# 科学データの公開・利用について

宇宙航空研究開発機構(JAXA)
宇宙科学研究所(ISAS)
科学衛星運用･データ利用センター(C-SODA)運営委員会
2011 年 9 月 20 日

## はじめに

1970 年に東京大学宇宙航空研究所（宇宙研）が日本初の人工衛星「おおすみ」を打ち上げて以来、宇宙からの科学観測は宇宙開発の発展と共に歩んできた。1981 年の文部省宇宙科学研究所の発足、2003 年の宇宙三機関の JAXA への統合を経て、今や JAXA は、世界最高レベルの巨大衛星プロジェクトを実現する力を持っている。それに伴い、衛星が取得する科学データは大量かつ複雑化・高度化し、各研究ドメインにおいて、JAXA の科学データを扱うコミュニティーが国内外に大きく広がりつつある。

宇宙科学の観測研究は人工衛星を打ち上げてデータを取得することから始まる。衛星の寿命は有限であるがデータの寿命は無限である。そのような意味で、衛星が生み出す科学データは人類の普遍的知的財産として、恒久的に保存されるべきである。しかし、衛星からの受信データはそのままでは科学的に解釈することはできない。また、少数の衛星プロジェクト関係者だけでは大量のデータを十分に活用しきれない。データが使いやすい形にまで処理され、さらに長期間にわたって保存され、より広い範囲の研究者によって使われてこそ、衛星による科学的成果が最大化されるのである。すなわち、衛星が優れた科学的成果をあげるためには、そのデータ公開・利用は、衛星の製作・運用に匹敵する重要さを持っていると考えてしかるべきである。しかし、実際には、近年の衛星本体や観測装置の著しい進歩に比べ、現時点における JAXA の科学データの整理状況は、満足できるものとは言い難く、そのために衛星が本来出し得る科学的成果を出し切れていない場合が多い。このような現状は改善されなくてはならない。

日本の衛星による科学データの公開・利用は、1980 年代から宇宙研の衛星プロジェクトとデータセンターが中心となって進められてきた。しかし、データ公開・利用の方針は定まっておらず、その実態は研究分野や衛星プロジェクトによって大きく異なっていた。そこで、当委員会では日本の衛星による科学データ公開・利用の現状を調査した上で、それらを総合的に分析し、望ましいデータ公開・利用のあり方についての議論を続けている。その第一段階の報告として、特定の分野や組織には依らない、科学データの公開・利用に関する原則と指針をまとめたので、本レポートにて報告する。広い範囲の読者からの御意見を期待すると共に、本提案によって、日本の衛星データが生み出す科学的成果が最大化され、それが宇宙と自然のより深い理解につながることを期待する。

目次

1 科学データアーカイブとその目的..... 4
2 科学データ利用に関する原則..... 5
3 プロセッシング、保存、サービスの関係..... 9
補足：用語の定義 ..... 10
付録： データプロダクトレベル定義の例..... 14
執筆者一覧: ..... 15

## 1 科学データアーカイブとその目的

JAXA の衛星が取得したすべての科学データは、公知の知識さえあれば使える形にまで処理（プロセッシング）された後、最終的には永久保存され、適切なサービスの提供によって世界中の誰もが学術目的のためにそれを無償で取得でき、得られた成果を自由に発表できるようにするべきである。それを可能にするための、データプロセッシング、データ保存、サービスの集合体を、本文書では「データアーカイブ」と呼ぶ。科学データをアーカイブ化する目的は以下の通りである。

### (1) 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。データを特定の手法で解析した結果が、第三者によって再現できなければ、それは信頼できるものとは言えない。アーカイブを用いて、誰でも同じデータから同じ結果が出せるようになって初めて、その結果は普遍性を得る。特に、宇宙観測においては同じ現象が再び観測される保証がないので、ある現象について複数の研究者による検証を可能にするためには、データのアーカイブ化が必須である。

