学術情報流通のための識別子とメタデータDBを対象とした融合研究シーズ探索

超高層物理学分野における観測データを例として

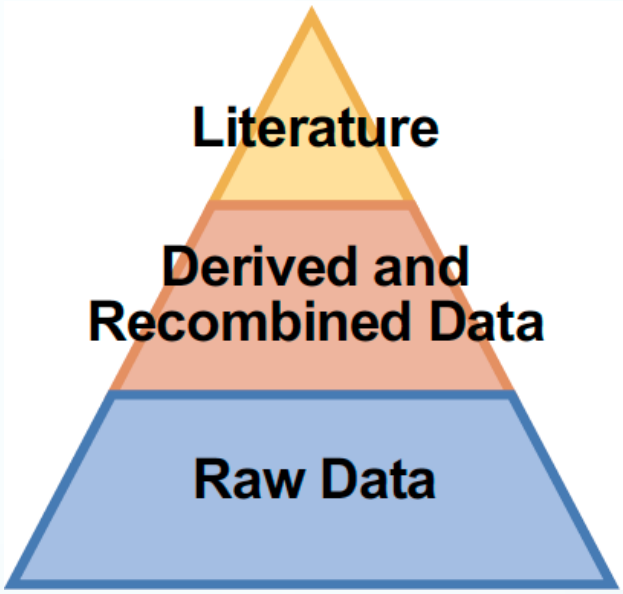
小山幸伸（京都大地磁気WDC）　蔵川　圭（NII）
佐藤由佳（NIPR）　田中良昌（NIPR）

データ集約科学における情報の組織化

- データ集約科学のビジョン
  - 第4の科学 (Fourth Paradigm) [Hey, Tansley, Tolle (Eds.), 2009]
    - 実験科学 (Empirical Science) (1st paradigm)
    - 理論科学 (Theoretical Science) (2nd paradigm)
    - 計算科学 (Computational Science) (3rd paradigm)
    - データ集約科学 (Data-intensive Science) (4th paradigm)
  - e-Science (UK)
- データ集約科学の基盤
  - e-Infrastructure (UK)
  - Cyberinfrastructure (US)
  - Cyber Science Infrastructure (JP)
- データ集約科学では、研究成果(論文)に至る一次データや計算結果を含む膨大なすべての情報をオンライン上で組織化してアーカイブし、再利用する

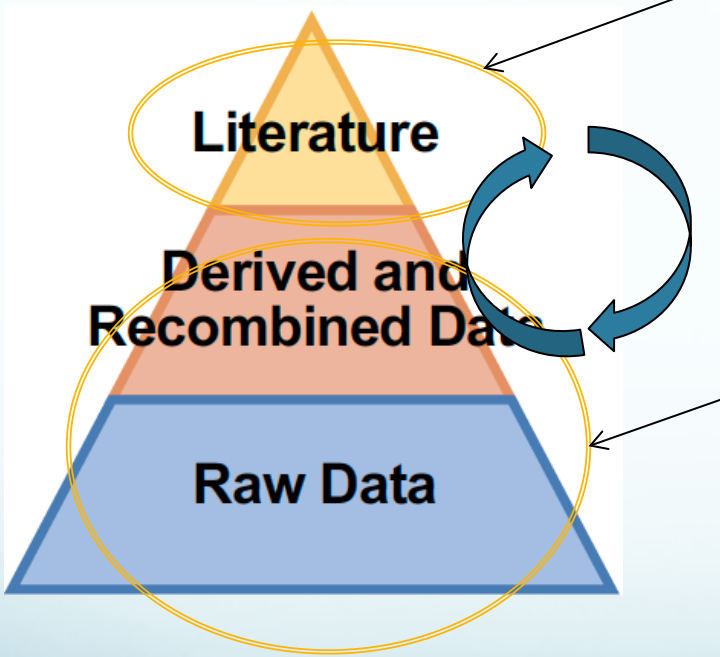
科学的データの階層

- 3つの階層
  - Literature
  - Derived and Recombined Data
  - Raw Data
- 階層の要素
  - 同一の階層の要素は互いに関係を持つ
  - 隣り合う階層を構成する要素が互いに関係を持つ
- ネットが分野をまたいだ要素の統合と関連を可能にする



