国立情報学研究所 学術情報システム総合ワークショップ，2016年6月30日，東京

オープンサイエンスとオープンデータ

武田英明
takeda@nii.ac.jp

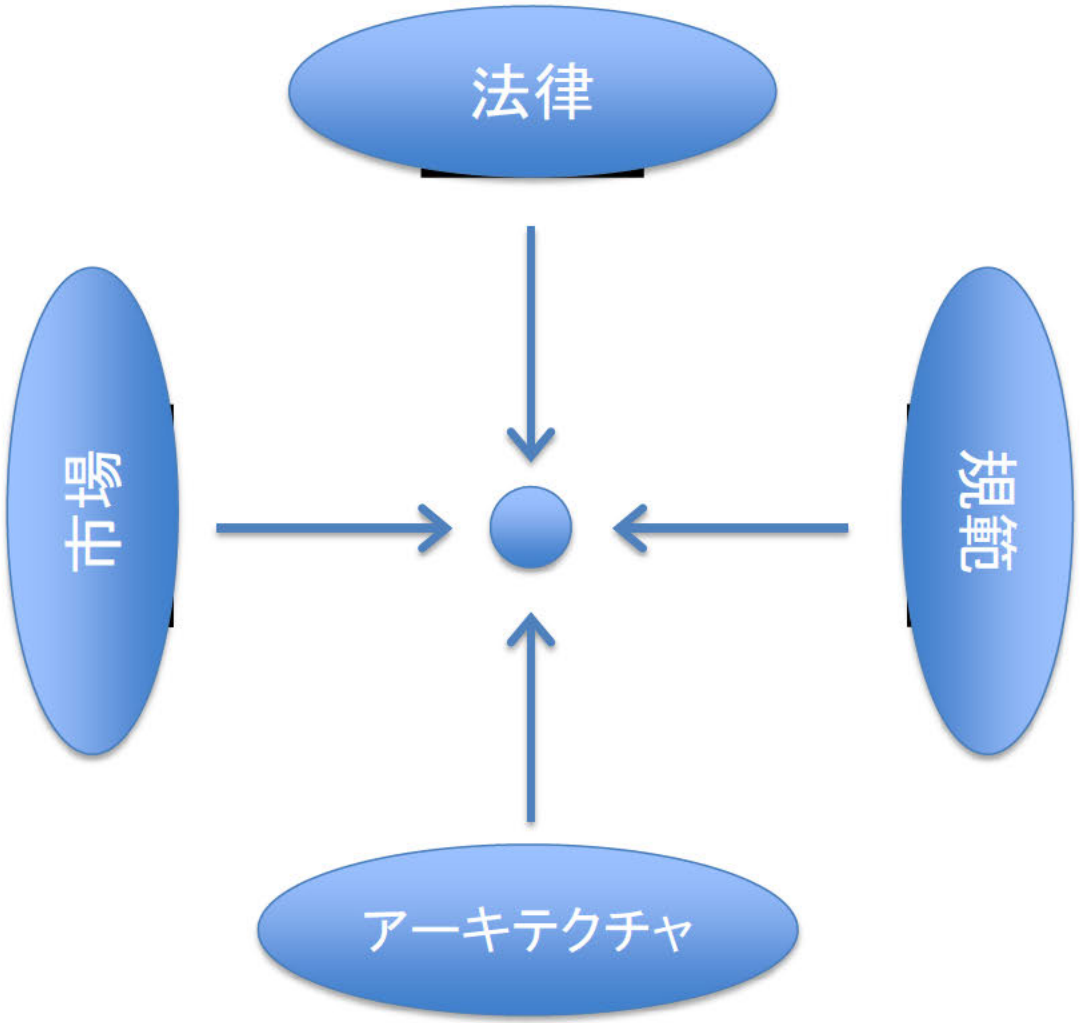
オープンサイエンスから
オープンデータへ

オープンサイエンス
- なぜオープンにするのか-

- 社会的要請
  - 社会での成果の共有、知識の共有
  - 公的資金の公共性のため
    - 例：研究資金助成機関のオープンデータポリシー
- 研究のオープン性
  - 研究の発展性のため
    - “巨人の肩に乗る”
  - 再現性担保のため