### (2) データの寿命を延ばす

衛星の寿命は短いものでは数年、長いものでも 10 年のオーダーである。一方、データをアーカイブ化しておけば、その寿命は原理的には無限である。すなわち、データアーカイブの存在によって、衛星の寿命が尽きた遙か後でも、その衛星から科学的成果を生み出すことが可能になる。また、多くの宇宙現象の時間スケールは衛星の寿命よりも長いので、研究のために過去の衛星データを利用することがよく行われる。すなわち、データのアーカイブ化は、未来における研究を支援することに直結する。

### (3) データが使われる範囲を広げる

衛星開発や運用に関わってきた少数の衛星プロジェクトメンバーだけで生み出すことのできる成果は限られているが、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。

### (4) 国際的な科学の発展に貢献する

衛星の開発や運用は、間接的に国民によって支えられているので、その成果は広く人々に還元されるべきである。衛星の取得物はデータなので、それはデータの公開とその帰結である科学的成果の産出によって担保

されるべきである。衛星による科学データは人類共通の財産であり、世界に向けて広くデータを公開することが、日本による国際的な科学の発展への貢献と考えられる。

## 2 科学データ利用に関する原則

現時点においてJAXAの衛星による科学データ公開・利用における状況は研究ドメインや衛星によって大きく異なり、それがうまく機能して成果が出ているケースもあれば、機能していないがために十分な成果が出ていないケースもある。その主な原因として、JAXA の科学データ公開・利用に関する大方針が定まっていないため、衛星プロジェクトの提案、選定、予算措置、実行等の各局面において、データ公開・利用が図らずも軽視されており、多くの場合、衛星データの公開・利用業務が「ベストエフォート」(=できれば望ましいが、できなくても構わない）になってしまっているという現実が挙げられる。

このような状況を改善し、JAXA の科学データが国内の大学等から成る宇宙科学コミュニティーを中心に、国内外の広い範囲に利用され、それによって最大の成果が得られることが保証されるよう、JAXA の科学データ利用に関する原則を、以下の通り提案する[1]。JAXAの衛星プロジェクト（小型科学衛星、ISSデータを含む）は、これらの原則をすべて満たすべきである[2]。衛星プロジェクトを選定・評価する側も実行する側も、プロジェクトの各局面において、これらの原則を肝に銘じておくことが望まれる。

**JAXAの科学データ利用に関する原則**

**1. データプロセッシングの原則 : すべての科学データについて、機器較正やデータ処理アルゴリズムを適用し、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段階に至るまでの処理（プロセッシング）を行う。**

**2. データ保存の原則 : 取得したすべての科学データは、使用できる状態で永久に保存する。**

**3. データサービスの原則 : データセンターは、データプロバイダを明らか**

---

[1] これらの原則は、世界中のデータセンターを包括する組織であるWorld Data System(http://icsu-wds.org/)が目指す方向性と合致している。

[2] ISS曝露部データ、および与圧部の実験データのうち、画像、動画などを対象とする。気球、観測ロケットは、とりあえず対象としないが、将来的にはこれらのデータのアーカイブ化も検討の余地がある。

**にした上で、そのデータが長期にわたってできるだけ広い範囲のユーザーに使われるようにするための基盤サービスを無償で提供する。**

以下、上記原則の補足説明である。

**1．データプロセッシングの原則**

衛星からの生（テレメトリ）データは、それだけでは物理量として解釈できない。衛星ごとのテレメトリフォーマットに応じてデータを変換し、検出装置ごとの特性を除いて（=データのキャリブレーション）初めて、観測データは物理量として解釈できるようになる。そのような高レベルまでデータを処理することには大変な労力を必要とするが、各衛星プロジェクトは、データに対してそこまでの責任を持つべきである。それによって、衛星固有の知識を持たない研究者でも、そのドメインにおける科学的知識さえあれば、それらのデータを利用可能になる。

