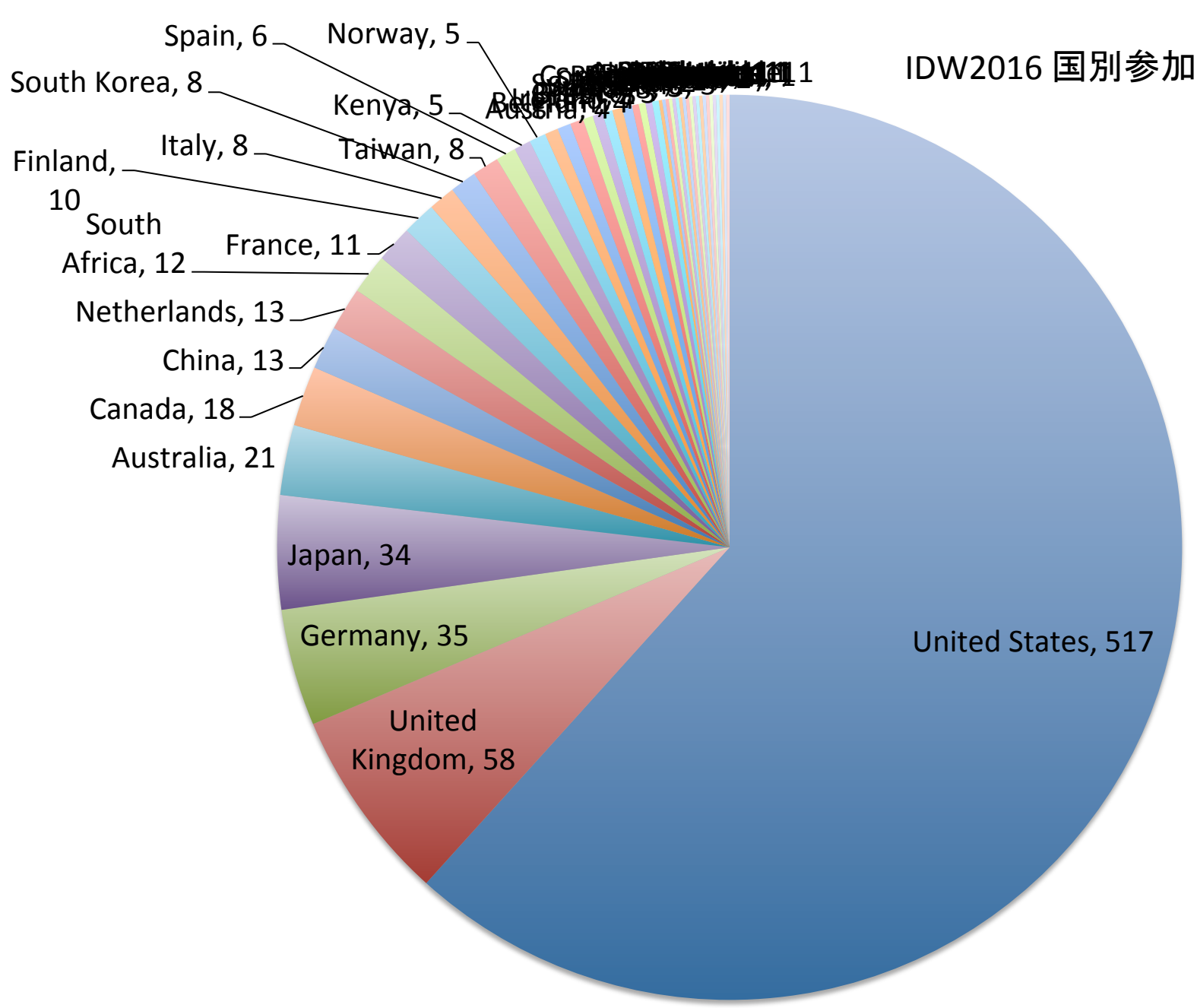
RDA 8th plenary 報告

蔵川圭
国立情報学研究所

International Data Week

- 2016年9月11日（月）から9月17日（土）まで
- 1週間を通して研究データを取り扱う関連会議が開催された
  - 9/11 ICSU CODATA General Assembly 2016
  - 9/11 WDS Members' Forum 2016
  - 9/11-13 SciDataCon 2016 (previously CODATA conference, organized by ICSU CODATA and WDS)
  - 9/14 International data forum
  - 9/15-17 RDA
- 録画アーカイブ
  - 3つの会議のメイン会場の録画
  - http://www.tvworldwide.com/events/idw/161117/

