

今後の日本のデータアーカイブ運用に ついての提案

データアーカイブワークショップの議論の紹介

2020.9.14 光赤天文連シンポジウム

古澤 久徳@国立天文台（天文データセンター・ハワイ観測所）

for データアーカイブワークショップ有志

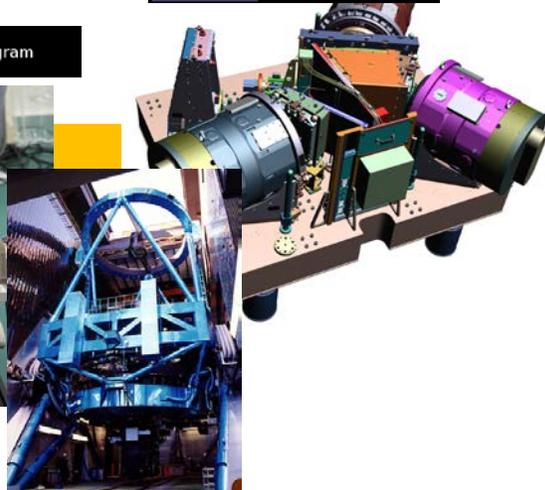
世話人：古澤久徳, 諸隈智貴, 表 泰秀, 川端弘治, 山村一誠

重要な話

- ▶ 日本の可視赤外観測データアーカイブの維持・改善を目指し、運用に対して提案を作ろうとしています
- ▶ 2021年3月に光赤天連の声明とすることを目指しています
- ▶ そのためのワーキンググループへの参加者を募ります

天文ビッグデータ時代とデータ運用

- ▶ HSC/PFS/Euclid/Roman/LSSTなど大サーベイ観測
- ▶ 処理済みデータを戦略的に活用する巨大データベースプロジェクト
- ▶ データをどう作り（生データ・処理）、どう保存し活用するのが、成功の鍵であるとともに責務 → データアーカイブ



動機とアクション（1）

- ▶ データアーカイブをとりまく内外環境の変化
 - ▶ オープンデータ・オープンサイエンス推進
 - ▶ 第4期中期目標：
大学共同利用機関→大学で整備・運用が難しい貴重な
学術データ等の保有提供、全国的な視点で共同利用..に供す
 - ▶ 大学と共同利用機関の関係の変化（4機構連合など）
 - ▶ 対して、こういった基盤に投入できる運営費交付金の継続的な減少
 - ▶ 国立天文台・すばるについてだけ見ても、
 - ▶ データ受け入れポリシー(2010), データポリシー(2014)、
データ利活用の提言・答申(2018) → 実施の困難。。
 - ▶ 装置とアーカイブのデータ管理の連携は十分とは言えない
 - ▶ 生データ管理（現在のSTARS/SMOKA等）と科学ポータル（HSC等）は
データ管理・データ運用という幹を通して連動し、
データ信頼性と利活用促進の両方を守らなければならない。

動機とアクション（2）

日本の天文学コミュニティ全体で考えるべき問題

→ 将来にわたる動機付けと目標設定が「今」必要！

人材の維持・育成の上でもクリティカル

- ▶ データアーカイブの意義と目的の再共有
 - ▶ 本当に必要だと思っているのか？
 - ▶ 何が必要なのか？
 - ▶ それを実現するための手段と役割の明確化
 - ▶ 誰がどこで何をすべきか
- データアーカイブWGでの議論 → 提案の作成
光赤天連声明としての合意を目指す

データアーカイブワーク ショップの開催

https://www.adc.nao.ac.jp/people/~furusawa/work/da_ws/

- ▶ 1回目 2020.1.29-30 約25人参加
 - ▶ 光赤外観測データ運用者等有志によるキックオフ
 - ▶ NAOJ/ADC/すばる、東大(TAO, Tomo-e-Gozen)、京大(せいめい)、広大(かなた)、宇宙研DARTS、西はりま、東工大 MiTSuME など
- ▶ 2回目 2020.3.30 約55人参加
 - ▶ データ利用者によるアーカイブによる成果と期待
- ▶ 提案書に向けて議論を継続している



ワークショップ開催主旨

目的

近年、大学・公開天文台の望遠鏡・装置開発が活発に行われ、同時に大規模観測プロジェクトの推進が進む中、日本のデータアーカイブの置かれる状況を整理して理解し、データアーカイブの意義・在り方を確認した上で、その維持・向上のため向かうべき方向について意見交換を行う。本ワークショップは、その議論のキックオフのための検討会の位置づけであり、各大学所有の観測装置やすばる望遠鏡から得られるデータの運用といった具体的な課題を持つ光赤外分野の有志を中心に課題をあぶり出し、次回以降のより全体的なデータアーカイブの在り方に向けた検討会で議論を深めたい。

議論のテーマ

データアーカイブを取り巻く背景整理

日本・国政策・大学共同利用体制・国立天文台の状況
世界の時流

データアーカイブの存在意義

何をするためのものか
それが日本・世界の天文学にとってどのような位置づけか

データアーカイブの維持・向上の在り方

何を・どこで保持し、どのように提供し、何をしたいのか、誰が行うのか

データアーカイブWSの議論

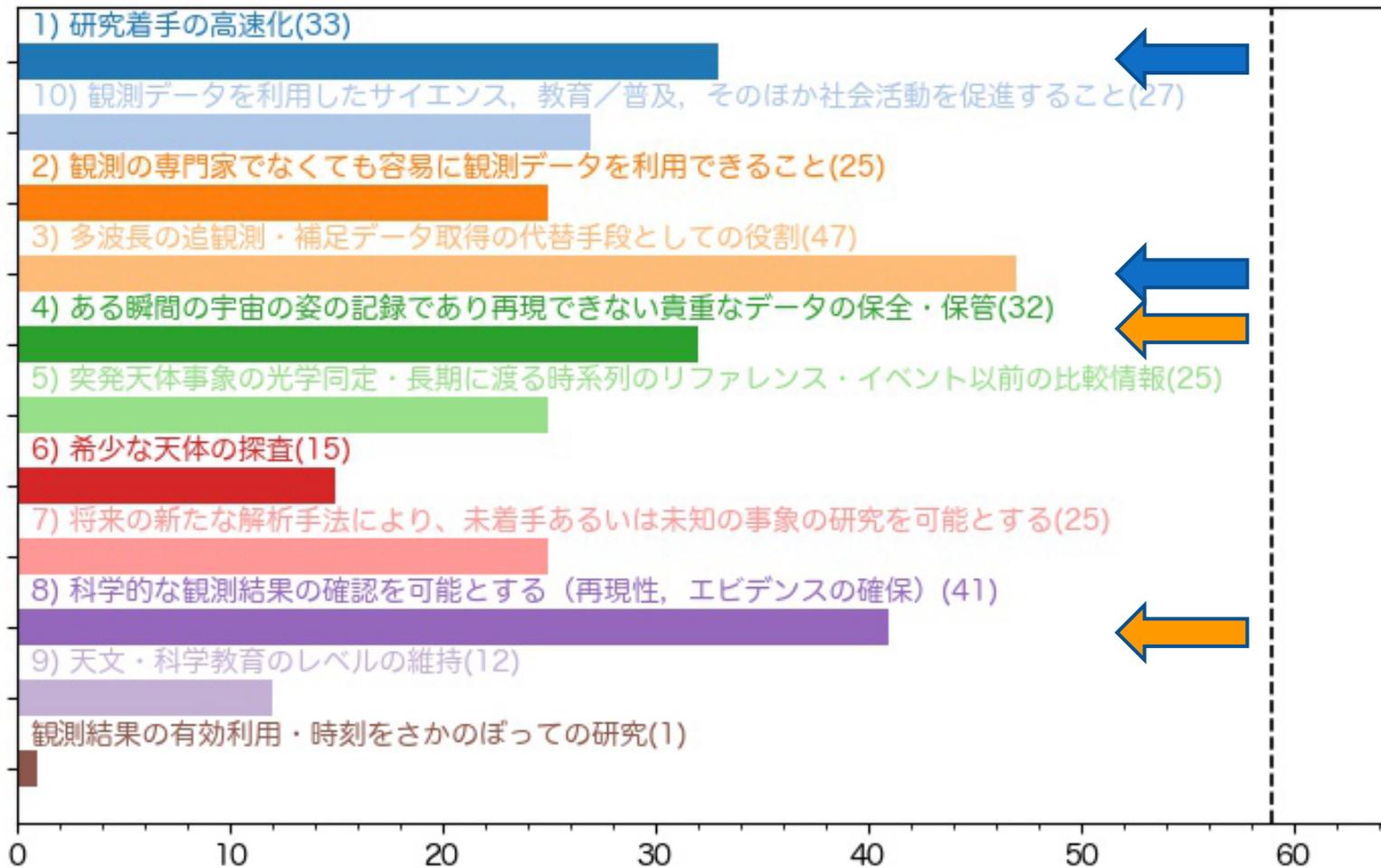
- ▶ 背景の理解、海外の状況調査
 - ▶ データアーカイブの意義と役割
 - ▶ 必要なデータと選定基準
 - ▶ 必要な機能
 - ▶ データ提供者とデータアーカイブの役割
 - ▶ データアーカイブへの理解を得るために必要なこと
 - ▶ データアーカイブが果たした成果・見込める成果
- 詳しくはウェブサイト上の資料をご覧ください。