人工衛星による観測データは地上観測データとは違い、観測条件が比較的均一で、あらかじめ定められたコマンドシークエンスに従って取得されるため、定形処理しやすい。そのような定形処理スクリプトに従って自動データ処理を行い、高次データプロダクトを生成する「パイプラインプロセッシング」は、すべての衛星に対して導入すべきものである。パイプラインプロセッシングの生成物であるデータプロダクトのレベルは、衛星プロジェクトによって様々である。各衛星プロジェクトは、データをどのレベルまで処理し、プロセッシングの最終成果物としてどのようなデータプロダクトを生成するかを明確に定義する義務を持つ。

パイプラインプロセッシングの実行はサイエンスというよりはエンジニアリングとして位置づけられるべきものである。パイプラインプロセッシングシステムの開発には検出器と衛星固有の知識が必要なため、衛星プロジェクトが担当する必要があるが、いったんそれができてしまえば、定常処理を実行することは衛星固有の知識を持たないエンジニアでも実行可能である。パイプラインプロセッシングの実行体制は、衛星プロジェクトごとに異なり、たとえば、JAXA 外で実行されることもありうる。衛星プロジェクトは、その実行体制を定義し、明確にする責任を負う。JAXA は、どこでパイプラインプロセッシングが実行されてデータプロダクトが生成されるにしても、ユーザーのニーズに応え、デ

ータの品質を明らかにする責任を持つ[3]。

## 2．データ保存の原則

すべての科学的発見は客観性、再現性を持たなければならないことは明らかである。衛星による科学成果も第三者によって再現できるものでなくてはならないが、地上の実験とは異なり、宇宙現象は対象をコントロールできないので、ある発見を別の独立な観測によって検証することは容易ではない。実際、検出装置を開発したグループが非公開データを用いて行った「発見」が学会を賑わし、それが何年も後に、より精度の高い別の観測によって否定された例が数多くある。そのような間違った発見に共通することは、データとそれから発見を引き出したアルゴリズム（ソフトウェア）が適切に保存されておらず、第三者による検証が不可能だったことである。

そのような事態を防ぐために、JAXA の衛星による発見、あるいは何らかの新たな知見をもたらすのに用いられた観測データ、機器較正データ、工学データ、データ処理のアルゴリズム（ソフトウェア）等は、第三者によってそれらの発見・知見を再現可能であるように保存されるべきである。それは、利用者だけでなくデータを出す衛星プロジェクト側にとっても有益である。また、保存された科学データは、それに知的財産権が付随すると考えるよりも、科学的な「公共財」と考えて、最終的には無償公開するべきである。

ただし、多くの科学者にとって、自分で開発した衛星や観測装置を用いて新たな発見を行うことが衛星プロジェクト実行の強いモチベーションであるので、衛星プロジェクトチームに優先的にデータ使用権が与えられることは当然認められる。また、工学データに関しては、技術的な優位性を保持するために、公開が難しい場合もある。そのような非公開データに関しては、機密性を保持することによって、守るべき価値が存在することを明らかにすべきである。

衛星観測がプロポーザル制で行われる場合、観測提案が採択された科学者がデータの優先権を持つことが多い。そのような場合、データの優先権、公開時期やプロポーザル制の運営方法については、衛星プロジェクト内で調整が必要である。衛星プロジェクトは、そのようなデータポリシーを決定し、明らかにする責任を持つ。

---

[3] エラーが大きすぎて科学解析に使えないこともありうるが、それも品質として明示する。

衛星データがアーカイブ化されることによって、世界中の多くの科学者によって長期間にわたる利用が可能になり、より多くの成果が生み出される。そのように、データがアーカイブ化されることにより、それが使われる範囲と時間が広がると言う「量的な」メリットだけでなく、データが検出器チームや当初の観測者が想定したものとは異なった使われ方をされるという「質的な」メリットもある。たとえば、アーカイブデータを第三者の目で見直すことによって当初見逃していた新たな発見があり得るだろうし、大量のアーカイブデータを統一的に解析することで、当初は不可能だった統計的な研究も可能になる。そのような科学的な視点からも、科学データのアーカイブ化は重要かつ必須の課題である。

## 3.データサービスの原則

データを作成するのはデータプロバイダの役目であるが、データセンターはその永久保存、公開のためのサービスをデータプロバイダに提供するべきである。これによって、データプロバイダがデータを公開するためにかける労力を削減し、データ公開への敷居を下げる。