IDW2016 国別参加者数（N=838）



RDA outputs

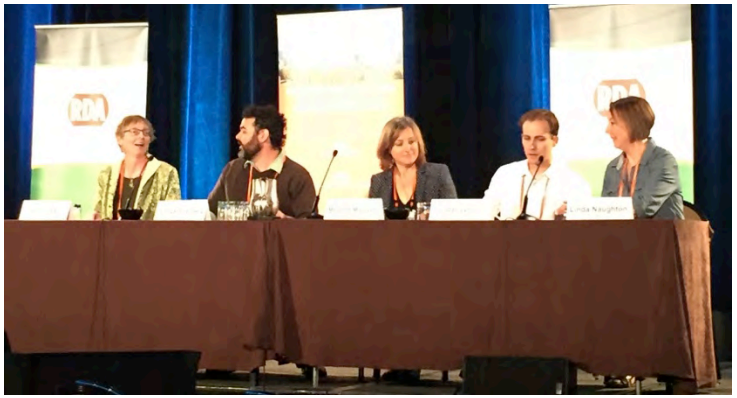
- Practice in pieces
  of librarians, researchers, professors,
  program managers, students, consultant,
  IT-specialist, CTO, CEO, and others
- Deliverables
  - Community
  - Specification
  - Recommendation
  - Adaptation

私のセッションへの参加

- RDA8のPlenaryセッション
  - 活動の持続可能性がテーマ
  - アフリカのオープンアクセス、研究データ共有
  - WG成果の適応事例の紹介
- 私のWG, IG, BoFへの参加の視点
  - 研究データ共有
  - ドメイン非依存
  - 情報システムづくり

パネルディスカッション Sustainability

- ディスカッションの目的
  - RDAの、R&D段階から持続可能(sustainability)な組織への移行を探る



- 議論
  - Sustainabilityとdevelopmentは違う
  - もちろん、sustainabilityにも、ソフトウェアや図書館運営などやり方もいろいろ
    - ビジネスはrevenueベースのモデル
    - ヘルシーコミュニティを作るという基準もありえる
    - オープンソースコミュニティのように運営するモデル
  - 研究データ固有のsustainabilityフレームワーク
    - ファンディングから考えると、その方法もsustainabilityを左右する
    - 研究では成功するといのは論文が出たとか研究助成を獲得したとかだが、研究データの場合はデータセットやツールがでたとかだと思われるが、組織のテニュア委員会でそういうことが議論されないといけない
    - ファンディングエージェンシーが研究コミュニティと長期のフレームワークを作っていく必要がある
    - 持続可能性を考えるのに経済モデルが有用だが、オープンデータの場合はフリーライダー問題を含めて考え直す必要がある

キーノートスピーチ
Dr. Kay Raseroka

- University of Botswana & RDA Council
- IFLA president 2003-2005
- 講演要約
  - アフリカのアカデミーの現状
    - アフリカとRDA
      - RDAにおいては低いプレゼンス
      - アフリカでは、エジプトは特別な存在
    - アフリカとオープンサイエンス
      - ここ5年でオープンサイエンスは議論されている
      - 研究データはそれぞれが保有管理されている
      - オープンアクセスやオープンサイエンスは受け入れられにくい
      - 退職するときに研究データを保管したいという要望があるが、大学にはその余裕はない
    - アフリカの先導的組織
      - アフリカの図書館コンソーシアムは、研究データやオープンサイエンスプリンシプルにも大きな影響を持っている
      - African Academy of Sciences (1985-): 科学技術を先導
  - アフリカにおけるアカデミー人材への機会
    - 若手の研究者をオープンサイエンス、オープンデータ、研究データへのアクセスと保護のバランスを議論する機会を与えたい

