データアーカイブ構築・維持の意義・重要性について

その意義や重要性と現状および近未来について

高田唯史(国立天文台天文データセンター)

今日の話題

- 天文データアーカイブの目指すもの
- 天文データアーカイブに要求されるもの
- 天文データアーカイブの現状について
- 今後のデータアーカイブについて
- 目指すべき方向性についての私見

天文データアーカイブが目指すもの

1. 観測者とは異なった発想・目的や解析方法、観測者が対象としなかった天体や波長、あるいは、複数の時間／波長／天体のデータの組み合わせなどによって新たな研究成果を創出すること。
2. 研究成果の検証を可能にすること（研究成果が画期的なものであればあるほど検証が可能であることが求められると思われる）。
3. 観測計画の立案、研究テーマの発案、ソフトウエアの開発・試験、教材の開発、データ解析の実習、演習や自由研究、など様々な活用によって研究・教育活動を進めること。
4. 観測者などの利用者への迅速なデータの提供やデータ保全（観測所アーカイブ、（サーベイ）プロジェクト等のアーカイブ）
5. 観測所内での利用による品質(データや観測そのものの）向上のためのデータサーバー(エンジニアリングアーカイブ)

1-3については　https://smoka.nao.ac.jp/about/mezasu.ja.jsp　からの引用

天文データアーカイブに要求されるもの

- 科学研究の推進に関するもの
  - ✓旬のデータなど、瞬発力が必要なデータについては処理済みや天体カタログなど、より、即座の研究に直結できるものの提供
  - ✓一方で、結果の再現性や測定精度改善のためには生データの管理保存は必要不可欠。
  - ✓なるべく早くデータを観測者に渡せるための仕組み
  - ✓観測所のスタッフによるデータを用いた様々な調査の補助具
- 社会（国民、政府等）からの要求に関するもの
  - ✓そもそもほとんどの観測所や観測装置は国税などの公的資金で構築されている事がほとんどであり、そこから生み出されるデータの可能な限りの有効(再)利用が求められている。
  - ✓得られた成果に関するエビデンスの提供
  - ✓これらを目的とした上での国策としてのオープンデータ・オープンサイエンスの推進が叫ばれているのが現状

・オープンデータ・オープンサイエンス

- オープンサイエンスとは（国情研オープンサイエンス基盤研究センター(https://rcos.nii.ac.jp/openscience/)より抜粋）

デジタル時代に鑑み、これまで以上にオープンで、多様な可能性をもって行うことができるようになった研究活動の諸側面の総称

  - ✓サイエンスはよりオープンであるべきという理念
  - ✓説明責任や透明性などの観点（主に行政サイドから来るが、自然科学においても当然要求されるものとなっている。）

- オープンデータとは？
  - ✓特定のデータが、一切の著作権、特許などの制限なしで、全ての人が望むように利用・再掲載できるような形で入手できるべきという考え方

