

DIAS における地球科学データ公開 GeoScience Data Release in DIAS

絹谷 弘子^{1*}, 清水 敏之², 吉川 正俊², 喜連川 優¹, 小池 俊雄¹

Hiroko Kinutani^{1*}, Toshiyuki SHIMIZU², Masatoshi YOSHIKAWA², Masaru KITSUREGAWA¹, Toshio KOIKE¹

¹ 東京大学, ² 京都大学

¹The University of Tokyo, ²Kyoto University

データ中心科学の時代となり地球科学データを利用する科学者が扱うデータは、衛星観測データやセンサーデータ、あるいはシミュレーションによって導出されたデータなどデータ量も膨大となっている。研究者が収集したデータ、作成したプロダクトデータ分野を超えた利活用を目的として我々のプロジェクト「データ統合・解析システム (DIAS: Data Integration and Analysis System)」は2006年から開発を始めた。DIAS開発の目的は、最先端の情報科学技術と地球環境に関わる様々な科学技術の連携によって、地球観測データや数値モデル、社会経済データを効果的に統合し、情報を融合するデータインフラを構築し、地球環境問題を解決に導く知を創造し、公共的利益を創出することである。DIASの研究者は、水循環、気象、海洋、農業、生物多様性、生態系、情報科学など多分野にわたり150名を超える。

DIAS開発開始から4年が経過し、この間様々な地球科学データが蓄積されている。我々は科学的・社会的に有用な情報への変換、付加価値の創出など利用を促進するためこれらのデータを一般に公開することにした。データ提供者、研究者とシステム開発者が協議を重ね、公開に向けたシステムを構築し、昨年10月データの公開に至った。

DIASが保有するデータは

1. DIAS研究者が作成したデータセット,
2. 前身プロジェクトや関連プロジェクトが作成したデータセット,
3. ミラーデータ (研究用衛星データ, 研究用モデル出力データ, 研究用気象データ) と
4. DIAS研究用ワークデータに分類される。

公開対象は1,2,3である。

データセットの利用対象分野は、農業、生物多様性、気候、災害、生態系、エネルギー、水循環、気象などである。

分野を超えたデータ利用のためにはデータ提供者が保有するデータに関する知識を文書化しデータと共に提供することが重要である。我々はドキュメントセントリックなメタデータ作成システムを構築している。公開対象データセットに対しデータ提供者にこのシステムを利用してデータセットを記述するドキュメントとメタデータの作成を依頼した。ドキュメントとメタデータを日本語と英語の2種類作成し、海外のデータに関する日本語ドキュメントや日本のデータに関する英語ドキュメントの充実を図っている。

データ公開に当たり重点をおいて協議したのは、1. データセット単位の管理と検索、2. 日本語と英語での表記である。何をデータセットとするのがよいかは、実際のデータごとに関係者と協議して決めることにした。その結果 数ファイルで構成されるデータセットから数百万ファイルで構成されるデータセットが共存する。ただし、同一データセット内のファイルについてはデータセット利用規約と参照規約はすべて同一のものとなっている。利用者ごとにアクセス権を与えるのもデータセットを単位としている。またメタデータとドキュメントに加え、検索インタフェースなどすべてのインタフェースを日本語と英語の表示を切り替えられるようにした。

データ公開に必要な機能は、1. DIASが保有するデータの俯瞰、検索、データダウンロードをシームレスに行う機能 2. 一般利用者が利用登録を行い、データ利用制約条項を順守する旨に同意してデータを取得する機能 3. データ、利用者ごとにアクセス権を設定する機能 4. データ提供者へのデータ利用報告を行う機能である。これらの機能を実現するためにユーザ管理システム、DIAS俯瞰・検索システム、アクセス管理システムを開発した。

このシステムは誰でも、<http://dias-dss.tkl.iis.u-tokyo.ac.jp/ddc/finder> から公開データの俯瞰、検索が行える。さらに利用者登録を行うことでデータのダウンロードができる。

キーワード: DIAS, データ公開, データ中心科学

Keywords: DIAS, Release of Geoscience data, Data Centric Science