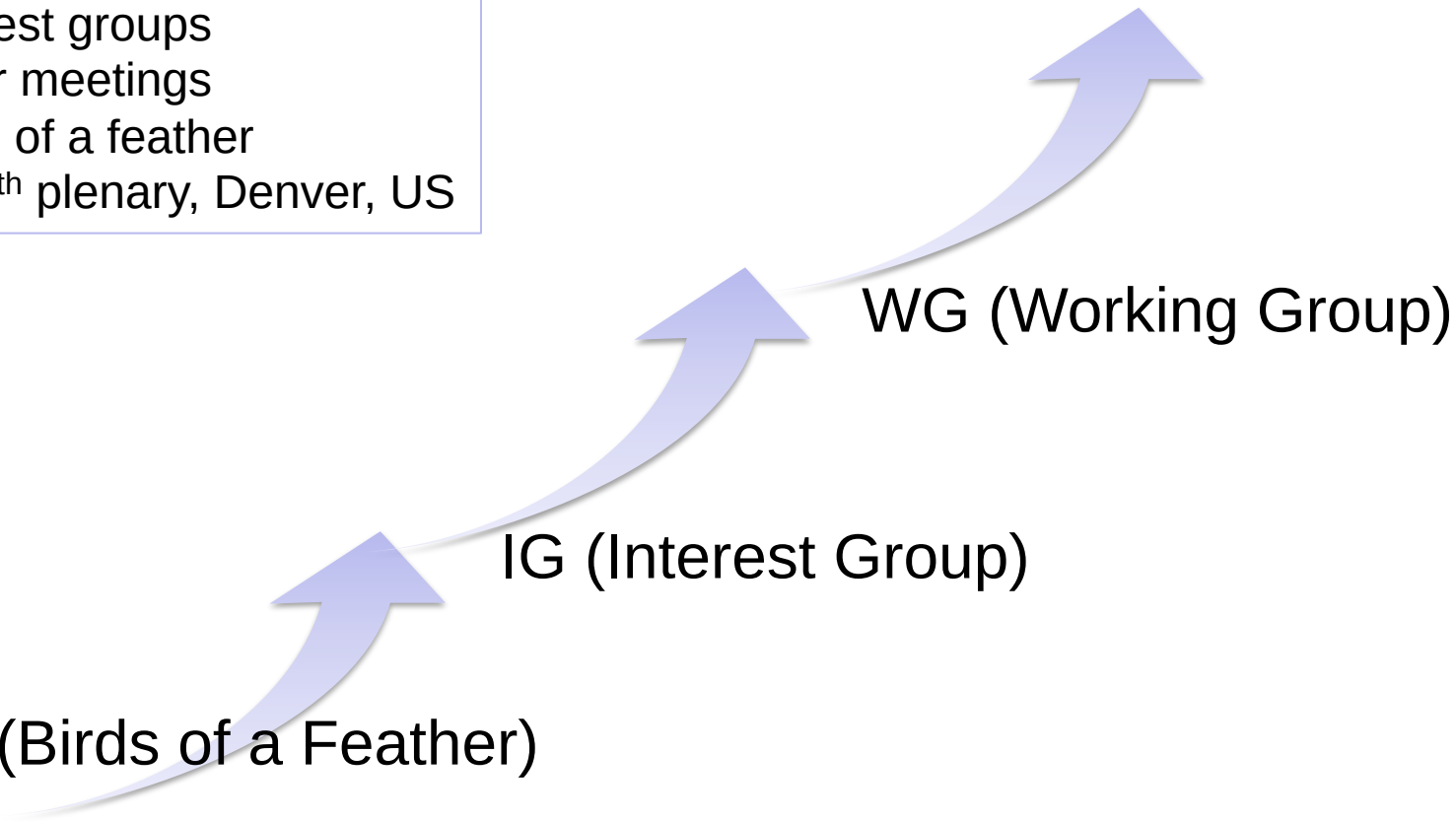
サブコミュニティの段階的発展と成果

71 breakout meeting includes:
- 10 working groups
- 31 interest groups
- 3 other meetings
- 13 birds of a feather

In RDA 8th plenary, Denver, US

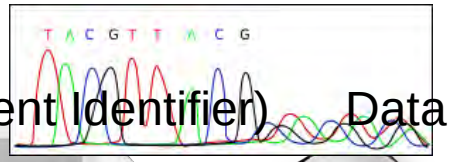
Specification 仕様　　Adaptation 適用

Recommendation 推奨



研究データの
インターネット公開のあり方

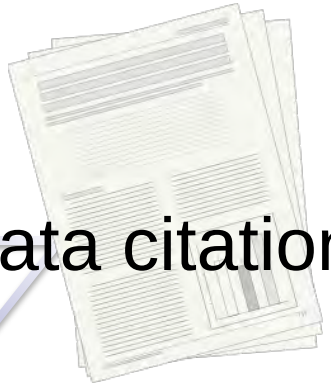
Data fabric

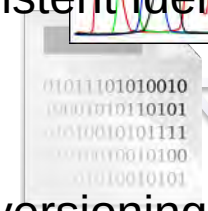


PID (Persistent Identifier)