科学データが適切に処理され、データセンターに保存されても、それが多くのユーザーにとって使いにくいものであったら、そこから科学的な成果は出にくい。データが最大限活用されるよう、データセンターはデータ利用のために使いやすいサービスを、無償で提供すべきである。ただし、そのような高度なシステム開発は科学者が片手間で行えるほど簡単なものではない。よって、データサービスの開発は、衛星の開発や科学研究とは独立の重要な業務として位置づけ、尊重されるべきである。

また、データ利用を促進するためには、ヘルプデスク等のユーザーサポートが必要であり、データサービスの開発とともにユーザーサポートにもリソースを投入すべきである。そのために必要なリソースが確保されているかどうかは、アーカイブ開発・運用体制のレビューの際に、チェックされるべきである。

公開されるデータには、データを作成したデータプロバイダ名を明示する。それによって、データプロバイダの責任を明らかにすると共に、データが利用される際には、データプロバイダに謝意が示されるようにする。

## 3 プロセッシング、保存、サービスの関係

データアーカイブに関連する議論を進めるにあたって、プロセッシング、保存、サービスの関係を明確にする必要がある。

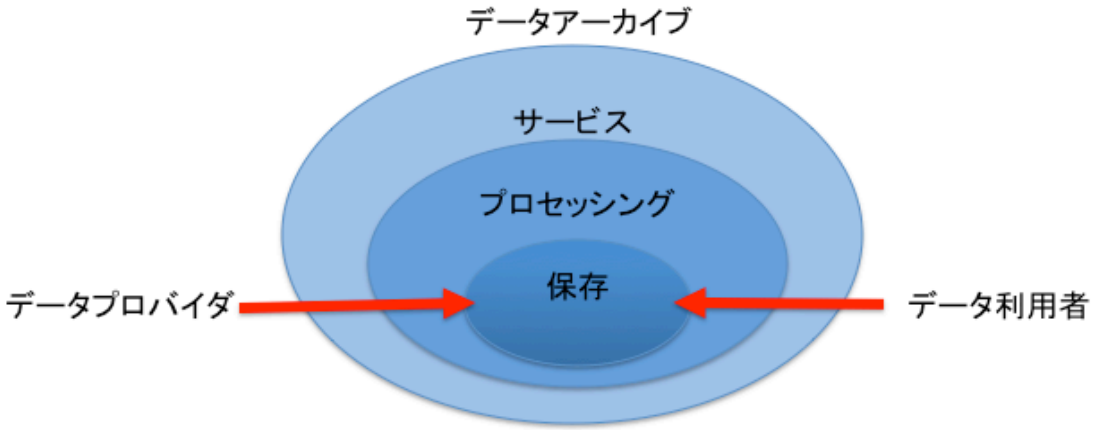


**図 1 プロセッシング、保存、サービスの関係**

次項で説明するとおり、ここではデータアーカイブをデータプロセッシング、データ保存、データサービスの集合体と考え、その関係は図 1 のとおりである。すなわち、データプロバイダ（衛星プロジェクトなど）からのデータはサービス、プロセッシングを経て、保存される。データ利用者はサービスを経て、保存されたデータを（場合によってはプロセッシングされた後に）引き出す。

データの保存、プロセッシング、サービスのそれぞれの要求条件と、その担当者に必要なスキルが異なることに注意が必要である。データ保存の担当者には、研究分野や衛星検出器固有の知識よりも、長期的な計算機インフラ整備の知識が必要とされるであろう。プロセッシングに関しては、データのキャリブレーションを実行するために、それぞれの研究分野や衛星検出器固有の知識が必須である。サービスには、データ検索機能の実装、データの早見機能の提供などが含まれるが、それにはミッション横断的に可能なものと、ミッション個別の知識が必要なものがある。サービスの開発者には、情報システムの幅広い知識が必要とされる。

