NII Open Forum 2025
「ガバナンスで進化する研究データ管理」トラック
Slido Q&A

| 番号 | Slido投稿内容 | 回答 |
|---|---|---|
| 1 | デジタルデータは恒久的に保存すべきと考えますが如何でしょう？例えば、数百年前のガリレオやメンデルの実験ノートなどは現在でも古典として参照できますが、現在のデジタル保存データが数百年後に参照できるかは難しいと思います。大袈裟に言えば、蓄積されている人類の叡智を後世に残すにはデジタルデータ管理・保存は適していないと感じていますが如何でしょう？（本セッションの論点とズレているような気もしますが） | ご質問ありがとうございます。この問いは「**何を保存すべきか（選別の問題）**」と「**どうすれば恒久的に保存できるか（方法の問題）**」という二つの論点に整理できると考えております。以下、順に回答させていただきます。<br>【あらゆるデータを恒久的に保存すべきか？】<br>**いいえ、膨大なデータの中から何を廃棄し、何を残すかが重要であると考えております**。数百年もの時を経て現存する実験ノートがある一方で、数多の物理的媒体が現存していない可能性があります。その背景には、災害などのによる喪失の他に、「残さなくてよい」という判断の下で淘汰されたものもあると考えられます。また、デジタルデータであれ物理的媒体であれ、情報を保存する場所（ストレージ）は有限です。したがって、データの廃棄と維持を意識していく必要があると考えております。<br>【デジタルデータが恒久的な保存に適しているか？】<br>**確かにノートなどの物理的記録と比較すると、デジタルデータは世紀をまたぐような保存に適さないかもしれません**。一方、物理的記録と比較して、デジタルデータは劣化なしで複製が可能であるという重要な特徴を有しています。**物理的媒体とデジタルデータを、互いの弱みを補うように活用するというハイブリッド戦略をとる**ことで、それら単体だけでは達成できないような、より恒久的な保存ができるのではないかと考えております。 |
| 2 | 藤原先生に質問です．技術的なことではなくズレた質問かもしれませんが．．．メタデータ・データとも公開されるものが公開，データ公開に条件が付くものを制限公開とすると，秘匿解析機能のようなことは想定外で，メタデータ公開，データ非公開，計算結果を得られるは，公開でも制限公開でもなく，共有や非公開でもなく，何と呼べばよいのでしょうか？ | ご質問ありがとうございます。**秘密計算技術の登場により、「情報の濃淡（グラデーション）」を制御する新しい公開概念を考える時期が来ている**のだろうと考えております。従来の「公開／非公開」「共有／非共有」といった分類は、情報のアクセス権を「0か1」で捉えるものでした。しかし、秘匿解析や秘密計算技術の登場により、「メタデータは公開されるが、データ本体は非公開」「計算結果のみが取得可能」といった、従来の枠組みに収まらない新しい情報の扱い方が可能となります。このような、情報の「濃さ」を調整しながら提供するという考え方は、社会制度や法制度においてもまだ十分に整備されていない新しい領域です。<br>また、こうした形態を何と呼ぶべきかという問いに対しては、現時点では明確な定義や用語は存在していないのが実情です。この「何と呼ぶか」は、技術の社会実装において非常に重要な要素です。適切な名称が与えられることで、制度設計や合意形成が進み、技術の受容が促進される可能性があると考えております。 |
| 3.1 | 研究室でのデータガバナンスは、初学者から専門家までデータリテラシーの幅があると思います。その中で、研究データをクローズな状態から公開する状態にするときに、<br>①データを一意に特定するPID付与のルールはどのように決めているのでしょうか？ | ご質問ありがとうございます。例えば自然言語処理分野では、GitHubなどを活用した迅速な共有文化が根付いており、**PIDの付与が十分に進んでいるわけではない**という傾向があります。この背景には、**PID付与に伴う手続きやコスト（メタデータ整備、登録作業など）**に対する意識がまだ十分に醸成されていないという状況があります。<br>また、ソフトウェアやアルゴリズムのソースコードに関しては、**Software Heritage ID（SWHID）をPIDとして活用する**という実践例があります。GitHubにパブリックリポジトリとして公開することで、Software Heritageが自動的にクロールし、SWHIDを付与してくれるため、特別な手続きを踏まずにPIDが得られるという利点があります。 |
| 3.2 | （研究室でのデータガバナンスは、初学者から専門家までデータリテラシーの幅があると思います。その中で、研究データをクローズな状態から公開する状態にするときに、）<br>②データを後で修正や差し替えをする必要が生じた時の運用はどうされているでしょうか。 | ご質問ありがとうございます。例えばパネラーである松原先生の研究室では、**修正されたデータを元の公開場所と同様の場所に置き、ファイル名やメタデータ（属性情報）に「リバイズ済み」や「バージョン番号」などを明記する**ことで、利用者が変更履歴を把握できるようにするという方法が採用されています。また、藤原先生がコメントされたように、**GitHubなどのバージョン管理システムを活用**し、コミットIDに基づいてPIDを付与することで、変更履歴を厳密に管理する方法が採用されていることもあります。 |

| | | |
|---|---|---|
| 3.3 | （研究室でのデータガバナンスは、初学者から専門家までデータリテラシーの幅があると思います。その中で、研究データをクローズな状態から公開する状態にするときに、）<br>③自然言語処理分野では、データ（アルゴリズムの場合）の再現性を担保するためのバージョン管理や用いるデータセットの信頼性については、どのような議論があるのでしょうか？ | ご質問ありがとうございます。自然言語処理分野において、研究データをクローズな状態から公開する際には、再現性とデータセットの信頼性の両面で慎重な配慮が求められます。<br>まず、再現性の確保については、特に近年のAI研究では、GPUサーバーなどの実行環境に依存する部分が大きく、「特定のバージョンの特定の環境でのみ動作する」という状況が頻繁に発生しています。そのため、**研究者は自身で実行可能な環境での検証を行い、「少なくともこの環境では動作した」という情報を明記する**ことで、最低限の再現性を担保するという運用がとられています。<br>次に、データセットの信頼性については、完全に信頼できる状態での共有は困難ですが、**広く使われているデータセットに関しては、問題点や得意・不得意な点が研究成果として蓄積されており、「皆で信頼度のレベルを共有し合う」という文化**が形成されています。 |
| 4 | GitHub のコミットIDは厳密なPIDではないかもしれませんが、プロビナンスを管理する点では有用かと思います。 | ご意見ありがとうございます。ご指摘の通り、Git（GitHub）がデータの変更履歴・来歴を記録する優れた手段であることを鑑みると、**GitHubのコミットIDがそれらの来歴を追跡できるIDとして非常に有用である**と考えられます。 |