光赤天連メンバーへのアンケートの結果

2020.8実施 59人が回答 ご協力ありがとうございました

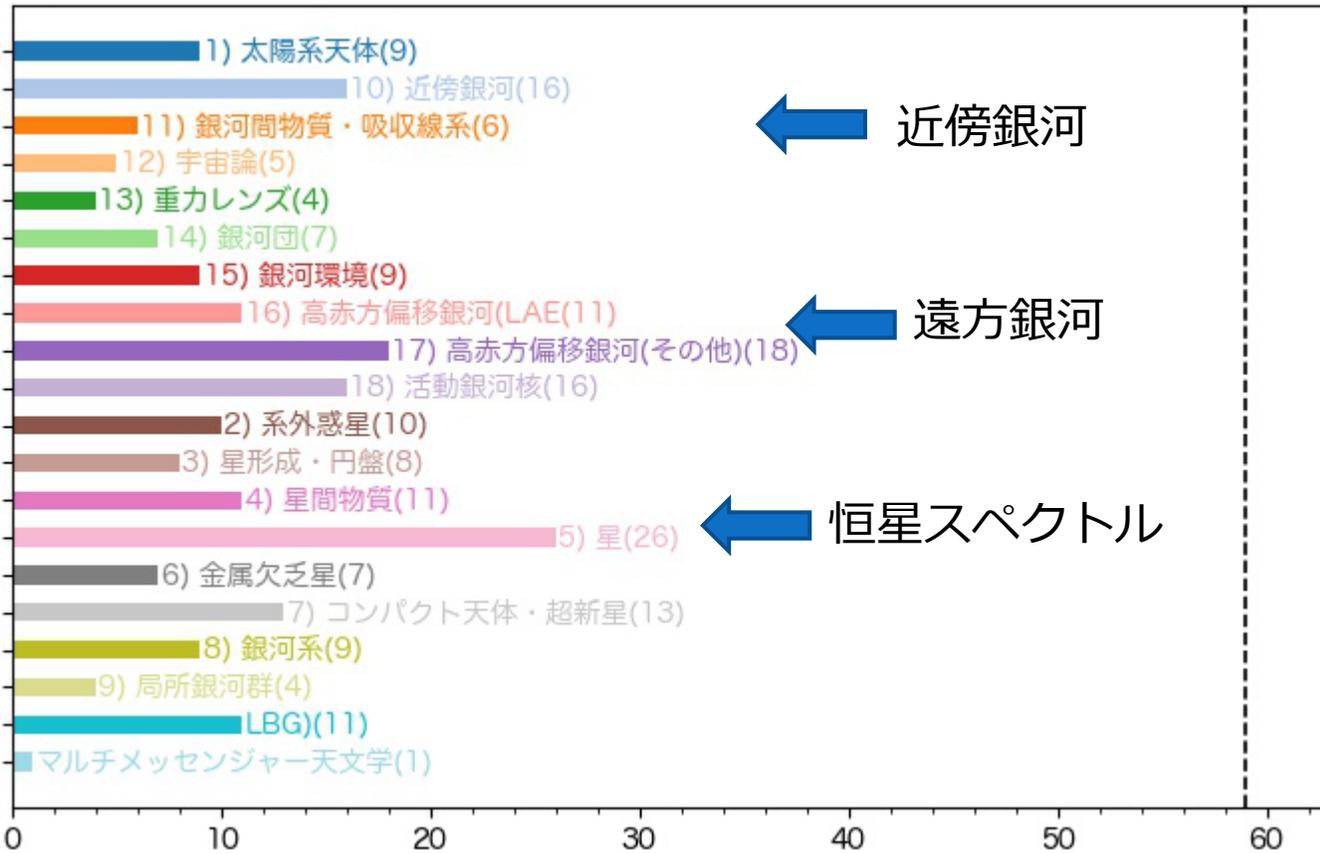
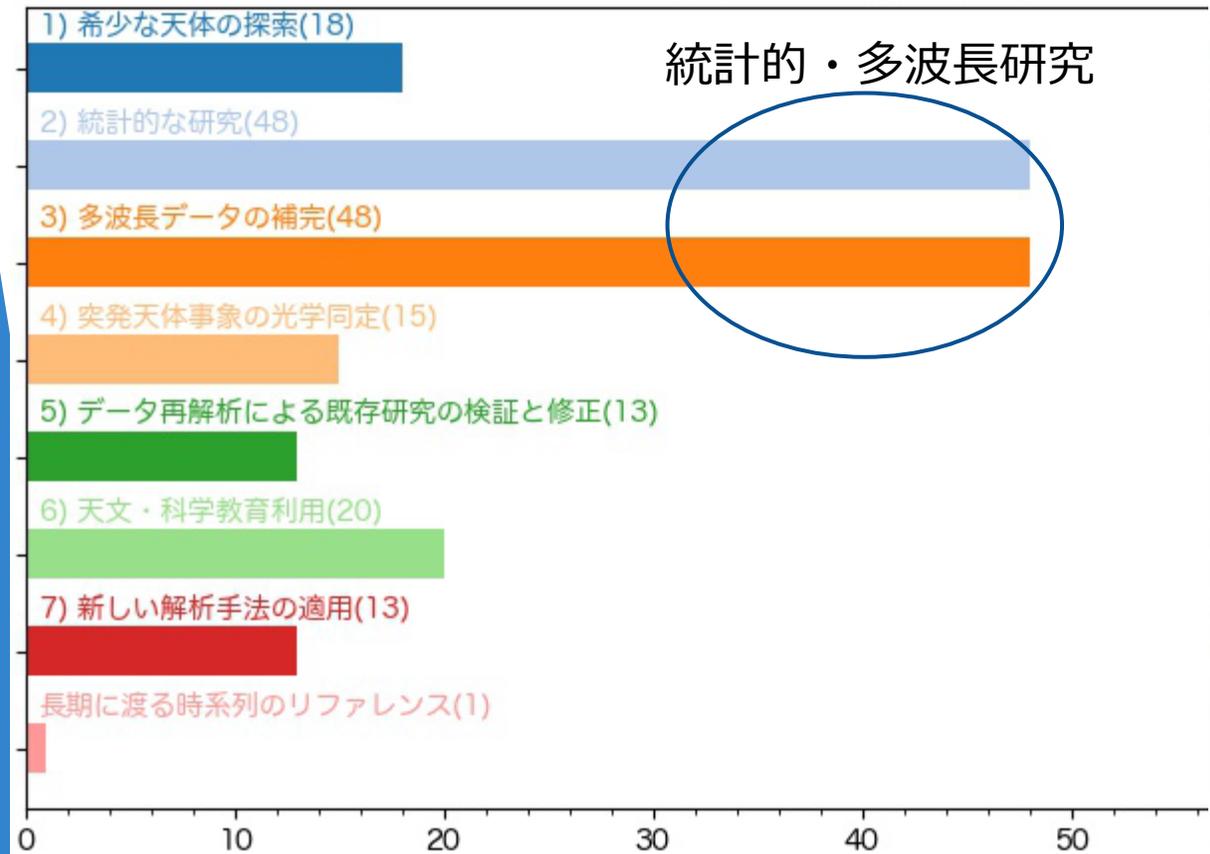
▶ 存在価値とは

▶ 多波長・追観測代替, 結果確認, **貴重なデータの保全保管**



光赤天連メンバーへのアンケートの結果

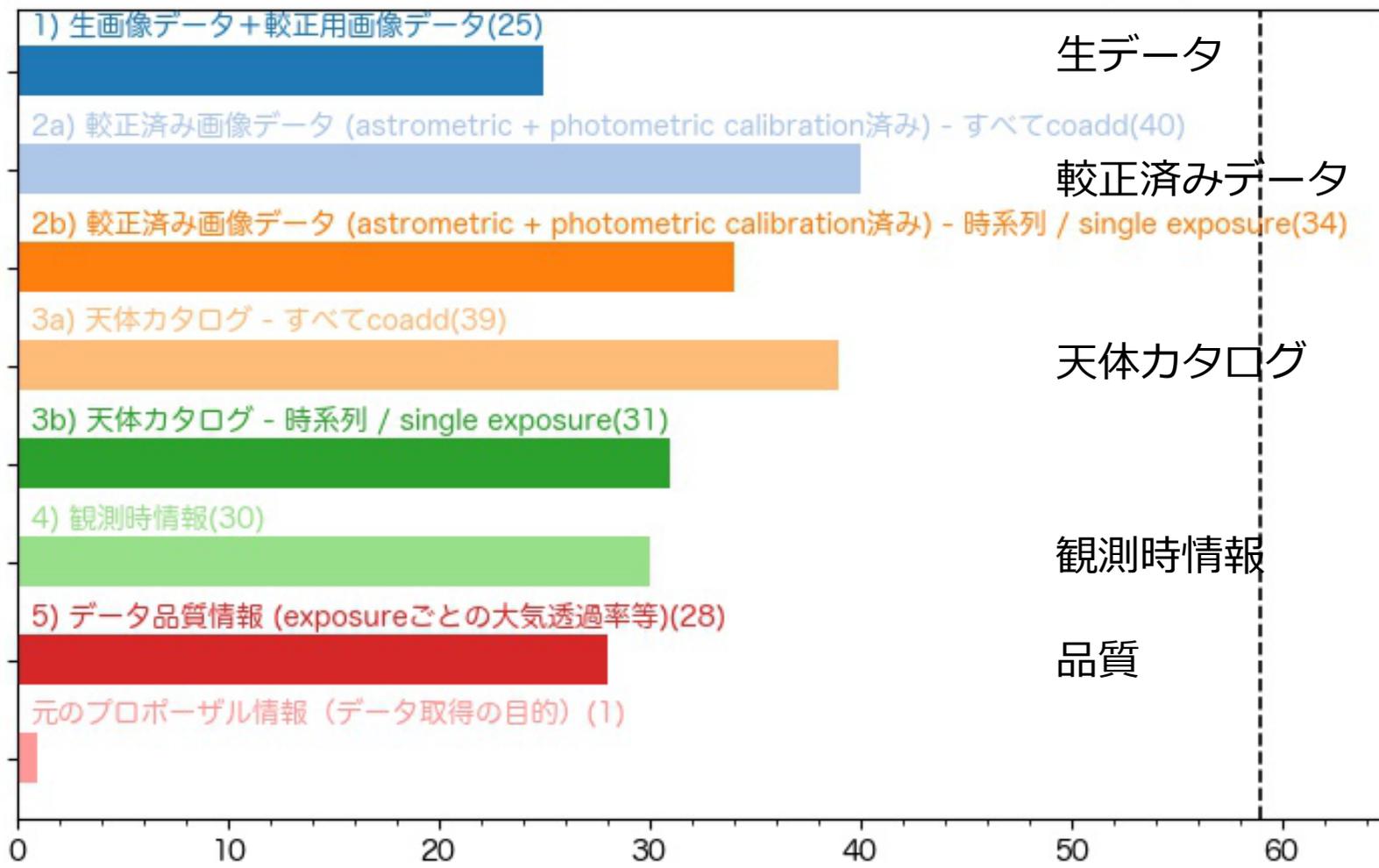
- ▶ データアーカイブを用いた研究目的と対象は
 - ▶ 多波長データ補完・統計的サンプルによる研究
 - ▶ 近傍・high-z 銀河, AGN, 恒星スペクトル