Tony Hey, Stewart Tansley, & Kristin Tolle (Eds.). (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research.
Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx

インターネット上の学術情報流通の飽くなき展開



NIIのような文献事業からの展開

ネット上の技術トレンドの上に繰り広げられる
相互波及

IUGONETのような
データサイエンスからの展開

Web上の学術情報の識別子

- オブジェクトの識別子
  - DOI (Digital Object Identifier)
    - CrossRef (1999 – , PILA)
    - DataCite (2009 – , BL and library related)
    - JaLC (Japan Link Center)
- 人の識別子
  - ORCID (Open Researcher and Contributor Identifier) (2010 – , ORCID. Inc.)
  - 研究者リゾルバーID(科研費研究者番号)
    - NIIによるプロトタイプシステム(2008 – , NII)

DOI (Digital Object Identifier)

- インターネット上のオブジェクトへのアクセス可用性を高品質に担保する仕組みと管理体制
- 論文ごとにDOIを付与するのが基本
- CrossRefは、ジャーナル論文、本、プロシーディングス論文にDOIを付与している
- 対象の詳細化
  - 論文内の図、表にDOIを付与する
  - 論文内の章、節にDOIを付与する
  - 本の章にDOIを付与する
- 対象の拡大
  - 論文に引用される形式のデータセットにDOIを付与する

ORCID
(Open Researcher and Contributor ID)

- 論文著者の名寄せを解決したい
- 学術論文のデータベースでは、2つの方法がとられてきた
  - 計算機による名寄せ
    - 例
      - Scopus Author Identifier
        (Elsevier社のScopusに実装)
      - Distinct Author Identification System
        (Thomson Reuters社のWeb of Scienceに実装)
  - 手動で登録
    - 例
      - ResearcherID (Thomson Reuters社)
- ORCIDは、学術コミュニケーションに関与するすべてのステークホルダーを包含した、研究者に識別子を付与するコミュニティを形成する

DOI, ORCIDとURI

- 学術情報の識別子
  - DOI
    - prefix / suffix
  - ORCID
    - 16 digit numbers
- インターネット上の識別子をURI(Uniform Resource Identifier)という
- HTTPでアクセスできるよう、学術情報の識別子をURIで表現する

