背景理解と本日の検討内容

2020.1.30

日本の局面・NAOJの局面

- 国
  - オープンデータ・オープンサイエンスへの要請
    - 産学巻き込んだ議論，具体的にはあまり進んでいない
  - 4機構連合体への動きと共同利用機関の位置づけの変化
    - 4期中期目標期間：個々の大学では整備・運用が困難な最先端の大型装置、貴重な学術データ等、全国的な視点に立って共同利用・研究に供する
- NAOJ
  - 運交金削減（＋使途の考え方の変化）・フロンティア経費化で運用費削減の要請
  - データアーカイブ効率化の要請
- 効率化の手段
  - データ階層化，優先付け-> なかなかアーカイブになじまない
  - 計算機内製 -> マンパワーがいるが非常勤職員中心で不十分，HPC経験不足
  - クラウド -> 旬のデータほど大きく良い解がない

各データ拠点の情報提供

- SMOKA）
  - 生データ公開の意義）異なった発想や方法での利活用，検証，観測立案・教育，論文化の動機，ドキュメント整備の動機
  - データの保全と公開
    - 観測装置が多く（>30）統一されていない
    - 生データは保全
  - データ修正・テスト
    - 装置は変化する＋トラブルは起きる＋限界がある
    - ヘッダ修正，補填，品質評価
    - 装置とのインターフェース：各所でデータ整備に対する事情が違う
  - 人員
    - 5人->2人->4人（ただし新人多い）
  - 状況によらず観測データは公開すべき
  - 観測前から準備すべき．さもなくば膨大な手間と時間がかかる．

続き

- VO)予算措置ー外部資金？コミュニティ全体の予算取りの枠組み
  - Asterisk, Escape?
- すばる）
  - データは観測所に帰属する（慣習的合意）
  - 処理済みデータも本来は公開できると科学成果が増える（例．HSTなど）
- 宇宙研DARTS・AKARI）
  - データポリシー
    - 自ら取得・整備するデータに適用
    - 結果を再現するための情報，エビデンスの公開
  - プロジェクト：期間・目的がある．国民への約束.
    - データ整備）適切なデータ処理，データの説明
  - アーカイブ：定常業務。プロジェクトと責任分分界の覚書
    - データ保存，容易な検索，識別子，外部機関との連携
    - 再現できないデータを長期間(30年），再現困難なデータは出来る限り保存.
  - AKARI）プロジェクト終了後にデータ処理・整備(1.5億円/5年)
    - 系統的な処理の体制づくりが課題（プロジェクトの進行と適切な人員など）

続き

- 京大・せいめい）
  - 装置が増える＞6-9
  - 半分はNAOJ共同利用でありその方針に巻き込まれる
  - しかし大学設備にはストレージなどお金・置き場もない->NAOJに期待？
  - データ整備に向けた検討が進んでいない
- 東大TAO）
  - 一番大きいデータレートは600GB/night程度
  - 共同利用でもある．データは保存公開したい
  - できれば一次処理データの公開は有用
- 東大Tomo-e Gozen）
  - データ駆動型天文学で科研費やインフラを誘致
  - 30PB/10年-> 3-10％の（ピックアップ・精製）データを長期保存する
  - 短期間のデータ利用は東大データセンター（アーカイブはしない），長期保存はNAOJへ
  - 社会・教育への利用，産学連携

続き

- HSC)
  - 装置，解析，アーカイブの連動をプロジェクト初期から行う
  - データ・運用を作りこんでから観測を開始したい
  - プロジェクトの進化->拡張と慢性的に人不足
- 岡村さん)
  - データは社会の一部・文化遺産という考え方が必要
    - DBを活用して天文研究を進めることと文化遺産としてデータを残すことは別のこと
      - どのように理解してもらいリソースを確保するか
    - 何をどのように保存するのか
  - ビッグピクチャーを描く
  - 30年後のビジョンを見通し（技術，体制，コスト），ロードマップを描く
  - 成果が上がる・上がった具体例が欲しい（経済界のSDSS，LSSTへの関心）