Data type registries

Data citation



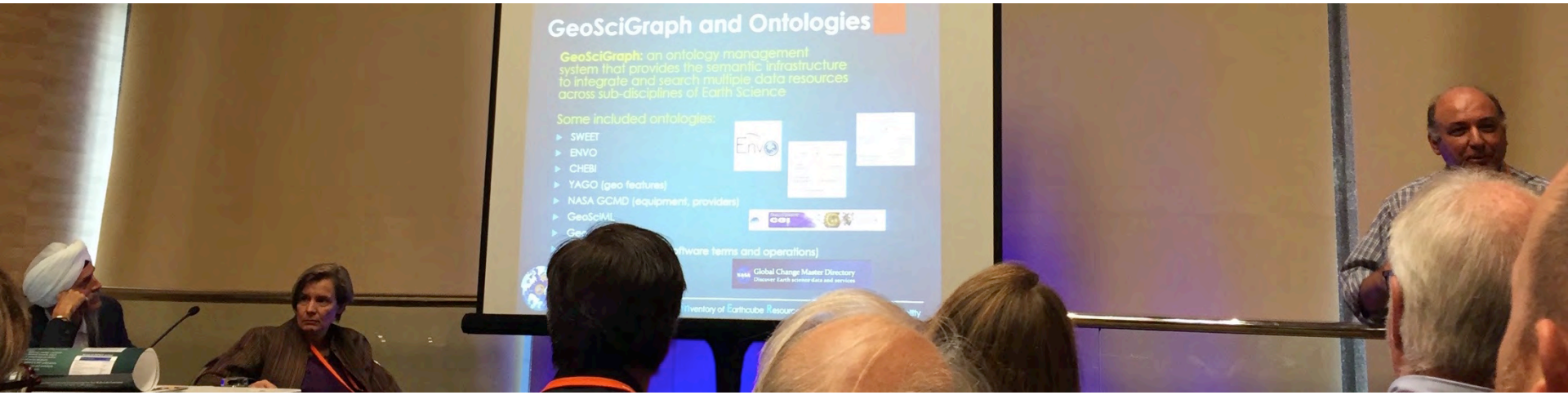


Data versioning

Data provenance

Data search

Data discovery paradigms IG



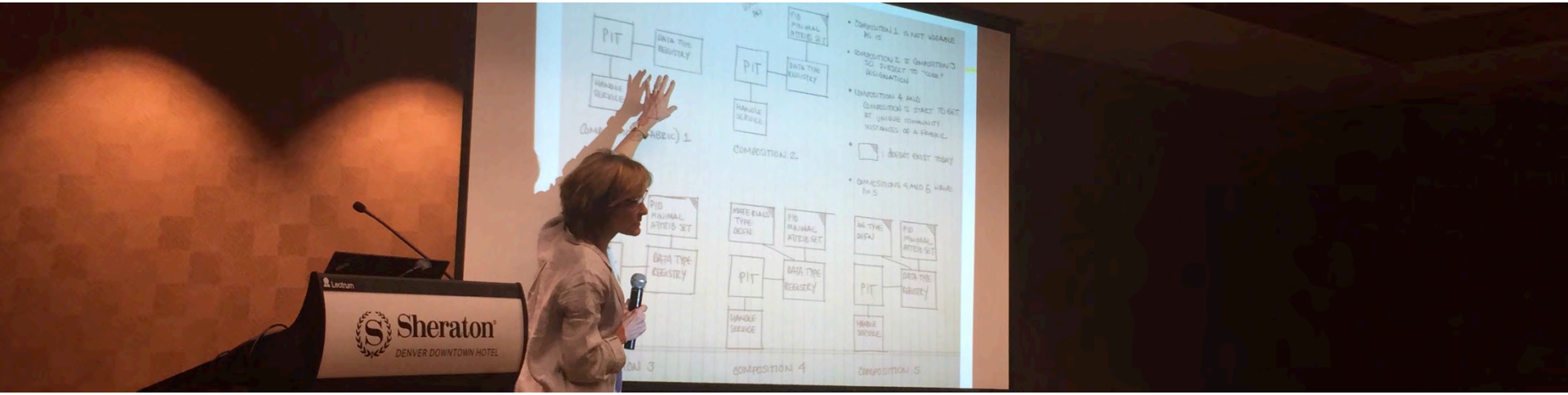
- データの発見（データ検索）に着目したグループ
- 7th plenary Tokyoの時はBoFとして開催し、今回IGとしてキックオフ
- データ検索に関する情報モデルとデモンストレーション
  - A Data discovery model: DESIRE（データへの付加価値、検索、ユーザーの意図）
  - デモ
    - Content enrichment and faceted search - Ilya Zaslavsky, San Diego Supercomputer Center
    - Relevance ranking – Jeff Grethe, Biocaddie/UCSD
    - User Characterisation and Search Personalisation - Siri Jodha Singh Khalsa, NSIDC
- データ検索でとりあげるべきテーマをブレインストーミング
  - 23のフォーカスエリア

データ検索 23のフォーカスエリア

1. Deduplication and cross-repository issues
2. Identifiers and how they help in search
3. Data citation: how do we access/use?
4. Relevancy ranking for structured data?
5. Enrichment tools for faceting and ranking
6. Domain-specific vs. generic issues: interfaces and enrichment
7. Different discovery platforms for Open Search, science-focused OS profile?
8. Metadata standards to enhance data discovery, e.g. schema.org and such
9. Models and methods of personalization
10. Identify core elements of Findability
11. Automated integration of records; granularity and findability
12. Common APIs (e.g. OpenSearch)
13. Upper-level ontologies for search
14. Creating test collections for search evaluation and methods of evaluation
15. Collections and granules: build tool that enables guidance for data submitters on how data is organized
16. Guidelines for making your data findable! Best practices based on experiences.
17. Identify collections of use cases for users: e.g. browsing vs search
18. Measures of data quality: and impact of findability
19. Define series of reference datasets – can be used to do these metrics
20. Identify list of prototyping tools, use by WG!
21. Cross over between domains: how to enable cross-walk between domains
22. “Return to the semantic”: schema has been populated by crowdsourcing rather than 1 researcher.
23. Implementing schema.org as it exists! How does it apply to science?

