正されたか，修正前はどのような情報が公開されていたかなどが分かることを示す．
- Git を用いてメタデータのバックアップや履歴管理を行うため，メタデータ・データベースは検索システムとして簡素化できる．

DSpace へのメタデータの登録の流れを図 2 に示す．メタデータ提供者が作成した XML 形式のメタデータファイルは，まず Git リポジトリに登録される．Git リポジトリではメタデータファイルの変更履歴を管理しているため，その履歴情報から DSpace に登録するファイル (新規提出・更新・削除されたメタデータファイル) のリストを自動生成することが可能である．そして，登録対象のメタデータファイルは，DSpace 登録フォーマットに変換後，DSpace へ登録される．DSpace への登録に先立ち，削除リストのファイルを DSpace から削除する．最後に，DSpace に登録したログから必要な情報を再び Git に登録し管理する．コミット ID は，Git リポジトリで変更履歴を表す ID であり，この ID を保存することでどの履歴までの変更を DSpace に登録したかがわかるようになっている．アイテム ID は DSpace に登録されているメタデータファイルの内部 ID で，更新・削除の時に利用する．

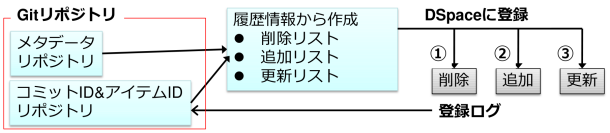

図 2　Git リポジトリから DSpace へのメタデータ登録の流れ．削除，追加，更新の順番で，メタデータを DSpace に登録する．

3.4 メタデータの追加登録速度

我々は，様々な観測データを日々収集しているため，メタデータ・データベースへのメタデータ登録速度は，1 日に登録可能な数が 1 日に作成されるメタデータ数より多いことが求められる．さらに，すでに終了した過去の観測データの登録や観測プロジェクト終了時にまとめて登録されるような一度に 10 万件以上のメタデータの新規・更新登録を扱う可能性がある．そして，IUGONET プロジェクトでは，将来的に 100 万件以上のメタデータを登録・管理することを想定している．このようなことから，DSpace のメタデータ登録速度の性能を調査することは重要である．

本論文では，IUGONET 向けにカスタマイズした DSpace を用いて，実際のメタデータの登録時間を測定することでメタデータの登録性能を評価した．性能評価に用いたメタデータは観測データファイルに対応するメタデータで，一つのメタデータ中の要素数は約 20 個と少ないが実際の運用では大量に登録する必要がある．次に，メタデータ登録性能の評価環境を表 1 に示す．比較として DSpace 1.6.2 環境と 1.7.0 環境の二つバージョンの登録性能を測定した．DSpace 1.7.0 は 2010 月 12 月 17 日にリリースされた最新版であり，以前のバージョンに比べ，メタデータ登録の性能と登録数に対するスケーラビリ

ティの向上が報告されている [11]．検索エンジンである Lucene は，DSpace に組み込まれているものである．そして，DSpace はそれぞれ IUGONET 向けのカスタマイズをしたものを利用した．IUGONET 向けのカスタマイズをした DSpace とは，SPASE メタデータを登録するために登録フォーマットを追加したことと，SPASE メタデータ要素を詳細検索するために検索フィールドを設定したものである．メタデータ登録時間の測定は，DSpace の import コマンドを用いて 1000 件ずつ追加登録したときに，各 1000 件の登録に掛かった時間を測定した．

表 1　メタデータ登録性能の評価環境

項目	DSpace1.6.2 環境	DSpace1.7.0 環境
CPU	Intel Xeon W5590 3.33GHz × 2	←
Memory	48GB	←
HDD	1TB (Software RAID1)	←
OS	CentOS 5.5 x86_64	←
データベース	PostgreSQL 8.4.5	←
DSpace	1.6.2	1.7.0
Lucene	2.3.0	2.9.3
Java	1.6.0_23-b05	←

メタデータ登録時間測定の結果を図 3 に示す．横軸は DSpace に登録されているメタデータの積算件数であり，縦軸は積算件数が登録された状態からさらに 1000 件のメタデータを追加登録するのに掛かった時間 [秒] を示す．DSpace1.6.2 環境では，既に登録されている件数が増加するに従って線形的に登録時間が長くなる結果が得られた．DSpace1.6.2 環境の場合は，登録時間の事情から 4 万件を登録したところで測定を終了した．登録が時間がこのまま一定増加すると仮定した場合，100 万件が登録された状態において，さらに 1000 件の追加登録をする場合約 7 時間掛かる見積りになる．一方，DSpace1.7.0 環境の場合，同じ 1000 件の追加であっても，2 万件の登録までは約 0.07 秒/件であり，その後登録されている件数が増加した状態でも約 0.08 秒/件で一定の登録速度を維持したままであった．さらに，100 万件まで登録を続け時間を測定したが登録時間に増加傾向は見られずほぼ一定の速度を維持する結果であった．