光赤天連メンバーへのアンケートの結果

▶ どのようなデータが必要か

- ▶ **生データ、処理済みデータ、品質情報、観測時情報、まんべんなく必要**



提案の骨子（1）

▶ 考え方

- ▶ **未来志向！**：必要な機能の維持 & 日本のデータ運用の改善
（生データ・処理済みデータの保管・戦略的利用の双方）
 - ▶ 既存プロジェクトに協力を強いるとか優先データを供出せよという話ではありません
- ▶ 利活用推進のため、**日本全体の観測データの集約と保全**を目指す
- ▶ 30年規模の**長期ビジョン**で全国的な視点に立って
データ運用拠点を（国立天文台）

▶ 意義と役割

- ▶ **日本の天文学にとって必要である**
- ▶ **データ保管と再利用（利活用）を進めるべき**
 - ▶ 科学の促進とエビデンスの保全
- ▶ **可能な限りデータを公開すべき**
- ▶ データは人類の文化遺産である。
ルールを決めて必要な情報を残す！

提案の骨子（2）

- ▶ データの選択について
 - ▶ データ提供者が最終的には決める
 - ▶ コミュニティの意見を聞いて決定する
- ▶ データの内容・状態について
 - ▶ 使える（解析できる）こと
 - ▶ 観測時情報・品質情報（公開を遅らせない）を付ける
 - ▶ 生データに加え、処理済みデータ・解析方法の提供を
 - ▶ 必要なデータを探しやすいこと

提案の骨子（3）

▶ 役割分担

▶ データ提供者（望遠鏡・装置、観測プロジェクト）

- ▶ データ・ヘッダの検証・整備（時間・座標・観測条件等メタデータ）
- ▶ 品質確認（公開を遅らせない）
- ▶ 解析実施・解析ツール・解析情報の提供

▶ データアーカイブ

- ▶ 検証の追認、修正の依頼
- ▶ 保管と公開
- ▶ データ利活用のための機能・情報・価値の付加

▶ ルール・インターフェースを決めて守る

- ▶ 設計段階からデータをどう作り保存し公開するのか、装置・アーカイブ間でコミュニティを含めて議論して決める
- ▶ 装置チームとも議論を始めた

使える状態のデータを作る
品質を評価する

データを使いやすく公開する
利用価値を高める

まとめと今後の目標

- ▶ 日本の天文学コミュニティのために必要なデータアーカイブを維持し、データ科学時代にむけて運用方法を改善することが必要
- ▶ 将来のデータアーカイブ運用について提案を作るためのワーキンググループを立ち上げます
 - ▶ 参加希望者は古澤まで (furusawa.hisanori_at_nao.ac.jp)
- ▶ データ提供者・データアーカイブ・光赤コミュニティで調整
- ▶ 2021年3月の光赤天連総会での声明採択を目指す

補足資料



望遠鏡グループへのアンケート（中間報告）

口径0.5-6.5mの望遠鏡を運用・建設中の15グループへ送付 9/13現在の回答10件

データ公開について（複数回答）

公開予定無し	0 %	
公開ないし提供したい	60 %	条件付き40%（サポートがあれば公開 20%、リクエストがあれば提供20%）
自前システムで公開	0 %	
外部システムで公開	60 %	既に公開 20%
外部システムの運用に協力の意思あり	20 %	
リダクション済みデータやパイプラインも公開ないし提供	20 %	条件付き20%（サポートがあれば公開10%、リクエストがあれば提供10%）