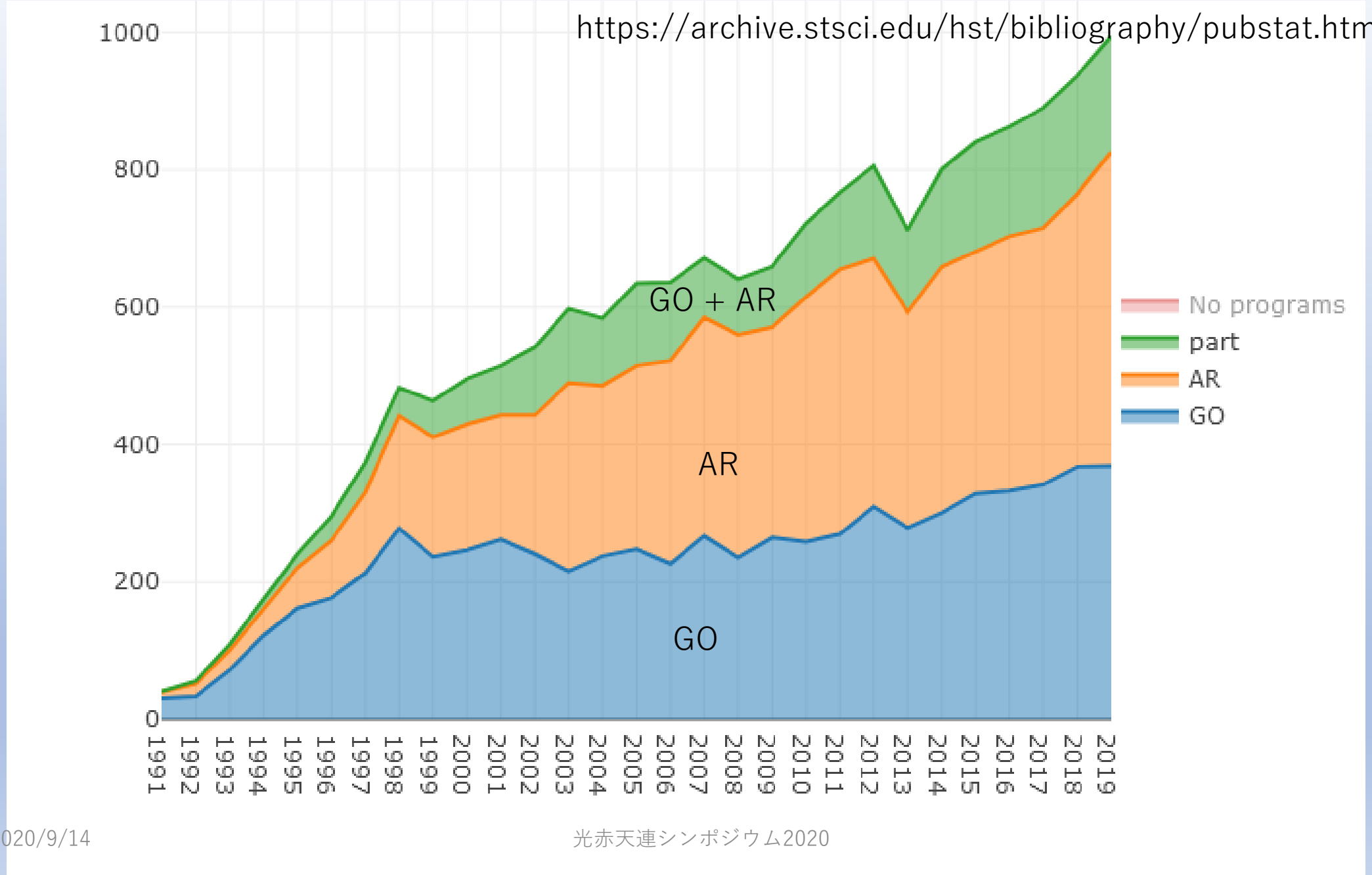
◆国の施策としてデジタル時代に備えた対応を組織的にしていこうとする動きは着実に進み、様々な方面での議論が行われているとともに、組織体系にも影響が出始めている。天文データアーカイブもその流れの中に確実に入っていくことになる。

  - ✓オープンサイエンスの推進について（文科省・科学技術・学術審議会総合政策特別委員会 H28.11.24）(https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu22/siryo/__icsFiles/afieldfile/2016/12/08/1380241_04.pdf)
  - ✓内閣府及び日本学術会議におけるオープンサイエンスに係る検討状況（https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/040/attach/1413786.htm）
  - ✓オープンサイエンスの深化と推進に関する検討委員会(日本学術会議)（http://www.scj.go.jp/ja/member/iinkai/openscience24/openscience.html）

天文データアーカイブの現状について

- 昔（例えば20年前）に比べれば、公開データを使って科学的活動を行う事に関する認識は変容して来た。
  - ✓アーカイブデータを主軸もしくは補助的に利用した論文数も着実に増えてきた（天文学だけでなく教育などの分野においても）
  - ✓論文にならないまでも、研究開始の事前準備の材料としてデータを調べる等の使い方
  - ✓観測の効率化を考えたduplication-checkのエンジンとしての役割
- 社会的な要求も変化してきている（米国（NSF)に似てきた？）
- データの巨大化とIT技術の急速な進化（技術的に最新レベルに追いつくにはかなりの投資が必要なものも増えてきている。）
  - ➢日本の例としてはすばるHSCはその典型例。今後の（広視野）CMOSセンサーカメラなどのデータについても、（一定数の利用者が見込まれるのであれば）その対策が必要。
- 予算的な問題（特に継続性について先が見通しにくい。予算の中長期的な見通しがないと難しい。）

HSTのデータを使った論文のアーカイブデータ利用の割合



Bordelon+2016 in SPIE Conf.