続き

- 広島）
  - SMOKA：データ・観測時の状況把握に便利，データ保管場所としての意義
  - 再利用による研究成果が課題
    - データの複雑さ，キャリブデータの紐づけ，パイプライン・観測メモの整備（カレンダー掲載の努力）
  - 大学での研究データ保管・公開の環境整備の検討がある
  - 個別大学でのアーカイブは非効率，引き合いのあるデータを戦略的に公開しては？
  - 光赤外線大学間連携の次期2022.4-の主テーマにデータアーカイブ・公開？
    - NAOJへのデータ集約の代替案として大学でコンソーシアムを組んで管理公開
  - 商業論文誌のオープンサイエンス化が強い：コード公開の要請
- 世界の状況）
  - 20人規模で開発運用
  - 機関・コミュニティからのサポートが得られている
  - コミュニティの中でのデータ拠点としての位置づけと予算措置
  - 高い論文文生産率

議論の話題

- （日本の）データアーカイブの意義と役割
- （日本の）データアーカイブの進むべき方向
  - 課題・目標
- 議論したいこと
  - プロセッシングとアーカイブの目安となる予算規模

意義・役割

- 必要性
  - 世界のトレンドはチェックするためのデータ情報を残すこと
  - 使えるデータの公開
- 保管
  - 生データ（整備済み）
    - ある観測の天体現象は再現できない．気づかなかった情報，将来の解析の可能性
    - エビデンス，コード・環境，付随文書の保存
    - 中期保存のニーズ
  - 処理済みデータ
    - 時間がたってソフトが動かなくなることもあるので処理済みデータは重要
  - 解析の再現ができること
  - 運用経費の期間の問題・予算規模の目安による
  - 装置・観測プロジェクトによる違いがある
    - サーベイは再利用できることが重要・捨てるデータ
- データの再利用・処理済みデータによる科学促進
  - 生データだけを置いておくのでは難しい → 処理済みデータ

サーキュラーに載せた話題

- 残さないといけないデータとは何か
- データの順位付けはどう行うべきか
- データはどのような状態にして保存・公開すべきか（保管場所、利用しやすさ）
- 共同利用機関の位置づけ、役割とは
- 大学・データ生成側チームとデータアーカイブ側の役割とは
- （特に外部経費による）装置開発におけるデータ公開について
- どう要請されているのか、チェックがどう行われるか、今後どう対応していくか

上以外に議論したい話題

- データアーカイブの役割は何を重視すべきか
- 生データと処理済みデータの優先度、重要性の考え方
- プロジェクトやデータの規模の違いによるデータアーカイブの扱い方
- 現在のデータアーカイブの使用感と今後データ・データアーカイブをどのように使っていくべきか
- 技術的な見通し
- 評価方法について
- 体制的な課題
- 現実的なアクション

課題

- どのようなデータを公開すべきか，それをどのように決めるか
  - 順位付け？どのように決めるか（生データ，処理済みデータ…）
  - 論文を書いたデータ
- どういう状態でデータを公開すべきか
  - 論文が書ける状況
  - データ一覧・リダクション済みデータ公開？品質評価
- データアーカイブの体制
  - 国全体の観測データの集約？
  - 共同利用機関の位置づけ，大学等の役割
  - 技術，コストの見通し
- 観測所/装置チーム（データプロバイダー）とデータアーカイブの役割
  - 検証，修正，データ処理，公開
  - その進め方，ルール，インターフェース，ドキュメンテーション，MOU
- アーカイブ/解析をするための準備，データコミッション
- コミュニティ・組織・予算措置者への説明と正当な評価をどう受けるか
  - 科学雑誌・グラントのデータ/処理情報の公開の要請
  - 社会・教育利用，産学連携

Q. どのようなデータを公開すべきか，それをどのように決めるか

- データの種類
  - データプロバイダーがそのデータに関わるコミュニティの意見を聞いて決める
- 順位付け？
  - 上と同じ．ただしアーカイブとも相談
- 誰がどのように決めるか
  - 観測所・データプロバイダー
  - 予算プロバイダー
  - コミュニティの要求は入力する
- 生データ，処理済みデータの重要性は
  - サイエンスに使う生データ，プロバイダーが決めて整備した生データは残す
  - アーカイブによるデータ再利用のためには処理済みデータ重視
- 論文を書いたデータは公開？

