平成25年度学術情報システム総合ワークショップ
2013.7.12 国立情報学研究所

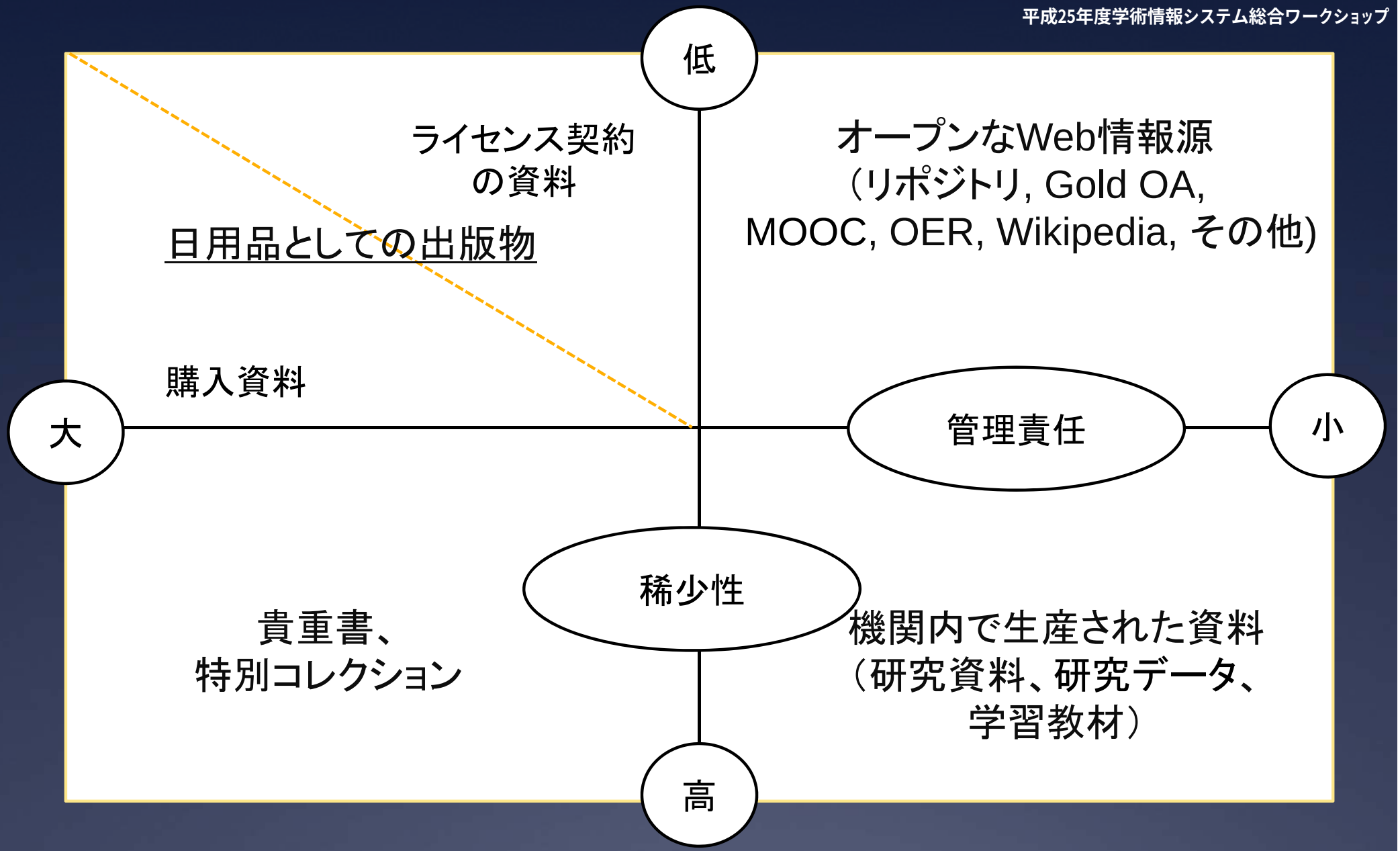
デジタル化資料をめぐる世界の動向

佐藤　義則
東北学院大学文学部

目次

1. 図書館とデジタル化資料の範囲の拡大

2. 新たな発見環境と書誌フレームワーク
   - ディスカバリーと図書館システム
   - 新たな書誌フレームワークへ向けた展開



引用元: Malpas, Constance. "Scarcity and Abundance: the Cooperative Imperative in Special Collections," 53rd Annual RBMS Preconference, 20 June 2012. available at http://www.oclc.org/research/presentations/default.htm (2012-07-04).