10.1007/s00163-004-0050-z

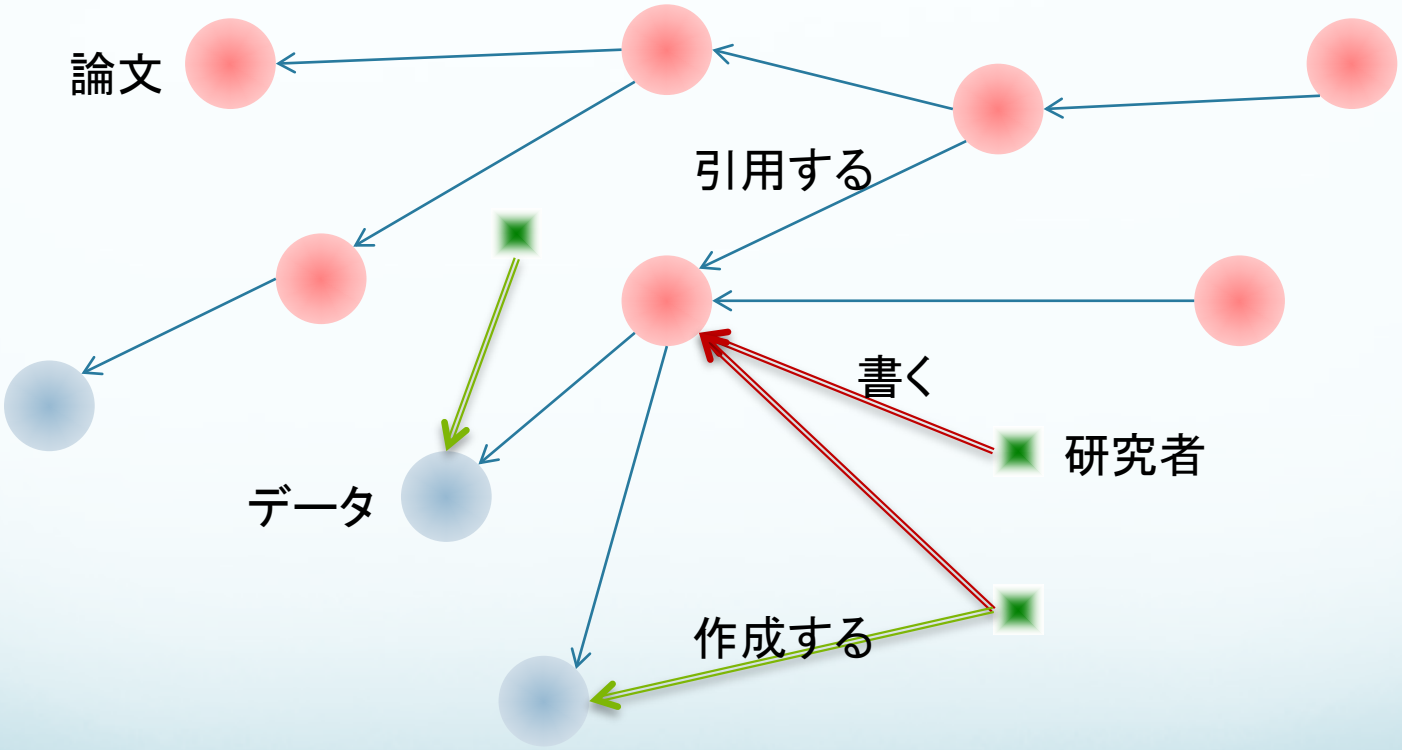
0000-0002-7031-1846

doi:10.1007/s00163-004-0050-z
または、
http://dx.doi.org/10.1007/s00163-004-0050-z

http://orcid.org/0000-0002-7031-1846

出版レイヤー

サイテーションメカニズムを構成する世界



サイテーションメカニズムを利用したサービス例

- 論文の引用
  - Citation Index
  - Impact Factor
  - H-index
- 論文引用サービス例
  - Web of Science (TR)
  - Scopus (Elsevier)
  - CrossRef (PILA)
  - Google Scholar (Google)
- データの引用
  - Data Citation Index (TR)
- データ引用サービス例
  - PANGAEA (Alfred Wegene Institute for Polar and Ma Research, Center for Mari Environmental Sciences, and etc.)
  - DataCite (BL, and etc.)
  - Dryad (National Evolutiona Synthesis Center and the University of North Carolina Metadata Research Center)

