DARTSが考える長期アーカイブ

篠原 育，松崎恵一，山本幸生，海老沢 研
JAXA宇宙科学研究所
科学衛星運用・データ利用センター

概要

- 議論の背景
- DARTSとは
- アーカイブの意義と機能
- DARTSの開発・維持の考え方
- まとめ

システムの長期的な維持・運用という問題意識でのコメント

議論の背景

- 科学衛星によって取得されるサイエンス・データの総量は膨大なものになった.
- しかし，将来にわたって有効に利用できる状態を保てるだろうか？

  （衛星プロジェクト・チームはミッション終了とともに解散するが，データはその後も残り続ける．）

- ISAS/JAXAとしての役割は何か？世界との差別化
  - インターネット公開時代には，公開データはあらゆる場所から入手できる

- データ“アーカイブ”のあるべき姿をあらためて議論することで，データ整備・システム開発方針を見直そう！

DARTSとは

科学衛星サイエンスデータ・アーカイブ・システム

- 公開データ・アーカイブ DARTS
  http://darts.isas.jaxa.jp
- 1998年から正式なサービス
  - 「ぎんが」(1987年打ち上げ)以降のミッションについて公開データをアーカイブ化
  - 様々な宇宙科学ドメインのコンテンツ
    X線，赤外線，電波天文学，太陽物理学，STP，惑星科学，ISS実験
- 総データ量は 15 Tbyte 以上(「かぐや」含まず)

現役ミッションを中心とした利用

データベースの所属	大分類	小分類	衛星(観測装置)	ミッション期間	備考
DARTS	天文学	X線天文学	てんま	1983-1989	観測ログのみ
			ぎんが	1987-1991	
			あすか	1993-2001	
			すざく	2005-	
		赤外線天文学	IRTS	1995-1996	
			あかり	2006-	
		電波天文学	はるか	1997-2005	
	太陽物理学		ようこう	1991-2001	
			ひので	2006-	
	STP		あけぼの	1989-	
			Geotail	1992-	
			れいめい	2005-	
	月惑星科学	小惑星	はやぶさ	2003-2010	
		金星	あかつき	2010-	
		月	かぐや(HTV)	2007-2009	
JSPEC	月惑星科学	月	かぐや(HTV 以外)	2007-2009	
理研	天文学	X線天文学	ISS(MAXI)	2009-	
プロジェクト	大気科学		ISS(SMILES)	2009-	
なし	微少重力		ISS(与圧部実験)	2008-	サンプルを除く画像、動画など。まだアーカイブされていない。

アーカイブの意義と機能

データアーカイブの意義

- **データから得られる結果の再現性，普遍性を保証する**
- あらゆるサイエンスの結果は再現性，普遍性が保証されていなければならない
- **データの寿命を延ばす**
- 衛星の寿命が尽きた遙か後でも，そのデータを使って科学的成果を出すことが可能になる
- **データが使われる範囲を広げる**
- データを世界に広く公開することによって，より多くの科学的成果が生み出される
- **国際的な科学の発展に貢献する**
- 科学成果は人類共通の財産．データを公開することで，人類の科学の発展への貢献

用語の使い方のぶれ

同じ単語でも，思い浮かべるイメージはさまざま

- **データアーカイブの機能**
- データプロセッシング，データ保存（保管），データサービスの集合体．<u>データプロダクトだけではなく，その利用に必要な周辺情報，ドキュメント，アプリケーションソフトウェアも含む</u>．