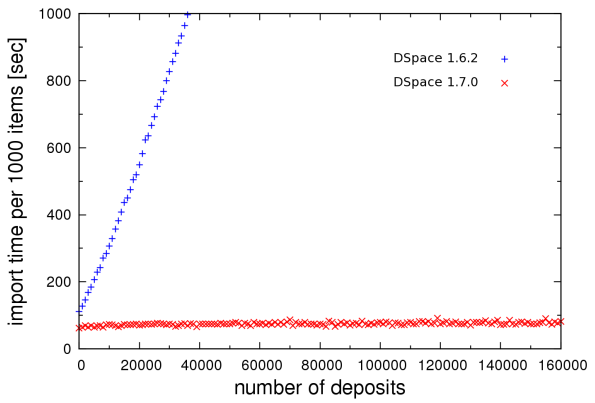

図 3　DSpace1.6.2 環境と DSpace1.7.0 環境におけるメタデータ登録時間の比較．横軸は追加登録時に既に登録されている件数．縦軸は 1000 件を追加登録した時の経過時間．

この結果から，将来 100 万件以上のメタデータを登録することを考慮した場合，DSpace 1.6.2 を使うことは実用的でないことが分かった．そして，現在最新版である DSpace 1.7.0 を使うことにより，大量の登録の場合であっても実用的な時間で行うことができ，地上観測データのメタデータ・データベースを構築することが可能であることがわかった．

4. ま と め

IUGONET プロジェクトは，超高層大気分野における地上観測データに関するメタデータ・データベースを構築中である．観測データのメタデータ・フォーマットは SPASE メタデータ・フォーマット，メタデータ・データベースにはリポジトリソフトウェア DSpace を採用した．DSpace がデフォルトで対応している Dublin Core メタデータ・フォーマットよりも複雑な超高層物理学における地上観測データのメタデータを扱うために，時間・地理空間の範囲検索を DSpace に実装した．さらに，XML で記述された SPASE メタデータファイルを，保管した Git リポジトリから DSpace へシームレスに登録するためのスクリプトを整備したことにより，登録管理を容易にした．そして，DSpace1.7.0 のメタデータ登録性能を評価した結果，超高層物理学のメタデータ・データベースの運用は問題ないことが分かった．

5. 謝　　辞

大学間連携プロジェクト「超高層大気長期変動の全球地上ネットワーク観測・研究」は，文部科学省特別教育研究経費 (研究推進)[平成 21 年度] および特別経費 (プロジェクト分)[平成 22 年度から] の交付を受けて，平成 21 年度より 6ヶ年計画で実施している事業である．

文　　献

[1] Ana G. Elias, Marta Zossi de Artigas, and Blas F. de Haro Barbas, Trends in the solar quiet geomagnetic field variation linked to the Earth's magnetic field secular variation and increasing concentrations of greenhouse gases, Journal of Geophysical Research, Vol. 115, A08316, doi:10.1029/2009JA015136, 2010
[2] IUGONET Web サイト http://www.iugonet.org/
[3] SPASE Web サイト http://www.spase-group.org/
[4] Todd King, James Thieman and D. Aaron Roberts, Spase 2.0: a standard data model for space physics, Earth Science Informatics, 1865-0473
[5] J. R. Thieman, D. A. Roberts, T. A. King, C. C. Harvey, C. H. Perry, and P. J. Richards, SPASE AND THE HELIOPHYSICS VIRTUAL OBSERVATORIES, Data Science Journal, Volume 9, March 6, 2010
[6] SPASE データモデル http://www.spase-group.org /data/doc/ spase-2_2_0.pdf
[7] 機関リポジトリ一覧 http://www.nii.ac.jp/irp/list/
[8] DSpace Web サイト http://www.dspace.org/
[9] Dublin Core Web サイト http://dublincore.org/
[10] 学術機関リポジトリ構築連携支援事業 Web サイト http://www.nii.ac.jp/irp/
[11] DSpace 1.7.0 ドキュメント http://www.dspace.org /1_7_0Documentation/ DSpace-Manual.pdf
[12] Apache Lucene Web サイト http://lucene.apache.org/