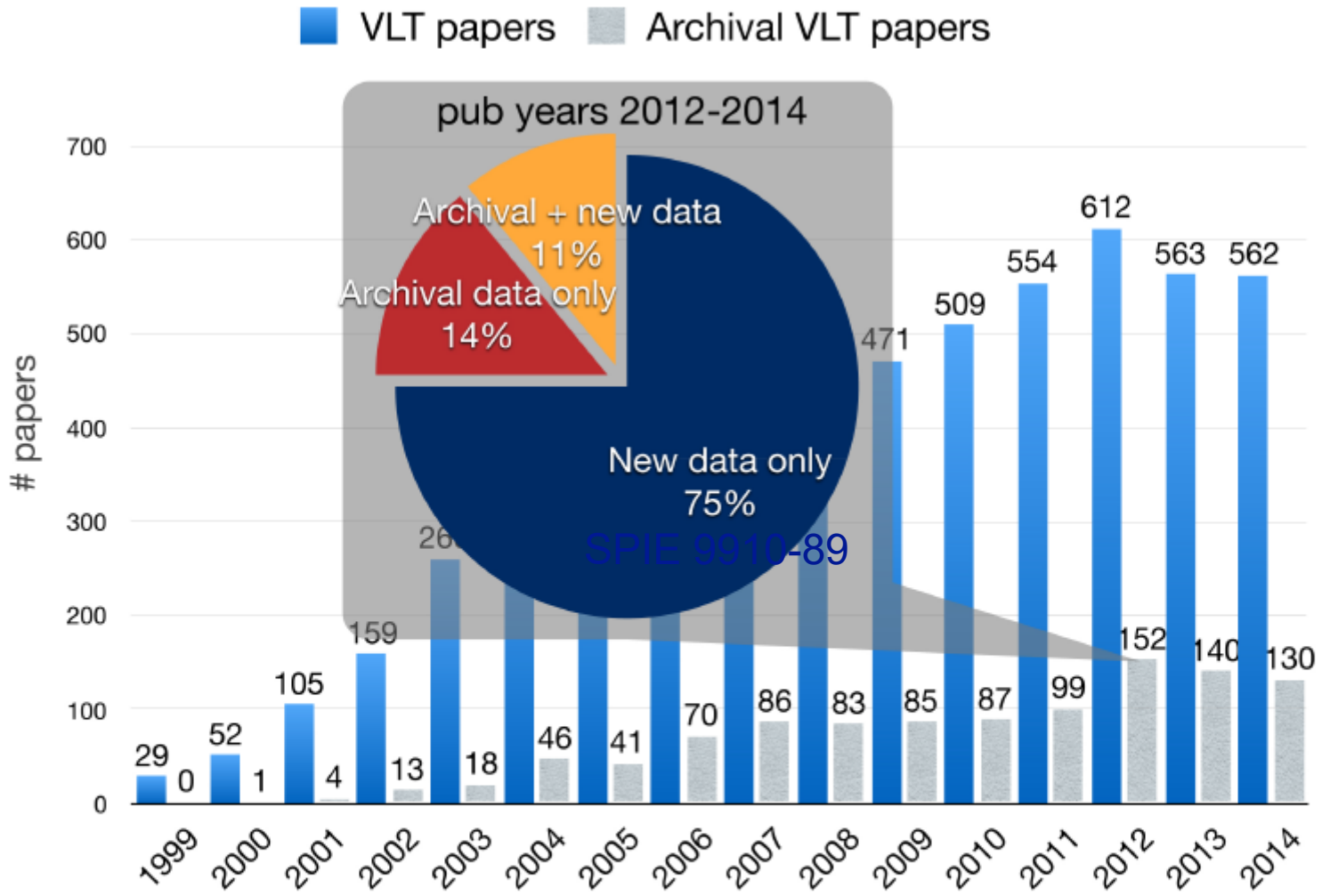


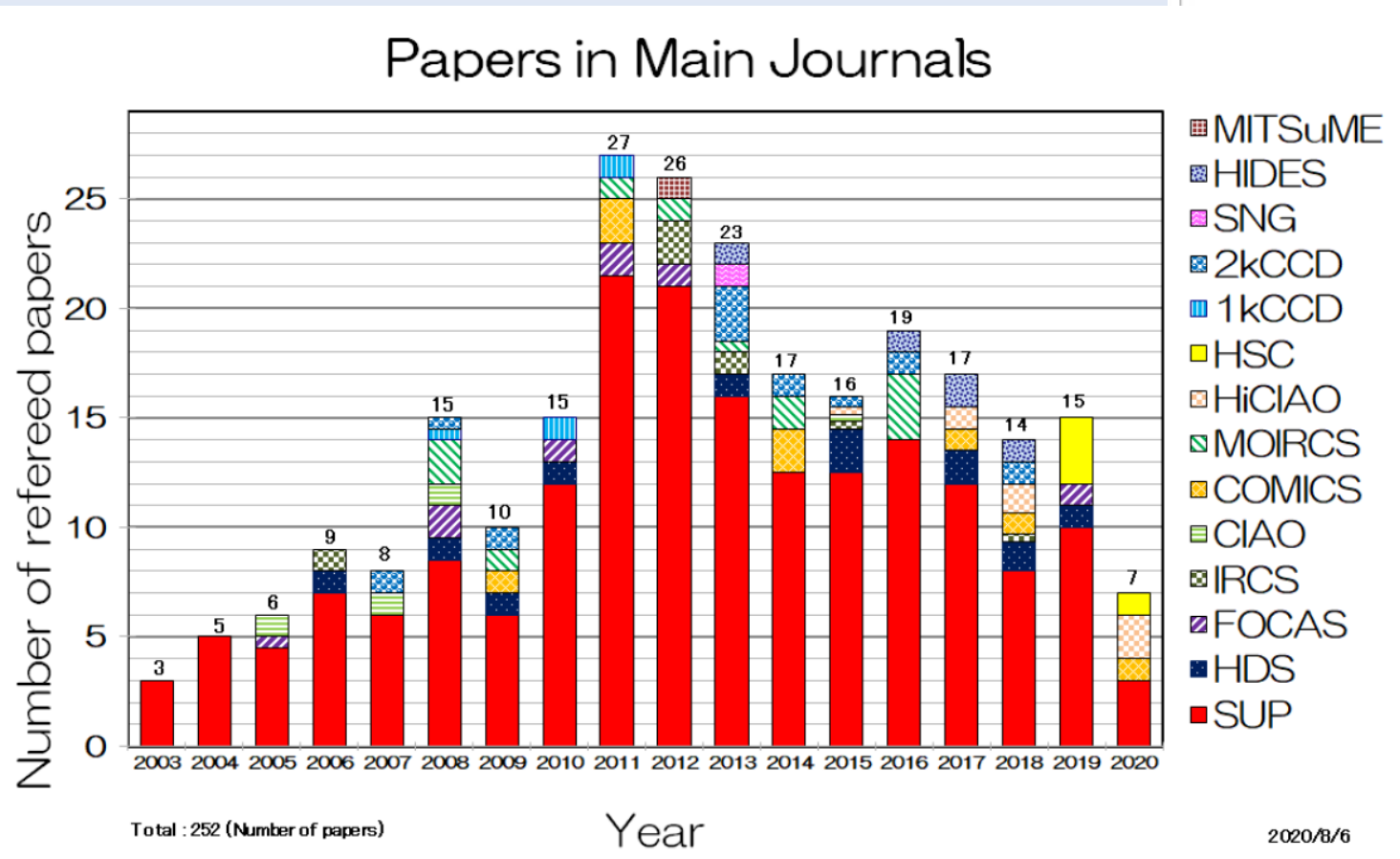
Fig. 2: Number of papers using archival data