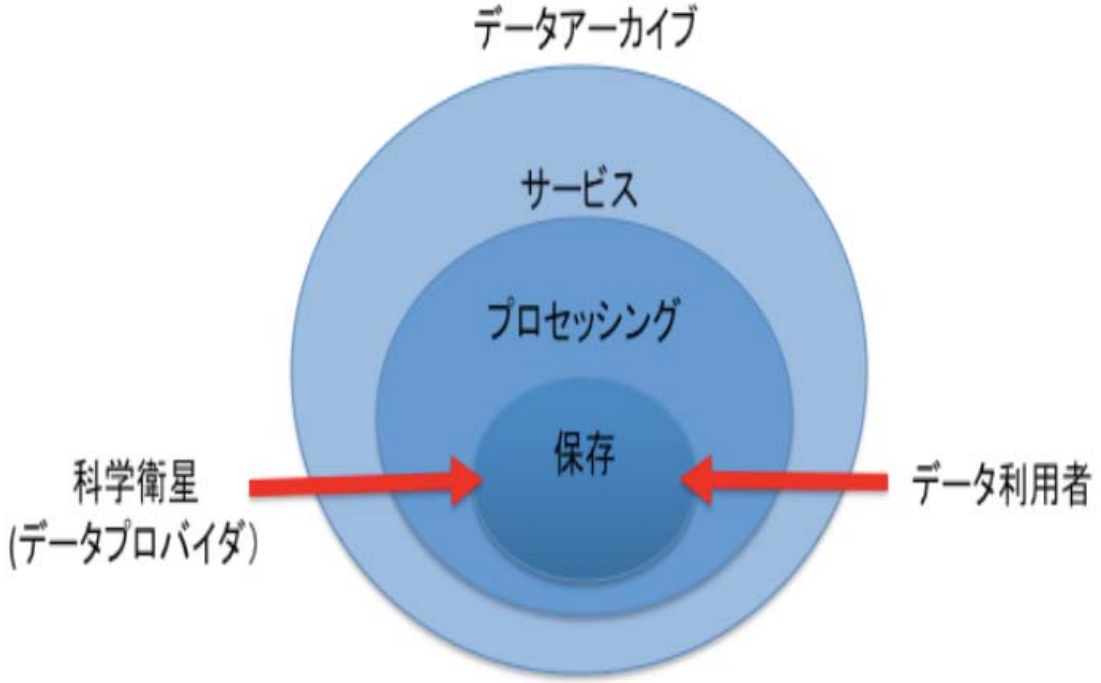


図 1 プロセッシング、保存、サービスの関係

保存，プロセッシング，サービス

- 衛星プロジェクトが思い描くデータベースシステムと長期アーカイブは必ずしも一致しない．
- プロセッシング，アーカイブ，サービスを区別して議論する必要（各要素や短期・長期で実施主体が異なる）
  - 研究者の都合のよいところだけでは実現できない
  - 開発者は何を実装しているのか，明確に意識する必要がある．
- 各研究コミュニティ毎にプロセッシング，アーカイブ，サービスの要求用件やその実現に必要なスキルレベルは異なる．
- 情報技術でカバーできること，そうでないことの区別が大切
  先端的な情報技術の寿命は短いことにも注意が必要

科学衛星データの公開・利用に関する原則（案）

データプロダクションの原則

科学衛星が取得したすべてのデータについて，機器較正やデータ処理アルゴリズムを適用した高次処理を行い，公知の知識だけでそこから科学的成果を引き出せるような高次データプロダクトを作成する．

データ保存の原則

科学衛星が取得したデータは，テレメトリデータも含め，使用できる状態で永久保管する．高次データプロダクトは，プロジェクトごとに定める優先期間の後，全世界に向けて無償公開する．

データサービスの原則

データセンターは，公開高次データプロダクトをできるだけ広い範囲のユーザーが使えるようにするために，データプロバイダに対してデータ保管等のサービスを提供するとともに，ユーザーのための基盤サービスを整備し，ユーザーサポートを行う．

あたりまえの文言のようだけれども...

短期と長期のアーカイブ

アーカイブの役割の質的な差

- **短期アーカイブ**
  - 進行中のプロジェクトが，プロジェクトの遂行に必要な情報とデータプロダクトを管理・利用するために必要なアーカイブ．（主な利用者はプロジェクト関係者と，その研究コミュニティに属する研究者）

現状の国内のアーカイブはほとんどすべてが短期アーカイブ
プロジェクトの寿命＝データの寿命　ならば議論はここまで

- **長期アーカイブ**
  - プロジェクト終了後も，プロジェクト関係者以外でもデータを利用するために必要な情報を集めた，データプロダクトの長期利用を可能とするアーカイブ．（研究コミュニティ外の学問分野に属する研究者も利用者として想定）

長期アーカイブを実現する価値にコンセンサスはあるのか…？
プロジェクトの成果で元がとれればよいのか

DARTSの維持･開発の考え方

システムの開発体制と維持

- 少ないリソースをどこに投入するべきか？
  - 上流のデータ処理はできるだけ合理的に共通化
  - 情報が散逸しないように最低限の文書整備は必須
  - 短期的な機能　と　長期的な機能　はわけて考える
  - やるべきこと　と　あったらいいこと　の要求はわけて考える
  - プロジェクト／データセンター／外部機関との適切な役割分担

- どのような技術を採用するべきか？
  - 個人・特定メーカーに依存する開発をすべきでない
    - 人員の異動にも耐えられる開発体制を考える
  - 適材適所の開発体制をとる
    - 長期間使う基盤は手堅く枯れた技術を採用
    - 機動性の要求される部分は短寿命にスクラップ＆ビルド