https://rd-alliance.org/ig-new-paradigms-data-discovery-rda-8th-plenary-meeting

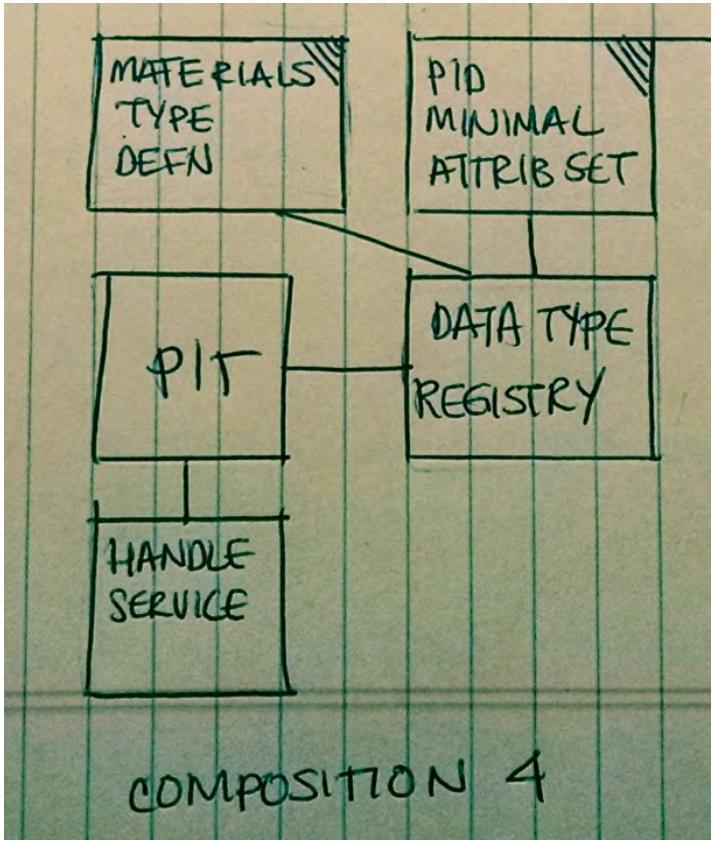
Data fabric IG



- データの基本構成
  - データの生産と消費のサイクルに着目
  - グローバルデジタルオブジェクトクラウド(DOC)の構想
- PIDを中心としたデータマネージメントとアクセスの実現を目指す
  - PIDに関連する最小のメタデータ構成（PID プロファイル）を探る
  - RDA P9までにサブグループを作ってプロファイルを定義する

PID プロファイルを探る

Data fabricの構成



- 簡単のため、PIT (PID information type)は、Handle serviceに限定しておく
- 分野によらないプロファイル最小属性
- 分野ごとのプロファイル属性

- 4つのサブグループを作ってプロファイルを探る
  - Digital Humanities
    - Data provider
    - Data consumer
  - Natural/physical science
    - Data provider
    - Data consumer

https://rd-alliance.org/group/data-fabric-ig/wiki/df-configuration-pid-centric-data-management-and-access.html

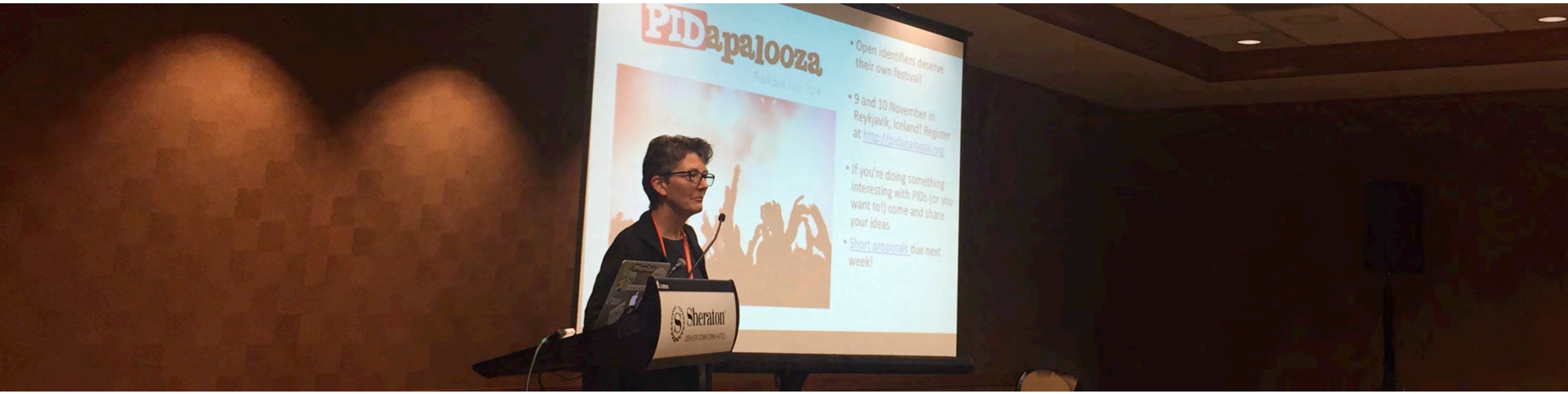
Data citation WG

- データサイテーションの仕組みを考案
- データサイテーションを実際のデータセンターの活動に適用する

Data citation WG アウトプット

- DCの目的
  - 研究で利用する(ダイナミックな)データのサブセットを正確に示したい
  - 任意の時刻におけるデータを正確に指し示したい
- DCの仕組み
  - データとそのアクセス手段を提供すること
    - データ
      - タイムスタンプとバージョン
    - アクセス
      - クエリーとタイムスタンプの組みにPIDを付与する
- アウトプット
  - 14の推奨
    https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf
  - より詳細な報告
    Andreas Rauber, Ari Asmi, Dieter van Uytvanck and Stefan Pröll, Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use, Bulletin of the IEEE Technical Committe on Digital Libraries (TCDl), Vol. 12, Issue 1, May 2016
    http://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf
  - パイロット適用プロジェクト
    - DBMI @ WUSTL
    - BCO-DMO
    - Argo
    - 他多数