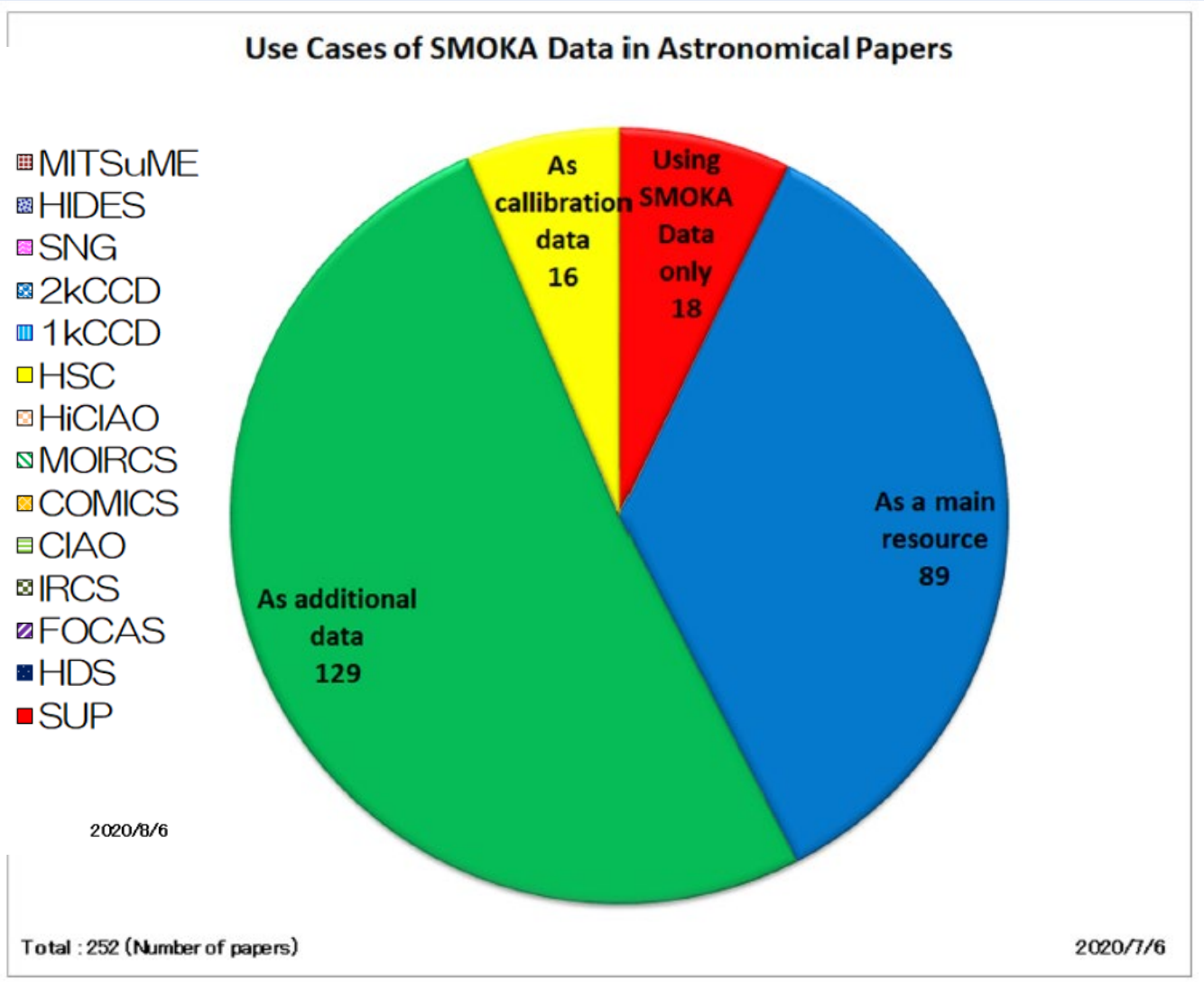
The fraction of archival VLT papers has steadily increased to a level of approx. 25% in publication years 2012-2014. Of these, approx. 13% use exclusively archival data, while approx. 11% use archival as well as proprietary ("new") ESO observations. (Total: 5,970 papers)