デジタル化の進展

- 電子ジャーナル
  - OAジャーナル、OAメガジャーナル、カスケード出版、、、
  - 機関リポジトリ – JAIRO Cloud…
- eBooks
  - 欧米の出版社プラットフォーム
  - 出版デジタル機構
  - Kindle, iPad, kobo, e-Book reader,,,
- マス・デジタイゼーション： Google Books, Internet Archive – Open Library, HathiTrust, DPLA,,, 国立国会図書館

出版社は，しばしば酷評される「ビッグディール」パッケージを通じて，雑誌文献へのより大きなそしてより平等主義的アクセス，真のオープンアクセスへの近似を提供している。また，こうした過程で出版社は図書館を追いやり，学術コミュニケーションに投入される資源のより大きな共有を実現している。このことは，出版社の利益の継続とともに，「持続不可能な雑誌価格上昇」と呼ばれてきた数十年を可能にしてきた。また，オープンアクセスの拡大を阻害しており，大規模なライセンス契約を通じた流通管理により出版社の寡占を導く可能性がある。

Andrew Odlyzko
Open Access, library and publisher competition, and the evolution of general commerce.
http://arxiv.org/abs/1302.1105

Google Books Settlement

2005.9- 10	The Authors GuildとAAP(Association of American Publishers)がGoogle を著作権侵害で提訴（クラスアクションとして）
2008.10.28	訴訟当事者3社が和解を申立て
2009.9.18	米国政府、裁判所に対し旧和解案の成立に反対する意見書提出
2009.10.7	裁判所、訴訟当事者に対し和解条項の修正を命令 （Googleはこの時点までに書籍1000万冊以上をデジタル化）
2009.11.13	訴訟当事者、裁判所に対し修正和解条項提出（修正和解案）
2009.11.19	裁判所、修正和解案を予備承認
2010.2.4	米国政府、裁判所に対し修正和解案の成立に反対する意見書提出
2010.2.18	裁判所、修正和解案に関する公聴会を開催 （Googleはこの時点までに書籍1200万冊以上をデジタル化）
2011.3.22	裁判所、修正和解案を認めない裁定を下す

国立国会図書館

- 2009年度補正予算：127億円
- スキャンのみ（デジタル化テキストは作成しない）、館内公開のみ；ただし、許諾の得られたものについてはネット上で公開
- 著作権法の改正（2010.1施行）
- 国立国会図書館サーチ（2012.1 - ）

デジタル化資料の概要

資料群	年代等
古典籍	貴重書・準貴重書、江戸期以前の和漢書等
図書	明治期以降、1968年までに受け入れた図書
雑誌	明治期以降、2000年までに発行された雑誌
新聞	東日本大震災直後に石巻日日新聞社が発行した壁新聞
歴史的音源	1900年初めから1950年頃までに国内で製造されたSP盤等に収録された音楽・演説
官報	1883（明治16）年7月2日（創刊）～1952（昭和27）年4月30日に発行された官報
博士論文	1991～2000年度に送付を受けた論文
憲政資料	幕末から昭和までの日本の政治家・官僚・軍人などが所蔵していた書簡・書類・日記等のうち、電子展示会に出展した資料
日本占領関係資料	米国の国立公文書館が所蔵する戦後の日本占領に関する公文書のうち、米国戦略爆撃調査団文書、極東軍文書の一部
プランゲ文庫	プランゲ文庫（戦後GHQが検閲のために集めた日本国内出版物）のうち一般図書の一部

国立国会図書館「資料デジタル化について」
http://www.ndl.go.jp/jp/aboutus/digitization.html

デジタル化資料提供状況（平成25年6月末時点）

資料種別	デジタル化資料提供数（概数）		
	インターネット公開	館内限定提供	合計
古典籍	7万点	2万点	9万点
図書	34万点	55万点	89万点
雑誌	0.5万点	104.5万点	105万点
新聞	6点	－	6点
歴史的音源	765点	4万点	4万点
官報	2万点	－	2万点
博士論文	1.5万点	12.5万点	14万点
憲政資料	140点	－	140点
日本占領関係資料	1.7万点	700点	1.8万点
プランゲ文庫	－	0.3万点	0.3万点
合計	47万点	178万点	225万点

国立国会図書館「資料デジタル化について」
http://www.ndl.go.jp/jp/aboutus/digitization.html

新たなタイプの電子的情報資源共有

- HathiTrust
  http://www.hathitrust.org/
- Internet Archive “Open Library”
  http://openlibrary.org/
- DPLA (Digital Public Library of America)
  http://dp.la/
- Europeana
  http://www.europeana.eu/

HathiTrust

- 共同の研究用コレクションの保存とアクセス提供
  - Google Books、Internet Archive由来 + 自前のコンテンツ
- 約1,100万の電子化資料（3割弱がパブリックドメイン）
- 訴訟問題（Authors Guild等による提訴）
  - フェアユース
- 『研究用コレクションのクラウド・ソーシング：大規模電子化後の図書館環境における印刷体の管理』※
  - 利用頻度の低い図書の管理をHathiTrustのようなデジタル・リポジトリおよび共有の印刷体保存リポジトリへ外部委託することの可能性と得られる効果（図書館スペースの節約、コスト削減）
  - “大学図書館がHathiTrustと連携し、大規模電子コレクションへのパブリックアクセスの拡大を進めること”（2011.1）