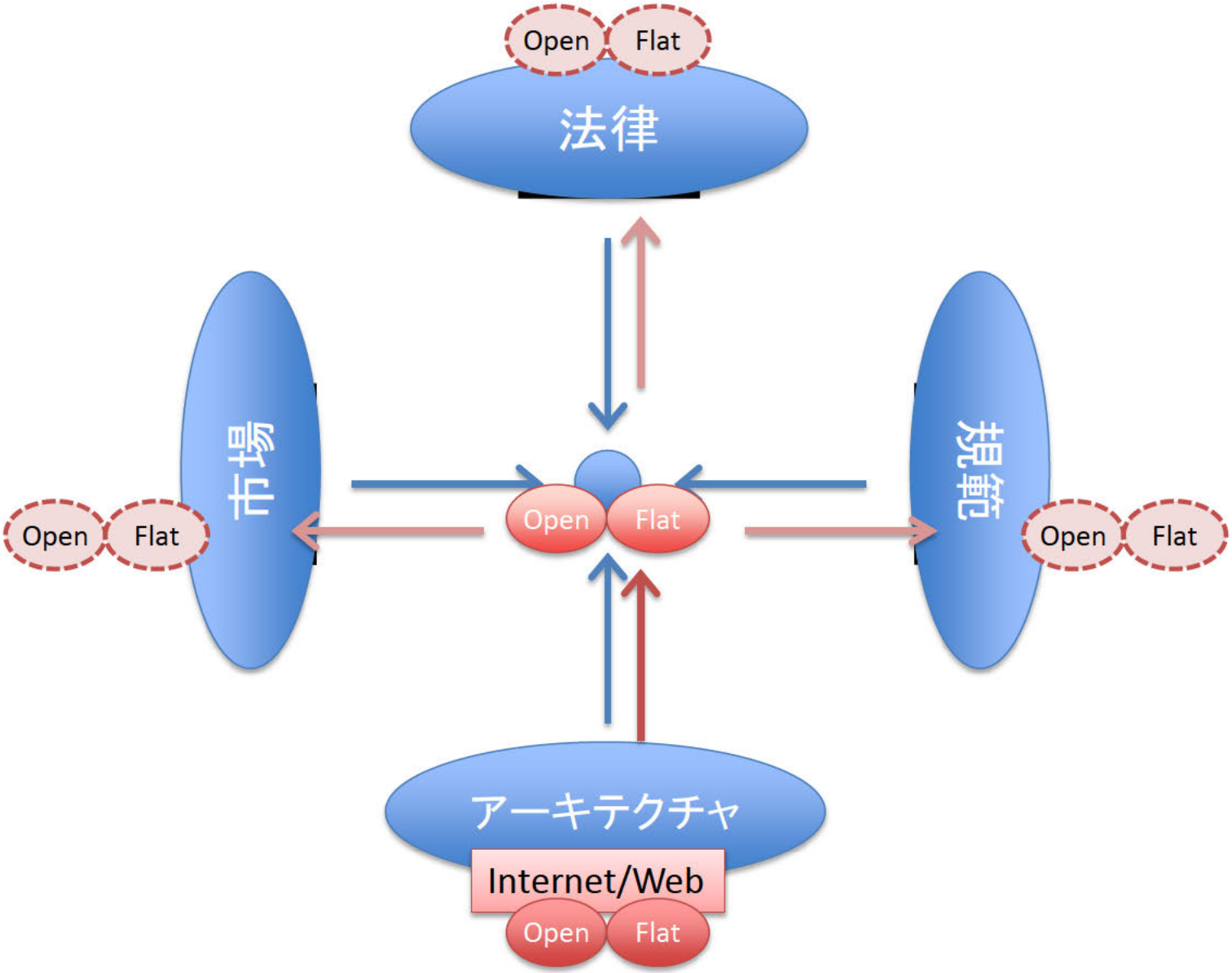
なぜいまオープンサイエンスなのか

社会における個人に対する4つの規制の様相

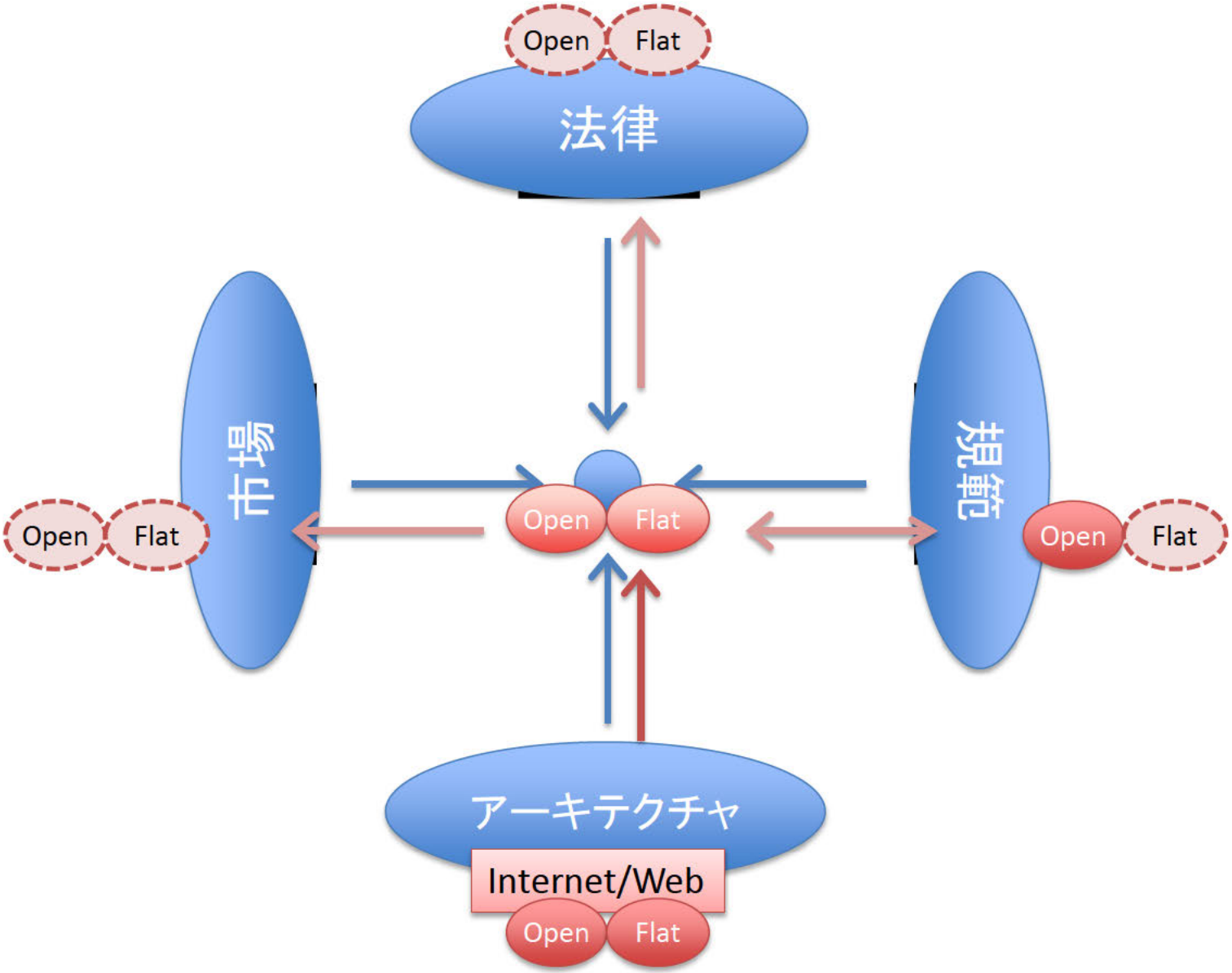


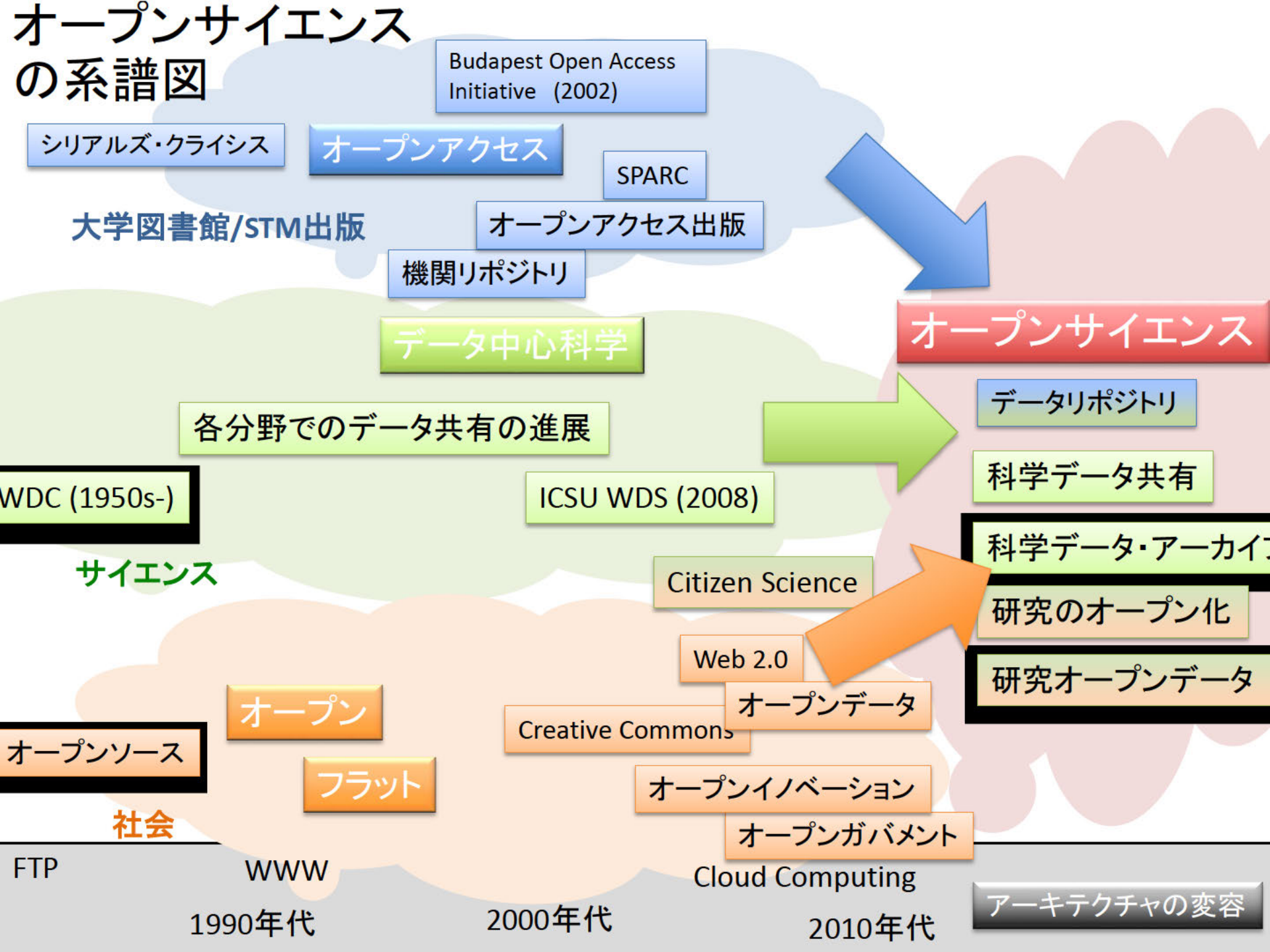
ローレンス・レッシグ：CODE VERSION 2.0, 翔泳社, 2007 (Lawrence Lessig: CODE Version 2.0, Basic Books, 2006)