SPIE 9910-89

日本の光赤外線天文アーカイブの例　　SMOKA（Subaru Mitaka Okayama Kiso Archive system）

https://smoka.nao.ac.jp/





今後（しばらく）のデータアーカイブの行く末

- **科学的な要求：** (解析)結果の再現性の確保、エビデンス提供の必要性は、今後さらに重要度を増すであろう。
- **社会的な要求：** 公的競争的資金による研究にデータ公開の条件がついたり、外部による監査などでもエビデンス確保のためのデータの管理については毎年のように質問されるようになってきている。
- 日本の天文コミュニティーから世界に向けてしっかりとオリジナル・データを提供し続けることで、世界の天文コミュニティーへの貢献、世界中のデータを使うことに対する正当性も保持できる**（give＆take）**。
- 一方で、研究者たちの研究活動環境も変化を続けており、装置開発者が継続的(永続的？）にデータを自分たちだけで保持・管理・提供し続けることも困難。
- データアーカイブに関する、国立天文台および関係する各機関の予算は厳しさを増している（運用費なので競争的資金には不向き、運営費交付金の減少も関連）
- アーカイブの開発者・運用者の育成も継続性が必要（システムの担い手にも魅力あるものにする必要（技術またはサイエンスにおけるメリットややりがい））

- 今後、どのように**「それぞれの関係者に多大な負荷は負わせない形で」**データアーカイブを天文研究の一道具として構築、維持、利用していくのかについては、コミュニティーにおいてもある程度の意識のすりあわせ、その後の協力体制の構築は必要不可欠であろうとの認識を個人的には持っている。

目指すべき方向性(私見)

- 各観測データについて、**データをアーカイブするために最低限装備するべきものは何なのかの明確化**とその実現（アーカイブデータのシステムへの取り込みの効率化（各観測所でのルール作りやそのテンプレート化））
- 各観測所の運用においての**データ保全(ファイルとして維持以外に、中身の保全（品質向上など）についても）の重要性**の再認識と体制維持
- とりあえず何年間データをオンラインで維持するべきなのか、またその後の扱いをどうするのか？（たとえ永続的に持つにしても、一定期間後に新しい装置のデータで大部分の用途が置き換え可能なものは、よりコストの低いデバイスでの管理を行うなど）**これをアーカイブを始める前（観測が開始される前）に決めておき、**数年単位などで必要に応じて見直す等の対応をする。
- **天文学者を中心にしたデータコンテンツの強化は必要**。（サーベイ）データ品質の改良の可能性はいつも模索し、必要に応じて戦略的な取り組み。(近年の例で言えばすばるHSCの戦略枠観測など)
- **日本の光赤外線データアーカイブの中心的な役割を担う機関（またはグループ）とその支援やその活動方針の議論などのための体制の確立**（コミュニティーとしてその必要性や存在意義を常日頃から考えて行くことは今後より一層重要となる）。それは世界の天文学に対する貢献にもなるという意識・自負は持ち続ける。

観測装置設計の段階からアーカイビングを意識したデータ生産を考えておく事の重要性

例えばKeckアーカイブ（KOA)の関係者もこんな事を言っている。

Complete keywords to support archiving.
Perhaps the best lesson is that in designing the instrument and its data product, the data should be acquired with the archive in mind. As such, it should be ensured that the full complement of KOA keywords conform to proper standards and are incorporated in the FITS headers well before the instrument enters operations on the telescope.

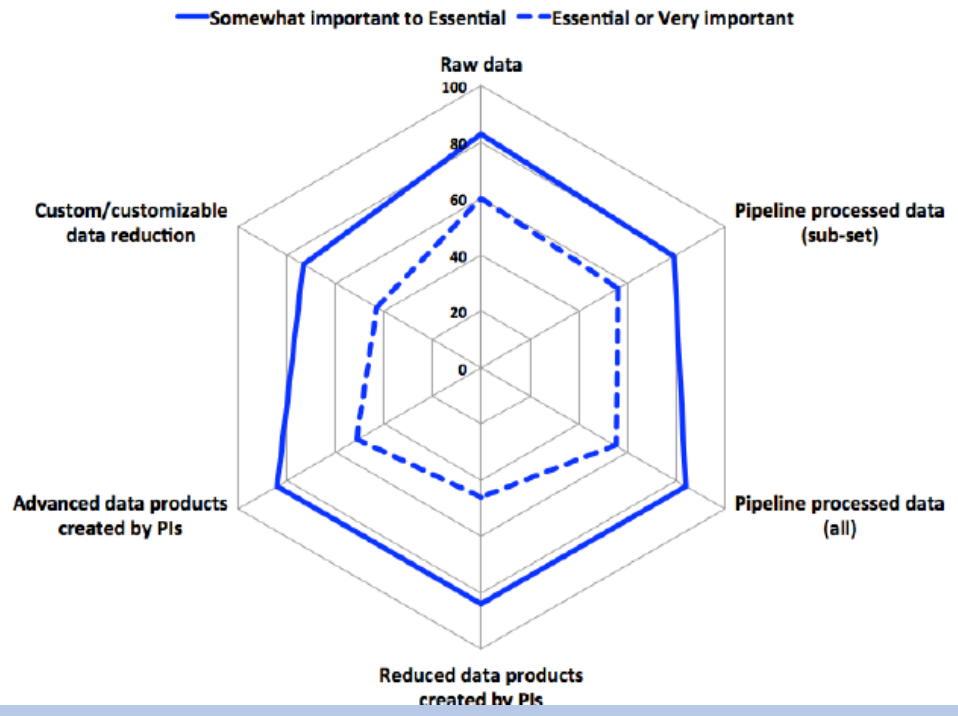
Berriman et al. “Data and Metadata Management at the Keck Observatory Archive” in ADASS XXIV @ Calgary, Canada (ASP Conference Series Vol 495)