※Malpas, Constance. *Cloud-sourcing Research Collections: Managing Print in the Mass-digitized Library Environment*. OCLC Research, 2011.1, 76 p. http://www.oclc.org/research/publications/library/2011/2011-01.pdf

電子情報資源の確保と共有コレクション構築

- JISC
  - 「コンテンツと電子化・プログラム」(2004 – 2009)
  - 「eコンテンツプログラム」(2009, 2011)
  - 「コンテンツプログラム」(2011 - 2013) - £350m
    A. 電子化とオープン教育情報源(9プロジェクト)、B. 大規模電子化(8プロジェクト)、C. 電子コンテンツのクラスタリング(7プロジェクト
  - Manuscript Online
  - Wikipediaとの連携(2013.6.27)
- JISC Collections
  - eコレクションズ (2011 -)
    - ✓ 商業出版社や商業プロバイダ等による電子資料の調達と提供
    - ✓ 有料制(機関による支払い)

コレクション構築の変化

- 紙媒体でのコレクション構築
  - 資料は「所蔵」
  - 利用者が図書館にやって来ることを前提
  - 事前の選書と受入（＋整理）が重要
    - 資料の検索（特定と配置場所への案内）
- 電子情報資源：　利用者（利用）中心のサービス構築
  - 資料は「ネットワーク上」
  - 利用者も「ネットワーク上」；利用者の必要に合わせた資料の調達も可能　アクセスの永続性に対する不透明さ
  - 資料のグローバルな発見可能性と入手可能性の保証が重要
  - 連携協力が不可欠

	印刷体資料	電子情報資源
資料の在処	図書館内	図書館(内)外
利用対象	(多数ある)コピーの一つ	単一(唯一)の情報源
アクセス	物理的所蔵に基づく	契約や協定に基づく
作成方式	人手による確認、入力	(プログラム等による) 既存データの有効活用
目録処理	共同分担目録 (書誌データと資源の共有)	集中的作業 (典拠データ、リンク形成)
課題	データの品質レベル	データの品質レベル 永続的アクセスの管理 情報源間の関係性の整理

MOOCと図書館

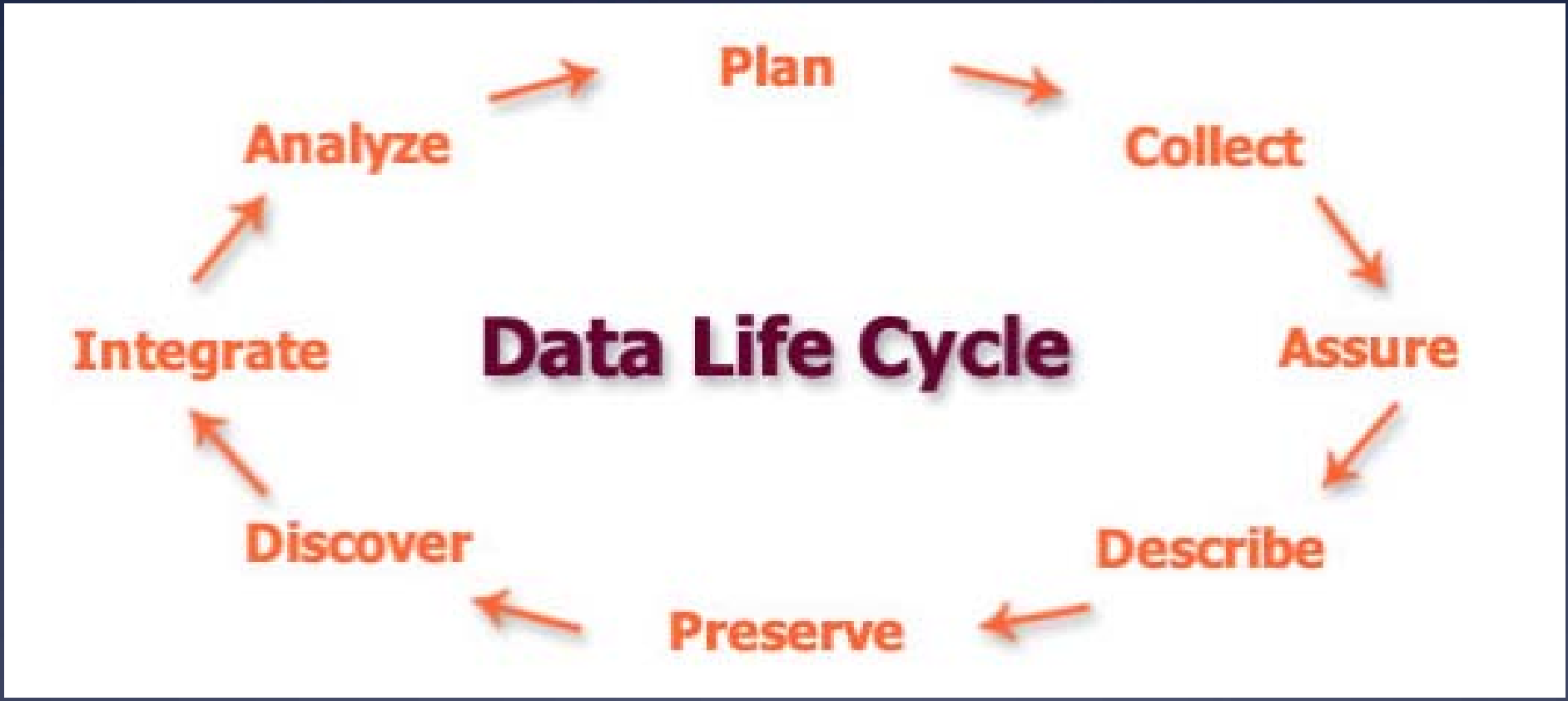
- MOOC: Massive Open Online Course
  - プラットフォームとしての，Udacity, Cousera, edX,,,
  - 講義コンテンツ提供者としての大学
  - 低価格: 非効率解消への期待と不安
- 図書館との関わり合い
  - 著作権，ライセンス関連の相談
  - 学生の関連資料の提供（ライセンス契約，オープンアクセス文献，フリー・テキストブックス，フェアユース?，障がい者サービス）