WS・提案書・声明にむけた議論への参加意思

控えたい／現時点では判断困難	30 %
可能な範囲で自分ないし自グループの誰かが参加	30 %
ぜひとも自分ないし自グループの誰かが参加	40 %

自由記載回答の一部

- アーカイブを前提として予算措置されたプロジェクトとそうでないものは分けて議論・対応すべき
- マンパワー不足の一言に尽きる

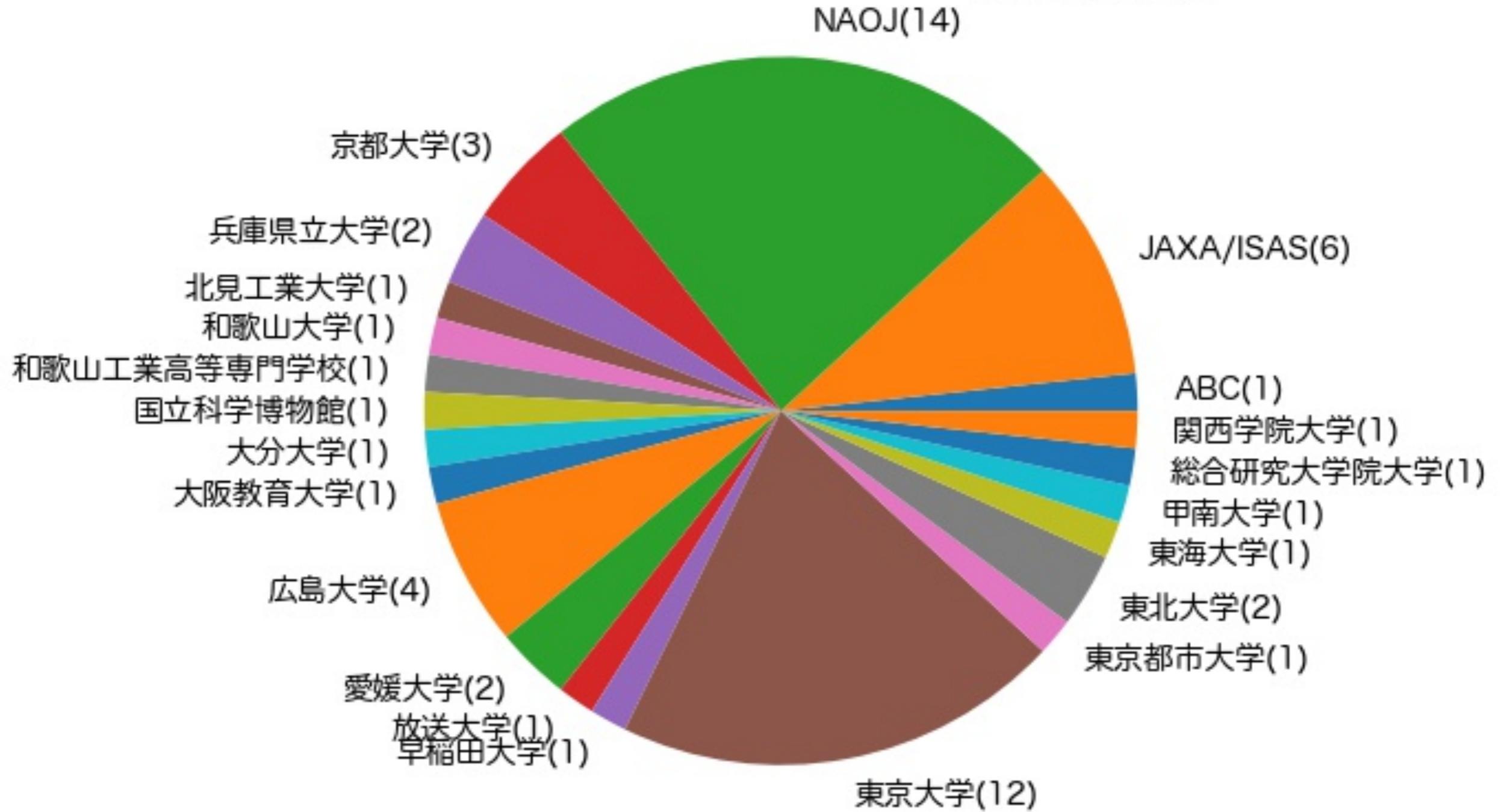
集計結果: データアーカイブに関するアンケート

ご協力ありがとうございました

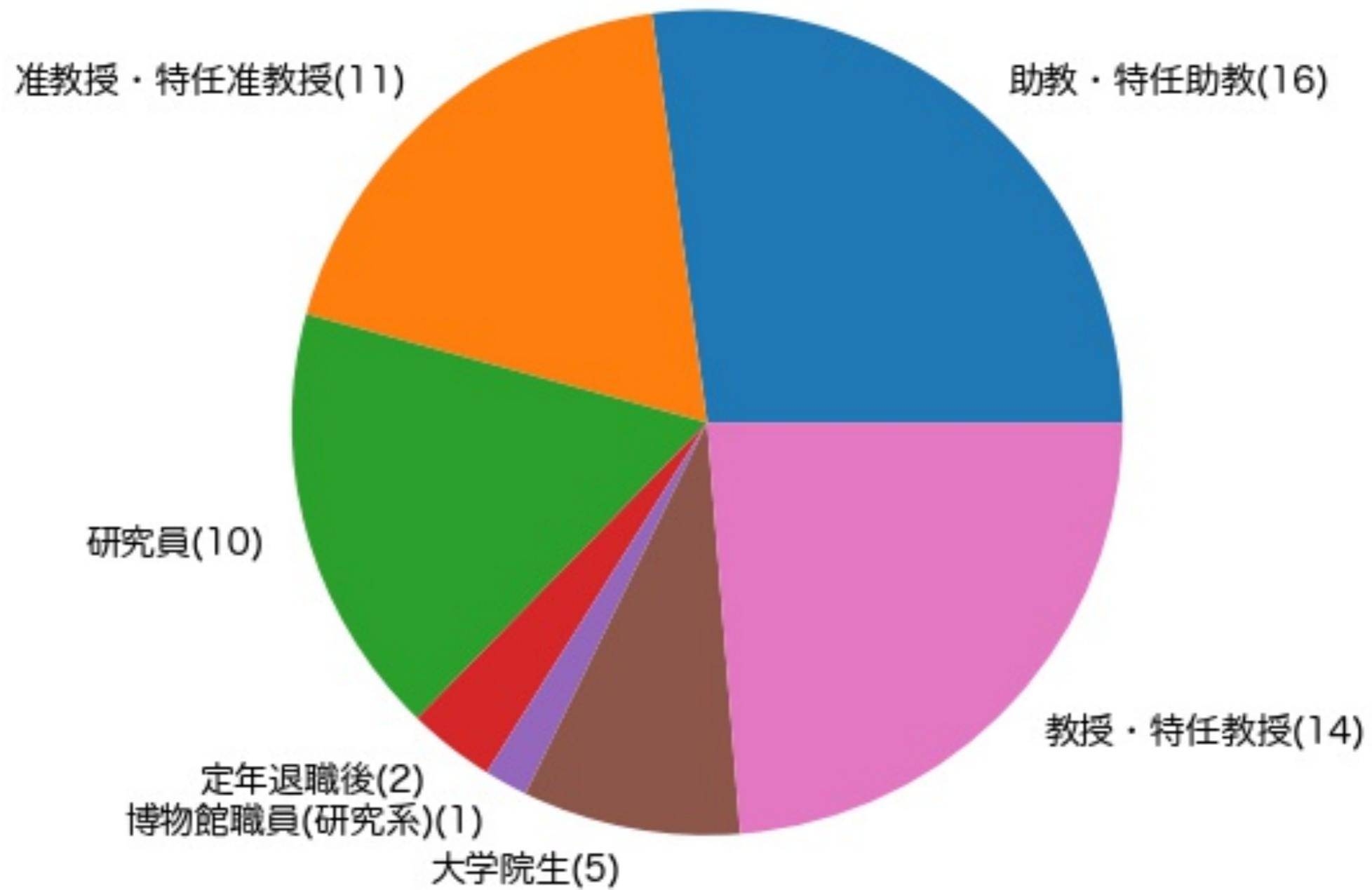
- [gopira:01176] データアーカイブに関するアンケートのお願い (2020/08/18)
- 集計期間: 2020/08/18-2020/09/02
- 回答者数: 59 (22機関)

Q1-2. 所属 (59)

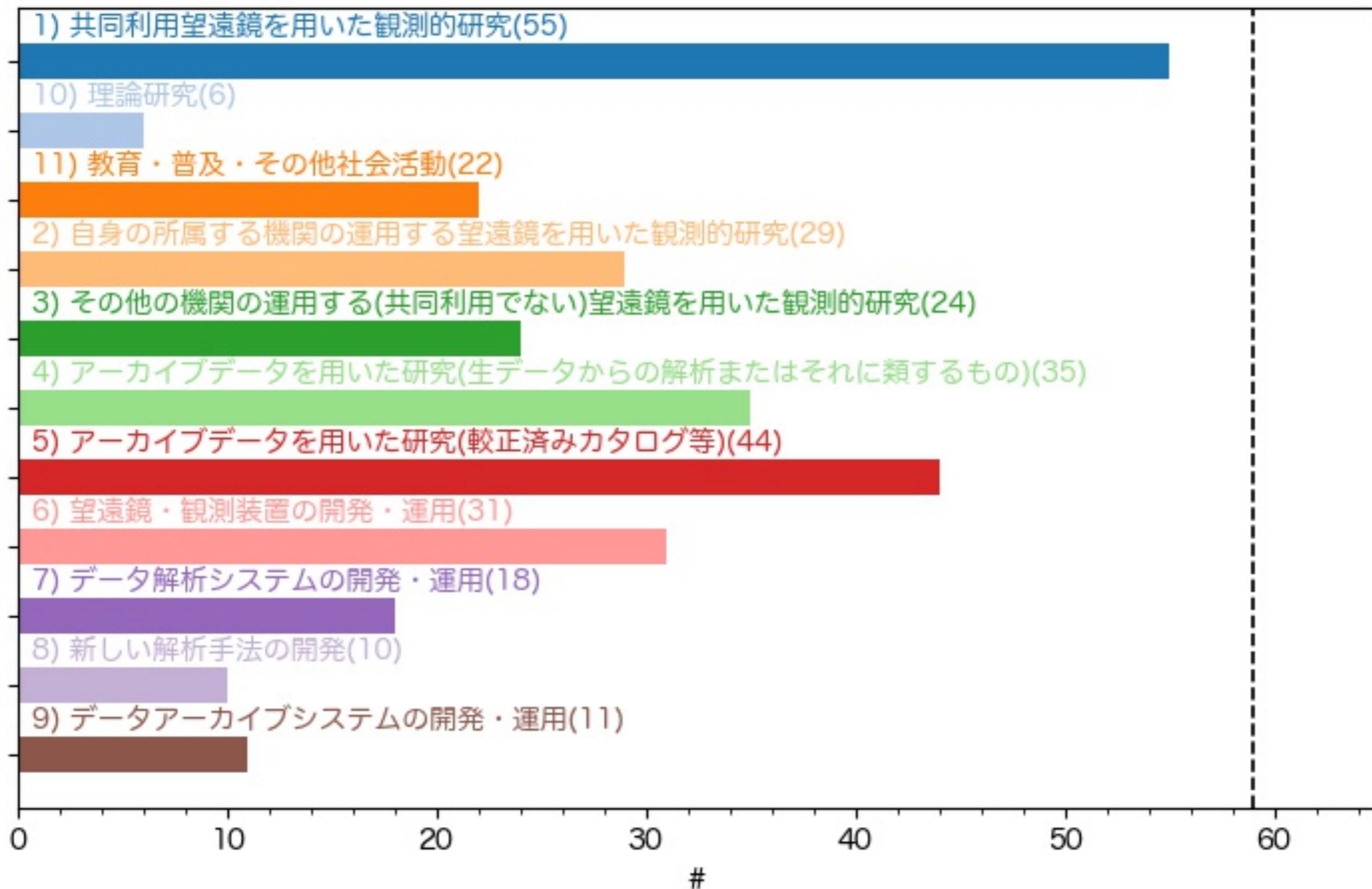
22 institutes



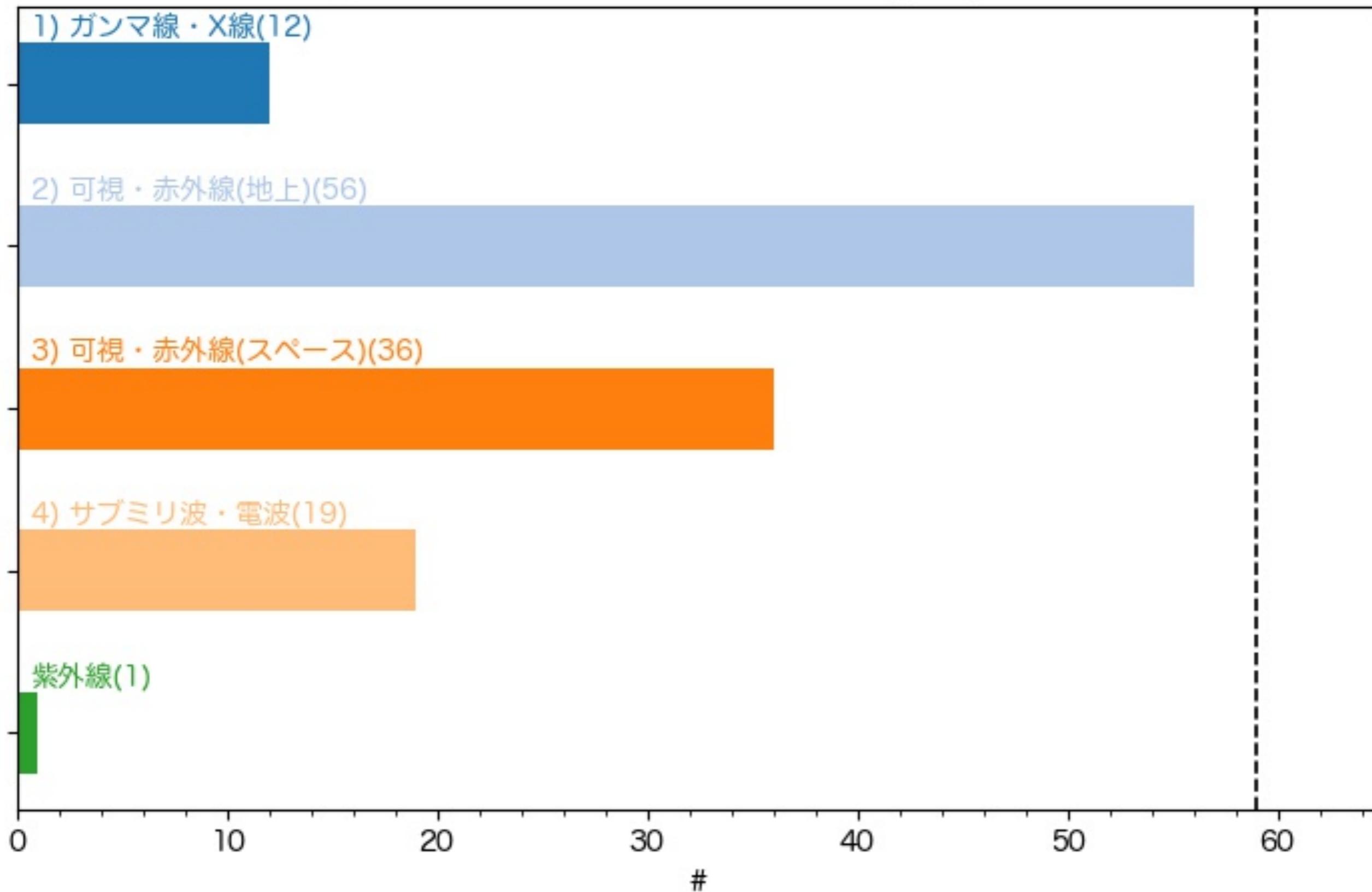
Q1-3. 身分 (59)