Internet/Webによるアーキテクチャの変容



サイエンスの変容


Open
Flat
法律
市場
Open
Flat
Open
Flat
規範
Open
Flat
アーキテクチャ
Internet/Web
Open
Flat


オープンサイエンスの系譜図
Budapest Open Access Initiative (2002)
シリアルズ・クライシス
オープンアクセス
SPARC
大学図書館/STM出版
オープンアクセス出版
機関リポジトリ
データ中心科学
オープンサイエンス
各分野でのデータ共有の進展
データリポジトリ
科学データ共有
WDC (1950s-)
ICSU WDS (2008)
サイエンス
Citizen Science
研究のオープン化
Web 2.0
オープンデータ
研究オープンデータ
オープン
Creative Commons
オープンソース
フラット
オープンイノベーション
社会
オープンガバメント
FTP
WWW
Cloud Computing
アーキテクチャの変容
1990年代
2000年代
2010年代

オープンアクセス

- オープンアクセス以前
  - コミュニティ独自の論文共有の世界
    - Preprint共有の伝統（物理学など）
      - -> LANL preprint archive (1991) -> arXiv.org (1999-)
    - Technical reportの刊行
    - 紀要の刊行
    - 論文請求の手紙

オープンアクセス

- シリアルズ・クライシス
  - 大学図書館における購読雑誌の減少
    - 雑誌講読料が高騰
      - 1986年から1999年の間で、雑誌支払単価は207%増加、支払額170%増加、タイトル数6%減少（北米研究図書館協会の調査）[1]
  - 対策
    - SPARC (STM出版社の非競争環境を変える）
    - オープンアクセス

[1] Case, M. M. (2002). “Capitalizing on Competition: The Economic Underpinnings of SPARC” . SPARC.

オープンアクセス

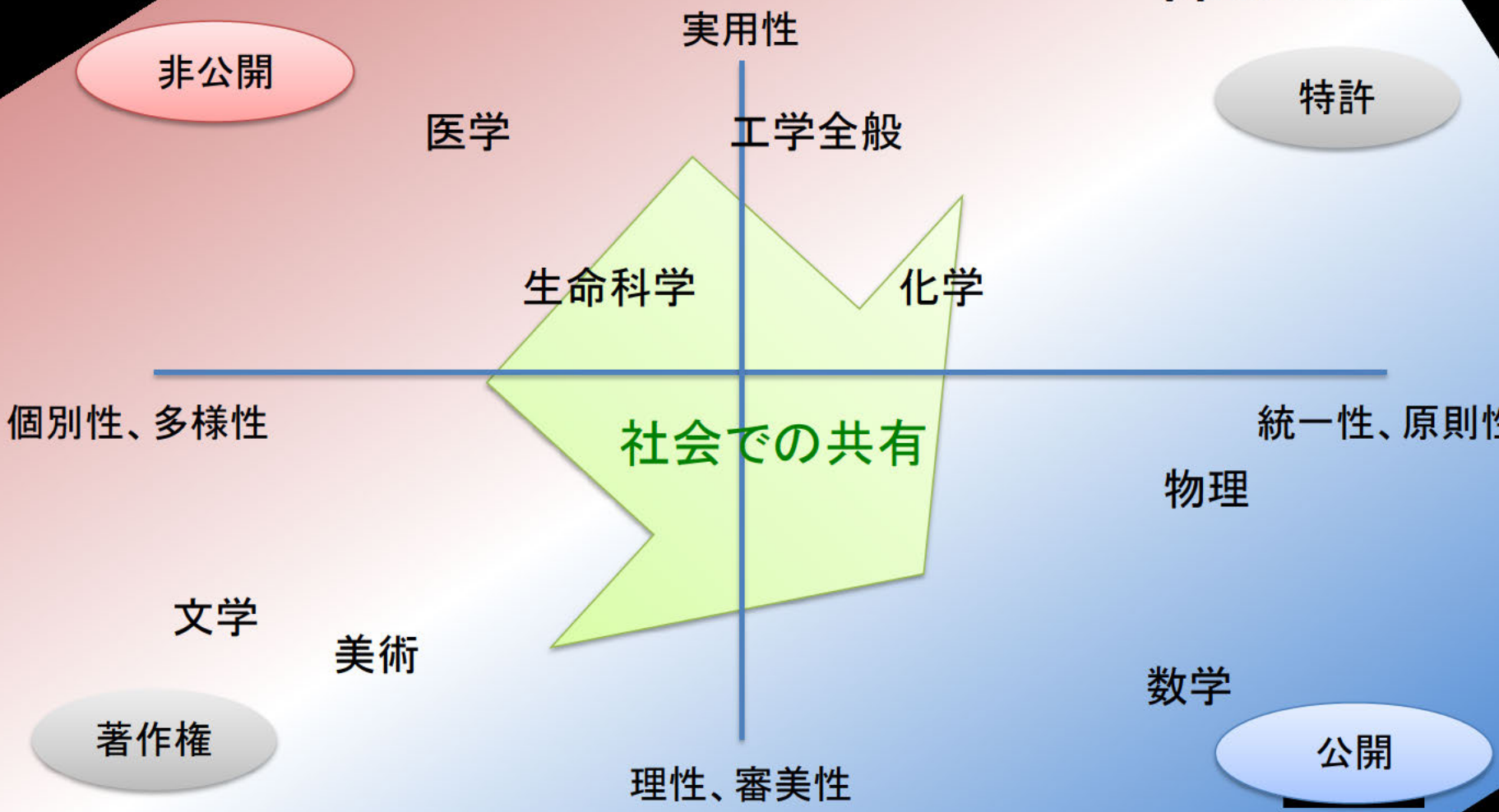
- 論文などの学術情報を無償で自由に利用できるようにすること
  - Budapest Open Access Initiative (2002)
- 理由
  - 論文の可視化とインパクトの最大化（他分野でも）
  - 広範な人からのアクセス（途上国の研究者でもアクセスできる）
- 実現方法
  - セルフアーカイブ（グリーン・ロード）
    - プリプリントサーバ
    - 機関リポジトリ
  - オープンアクセスジャーナル（ゴールド・ロード）
    - APC
- 学術情報
  - 査読付き論文
  - そのほか論文
  - ソフトウエア、データ
- 利用の程度
  - アクセスして閲覧できる
  - 利用や加工ができる

科学でのデータ共有

- 分野ごとのデータ共有
  - 天文学
  - 素粒子物理学
  - 生命科学
  - 地球惑星科学
  - 生物多様性
  - 社会科学
  - …
- 分野ごとにデータ共有の特性が違う
  - 共有／公開
  - データ量
  - 集中／分散

...る情報・データの共有可能...

[1]を元に筆者が...



[1] 有田正規：バイオインフォマティクスの現状とデータシェアリングの可能性について、「データシェアリングを利用...学技術」に関する勉強会　開催記録 第2回　2015/4/1, 文部科学省・科学技術振興機構

天文学

[1]を参考に筆者が作成

- データ量：巨大
  - すばる望遠鏡のデータ: 30TB – 150TB
  - ALMA: 200 TB/year
  - LSST: 6.8 PB/year
- データ保存・公開：各機関ごと
- 公開／非公開：専有期間のあと公開
- 公開の理由：
  - 科学的成果の最大化
  - 高い観測時間獲得競争に対応
- 課題：
  - 分散されたデータへのアクセス（仕様、取得時間、探索）
  - 巨大観測データの処理
- 挑戦：
  - International Virtual Observatory Alliance
    - 天文データの共有を効率化するための標準仕様策定

[1] 大石雅寿: 天文学におけるデータ共有、「データシェアリングを利用した科学技術」に関する勉強会 第4回、 2015/4/21, 文部科学省・科学技術振興機構

生命科学

- 多種・多様・大量のデータ
  - 重要なデータは集約・管理
    - ゲノム配列 : NCBI/EMBL/DDBJ
    - タンパク質: PDB
    - …
  - さらに多くのデータ
    - NBDCによる管理の例
- 公開／共有 :
  - public fundをもののは公開
  - プライバシーに関わるデータは非公開
  - 製薬などの競争分野では非公開
- 量
  - 次世代シーケンサー等では大量データが生成

データ共有のメリット

- データの早期公開はよりよい成果が期待できる
  - エラーの早期発見、早いコミュニティ形成
- 一つのデータから多様な研究
- 再現可能性
- 他データとの結合
- 学際的研究の促進
- データの保全
- サイテーション
- 教育やアウトリーチ
- 社会や市民科学とのつながり