参照: Brandon Butler　ISSUE BRIEF - Massive Open Online Courses: Legal and Policy Issues for Research Libraries http://www.arl.org/storage/documents/publications/issuebrief-mooc-22oct12.pdf

オープンデータ

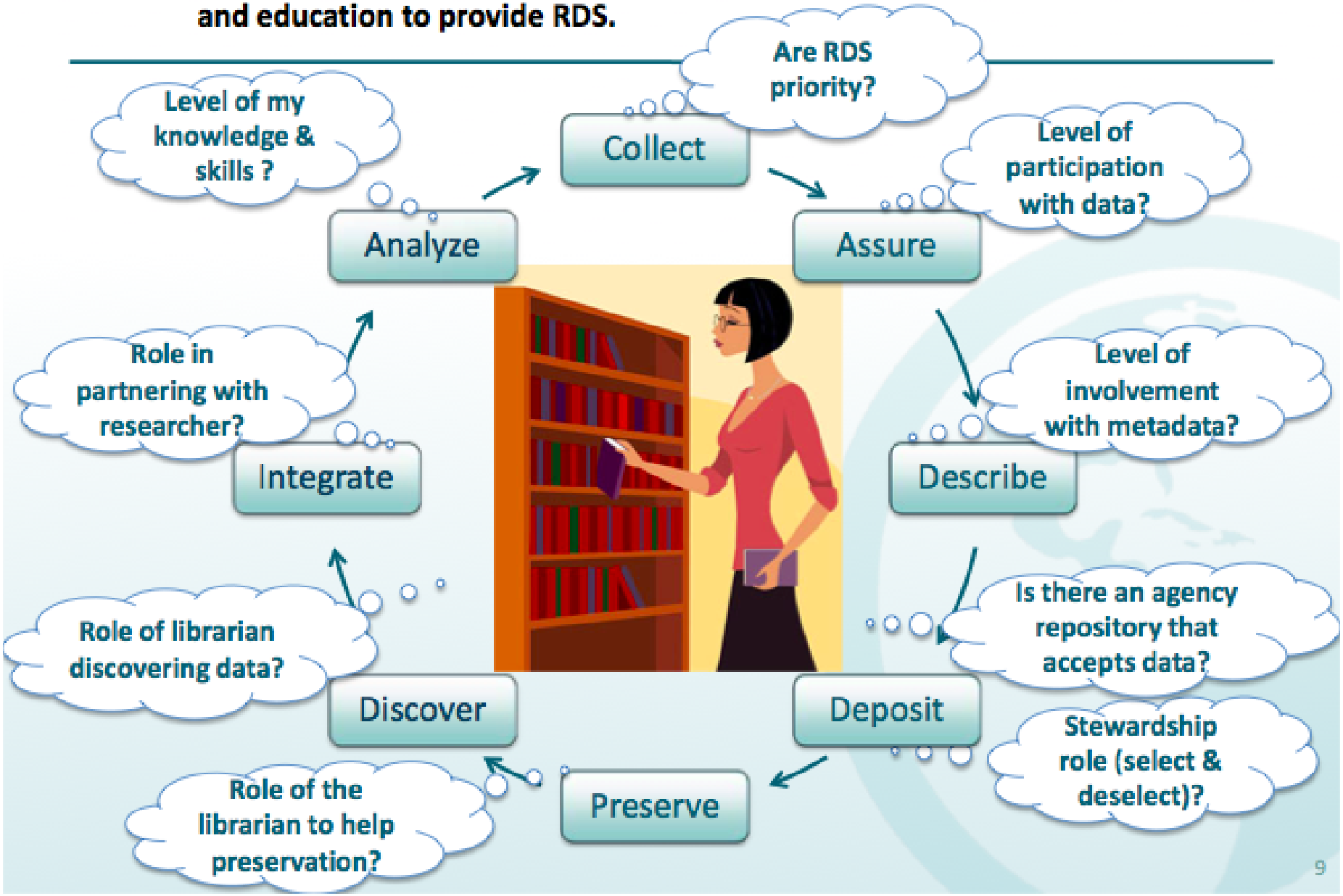
- RCUK Research Councils UKのオープンアクセス要求
- NSF（全米科学財団）における研究資金申請時のデータマネージメントプラン要求（2011.1）
- 米国大統領府科学技術政策局（OSTP）による，公的資金を受けた研究の成果物およびデータへのパブリックアクセスの計画案策定指示（2013.2）
- 欧米の図書館における研究データサービス（RDS）
  - ビッグデータおよびスモールデータ
  - 自然科学，社会科学，人文学（Digital Humanities）
- 日本「知的財産推進計画2013」