PID IG

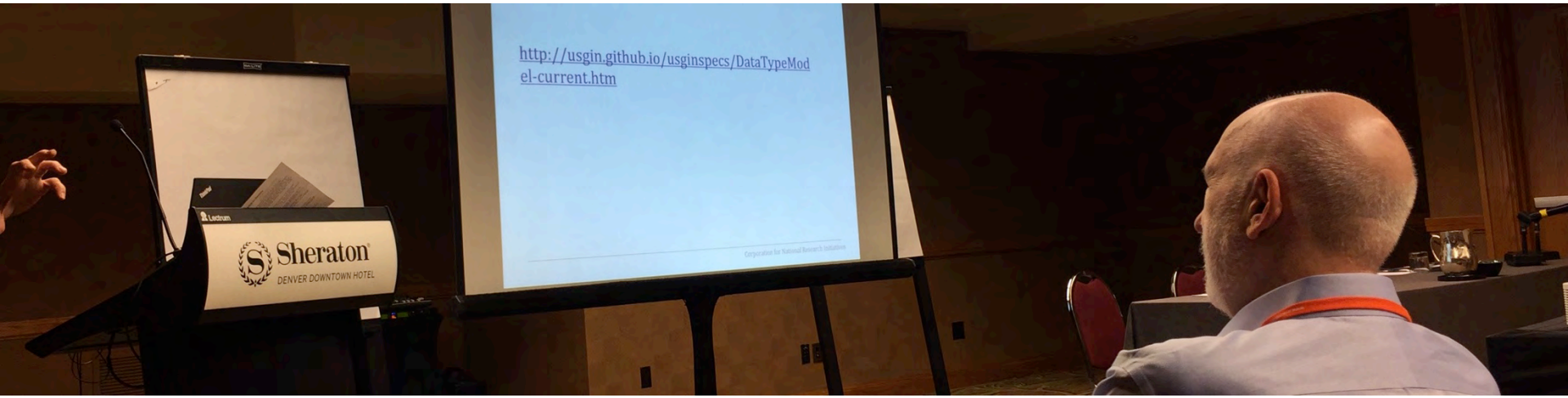


- 永続的識別子(persistent identifier)の重要性を認識
- 研究リソース（論文やデータ）の識別子と人の識別子
  - 例えば、CrossRefやDataCite
  - 例えば、ORCID
- それ以外の識別子も必要であると認識
  - たとえば、組織の識別子

PIDの進捗

- ORCIDの進展
  - Affiliation round trip
    - ユーザーの所属組織がORCIDメンバーなら、メンバー組織はユーザにデータの読み書きを許可をリクエストしたり、所属組織の承認をしたりできる
    - 2016年秋にリリース予定
- THORプロジェクトの進展
  - 30 month project on European Commission, Horizon 2020
  - 論文、データ、研究者の研究ライフサイクルを通じたシームレスなシステム間統合を進めています
  - 10の参加組織
    - British Library, ORCID, DataCite, CERN, EMBL-EBI, PANGAEA, ANDS, DRYAD, ELSEVIER, PLOS
- PIDのイベントをやります
  - PID apalooza, 2016, Nov, 9-10 @ Reykjavík, Iceland
  - http://pidapalooza.org