OAI-ORE

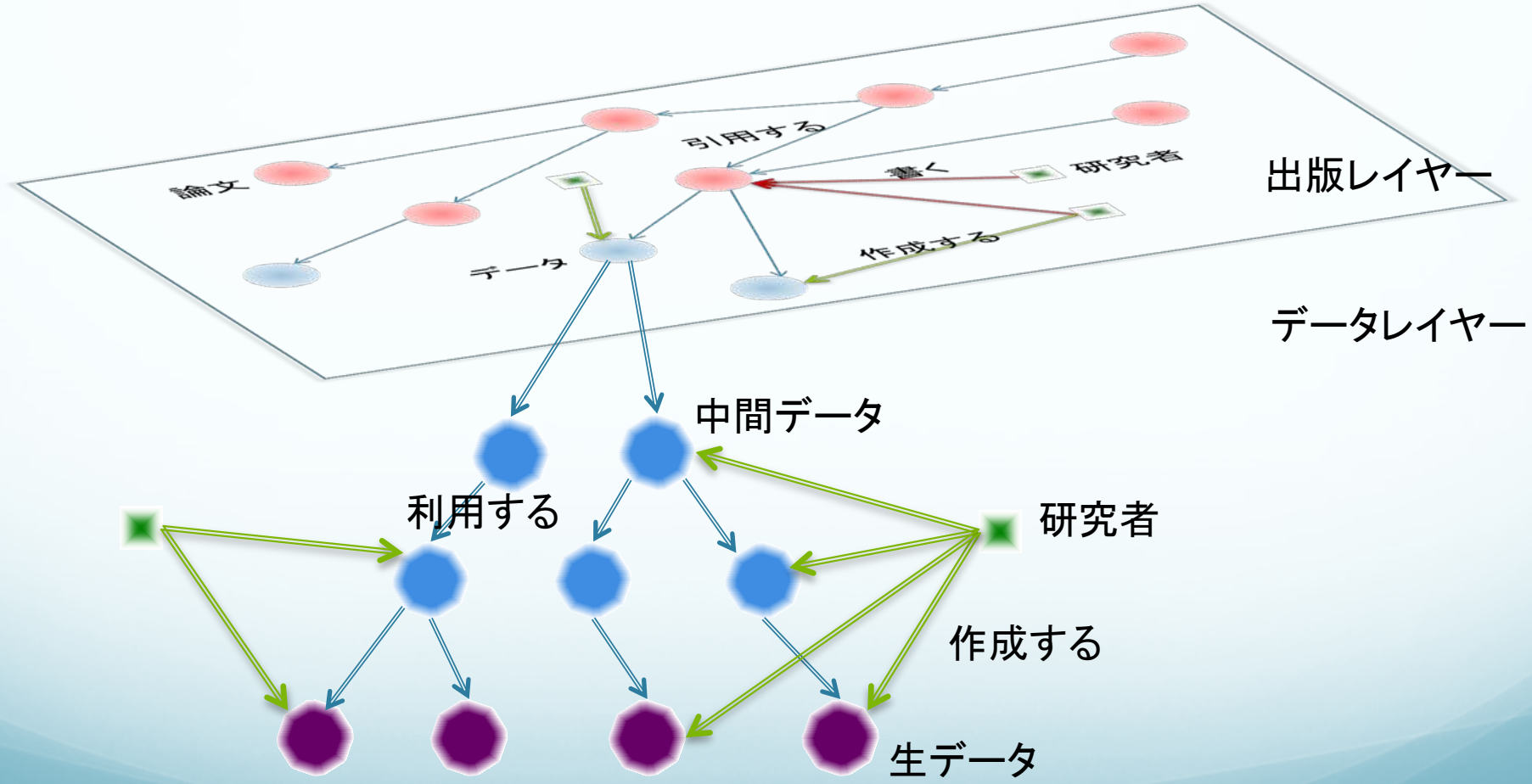
(Open Archives Initiative – Object Reuse and Exchange)

- Open Archives Initiativeが2008年に公開
- あらゆる学術情報リソースをURIで表現する
- 雑誌や論文、論文本体の包含関係を記述する
- 雑誌における論文の引用関係を記述する