データ利用者にとっては、アーカイブは「データが置かれている場所」と見えていれば良いので、サービス、プロセッシング、保存の違いを意識させる必要はない。ただし、アーカイブ設計においては、三者間のインタフェース（責任分界点）を明確にし、それぞれを実施する主体を定義することが必要である。

## 補足：用語の定義

今まで、データアーカイブに関連する用語は、研究分野や衛星によって統一がとれておらず、それが混乱の原因となることもあった。そこで、本レポートでは、データアーカイブに関連する用語を整理し、それらの定義を明らかにした上で用いることにする。これらの用語は本レポート中における定義であり、宇宙科学の分野に限ったものであるが、今後、これらの定義が国内の宇宙科学コミュニティーで広く用いられるようになることを想定している。

- データプロダクト：
  なんらかの科学観測の結果得られるデータを、利用しやすいように処理（プロセス）した生成物。
- キャリブレーション：
  「装置の」キャリブレーションは、装置の特性を調べ、それを特徴付ける物理量（キャリブレーションデータ）を求める作業。「データの」キャリブレーションは、キャリブレーションデータをデータに適用し、生のテレメトリデータを物理量に変換する作業。通常、「キャリブレーション」と言った場合には、この双方が含まれるので、曖昧さを除くためには「装置のキャリブレーション」、「データのキャリブレーション」と言うことが望ましい。
- データプロセッシング：
  データプロダクトを生成するためのアルゴリズムを確立し、それを適用する作業。それには、以下で定義するデータプロダクトレベルを変換する作業（データのキャリブレーションを含む）の他に、Ancillary データ、ソフトウェア、ドキュメントの生成作業や、装置のキャリブレーション、モデル化の作業等も含まれる。ただし、データプロセッシングに何が含まれるかは、衛星プロジェクトに依存する。
- データプロダクトレベル:
  データプロセッシングの段階に応じて、本レポートでは以下の通りデータプロダクトレベルを定義する[4]。ただし、実際には衛星によってデータプロダクトレベルの定義は異なるので、いくつかの具体的な例を、付録に示した。

---

[4] レベル 1 から 5 までの定義は、PDS (Planetary Data System) Standards Reference (http://pds.nasa.gov/tools/standards-reference.shtml)に準拠している(Version 3.8, p.6-6)。

なお、地上データ処理は必ずしもレベル 1 から始まるわけではないことに注意が必要である。たとえば、機上処理されたデータの場合は、いきなりレベル３，４が出てくることもある。

| レベル | タイプ | 内容 |
|---|---|---|
| 1 | Raw Data | 観測データが含まれているテレメトリデータ |
| 2 | Edited Data | テレメトリエラーを修正し、検出器の出力ごとに分割され、時刻付けされたデータ。画像の場合は DN 値（ピクセル値)。 |
| 3 | Calibrated Data | 機器較正を適用し、検出器の出力を物理量に変換したデータ。画像の場合は輝度値。 |
| 4 | Resampled Data | 検出器の出力を時間や空間方向に集積したデータ。データ生成者の恣意性が含まれない。ここから前のレベルに戻ることはできない。画像の場合は、一枚の画像を投影変換したもの。 |
| 5 | Derived Data | マップ、レポート、グラフィック、モデル化など、レベル 4 から導かれた結果。データ生成者の恣意性が反映される。画像の場合はモザイク処理したもの。 |

なお、上記以外に、アーカイブには以下のデータが必要である[5]。

- Ancillary データ
  レベル 3 や 4 のデータを生成するのに必要な、主に科学データ以外のデータ。たとえば、検出器のゲイン、オフセット、視野、スキャン方向など。ただし、あるデータプロダクトが、別のプロダクトにとっては、Ancillary になることもあるので注意が必要である。
- Correlative データ
  宇宙機から取得したデータを解釈するために必要な、別の科学データ。地上観測データなど。
- User Description
  データの必要性、データセットに付随する特異性など、二次ユーザーがデータを解釈するために必要な文書。