Data sharing in astronomy, Željko Ivezić, Department of Astronomy, University of Washington
http://www.astro.washington.edu/users/ivezic/Outreach/Talks/NAS2011_Ivezic.pdf

データ共有のデメリット

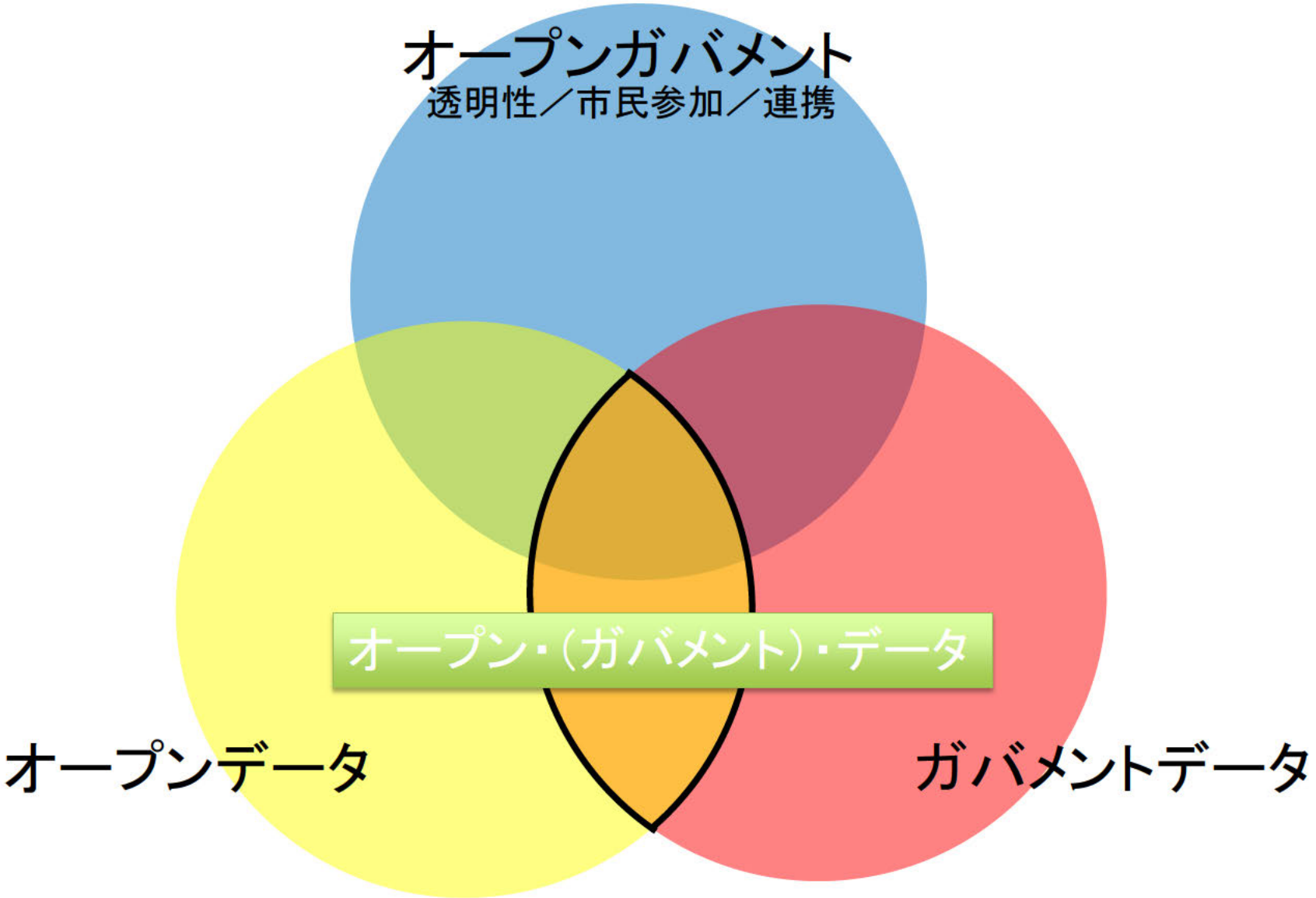
- 内部利用より高度な“標準化”の必要
- キュレーション
- 維持コスト
- 横取り研究の可能性

Data sharing in astronomy, Željko Ivezić, Department of Astronomy, University of Washington
http://www.astro.washington.edu/users/ivezic/Outreach/Talks/NAS2011_Ivezic.pdf

オープンデータとは

- なんのデータのこと？
- 単にデータをオープンにすればオープンデータ？

オープンデータとオープンガバメント



オープンで/フラットな社会への変化

- 国・統治のあり方
  - オープンガバメント
- 経済活動のあり方
  - オープンイノベーション
- 科学技術のあり方
  - オープンサイエンス

オープンガバメント

- 透明性 (transparency)
- 市民参加 (participation)
- 政府内および官民の協働 (collaboration)
  - President Barack Obama, Memorandum for the Heads of Executive Departments and Agencies, 2012/2

オープンデータとは

- なんのデータのこと？
- 単にデータをオープンにすればオープンデータ？

オープンデータとは

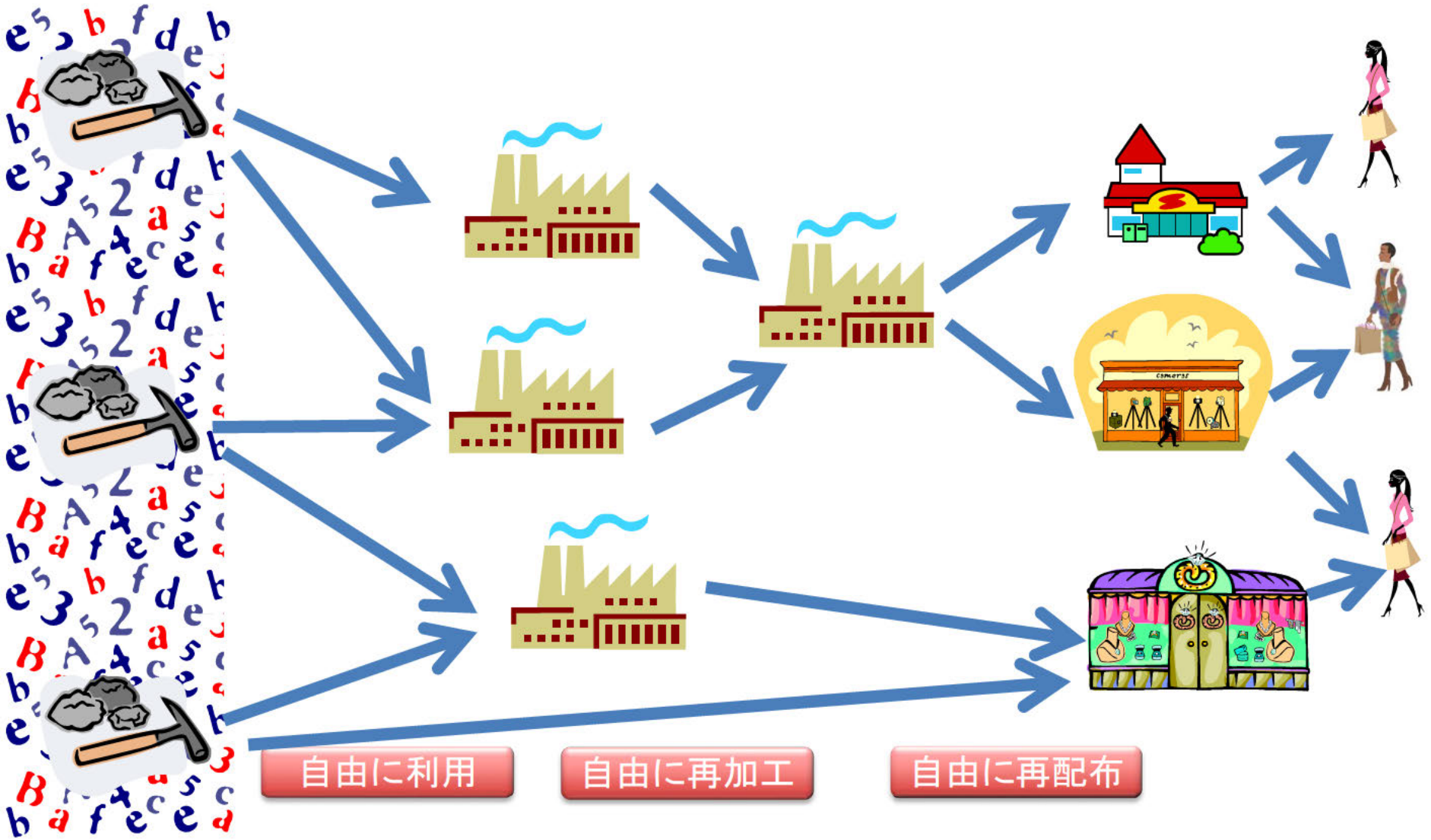
- オープンデータとは、誰でも自由に使えて再利用もでき、かつ再配布できるようなデータである。課すべき決まりは、たかだか「作者のクレジットを残す」あるいは「同じ条件で配布する」程度である。
  http://opendatahandbook.org/ja/what-is-open-data/
- “A piece of data or content is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike.” http://opendefinition.org/