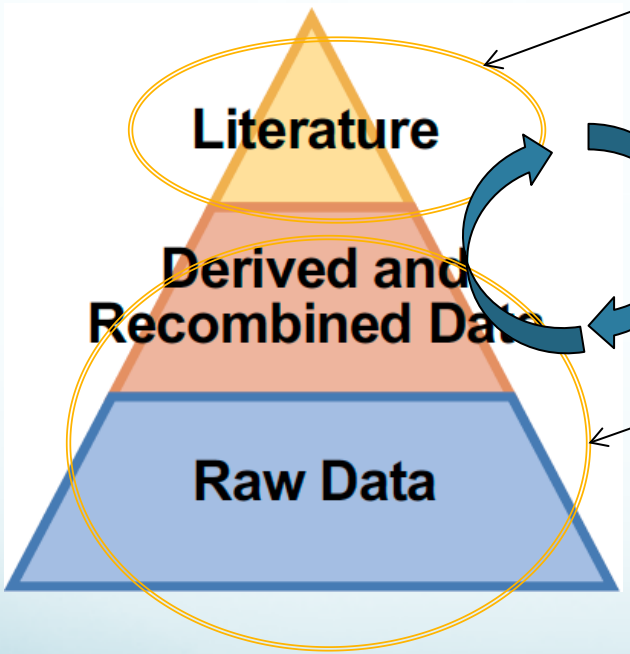
http://www.openarchives.org/ore/

- URIで表現された学術コミュニケーション上の概念に対して、最低限の関連性を規定する。リソースには、以下の4つの概念クラスが用意されている。
  - Aggregation（集合体）
  - AggragatedResources（被集合リソース）
  - ResourceMap（リソースマップ）
  - Proxy（プロキシ）
- 4つの概念クラスに分類されたリソースに付随して用意された語彙は以下のとおりである。
  - ore:aggregates（～を集める）
  - ore:isAggregatedBy（～に集められる）
  - ore:describes（～を記述する）
  - ore:isDescribedBy（～に記述される）
  - ore:similarTo（～に類似である）
  - ore:proxyFor（～のためのプロキシである）
  - ore:proxyIn（～にあるプロキシである）
  - ore:lineage（～をひとつ前とする）

データレイヤーとの相互展開



インターネット上の学術情報流通の飽くなき展開



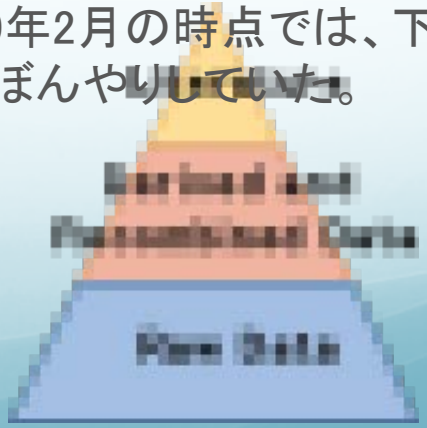
NIIのような文献事業からの展開

ネット上の技術トレンドの上に繰り広げられる
相互波及

IUGONETのような
データサイエンスからの展開

IUGONET

- 2009年 スタート
- 図書系のDSpaceをカスタマイズ
- 解析ソフトはIDL
(ドメイン研究者の大反対にあったが、当初はJython, Java, Java Web Startで書こうと提案していた…)
- 2009年2月の時点では、下図のようにぼんやりしていた。



IUGONETのメタデータ

- Raw Dataファイルと1対1で紐づく粒度で、メタデータを記述(Granuleリソースタイプ)。
- 知見情報の記述は、現在していない。(Annotationリソースタイプで記述可能)
- Derived Dataに紐づくメタデータは、現在記述していない。(例外: Dstインデックス等、専門家のコンセンサスが得られており、もはや一次データと同様に取り扱われるもの)

