R011-02
C 会場 ：11/25　AM1 （9:00-10:15）
9:20〜9:40

DARTS 宇宙科学データアーカイブにおけるメタデータ駆動型 Web コンテンツ整備の取り組み

#中平 聡志 [1)], 稲田 久里子 [1)]
(1 宇宙航空研究開発機構 宇宙科学研究所

Metadata-Driven Web Content Organization in the DARTS Space Science Data Archive

#Satoshi NAKAHIRA[1)], Kuriko INADA[1)]
([1]Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency

ARTS is a space science data archive operated by the Institute of Space and Astronautical Science (ISAS). DARTS provides over 300 datasets from approximately 40 missions dating back to before 1970, offering diverse data across various fields. However, due to changes in data archiving standards over the years and differences in policies among research fields, the data has become a ”hodgepodge”, making the site difficult to navigate and use for users without specific data goals.

To address this issue, we aggregated all information into machine-readable metadata and developed a website and web applications that autonomously operate and display using this metadata. The metadata was created using the schema.org Dataset format, referencing various forms of text and HTML data, with one metadata entry generated for each dataset and each observation mission.

The website was developed using a static site generator based on nuxt.js, which generates pages from the data, including the metadata. Additionally, we developed dynamic web applications: a ”Dataset Catalog Display” tool, which presents dataset contents in a human-readable format and allows for searchable access, and an ”Observation Time Coverage Display” tool, which visualizes the time spans covered by numerous datasets. These tools are organically integrated into the entire site, making it easier for users to explore data relationships using various keywords.

Furthermore, we are developing a data search tool that creates indexes for each data file and incorporates faceted navigation to speed up filtered searches. This tool is also designed for API usage, allowing for detailed table retrieval by connecting with a specific database based on filtered criteria.

With much of the site’s information now machine-readable, we have also considered leveraging AI. We are currently developing an interactive assistant system to support data exploration in DARTS. This system interprets user queries using AI, employing language models for searches and summaries, as well as API calls, to provide functions such as: 1) explanations of observation missions, 2) descriptions and suggestions of datasets, 3) information on data usage (e.g., code examples), and 4) data search and result provision, all through agent-based responses.

Through these efforts, DARTS is evolving into a data archive that is user-friendly for a diverse range of users and fosters new discoveries, while also enabling us to manage data more transparently.

DARTS は宇宙科学研究所が運営する宇宙科学データのアーカイブである。DARTS は、1970 年以前からの約 40 のミッションによる 300 以上のデータセットを提供しており、分野を超えた多様なデータを入手可能である。しかし、長年にわたるデータアーカイブの基準の変遷や、研究分野ごとの方針の違いにより、データは「寄せ集め」の状態となっており、特定のデータを目的としない利用者には見通しが悪く、使いづらいサイトであった。

この問題を解決するため、あらゆる情報を機械的に可読なメタデータとして集約し、メタデータを自律的に活用して稼働・表示される Web サイトおよび Web アプリケーションを構築した。メタデータは schema.org の Dataset を用いて作成し、様々な形式で書かれたテキストデータや HTML データを参考に、データセットごと、さらに観測ミッションごとに一つずつ生成した。

Web サイトは nuxt.js を用いた静的サイトジェネレータにより、メタデータを含むデータからページを生成する方式を採用した。また、データセットの内容を人間にとって読みやすい形式で表示し、検索可能にする「データセットカタログ表示」のツールと、多数のデータセットがカバーする時間範囲を可視化する「観測時間カバレッジ表示」ツールを、動的な Web アプリとして構築し、サイト全体と有機的に結合させた。これにより、利用者は多様なキーワードを用いてデータの関連性を調べやすくなった。

さらに、データファイルごとの索引を作成し、絞り込み検索を高速化するファセットナビゲーションを導入したデータ検索ツールを開発している。このツールは API からの利用も想定しており、絞り込んだ状態から詳細データベースと連携してテーブル取得も可能である。

サイト内の多くの情報が機械にとって可読な状態になったことで、AI の活用も視野に入れた。現在、DARTS のデータ探索を支援する対話型アシスタントシステムの構築を進めている。このシステムは、AI がユーザーの質問を解釈し、言語モデルを用いた検索や AI による要約、API 呼び出しにより、1) 観測ミッションの説明、2) データセットの解説・提案、3) データの使用方法（コードなど）の情報提供、4) データ検索と結果の提供、といった機能を持つエージェントを呼び出して応答するものである。

これらの取り組みにより、DARTS は多様な利用者にとって使いやすく、新たな発見を促進するデータアーカイブへと進化しつつあり、私たち自身もデータの管理を見通し良く行える状態に近づいている。