システムの寿命設定

- 長寿命であるべきもの：

  ファイル，データファイルのフォーマット，データベース，文書

- 短寿命でも仕方ないもの：

  Webアプリケーション，プロジェクトが進行中の時にのみ必要な情報 （データはどこかでファイナライズする必要）

  - 想定寿命に応じたバランスのよいリソース配分
  - 長寿命を想定した場合，定常的な運用業務の負担量，担当者の交代，などの要素はたとえ簡単なことであっても馬鹿にならない

アプリケーションは，セキュリティー対応，OSやミドルウェアのアップデート，etc，常に維持するためには作業は必要

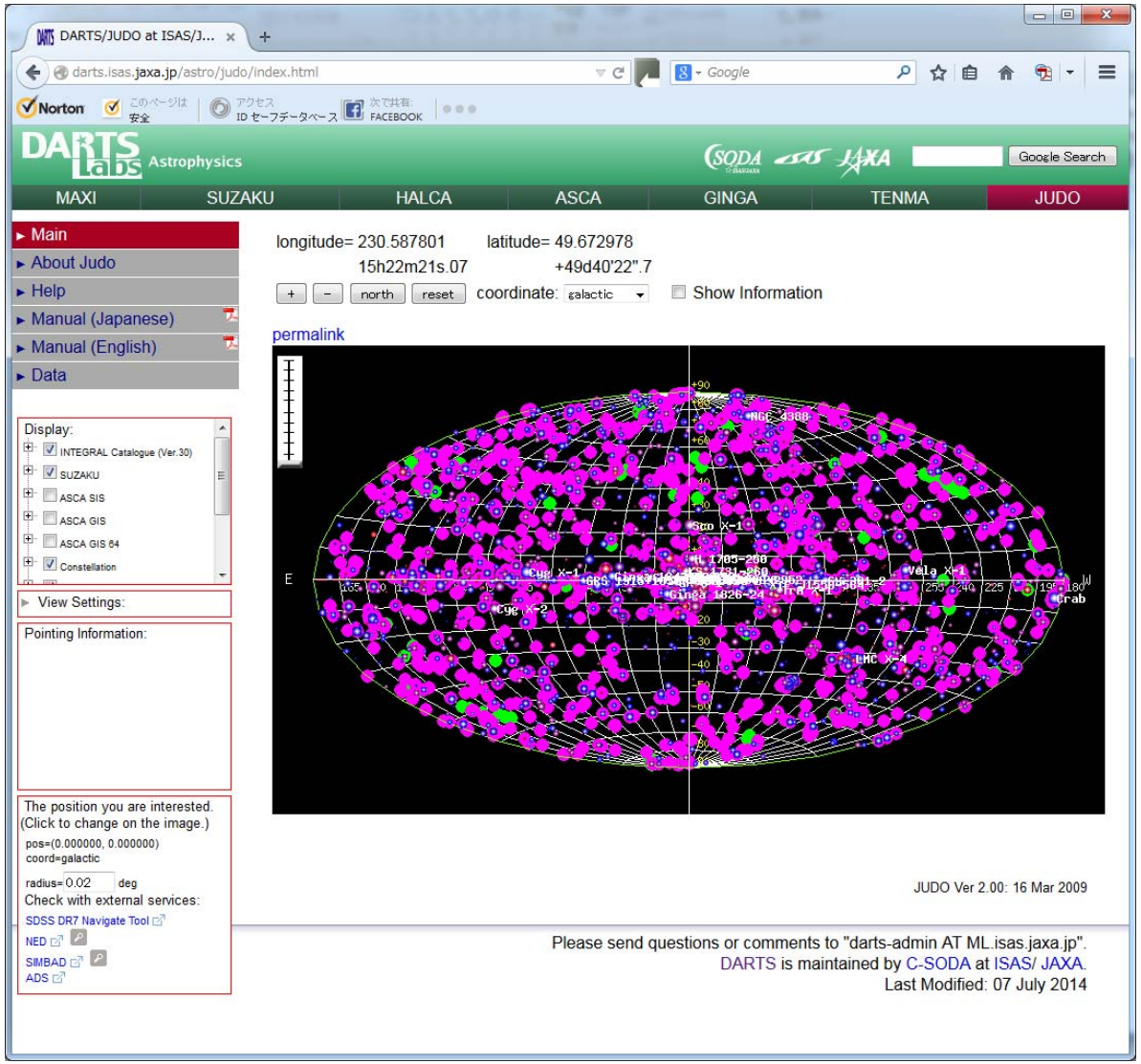
DARTSでの取り組み

- 既存アプリケーションの識別（3年以上かけた取り組み）
  - 設計書・インストールマニュアル・システム設定台帳，等，既存アプリケーションの維持に必要不可欠な文書はあるか？
    （相互レビューは必須）
  - 維持・管理に必要なコストを査定し，担当（職員か業者か）等の管理コストを明らかになっているか？
  - アプリ開発の共通ルールを定義　…　ソースコード管理法，利用パッケージの制限，等