オープンデータとは

- 利用できる、そしてアクセスできる
  - データ全体を丸ごと使えないといけないし、再作成に必要以上のコストがかかってはいけない。望ましいのは、インターネット経由でダウンロードできるようにすることだ。また、データは使いやすく変更可能な形式で存在しなければならない。
- 再利用と再配布ができる
  - データを提供するにあたって、再利用や再配布を許可しなければならない。また、他のデータセットと組み合わせて使うことも許可しなければならない。
- 誰でも使える
  - 誰もが利用、再利用、再配布をできなければならない。データの使い道、人種、所属団体などによる差別をしてはいけない。たとえば「非営利目的での利用に限る」などという制限をすると商用での利用を制限してしまうし「教育目的での利用に限る」などの制限も許されない。

http://opendatahandbook.org/ja/what-is-open-data/

データは情報流通社会の資源

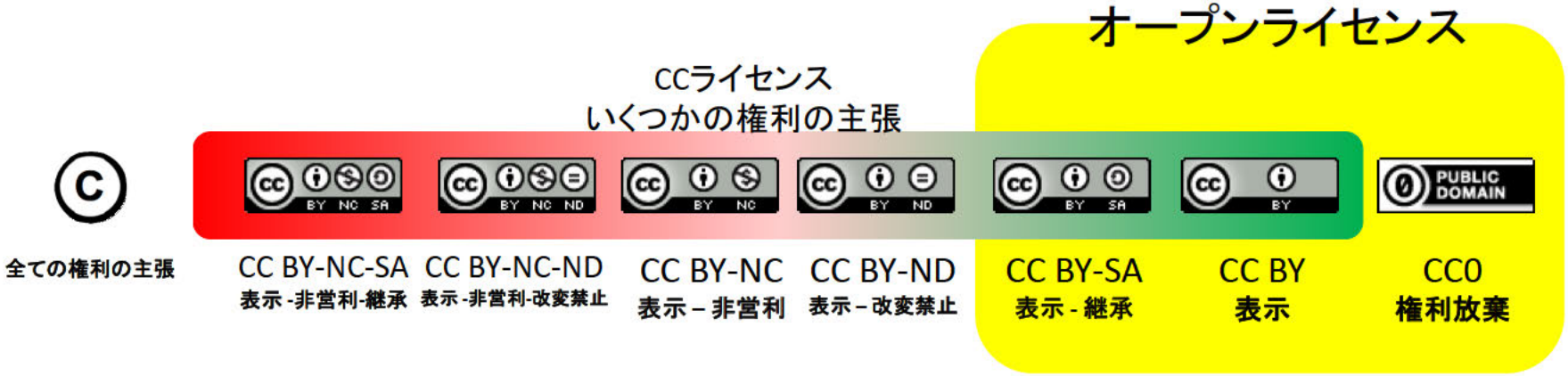


オープンデータに必要なもの

- オープンライセンス
- 機械可読フォーマット

オープンライセンス

- 情報を最小限の制約以外で自由に使うことを許すライセンス



機械可読フォーマット

- 再利用性を高める
  - 内容を切ったり、はったりできること

- 機械（コンピュータ）が内容を処理できる形式が望ましい
  - 特定のプログラムで処理できる
  - オープンなフォーマットで公開

オープンデータへの5つのステップ

★★★★★ OL RE OF URI LD

他へのリンクを入れたデータを公開

★★★★ OL RE OF URI

★★★ OL RE OF

RDF（とSPARQL）でデータ公開　例：RDFa, RDFストア

★★ OL RE

CSV

オープンに利用できるフォーマットで公開　例：csv