データのライフサイクルと データキュレーション



http://www.dataone.org/best-practices

Figure 2. The librarian ponders whether she has the background, skills, and education to provide RDS.



Carol Tenopir et al. *Academic Libraries and Research Data Services: Current Practices and Plans for the Future – An ACRL White Paper.* ACRL, 2012.6, p.12.

これまで，サブジェクトライブラリアンは，情報の発見，コレクション構築，情報管理の諸要素を中心に展開するどちらかと言えば伝統的なサービスによって，研究者のニーズを支援してきた。参加図書館のサーベイおよび文献からは，支援とサービスの性質の変化についての証拠がもたらされている。この変化はさらに多様かつ広範囲に及ぶようになりつつあり，単純な情報関連活動を越えて特に研究データ管理を大きく強調する方向に向かっている。

Mary Auckland, *Re-skilling for Research.* London: RLUK Research Libraries UK, 2012.1
http://www.rluk.ac.uk/files/RLUK%20Re-skilling.pdf

ギャップが確認された九つの領域

1. **研究成果の保存**に関し助言を行うための能力（49% - 2～5年の間に重要になる：10% - 現在保持）
2. 受入，発見，アクセス，頒布，保存，移植性を含む，**データ管理とデータキュレーション**に関し助言を行うための知識（48% - 2～5年の間に重要になる：16% - 現在保持）
3. **オープンアクセスの要求を含む資金提供者のさまざまな命令の遵守**に関し研究者を支援するための知識（40% - 2～5年の間に重要になる：16% - 現在保持）
4. 分野／主題で用いられる可能性がある**データ操作ツール**に関して助言を行なうための知識（34% - 2～5年の間に重要になる：7% - 現在保持）

5. **データマイニング**に関してアドバイスを行なうための知識（33% - 2～5年の間に重要になる：3% - 現在保持）

6. **メタデータ**の活用に関する提唱と助言を行なうための知識（29% - 2～5年の間に重要になる：10% - 現在保持）

7. **プロジェクト記録**（例．文書）**の保存**に関して助言を行なうための知識（24% - 2～5年の間に重要になる：3% - 現在保持）

8. 研究者が可能性のある資金提供元を見つけるのを支援するための**研究資金源**に関する知識（21% - 2～5年の間に重要になる：8% - 現在保持）

9. 個々の研究プロジェクトに対し，**メタデータスキーマ**の整備と**分野／主題の標準と実務**に関する助言を行なうためのスキル（16% - 2～5年の間に重要になる：2% - 現在保持）

「…研究者は、ますますこれらの活動をオンライン上で行うのであるから、当然、研究図書館のサービスはデジタルの研究環境に不可欠な部分となる必要がある。実際に、大学図書館、研究図書館はもうすぐそうなることを見込むべきである。」

「良いサービスとは、研究者が研究のすべての段階で必要とするデジタル情報を発見し、利用する能力によって規定されるだろう。」

Palmer, Carole; Teffeau, Lauren C., & Pirmannet Carrie M. *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development.* OCLC Research, 2009.1, p. 34

2. 新たな発見環境と書誌フレームワーク

- ウェブスケールの発見環境
  - Summon, EBSCO Discovery Service, WorldCat Local,,,
  - 参考： 図書館サービスプラットフォームへの移行 - WorldShare Management Service (OCLC), Alma (Ex Libris), OLE: Open Library Environment (Kuali), Open Skies (VTLS), … ＊マルチテナント（SaaS, クラウドコンピューティング）, セキュリティ保証
- 図書館界を超えたメタデータ利用
  - 発見のための多様なプラットフォーム（Google, Amazon, Open Library,,,）
  - 流通過程におけるメタデータ利用（ONIX, Amazon）