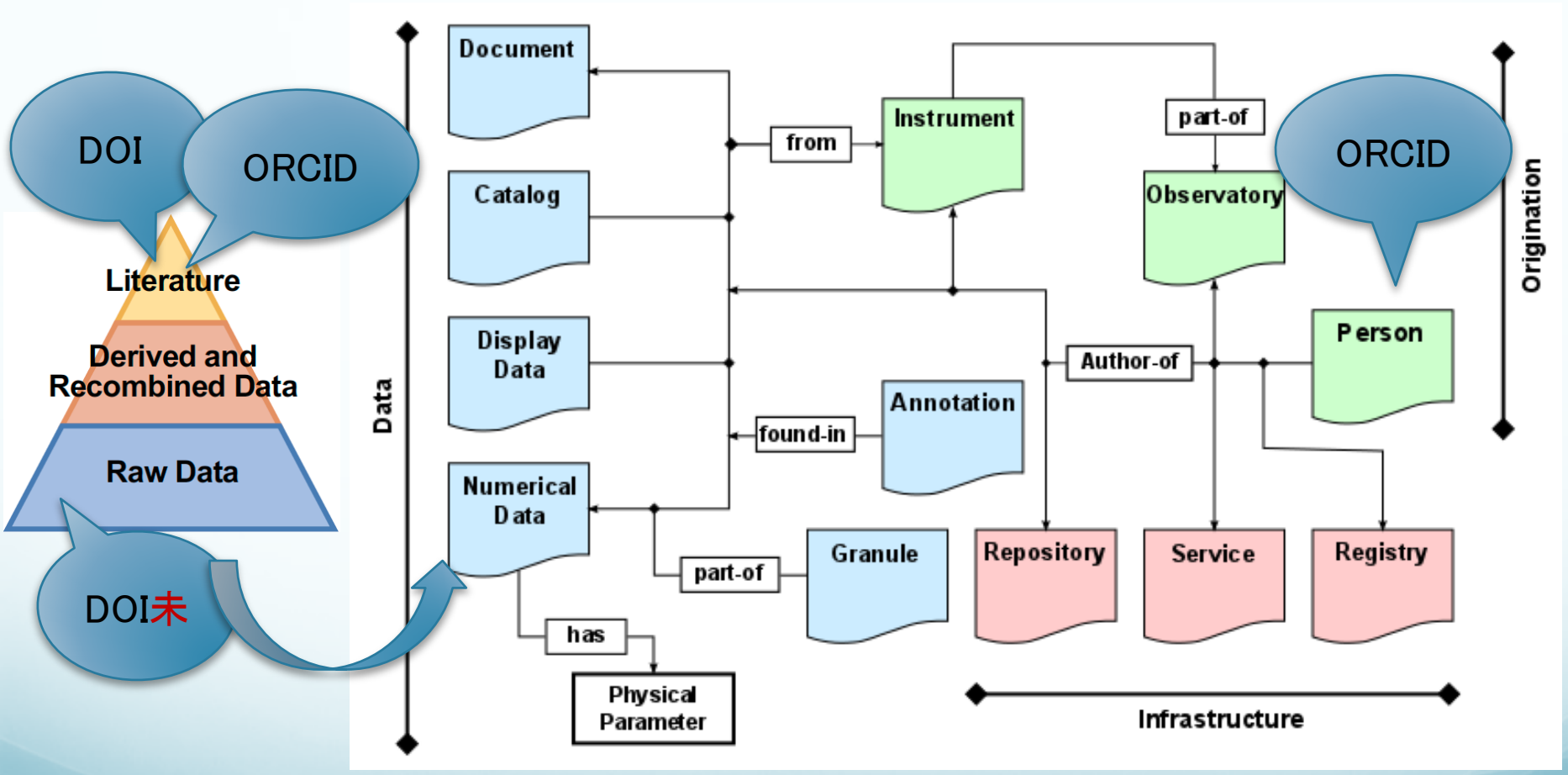
↓

Raw Data -> Derived Dataの変換過程(メタデータ)を、
データ解析ソフトウェアが自動生成する必要あり!?

異なる視点からの Data Citation

- データ提供者の視点
  1. データセット単位の粒度での参照により、貢献度が明示される。
  2. 実際に使用したデータの期間、利用者が分かる。
- 研究者の視点
  - データファイル単位の直接参照は必ずしも便利では無いはず。
  - 中間層である、Derived Dataを介してRaw Dataへ到達する必要あり。マシンリーダブルな変換過程記述の必要性。
- メタデータ提供者の視点(IUGONET)
  - メタデータ整備に尽力した、貢献を明示する必要あり。

データ/メタデータ作成者の収益構造



ORCID IDを検索キーとした“論文”と“データ/メタデータ”の横断検索によるバランスシート的な物の作成

構想

- 「超高層物理学分野における観測データのメタデータDBと著者IDの連携に関する調査」

から

- 「太陽地球系物理学分野におけるデータ集約型科学への検討（仮）」へ展開予定(2013年度)。

↓

- ほぼドメイン研究者による手製の無骨な仕組み(データベース、解析ソフトウェア)に対し、情報系の研究者を段階的に巻き込み、IUGONET2が出航する際の航海図を作る。

追記：第2層のイメージ

Githubそのもの！？

謝辞

- 「超高層物理学分野における観測データのメタデータDBと著者IDの連携に関する調査」は、情報・システム研究機構の新領域研究センターにおける、「融合研究シーズ探索提案」のサポートを受けています。

- (代) 佐藤由佳(NIPR)、田中良昌(NIPR)、蔵川圭(NII)、
  小山幸伸(京大)