★ OL

コンピュータが処理可能なフォーマットで公開
例：xls, doc

PDF

どんなフォーマットでよいからオープンライセンスでデータ公開　例：PDF, jpg

W3C RDF OPEN DATA HTML

WD

ライセンスをつけずにデータをWebで公開

研究オープンデータを支える情報基盤とは

Internet/Web時代の研究

- すべてがデジタルへ

デジタル化以前の研究者

研究と執筆

論文

文献調査

論文執筆

データ

対象

現在の研究者

研究と執筆とデータ生成

論文

文献調査

論文執筆

データ

データ利用

データ公開

対象

今後の研究者

論文・データの一体化

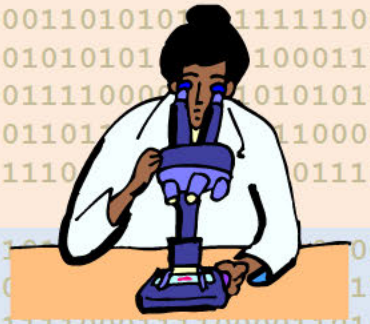
研究成果＝データ生成

論文

データ

対象

データ利用

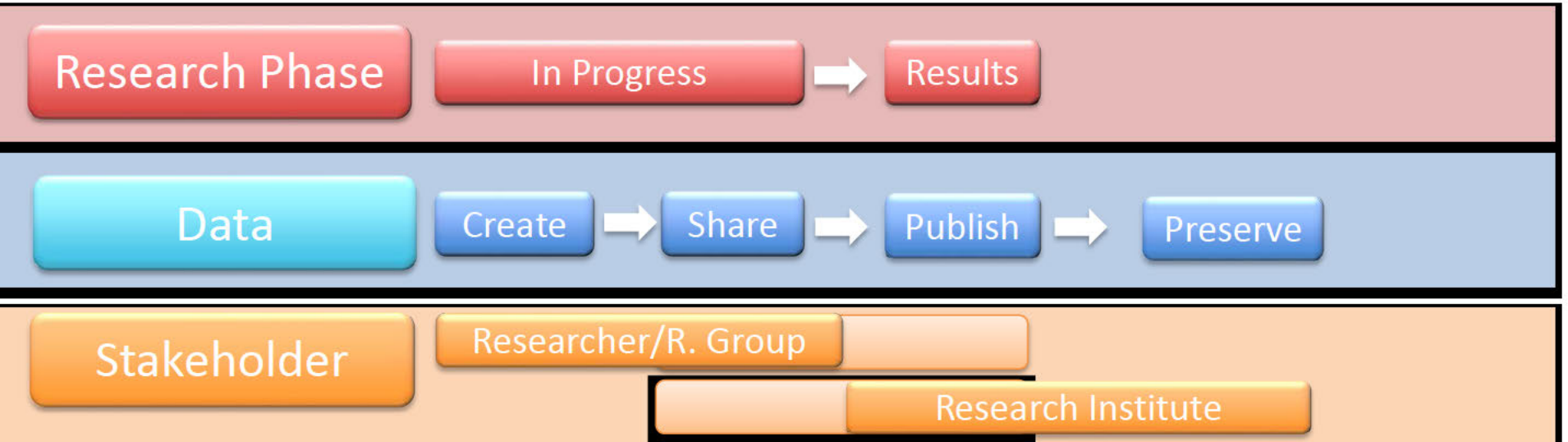


データ公開

研究＝データのサプライチェーン

Data Life Cycle

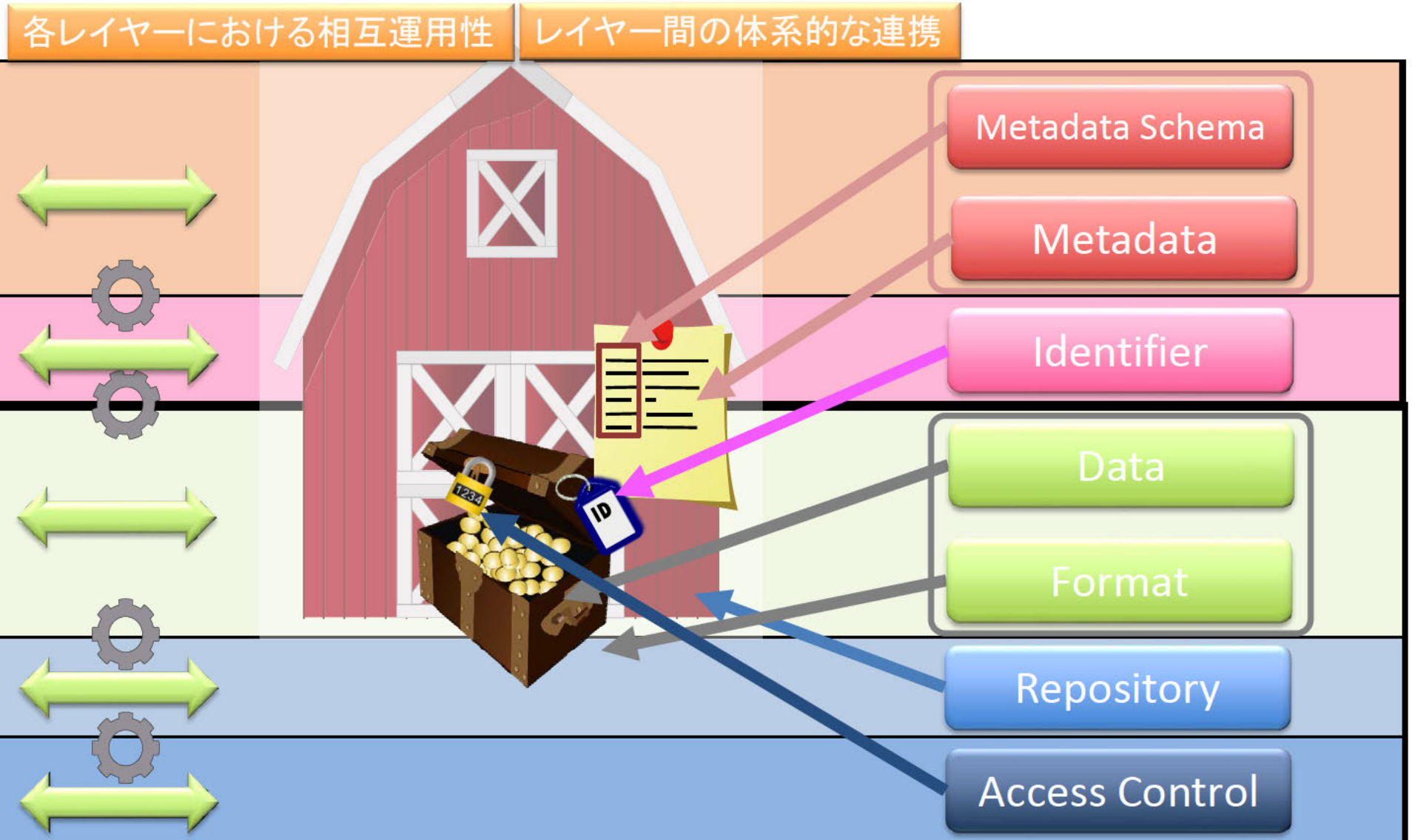
- データは作られ, 共有され, 公開され, 保存される
  - 多くは共有からオープンに公開へ
  - 一部(プラバシーデータ、セキュリティデータ)は共有のまま
- データのライフサイクルを通じたサポート



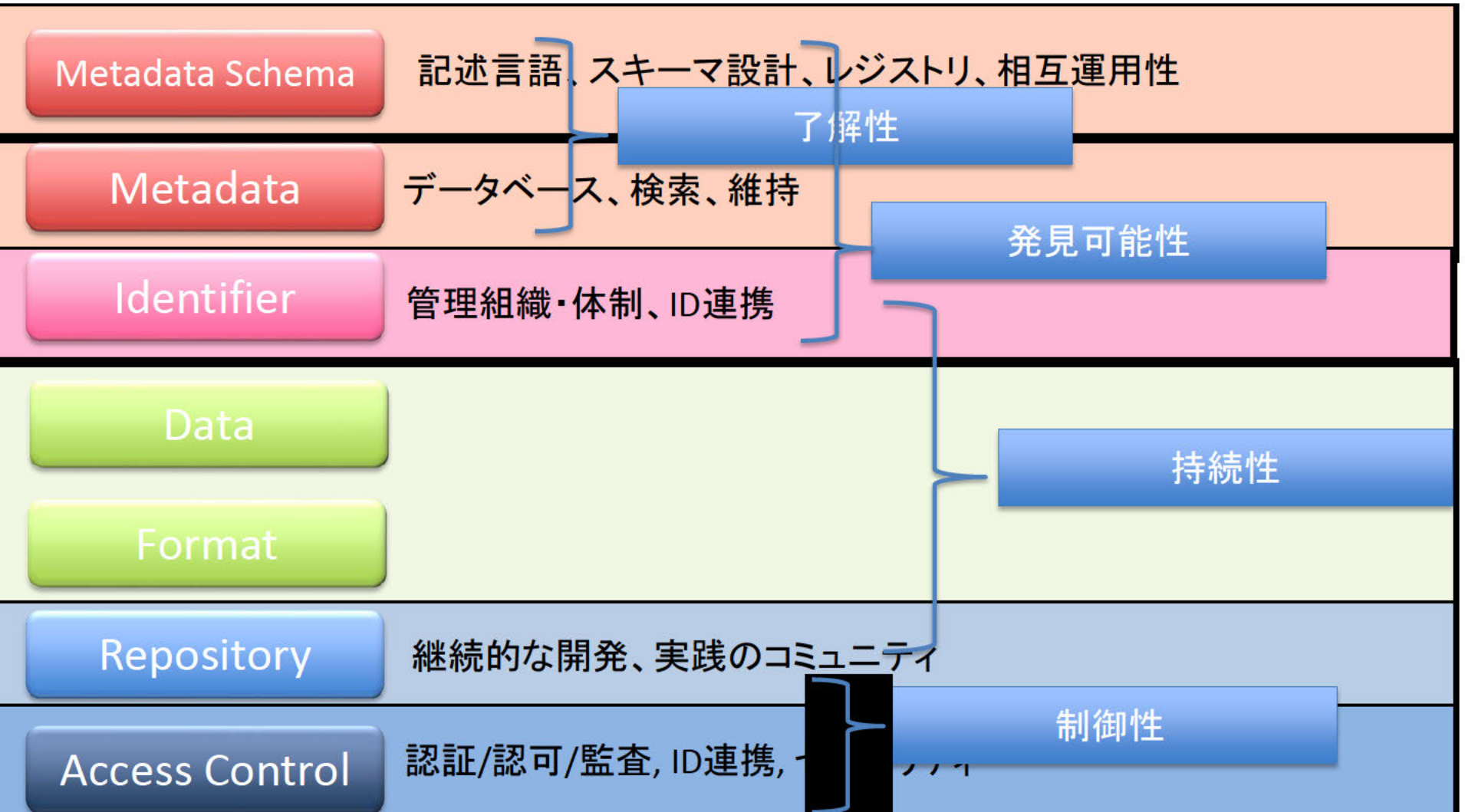
研究データ共有への要求

- 了解性(Understandability)
- 発見可能性(Findability)
- 持続性(Persistency)
  - データに対する持続性(Persistency for data)
  - メタデータに対する持続性(Persistency for metadata)
- 制御性(Controllability)