「情報カオス：　四つの要因」

1. 組織毎の、数多くの情報サイロ：　それぞれ別個の、検索／発見の仕組み、メタデータの「システム」、そして不完全なコンテンツ記述
2. 図書館／電子図書館の発見の仕組み：　低い精度、不適切な再現率
3. 図書館が作成／収集するメタデータ：　ウェブの世界と遠くかけ離れたメタデータ
4. Google等の検索エンジン：　使いやすい検索や発見の方法を提示。しかし、多くの情報資源がインデックス化の対象外であることは非周知

Keller, Michael A.　“Linked data: A way out of the information chaos and toward the Sematic Web,” EDUCAUSE Review, 2011.7.http://www.educause.edu/ero/article/linked-data-way-out-information-chaos-and-toward-semantic-web

Linked Open Data (LOD)の展開

- W3C（World Wide Web Consortium）Library Linked Data Incubator Group 最終報告（2011.10）
- LC*,BL*, Europeana*, Cambridge U.*, Harvard U.*, OCLC*,,,
- New York Times*, BBC*, Nature*,,,
- DBpedia*, GeoNames*,,,
- NDL Web Authorities*
- CiNii*
- OCLC – WorldCatのデータをSchema.org形式で公開（加えて、Dewey, VIAF and FAST headingsへのリンク） ODC-BY
  - ダウンロード用ファイル（250以上の所蔵館を持つ書誌データ、120万件=8,000万トリプル）

Linked Dataの4原則（Tim Berners-Lee）

① ものごとの名前としてURIを使うこと

② 人間やコンピュータ・ソフトウェアがものごとの名前を参照したり、調べたりできるようにHTTP URIを使うこと

③ URIを見に行ったとき、RDFやSPARQL（XML）のような標準方式によってそれに対する有用な情報を提供できるようにすること

④ より多くのものごとを発見できるように、データの中に他のURIへのリンクをいれること

RDF (Resource Description Framework)

RDF（Resource Description Framework） [p.5,6,12]

W3C（World Wide Web Consortium）が標準化しているメタデータの規格。コンピュータが理解可能な意味表現のための形式で，主語，述語，目的語の 3 つ組（トリプル）で情報の関係を表現する。主語はURIか空白，述語はURI，目的語はURI，文字列（リテラル），空白のいずれかで表現する。

たとえば，「シロクマ」というタイトルは以下のように表せる。

「電子的学術情報資源を中心とする新たな基盤構築に向けた構想」
http://www.nii.ac.jp/content/archive/pdf/content_report_h23_with_glossary.pdf

LOD化のメリット

- **データ品質の向上と作業の効率化**： 著作、場所、人物、出来事、主題、その他に対する識別子の利用による、信頼できる情報源からの補足データとのリンク形成によって、あるいは図書館ではこれまで作成できなかった粒度の外部データとのリンク形成によって
- **データの発見と利用可能性の向上**： 図書館の目録データとDBpedia、GeoNames、BBC、New York Timesといった他の領域のサービスとのリンク形成、あるいは実験のためのデータセット、データ処理に使用されたモデルとのリンク形成によって
- **図書館のウェブ上での存在の強化**： データ利用、再利用からもたらされる機関の可視性の向上によって
- **専用ソフトウェアからの解放**： RDFやHTTPの活用により、より一般的なツールの利用の道が開かれることによって

（引用元：「W3C Library Linked Data Incubator Group最終報告書」）

LOD対応の期待される効果

1. 発見可能性の向上
   - Schema.org
   - Google 等の検索エンジンの品質向上
     - ドキュメント間の関係に着目した検索システムの提供
2. 図書館システムの変革
   - MARCからの移行
   - FRBR -> RDA のプラットフォームとしての活用
   - 新たな連携可能性
     - 例: データ更新の際のアラーティング、双方向のデータ更新

新たな発見システム

- Open Library
- The Data Hub
- Freebase
- Archives Hub
- Google Knowledge Graph

目録システムの変革の動き

- 1995年　Dublin Core 開始
- 1998年　IFLA - FRBR (Functional Requirements for Bibliographic Records)
- 2005年　JSC - AACR2からRDA（Resource Description and Access）へ
- 2009年　IFLA - Functional Requirements for Authority Data

LCの新たな書誌フレームワーク

- 2008.1　書誌コントロールの将来に関する米国議会図書館ワーキンググループ最終報告書 “On the Record”
- 2010.7 – 2011.3　RDAの検証
  - 2011.6　検証結果と勧告 -> 2013.1以降の採用
- 2011.10　*A Bibliographic Framework for the Digital Age* --- MARC21からLODへの移行の宣言
- 2012.1　ALA Midwinter Meeting で移行計画の概要説明
- 2012.5　LODのためのモデリングおよび移行計画をZepheira社に委託
- 2012.11　“**BIBFRAME**” (*Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services*)