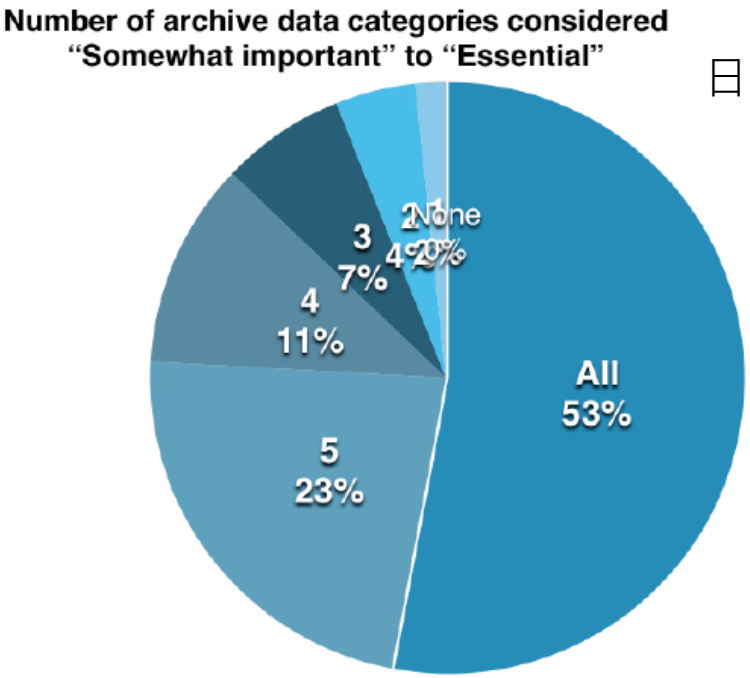
まとめ

- 科学的、社会的観点からみて、データアーカイブによるデータの有効活用はより一層重視されてくる。
- 生データのアーカイブは必須。品質改善や異なる解析法の導入を見越す。
- より活発な科学的成果のアウトプットのためには、品質管理された処理済みデータの提供は必要。但し、相応の戦略的取り組みが必要であると同時に、あくまでも、一つの解析方法での結果に過ぎないことも念頭に置く必要がある。
- 一方で、データの巨大化、予算状況、関係者の研究・運用環境の変化によって、個人もしくは小さなグループの努力だけでは、継続的なシステム構築、運用は難しくなってきている。
- 観測装置・もしくは科学プロジェクトを立案する段階からデータの保存管理を見越した計画が必要不可欠である。
- 継続的なデータ管理には、データ供給側・管理する側の合意のもとでの一定の責任分担が必要。
- ユーザーの要求をすべて満たすことは困難な場合もあるが、貴重なデータを紛失するような事のないように、データの保全のためのノウハウを如何に共有していくか、またその中心的存在としての国立天文台や大学共同利用や大規模な共同研究を行う機関について、その支援とともに、運用状況の確認や方向性の議論などを定常的に持ち続けることは必須。

Polling the community: ESO2020 questionnaire

- "How important is access to the following sorts of archived data products in order to maximize your scientific productivity?"
  - Statistically (very) significant sample: 1439 answers





アーカイブの存在意義について、ユーザーから見た意見についての一つの例

日本の場合はどんな感じ？

From Freudling @ ESO Calibration Workshop 2017