研究データ共有のアーキテクチャ



研究データ共有のアーキテクチャ



研究データ共有のアーキテクチャ

協調と競争

Metadata Schema

記述言語、スキーマ設計、レジストリ、相互運用性

Organization: DataCite, CrossRef, JaLC

Schema: Dublin Core, DCAT

System: CKAN

Technology: Linked Data

Metadata

データベース、検索、維持

Identifier

管理組織・体制、ID連携

DOI, ORCID, FundRef

Data

Format

Repository

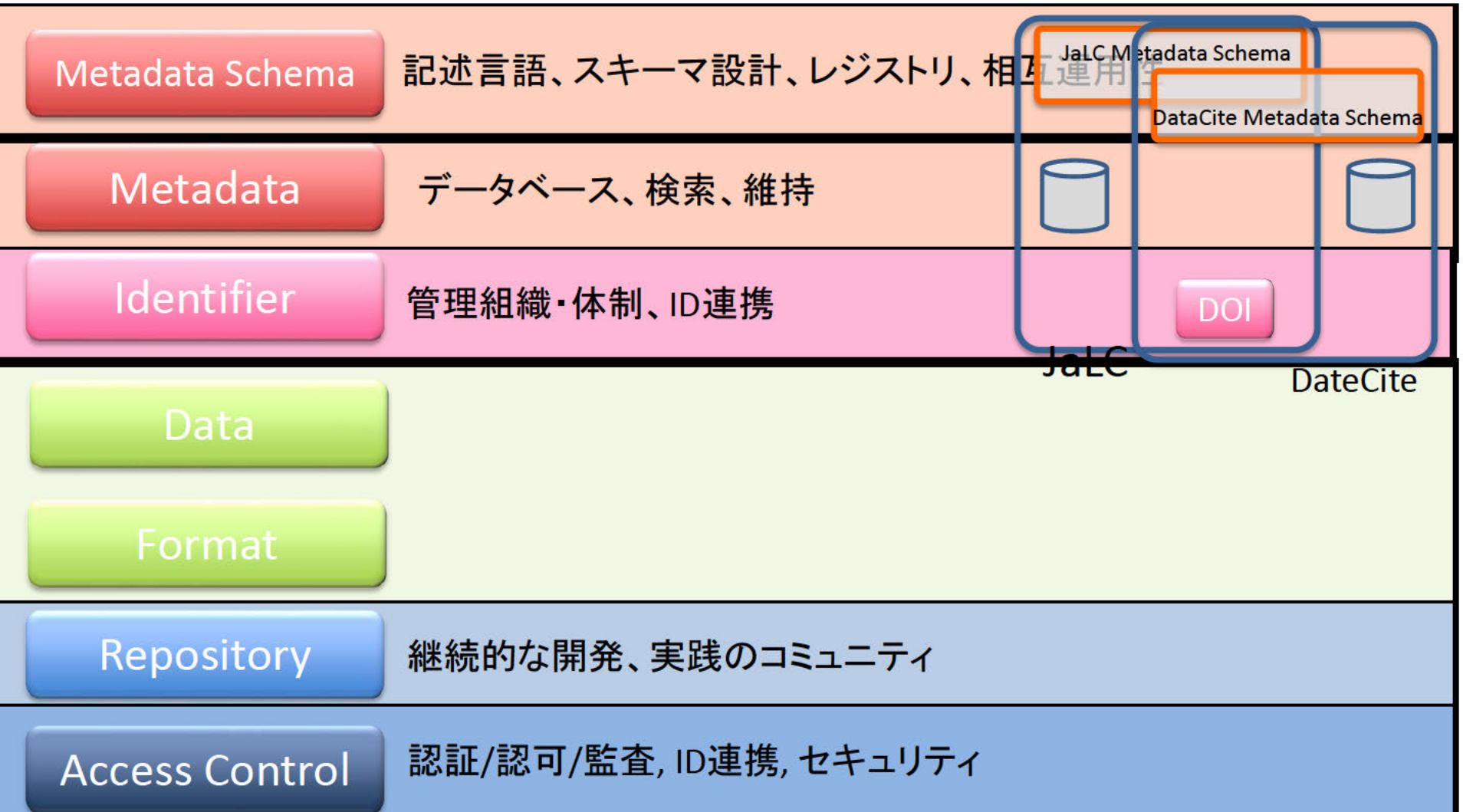
継続的な開発、実践のコミュニティ

Dspace, Fedora, Weko

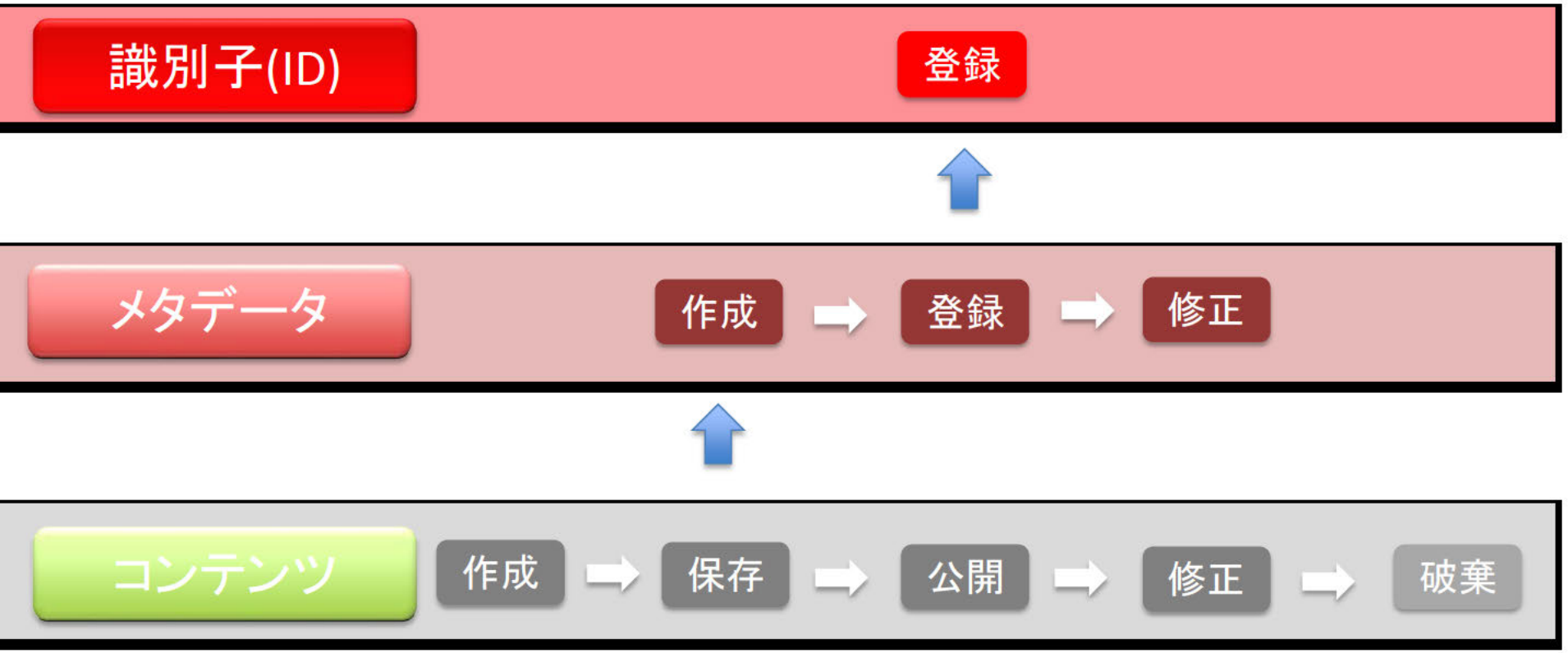
Access Control