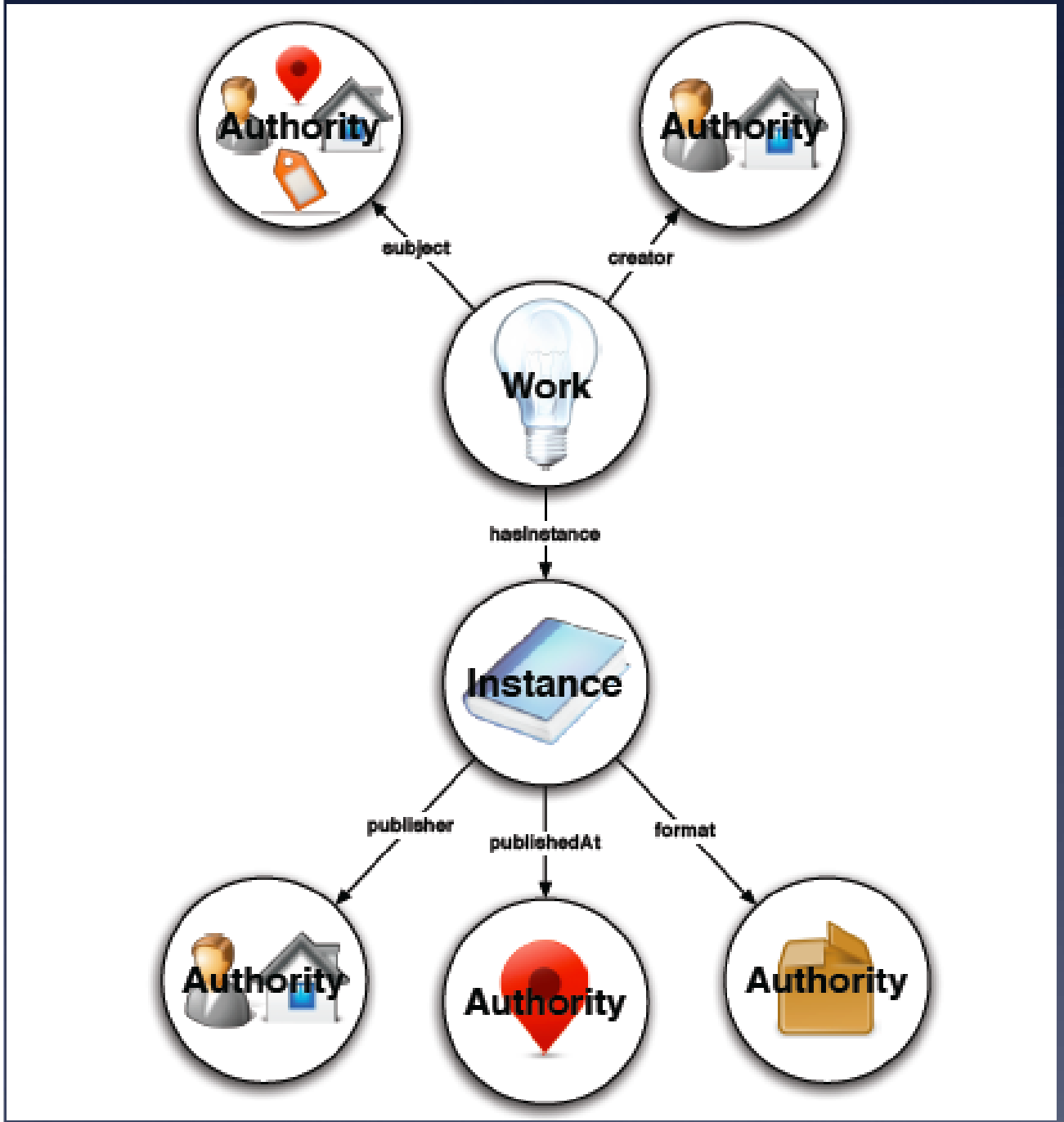
LCのLOD採用の影響

- 各国国立図書館への波及
  - 英国、カナダ、ドイツ、オーストラリア、その他
  - NDL – 平成24年度書誌調整連絡会議（10/12開催）において方針表明（RDAと新書誌フレームワーク）
- ただし、たんなる交換用フォーマットの変換から、書誌データ作成方式およびシステムの全面的再構築までの、さまざまな可能性が考えられる