- 識別結果を2つのサービス・カテゴリに分類
  - 長期的な維持・運用の見通しがあるのも　…　DARTS正式サービス
  - 維持にリスクを含んだもの　…　DARTS labs
  - DARTS laboに識別されている間は，担当者責任で維持・運用を行う．（維持できなくなった時点でサービス終了）

2013年には大規模なシステム換装があったが，正式サービスの移行は担当者の手をかけずにシステム移行契約の範囲で移行が完了できた．

DARTS labsの例



高機能で複雑なアプリケーションを外部業者に開発を依頼した場合，十分な情報をあわせて納入してもらわないとその維持は困難.

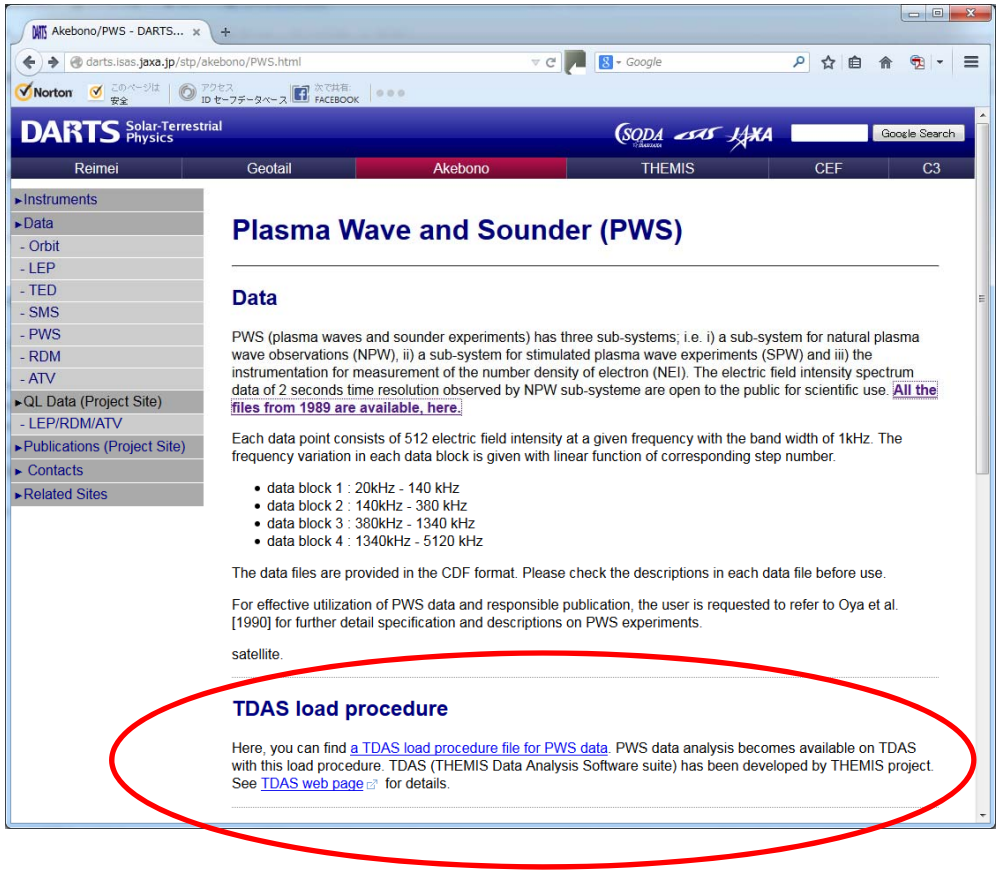
文書整備に時間をかけられないポスドクの開発による研究的なサービス実装についてはDARTS labsからスタート

