資料と公共性 ： 2018年度研究成果年次報告書

岡崎, 敦
九州大学大学院人文科学研究院 | 九州大学大学院統合新領域学府：教授

市澤, 哲
神戸大学大学院人文科学研究科：教授

石田, 栄美
九州大学附属図書館 | 九州大学大学院統合新領域学府：准教授

後小路, 雅弘
九州大学大学院人文科学研究院 ： 教授

他

https://doi.org/10.15017/2230688

出版情報：2019-03-14. 九州大学大学院人文科学研究院
バージョン：
権利関係：

東京大学デジタルアーカイブズ構築事業におけるオープンデータに関する取り組み

中村　覚

1.　はじめに

本報告では、人文学分野におけるオープンデータの動向について概観した後、東京大学が実施する学術資産のデジタル化事業「東京大学デジタルアーカイブズ構築事業」におけるオープンデータの利活用に関する取り組みについて述べる。本報告に関する図表等については、発表スライド[1]を参照されたい。

2.　人文学分野におけるオープンデータの動向

2.1 オープンデータ

近年、様々な分野において、オープンサイエンスの導入や実践が進められている。オープンサイエンスに関わる要素は多岐に渡るが、情報の共有に関する「オープンアクセス」「オープンデータ」「研究データ／データ出版／データリポジトリ」、研究の協働に関する「学際研究」「市民科学／クラウドファンディング」、および研究の透明性に関する「研究の再現性／透明性／研究データ保存」などがキーワードとして挙げられる[2]。

本報告では、本シンポジウムの名称にも取り上げられている「オープンデータ」を中心に述べる。本発表では、以下に示す Open Knowledge International による定義[3]を参考とする。

> オープンデータとは、自由に使えて再利用もでき、かつ誰でも再配布できるようなデータのことだ。従うべき決まりは、せいぜい「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。

また、オープンデータは公開のレベルによって 5 段階に分類することができる[4]。オープンライセンスでデータを公開する第 1 段階から、オープンに利用できるフォーマットでデータを公開する第 3 段階、Linked Open Data（LOD）を公開する第 5 段階までのレベルがある。LOD とは、オープンライセンスで公開された Linked Data を指す。Linked Data とは、インターネット上で機械可読な構造化データを公開する技術の総称であり、Linked Data の 4 原則に従ったデータを指す[5]。LOD の公開により、第三者／計算機によるデータの利活用が容易となる。

2.2 人文学分野におけるオープンデータの活用例

人文学分野においても、デジタルヒューマニティーズに代表されるように、研究のデジタル化とオープン化が進んでいる。特に、画像などの Web コンテンツを共有するための

国際的な枠組みである IIIF（International Image Interoperability Framework）が、日本国内においても広まりつつあり、IIIF の特徴を生かした活用事例も数多く報告されている。IIIF も Linked Data のひとつであり、その利用の容易さなどから、特に Linked Open Usable Data などとも呼ばれる[6]。

国内における IIIF とオープンデータの活用事例として、例えば「富士川文庫デジタル連携プロジェクト試行版」が挙げられる[7]。IIIF の利点を活用した分散コレクション仮想統合の一例として、京都大学と慶應義塾大学が分散して所蔵する「富士川文庫」を横断して検索することが可能となっている。また、永崎らは仏教研究等の特定の分野に特化した IIIF 準拠の資料を共有するためのサイト「IIIF Manifests for Buddhist Studies」を提供している[8]。さらに、人文学オープンデータ共同利用センターでは、IIIF 準拠の画像の一部切り出しや、収集／並び替え／保存といった「キュレーション」作業を支援するプラットフォームとして、「IIIF Curation Platform」を開発している[9]。本プラットフォームを活用し、例えば、国文学研究資料館、慶應義塾大学、京都大学附属図書館が公開する IIIF 画像から、顔貌だけを横断的に収集し、検索可能とするサイト「顔貌コレクション」などが構築されている[10]。