認証/認可/監査, ID連携, セキュリティ

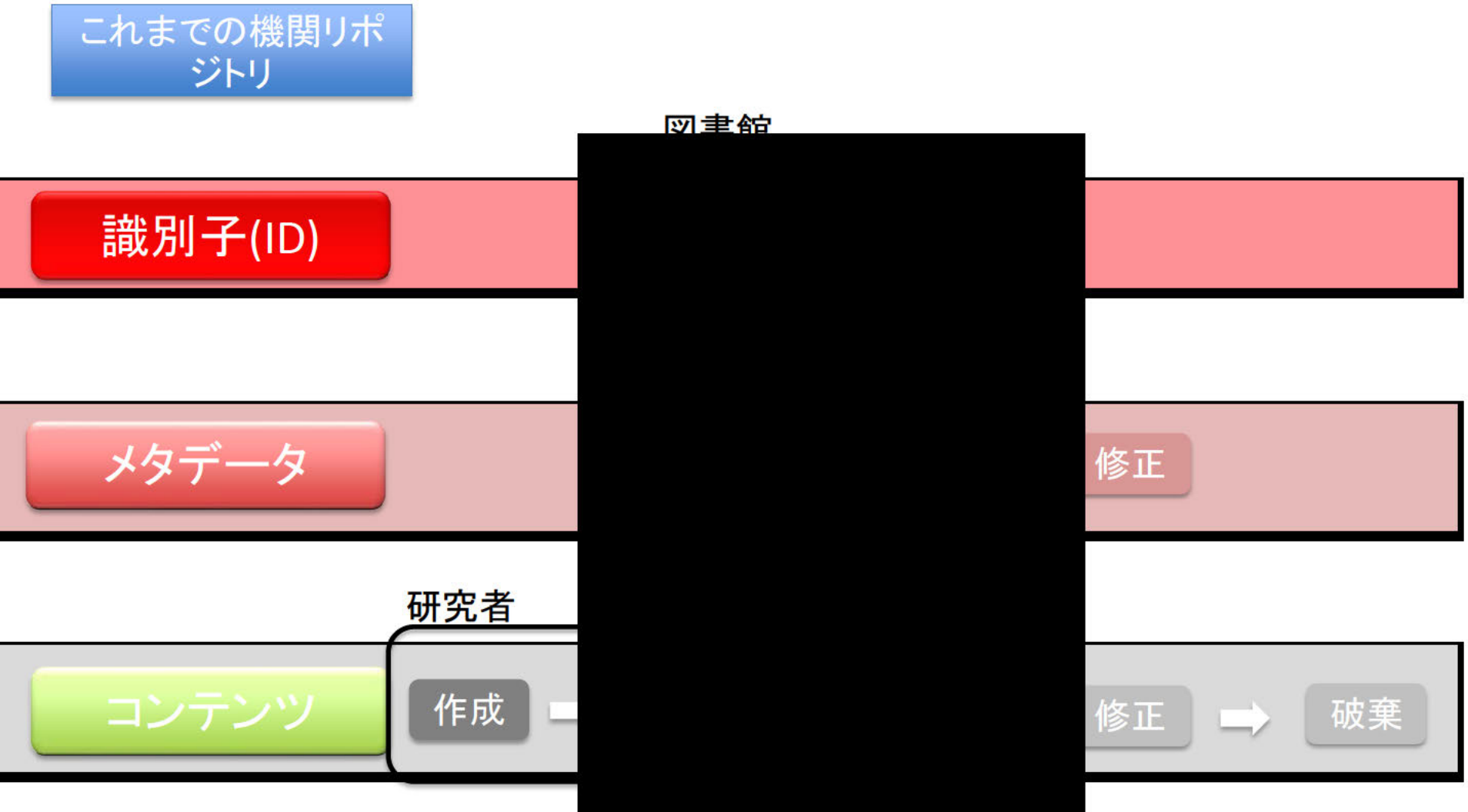
研究データ共有のアーキテクチャ



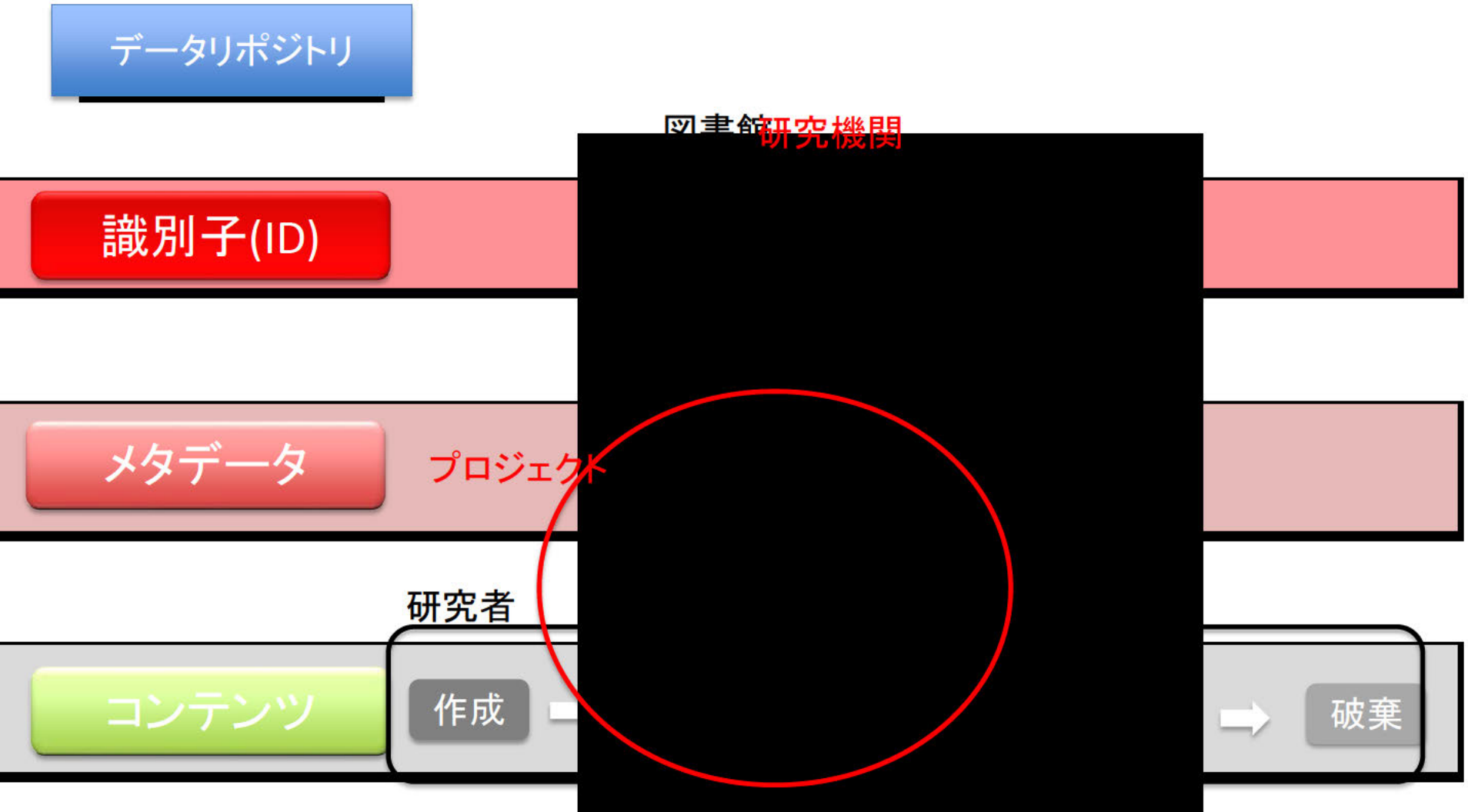
データのライフサイクルと担当者・担当機関



データのライフサイクルと担当者・担当機関



データのライフサイクルと担当者・担当機関



まとめ

- 「いまオープンサイエンス」なのはInternet/Webのおかげであるが、単に技術的な理由だけではなく、それが規範や市場、法律にも影響を与えてきた結果である。
- 研究成果はいずれ「データ」になる
- 研究データ流通基盤は必須の仕組み
- 研究データ流通はいくつかのレイヤー
  - 識別子、メタデータ・スキーマ、メタデータ、コンテンツ、フォーマット、リポジトリ
- 沢山のTO DO
  - 世界における研究データ流通のそれぞれのレイヤーでの**「協調と競争」**にいかに加わっていくか