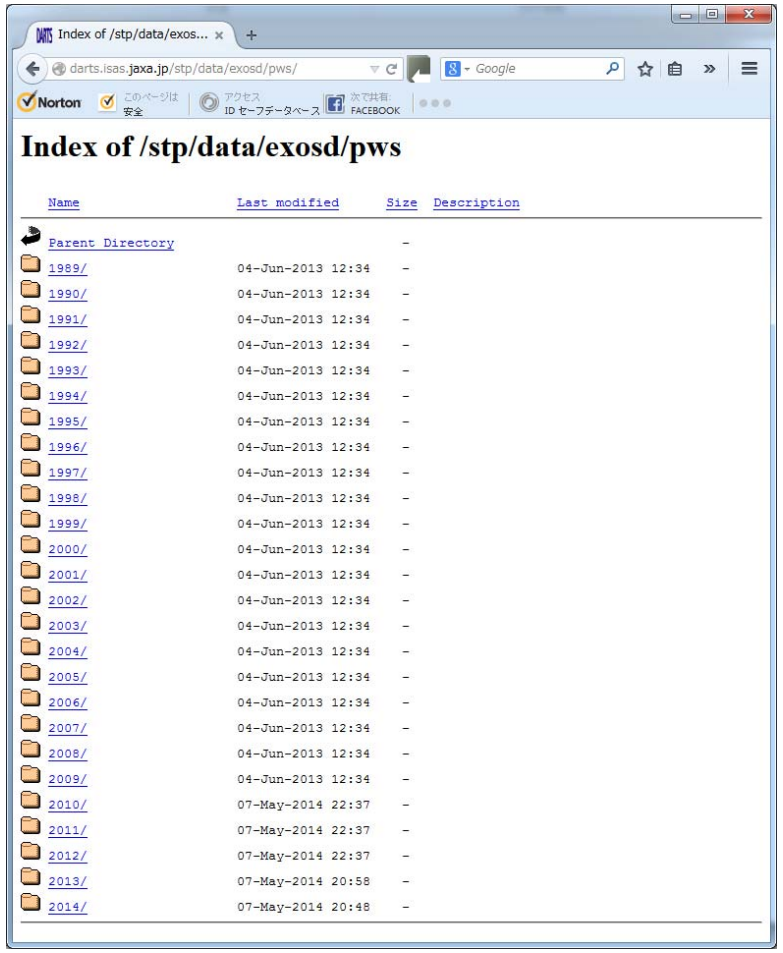
DARTSの今後の方向性

- プロジェクトの短期アーカイブは原則プロジェクトが開発する.
- ただし，長期アーカイブ化を見据えて，短期・長期を共存できる場合にはプロジェクトとC-SODAで共同開発する.
- C-SODAは長期アーカイブ・サービスとして，アーカイブの開発・維持・管理・運用を進める.
- サービス・レベルとしては，長期アーカイブの範囲で高機能のものは考えない．高機能化を考える場合も，長寿命化を阻むリスク要因を局所化させる.
- 機動的・短期的なサービスについてはDARTS labsの枠で開発を進める.
- 外部機関におけるデータ提供サービス（データのミラーリングを要しない）の実装を推進できるように，データ取得I/Fを公開する.　（IUGONETへの参加）

DARTS/STPの場合

- ほぼすべてのデータについて，httpでdirectory indexを直接公開している．→ TDAS load procedure に対応





まとめ

よいデータアーカイブの実現は，地味な作業の積み重ね

- “データアーカイブ”を構築・運用するに際して，ポリシーの明確化は重要
  - “あたりまえ”と思っていることも，よく考えるとそうではない．
- アーカイブを構成する3つの要素「プロセッシング」「保存」「サービス」を意識的に区別したシステム設計・運用設計
- 長期アーカイブの実現には，データの品質の確保，および，周辺情報の収集が本質的
  - 文書を整備・収集することは長期的な維持には欠かせない．
  - データ・サービスとデータ・保存のレイヤーはきちんと切り分ける．
  - NASA PDSのArchive Preparation Guideはとても参考になります．

    https://pds.jpl.nasa.gov/tools/

IUGONETの今後へのコメント

データサービスに関わる研究開発は短期的な成果よりも長い目でサポートして欲しい

- 新しいサービスを次々に追加するよりも，長期間・安定して一定のサービスを提供できるように，プロジェクト期間の終了後を見据えたシステムの設計書などを充実させて，引き継ぎ可能にする．（学術論文だけではなく，white paper等の技術文書の執筆も業績評価の対象とすべき．）
- サービス面だけでなく，上流のデータプロセッシング作業と連携して，新しい観測データを継続的に追加し続ける体制の強化も視野に入れる必要．
- いずれにせよ，関係協力機関が足並みをそろえて，長期的にデータアーカイブを維持する体制について考える必要がある．
- 近接分野の新しいデータ利用を高めるためには，システムの開発だけではなく，解析手法の入門文書など，教科書的な文書の収集・作成をする活動に期待．（解析講習会の経験をうまく蓄積することが重要．）