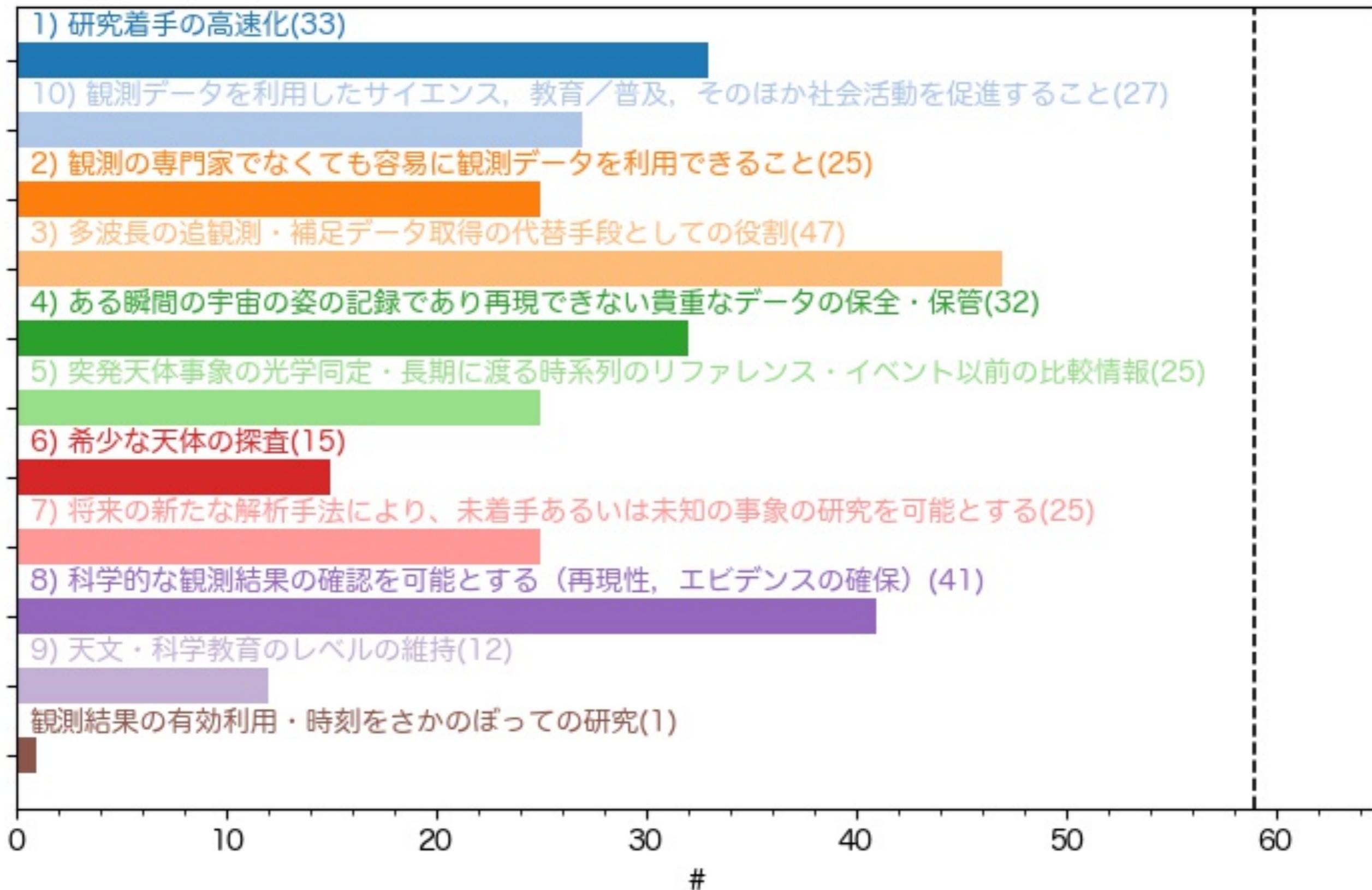
Q1-4. これまでに主に行ってきた研究・教育普及・そのほか社会活動の手法はどれですか？ (285/59)



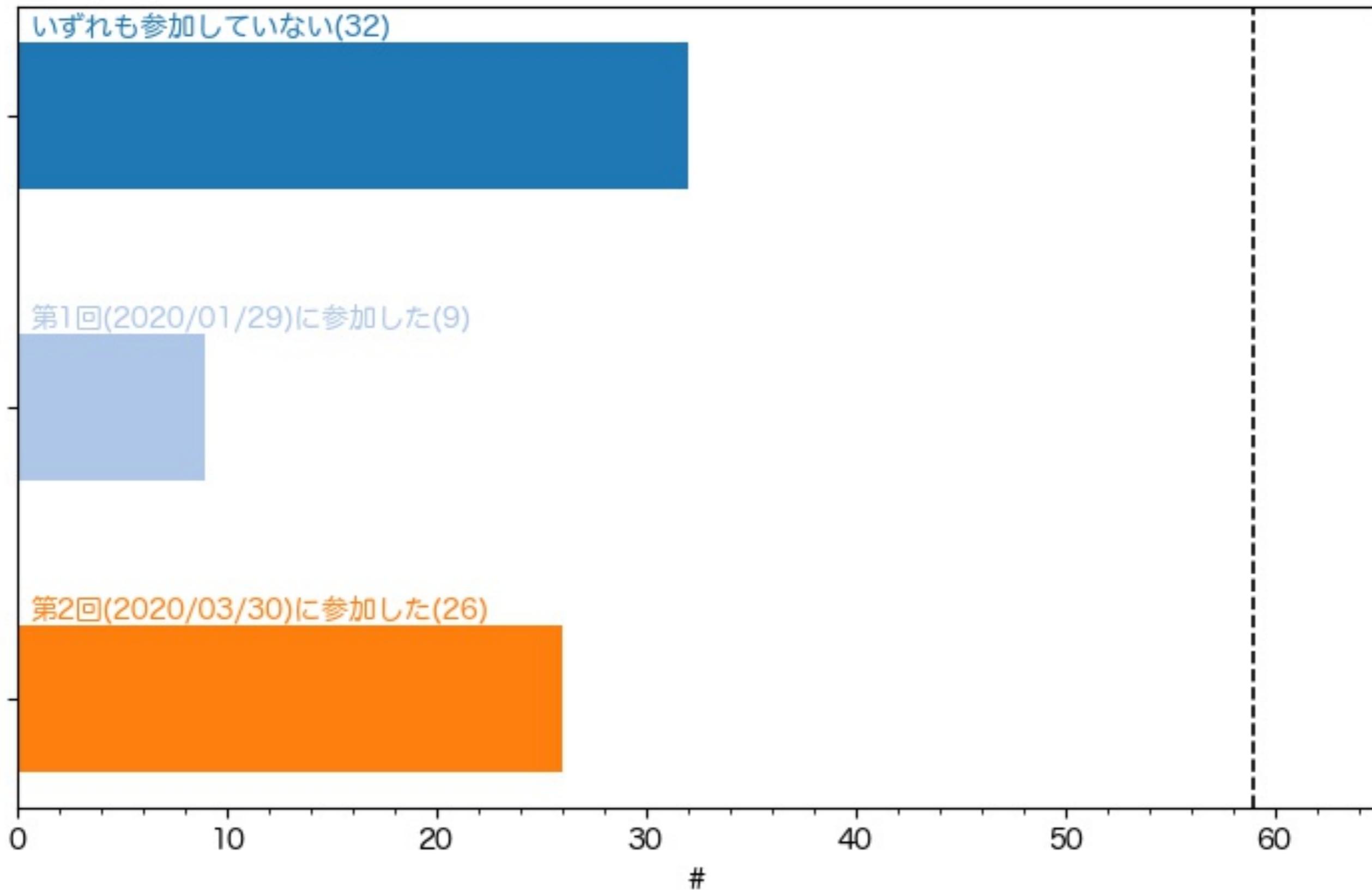
Q1-5. 扱ってきた観測データは以下のうちどれですか？ (124)



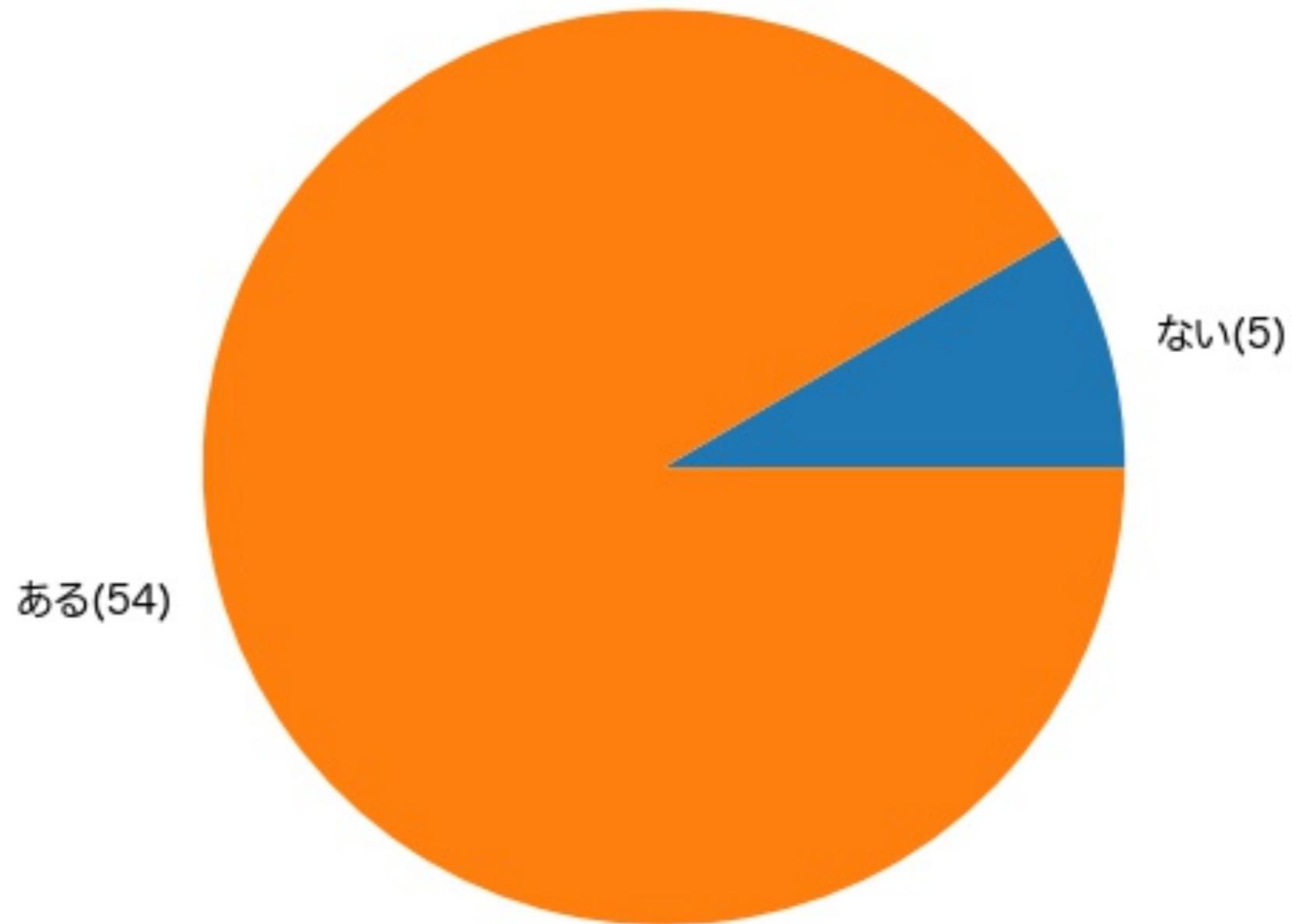
Q1-6. アーカイブデータの存在価値は何だとお考えですか？(複数回答可・最大5まで) (283/59)