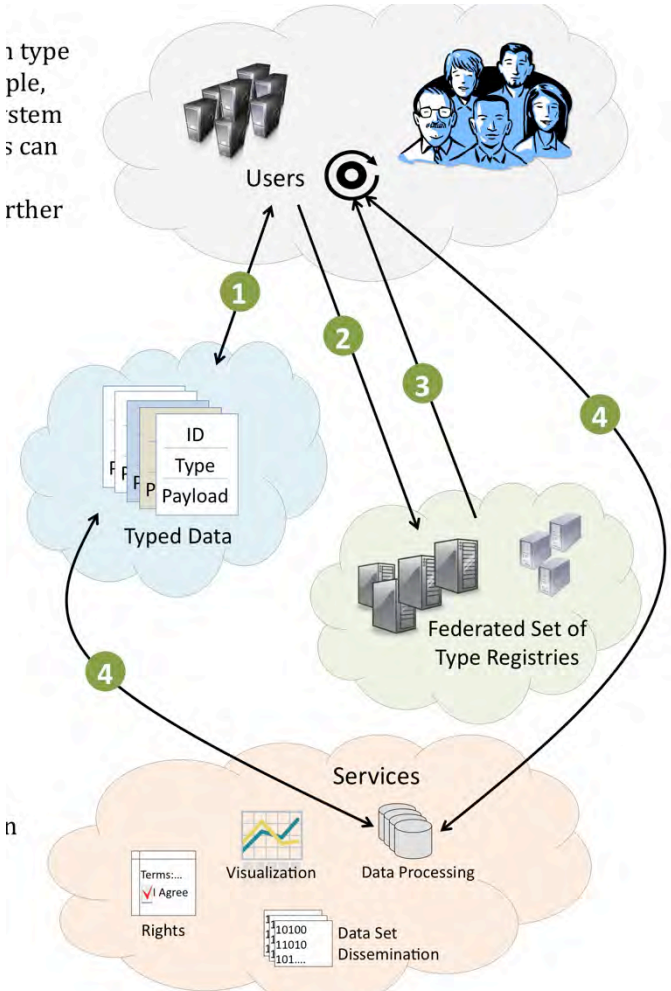
Data type registries WG



- データタイプレジストリ
  - データ作成に関わっていない研究者でも研究データを解釈し利用できるようにする仕組み
  - データのタイプ
    - フォーマット
    - データのアトリビュート
- 2回目のWGとしてキックオフし、初回のWGアウトプットを継承
- ISO-IEC/JTC1/SC32(Data management and interchange)/ WG2(Metadata)でデータセット記述のメタデータモデルを作成中

Data type registry

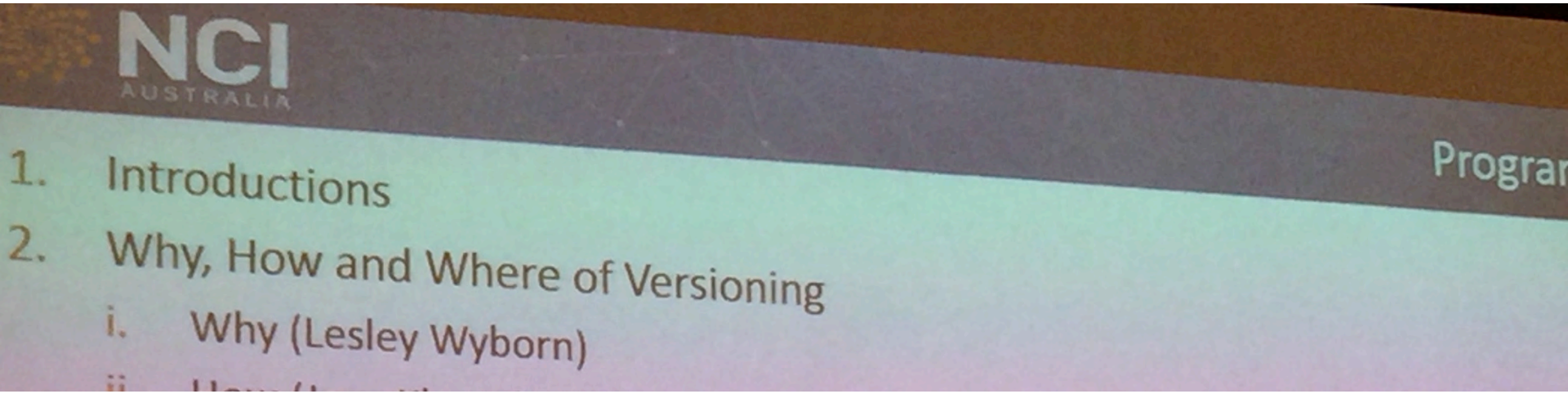
WG #1 output



- 個別活動報告
  - The German Climate Center (DKRZ)
  - Columbia U.
  - Open Knowledge International
  - Gesellschaft fuer wissenschaftliche Datenverarbeitung mbH Goettingen (GWDG)
  - Vermont Monitoring Cooperative