Q. どういう状態でデータを公開すべきか

- データ整備/validation
  - 基本は使える状態にする
  - データ取得時の天候が分かるようにする
- 論文が書ける状況
- データ一覧とリダクション済みデータ公開？
- 品質評価情報
  - 処理済みデータ・ライトカーブなどの品質・エラー-> ベストエフォート，Caveat付
  - 使う人が判断する・間違いは別の論文などで修正される
- 保管場所，使いやすさ
  - 必要なデータが探しやすいこと
  - （直にデータを置く？->利用者統計・評価が難しい）
- 必要な機能（現在，将来）

Q. データアーカイブの体制

- 国全体の観測データの集約？はしていきたい
  - 過去のデータの確認に有用．タイムドメインなど
- 共同利用機関/NAOJの位置づけ・役割
  - 継続性を維持すべき
  - 論文生産率・科学生産性を高めるためにアーカイブを積極的に使うべき
    - TMTなど
- NAOJと大学等の役割
  - データを集約する意味のひとつとして大学間連携で人を育てることがある
- 技術，コストの見通し(30年？)
  - 共同利用機関が継続性を持つのが自然．予算は問題．
  - ネットワークは高速になってきた．観測所をSINETなどに繋ぎたい

Q.観測所/大学等装置チーム（データプロバイダー）とデータアーカイブの役割

- データ整備（検証，修正），データ処理，情報付加，公開
- データプロバイダー（観測所）
  - 検証（なるべく上流でかけてほしい）
    - ヘッダ・メタ情報のフォーマット検査
    - 時間，座標などの値検証
  - 品質確認（ベストエフォート，ただし遅らせない）
  - 修正
    - 元のデータ修正？修正情報の提供（なるべく）<- アーカイブと協力して
    - 十分早い時間に完了する
- アーカイブ
  - 検証（なるべく上流でかけてほしい）
    - ヘッダ・メタ情報のフォーマット検査
    - 品質確認も支援
  - 修正はプロバイダーへ依頼
  - 保管と公開作業
- その進め方
  - ルールを作って守る
  - インターフェース，ドキュメンテーション，系統的に・MOU（期間・責任境界）

Q. アーカイブ/解析をするためのデータの準備作業，データコミッションの考え方

- 保存するデータ・どこにどう保存するのかを設計に含める
- アーカイブするためのデータフォーマット・ヘッダの策定
- 解析ツール，Pipelineの整備？
  - 共同利用観測所はやるべきー装置開発者の協力．レビューして残す．
  - 装置開発者だけで整備するのは大変
  - 少なくとも最新のデータ処理方法の情報を整理してアーカイブユーザが分かるようにする

Q. オープンデータ化の動きへの対応

- 装置開発など外部予算からの要請，それへの対応

Q.コミュニティ・組織・予算措置者への説明，訴え方

- 科学雑誌・グラントのデータ/処理情報の公開の要請
- 社会・教育・普及利用，産学連携，オープンデータ対応
- 活動への評価

- 研究会でのアーカイブデータによる研究成果発表を増やす
- 装置開発者の動機付け：データ公開が装置望遠鏡の価値を高める
- アーカイブ利用の成功例・成果が出るケースをまとめる
- 教育普及関係者にも有用さを理解してもらいサポートしてもらう
- アーカイブの有用性・成果を正式な文書として残す.
- 講習会でアーカイブデータを使う（アーカイブだから出来ること）＝教育

進むべき方向・目標

- 現実的なアクション

この集まりの次への展開

- 成果が出た・出ること，データ公開要請の理由付けの洗い出し
- 調査
  - ほかのアーカイブの予算規模？
  - これ自体は予算請求の根拠にはならない
- 見込まれる成果，コミュニティとの会話が重要
- コミュニティからの支持を得る
- より広い人を巻き込んだワークショップ？
  - (講習会)
  - 事例紹介
  - 予算取り・体制の突っ込んだ話
  - 3/26または3/30，1週間で人・グループの候補を挙げる

データ整備，保存，公開

- 扱うデータ種類：　生データ、解析済みデータ、それらを記述したり補足するメタ情報
- データ（永続）保管
- データ公開
- データ品質評価
- アーカイブの有用性向上のため
- 望遠鏡・装置へのフィードバック
- データ健全性確保・修正
  - 解析可能性の確保
  - データ素性（特性）の完全かつ整合的な記録
- 解析に資する情報の向上、解析結果（較正等）の向上