[5] これらは、PDS ではそれぞれレベル 6, 7, 8 と定義されている。

- パイプライン：
  人手を介さずに実行されるデータの自動処理。データプロダクトを自動生成するための「パイプラインプロセッシング」、自動処理のためのスクリプトとして「パイプラインスクリプト」という用語がよく用いられる。
- データプロバイダ：
  データプロダクトの生成者。衛星プロジェクト、PI チームに限らず、データの再処理などのデータプロダクト生成作業の主体も含む。
- データサービス：
  データプロダクトの作成者と利用者に向けた、データプロダクト保存や利用のための機能。
- データアーカイブ：
  最終的な公開を前提とした、データプロセッシング、データ保存、データサービスの集合体。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。一般的に、データプロダクトがアーカイブに保存されてから一定の期間の後、誰でもそれを取得し、得られた科学的成果を自由に発表することが許されている[6]。
- アーカイブ化(アーカイビング)
  データプロダクトを作成、長期保存し、サービスを提供してアーカイブを構築すること。
- 短期アーカイブ：
  進行中の衛星プロジェクトが，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[7]。主な利用者は衛星プロジェクト関係者と、その学問分野（研究コミュニティ）に属する研究者。
- 長期アーカイブ：
  衛星プロジェクト終了後に、衛星プロジェクト関係者以外でもデータを利

---

[6] 衛星プロジェクト関係者にのみ限定公開されているデータの保存庫が、テレメトリアーカイブ、エンジニアリングアーカイブなどと呼ばれることもあるが、本文書では、最終的には公開されることが定められているものに限ってアーカイブと呼ぶことにする。

[7] データレポジトリと呼ぶこともあるが、この用語の意味は曖昧であるため、推奨しない。

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。衛星プロジェクト以外、他の学問分野に属する研究者も利用者として想定する。

## 付録： データプロダクトレベル定義の例

上で PDS に準拠したデータプロダクトレベルを定義したが、JAXA の科学衛星で慣用的に用いられているデータプロダクトレベルの定義や名称は、衛星毎に異なる。ここではそのマッピングについて、いくつかの例を示す。

| PDS レベル | あすか | すざく | ひので | STP | かぐや[8] |
|---|---|---|---|---|---|
| 1: Raw Data | SIRIUS/First Reduction File (FRF) | SIRIUS/Raw Packet Telemetry (RPT) | SIRIUS | SIRIUS /Level 0 | Level0 |
| 2:Edited Data | frfread の出力(アーカイブされていない) | First FITS File | Level0 DARTS に登録される FITS ファイル | Level1 | Level1 |
| 3:Calibrated Data | Unscreened/Screened Event files | Second FITS File (Calibrated Event File) | Level1 各データセンターが検出装置毎に作成する較正済みデータ | Level2 | Level2 |
| 4:Resampled Data | 補正していない画像、スペクトルなど | 補正していない画像、スペクトルなど | Level2 ムービー、磁場マップなど | | |
| 5:Derived Data | Exposure 補正、スムージング、モザイク処理を施した画像など | （パイプラインでは自動生成されない） | | | |

[8] 搭載装置の一つである月面撮像/分光機器(LISM)では、より高次な処理を施したデータを Level2 データとは区別して「高次データ」と呼んでいる。

## 執筆者一覧:

H21-22, 23 年度科学衛星運用・データ利用センター運営委員；50 音順
稲谷 芳文（宇宙科学研究所）
海老沢 研（宇宙科学研究所）
尾中 敬 （東京大学）
鎌田 幸男（宇宙科学研究所）
阪本 成一（宇宙科学研究所）
篠原 育（宇宙科学研究所）
清水 輝久（宇宙科学研究所）
出村 裕英（会津大学）
中村 正人（宇宙科学研究所）
波形 寿英（宇宙科学研究所）
福田 徹（JAXA 地球観測研究センター）
松崎 恵一（宇宙科学研究所）
水本 好彦（国立天文台）
村田 健史（情報通信研究機構）
山内 茂雄（奈良女子大学）
山本 幸生（宇宙科学研究所）
横山 央明（東京大学）
依田 眞一（宇宙科学研究所）