dx.doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458

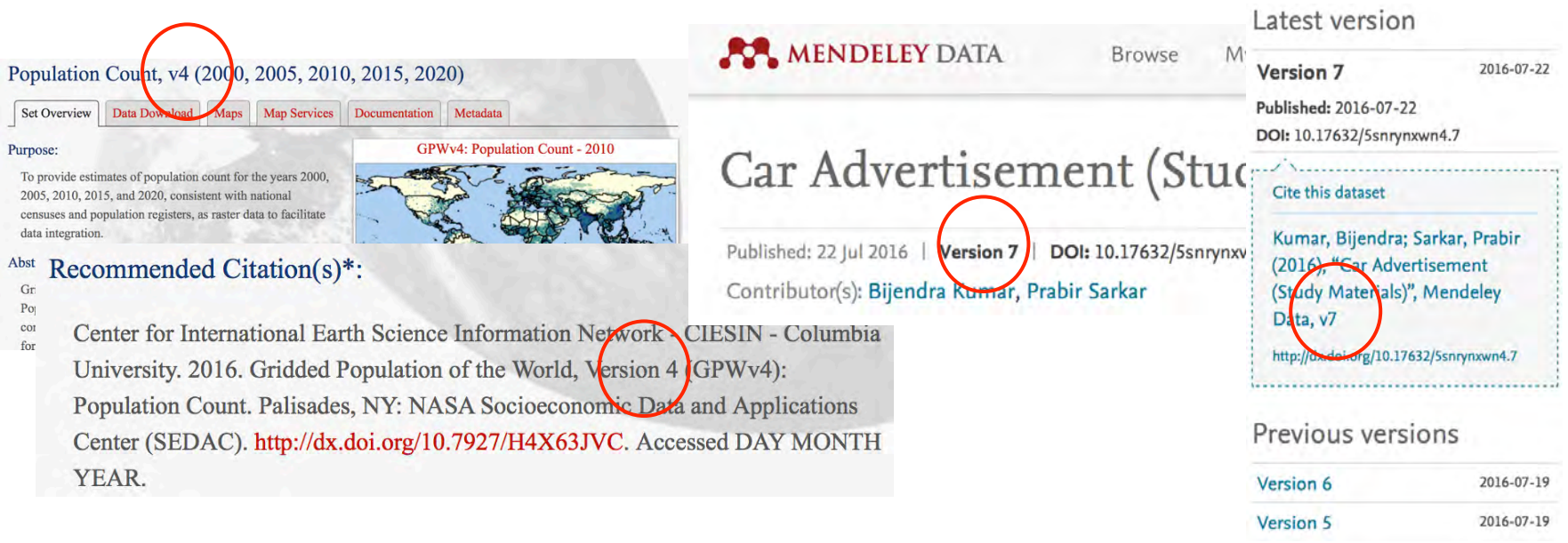
Data versioning BoF



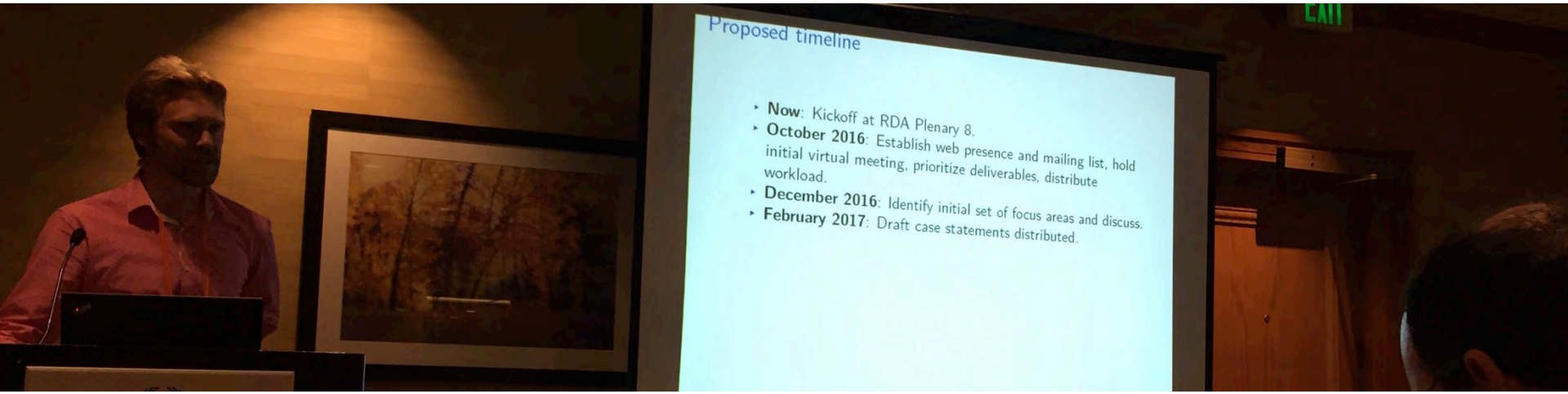
- データサイテーションにはデータのバージョニングを考慮する必要がある
- バージョニングに、標準やベストプラクティスが必要かを問う

Data versioning

- すでにデータのバージョニング管理は様々なところで行われている
  - NASA Socioeconomic Data and Applications Center (SEDAC)
  - Mendeley Data
  - DataONE
  - Dat Data project http://dat-data.com
  - W3C Recommendation https://www.w3.org/TR/dwbp/#dataVersioning
- ソフトウェアのバージョン管理も参考になるだろう



Research data provenance IG



- データのprovenance(出所)のモデルについて議論する
- 動機
  - データを利用するときに、そのデータはどこから来たのか？誰が修正したのか？自分が掲載したデータと同じか違うのか？というよくある研究者の疑問に答えたい
- 先行例があるがちょっと研究データのためには粗いのではないか
  - https://www.w3.org/TR/prov-overview/

Data provenance

- スケジュール: October 2017 from now
- アクションアイテム:
  1. Provenance patterns
  2. Sharing provenance
  3. Strategies for implementation
  4. Connecting to other groups
     - Dynamic data citation WG
     - PID information types
     - Reproducibility IG
     - PID IG
     - Archives and records professionals for research data IG
     - Data discovery IG
     - Preservation e-infrastructure

おわりに

- RDAは3年半を経た
- ボトムアップでオープンな活動の形態を維持
- RDAとしてやるべきことの明暗がはっきりし、成熟してきた感
- より中心的な関係者が継続してリード
- グラント期間後の活動の持続可能性が問われている