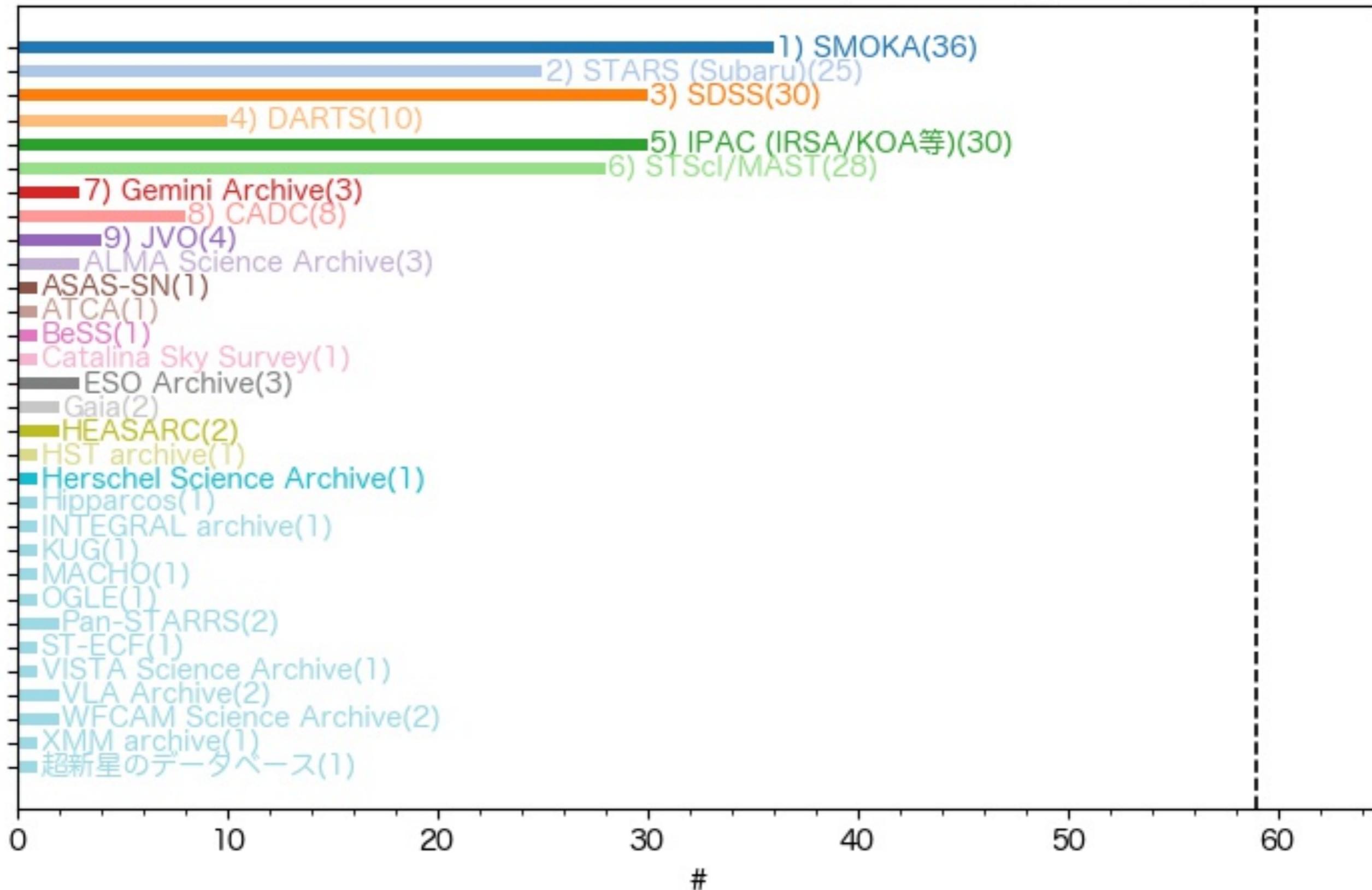
Q1-7. 2020年に開催したデータアーカイブワークショップに参加されましたか (67/59)



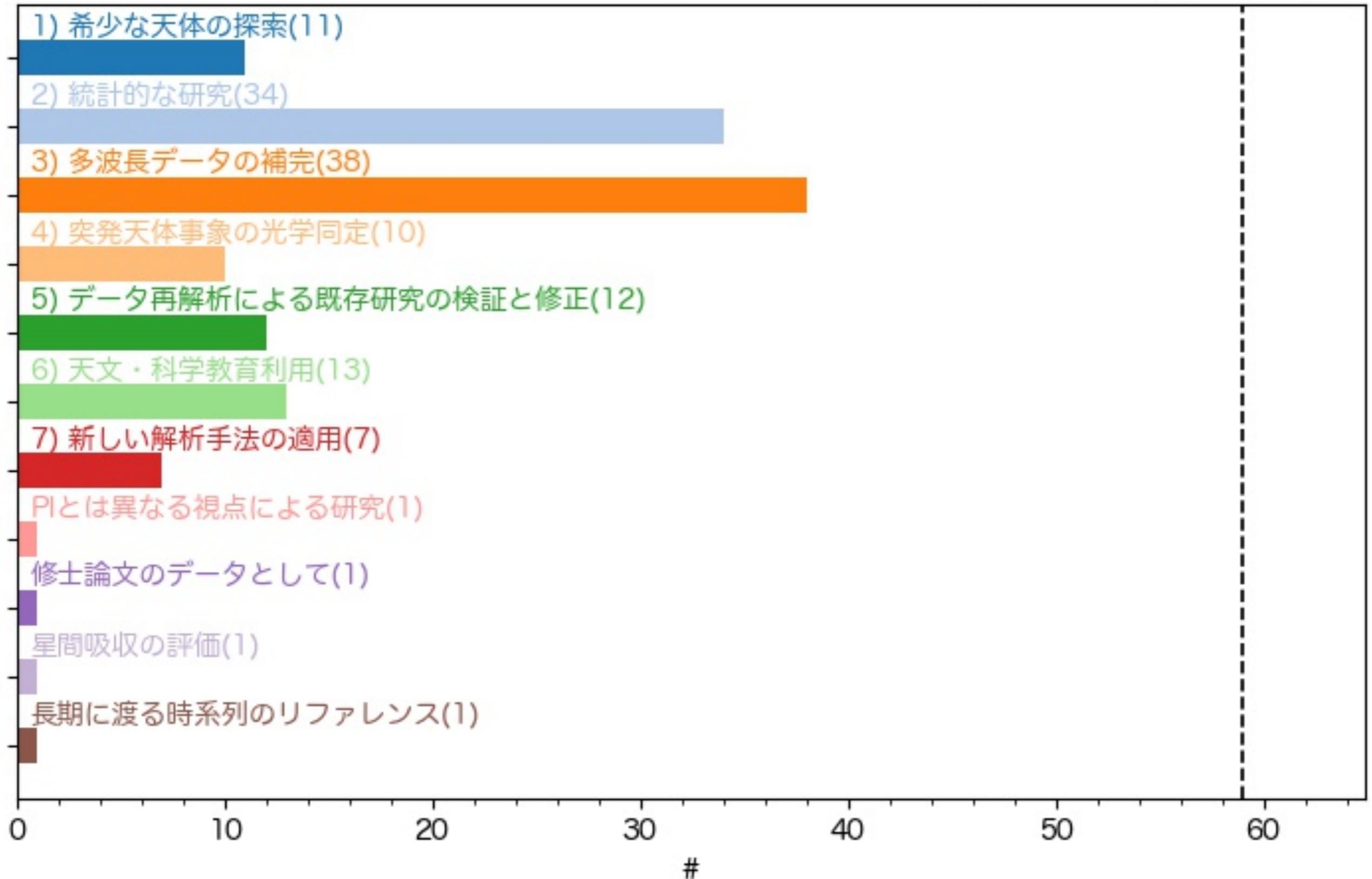
Q2-1. アーカイブを用いて研究・教育普及・そのほか社会活
を進めたことがありますか？ (59/59)