LC-BIBFRAMEの高次モデル（主クラス）

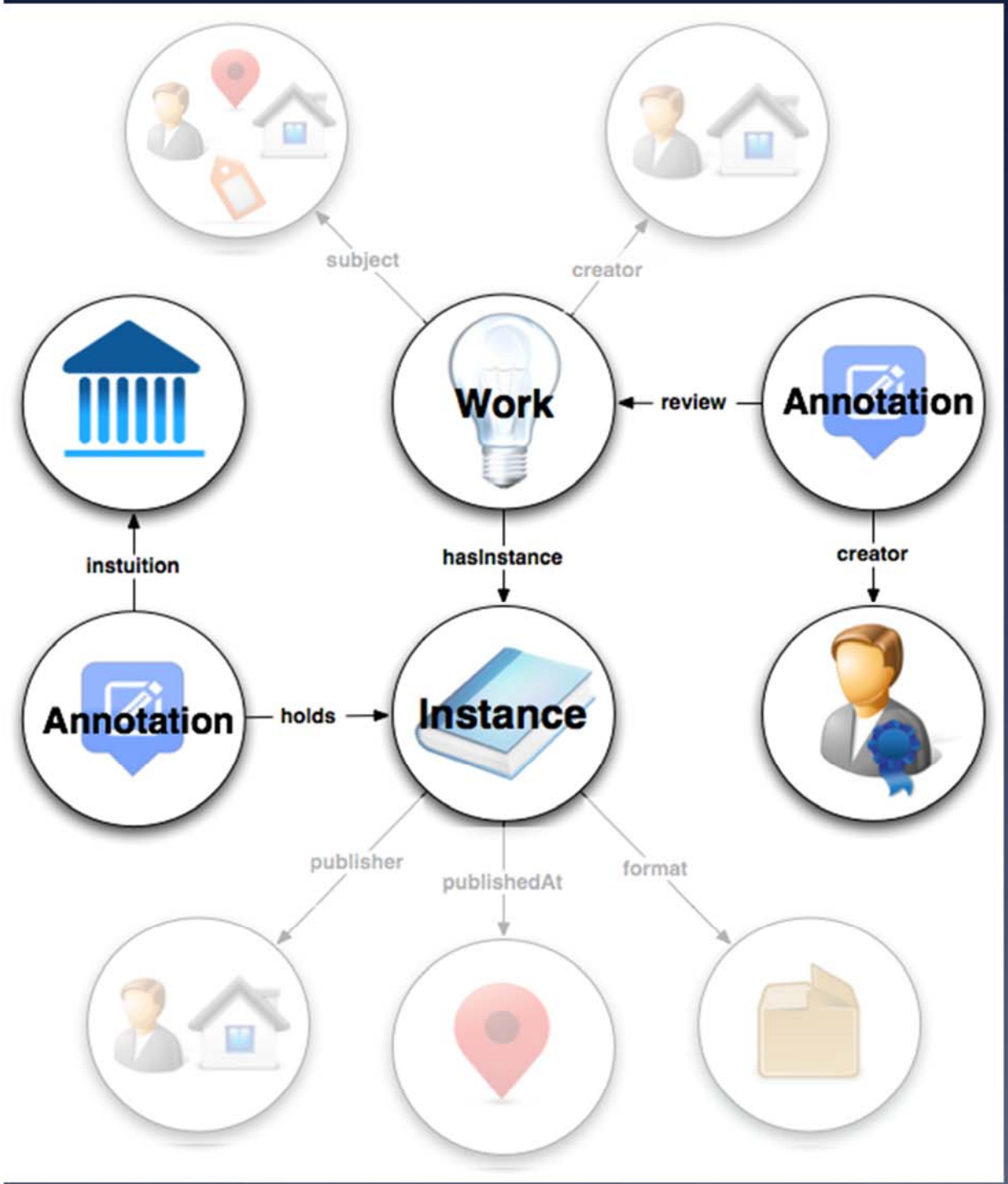
- 著作 (Work)
  - 目録対象資料の概念的本質を反映した資源
- インスタンス
  - その著作の、個別のものとしての具象化を反映する資源
- 典拠
  - 著作とインスタンスに現れた関係性を定義した主要な典拠概念を反映する資源。人名、地名、件名、組織名等
- 注釈
  - 追加の情報によって他のBIBFRAME資源を修飾する資源。図書館の所蔵情報、表紙の絵、レビュー等

Library of Congress. *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Service.* 2012.11, 42 p.
http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf



「著作資源とインスタンス資源間の関係性を定義するBIBFRAMEリンクトデータ・モデルとそのウェブによるアドレス指定が可能な典拠情報源への文脈化の図式表現」

Library of Congress. *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Service.* 2012.11, 42 p. http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf



「柔軟な注釈フレームワークの文脈でのBIBFRAMEリンクトデータ・モデルの図式表現」

Library of Congress. *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Service.* 2012.11, 42 p. http://www.loc.gov/marc/transition/pdf/marcld-report-11-21-2012.pdf

LOD対応の課題

- LODのメリットを享受するためには、対応するデータセットの公開が前提となるのは当然

1. データ形式の策定とこれに対応するシステム整備
2. データの同定識別のための識別子の設定および管理
3. 重複データ排除のための（海外を含む他機関との）連携
4. データの権利関係の定義（CC0 or ODC-BY？、参加機関との調整）
5. NIIと大学図書館の双方における、目録作成／検索機能提供の役割の再定義