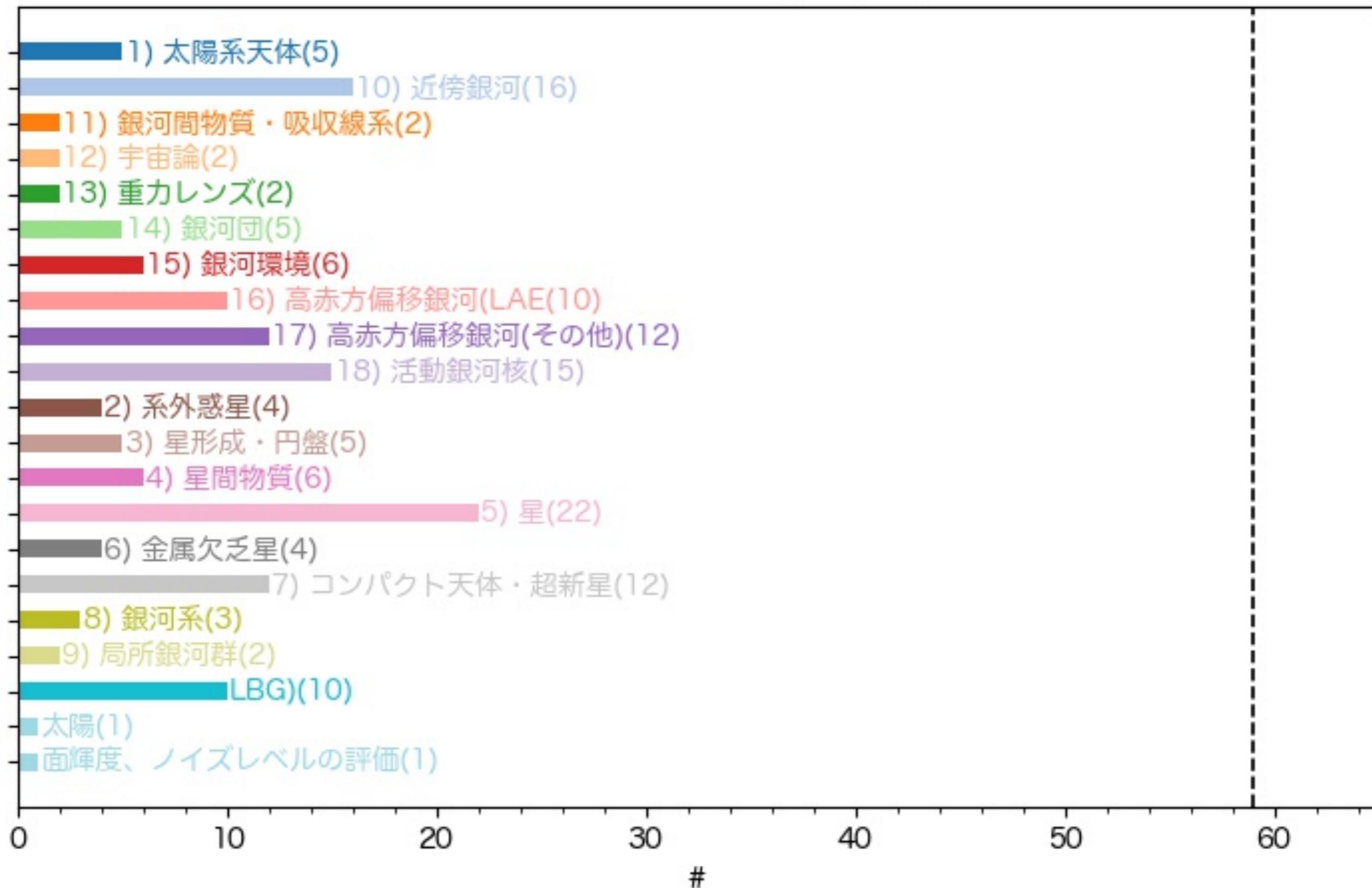
Q2-2. アーカイブを用いて研究・教育普及・そのほか社会活動を進めたことがあるにうかがいます。どのアーカイブを使用しましたか？(複数回答可) (205/59)



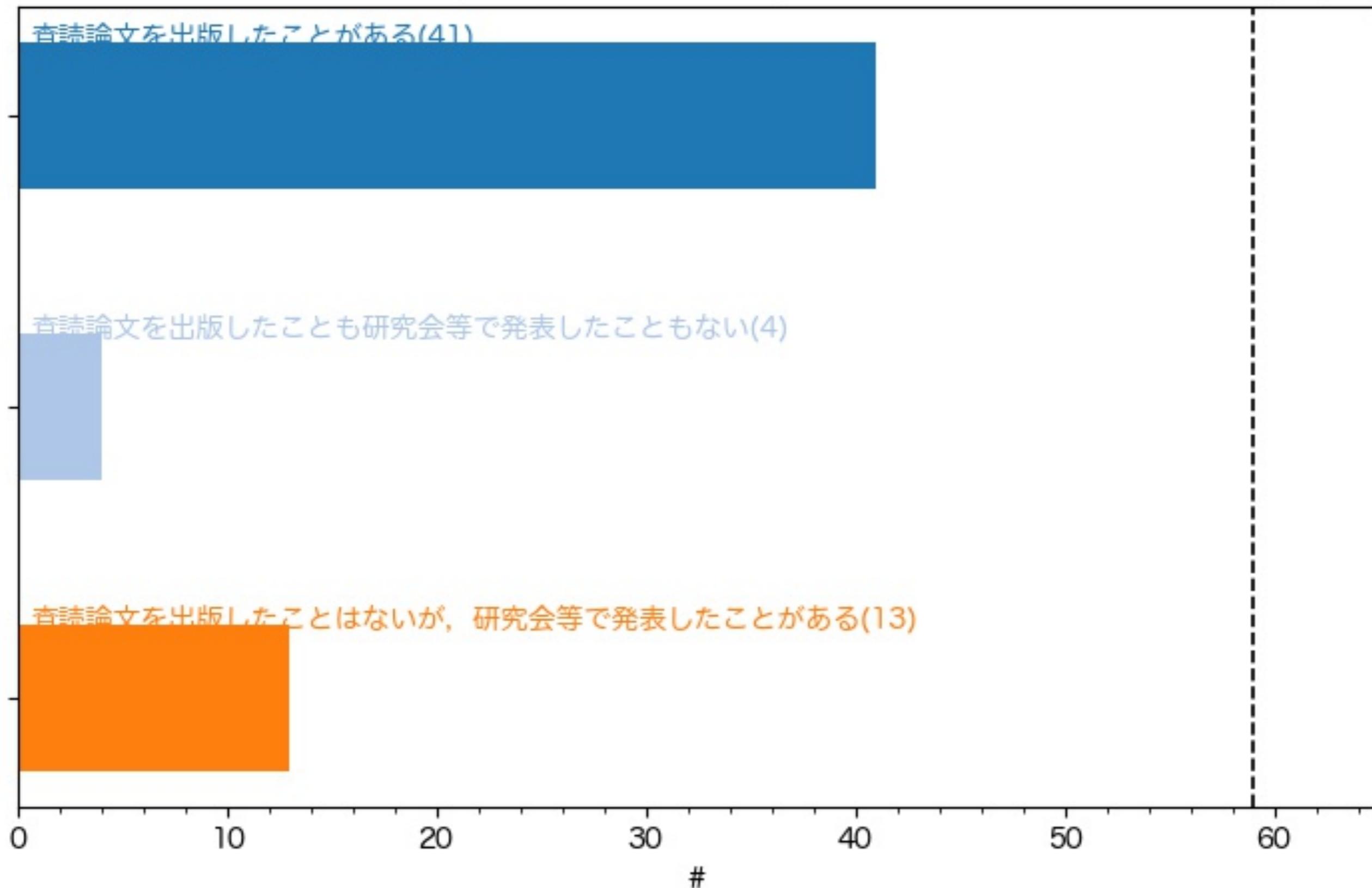
Q2-3. アーカイブを用いて研究・教育普及・そのほか社会活動を進めたことがあるにうかがいます。利用目的は以下のどれにあてはまりますか？(複数回答可) (129/59)



Q2-4. アーカイブを用いて研究・教育普及・そのほか社会活動を進めたことがあるにうかがいます。対象とした天体はどれですか？ (145/59)



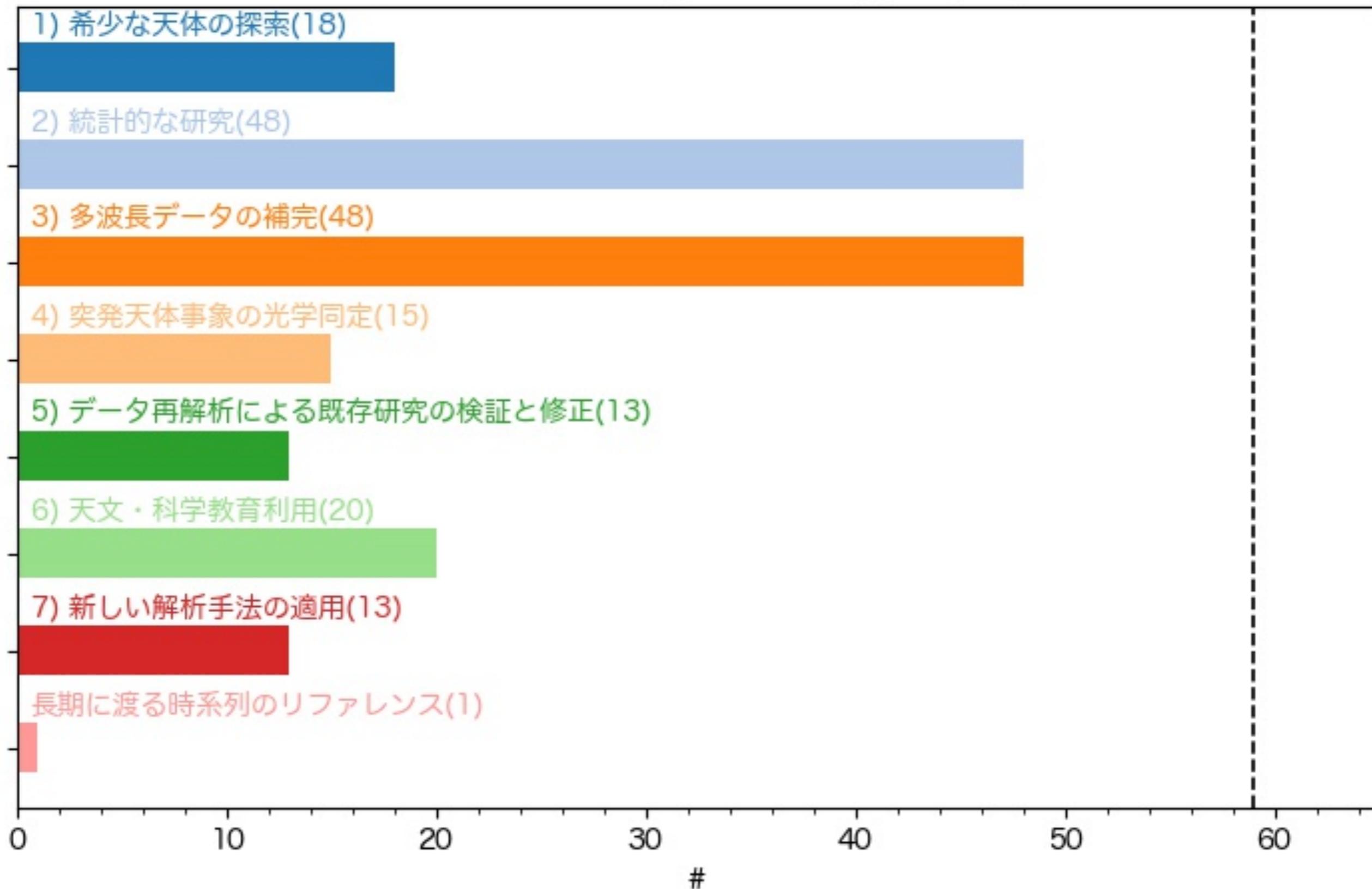
Q2-5. アーカイブを用いて論文出版・研究他の発表をされたことがありますか？ (58)



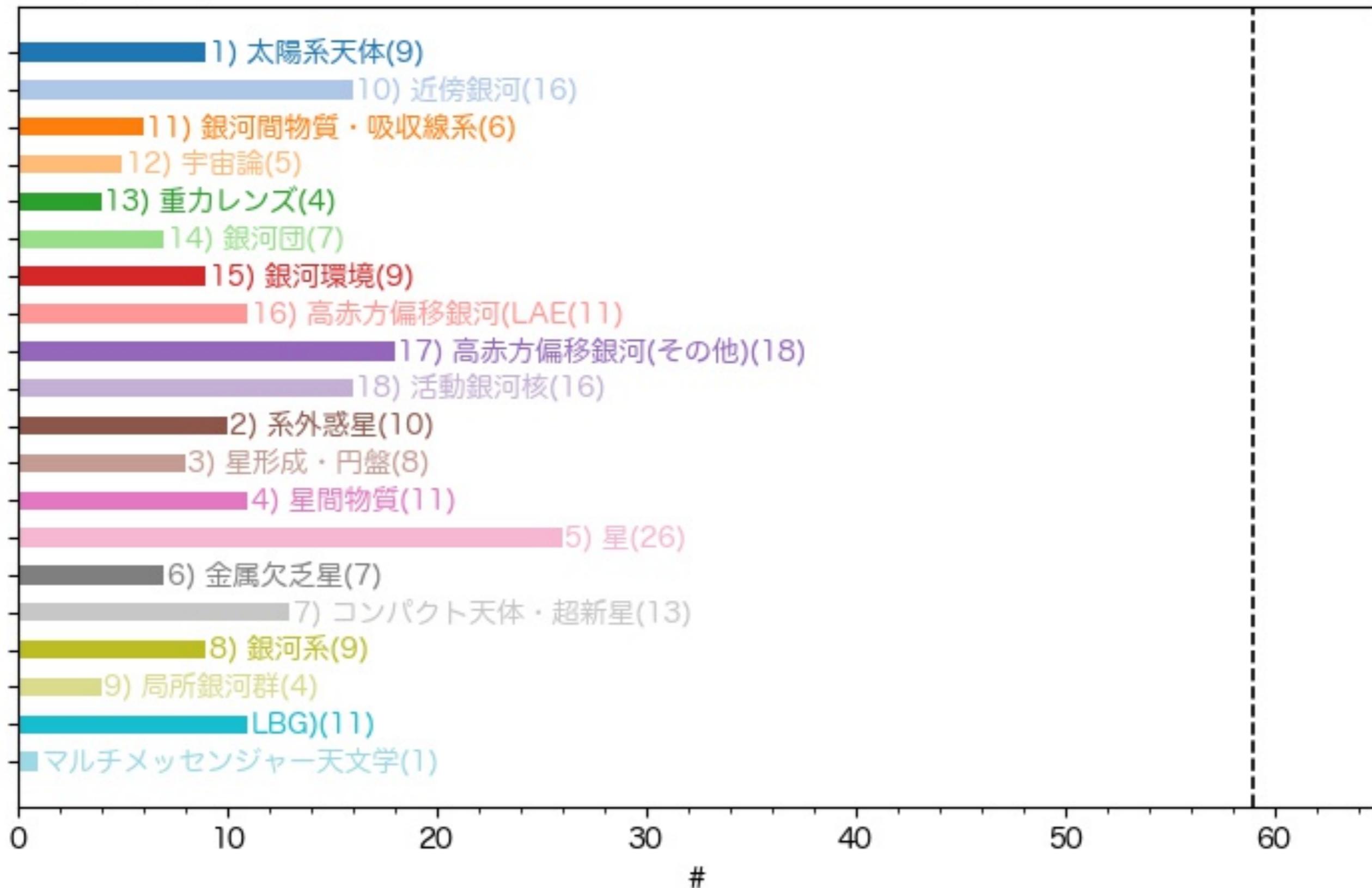
Q2-6. アーカイブを用いて論文を書かれたことがある方にうかがいます。出版した論文のDOIを教えてください。(論文ごとに改行してください)

- 125本
- 詳細を集計中

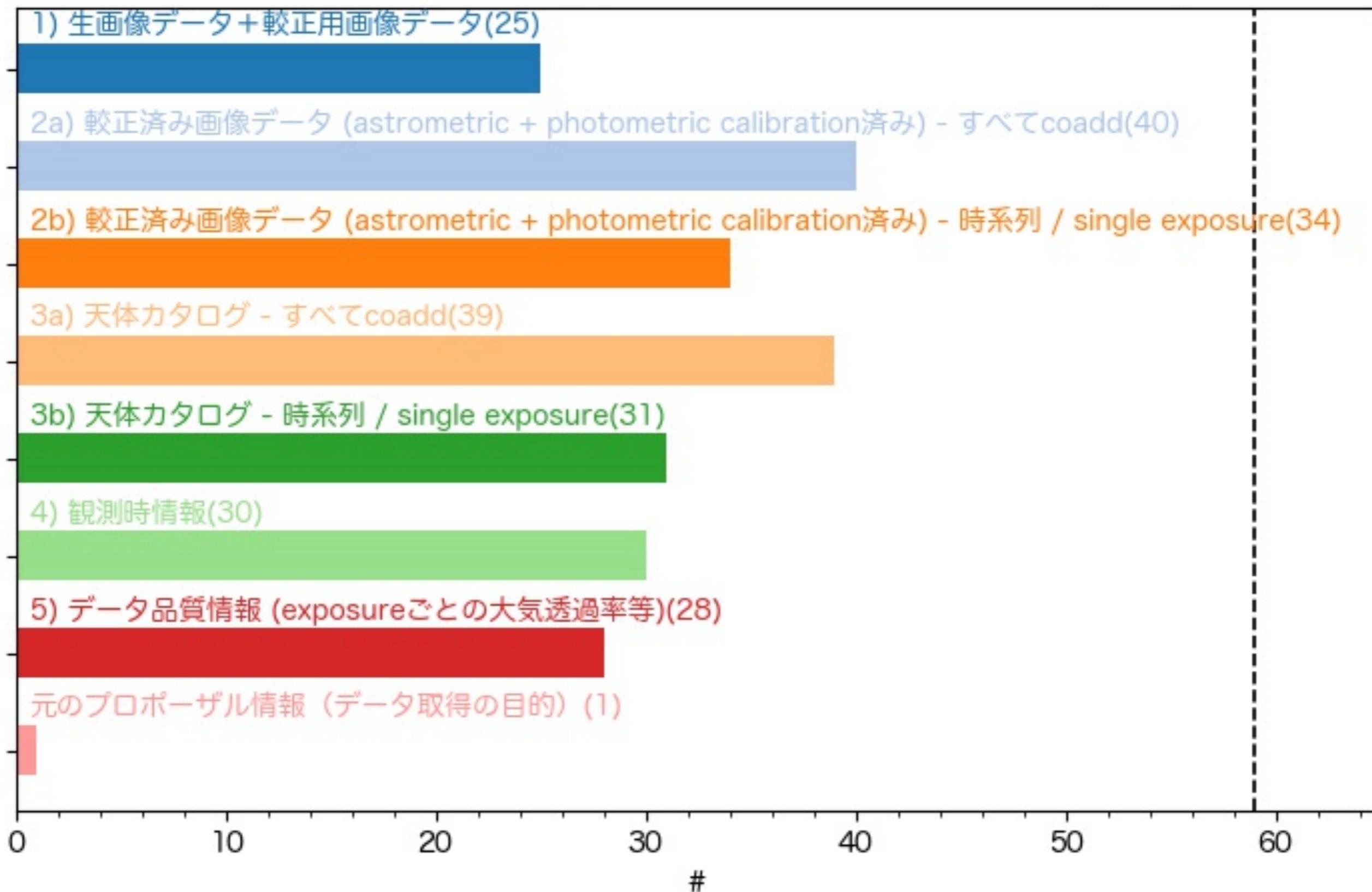
Q3-1. 今後アーカイブを用いて進めるようと考えている研究・教育普及・そのほか
会活動は以下のどれにあてはまりますか？(複数回答可) (176/59)



Q3-2. 研究・教育普及・そのほか社会活動に用いるため対象とする天体はどれです？ (201/59)



Q3-3. 上記研究を実現するために必要なデータの種類の種類はどれですか？ (228)



Q3-4. 上記研究を実現するために必要な検索取得機能は何ですか？ (467)