その他、人文学分野におけるオープンデータの活用例として、例えば、国文学研究資料館では、人文学オープンデータ共同利用センターと連携し、研究者のための「日本古典籍データセット」、機械のための「日本古典籍くずし字データセット」、市民のための「江戸料理レシピデータセット」を提供している[11]。また、オープンサイエンスに関わる代表例の一つとして、市民参加によるオンライン翻刻プロジェクト「みんなで翻刻」が大きな成功を納めている。「みんなで翻刻」は、地震に関する史料を多数の参加者が Web 上で翻刻するプロジェクトである[12]。2017 年 1 月に公開されて以来、2018 年 11 月時点において、469 点の翻刻が完了し、入力文字数の合計は 500 万文字を超えている。さらに、オープンアクセスに関する代表的な事例として、京都大学学術情報リポジトリ「KURENAI」が、スペイン高等科学研究院（CSIC）による世界リポジトリランキング（2018 年 11 月版）の機関リポジトリ部門において、世界第 5 位を獲得している[13]。

3. 東京大学デジタルアーカイブズ構築事業におけるオープンデータに関する取り組み

3.1 事業の概要

2015 年に公表された「東京大学ビジョン 2020」の一つに「学術の多様性を支える基盤の強化」が掲げられ、「東京大学が保持する学術資産のアーカイブを構築し、その公開と活用を促進することで、学術の多様性を支える基盤を強化する」という方針が示された。これを受け、2016 年 9 月に「東京大学学術資産等アーカイブズ委員会」が東京大学内に設置された。本委員会は東京大学附属図書館長、東京大学総合研究博物館長、東京大学文書館長、東京大学情報基盤センター長を中心として構成され、東京大学内の MLA（博物館 Museum／図書館 Library／文書館 Archives）機関が参画している点に特徴がある。委員会では、東京大学の保持する多様な学術資産等のデジタルアーカイブ化を行い、かつ国内外に向けて

広く公開し、その活用を促進するための「東京大学デジタルアーカイブズ構築事業」を 2017 年度から実施している。

具体的には、学内公募に基づく予算配分による学術資産のデジタル化の促進に加え、学内の各部局が公開するデジタルアーカイブの情報を集約し、学内に分散した学術資産を横断して検索可能なシステム「東京大学学術資産等アーカイブズポータル」を開発している。本システムは 2019 年度に一般公開することを予定している。また、デジタルアーカイブシステムの構築や運用が困難な部局に対しては、学術資産の公開を支援するホスティングサービス「東京大学学術資産等アーカイブズ共用サーバ」（以下、共用サーバ）の構築／運用を行っている。本サービスは、学術資産の公開支援だけでなく、IIIF に準拠した画像公開を行うなど、資料の利活用支援を目的としている。さらに、利活用支援の一環として、学術資産のオープンデータ化を推進している。例えば、東京大学総合図書館では、2018 年 6 月に利用規約を改定し、著作権の保護対象ではない公開画像を CC BY 相当の条件で利用可能としている。加えて、これらのオープンデータを GitHub 上で公開している[14]。この GitHub を用いたデータセットの公開について、データの利用者に対する利点として、データの一括取得が容易となる点、公開システムのユーザインタフェースの制約等に依存しないデータ利用が可能になる点、などが挙げられる。また、データの提供者（ここでは、東京大学学術資産等アーカイブズ委員会）に対する利点としては、データと公開システムを分離することによるデータの長期的な公開、データの変更履歴の管理・バックアップ等を支援することができる点が挙げられる。

3.2 学術資産の活用例

ここでは、本事業でデジタル化を実施した、または本事業が公開支援を行う学術資産の活用例について述べる。

例えば、U-PARL（東京大学附属図書館アジア研究図書館上廣倫理財団寄付研究部門）が共用サーバ上で公開する「漢籍・碑帖拓本資料」を主な対象資料として、IIIF を用いた法帖比較支援システムを開発している[15]。本システムは典拠データの Linked Data 化と，複数機関が提供する IIIF 準拠画像に対するアノテーション付与による個別作品の識別により，異版関係にある個別作品を検出可能なシステムを構築し、国立国会図書館，国文学研究資料館，U-PARL が提供する法帖画像を対象としたケーススタディを通じ，異版作品の異なる個別資料における再現例を検出できることを確認した。複数の機関が公開する IIIF 準拠の法帖画像の検索においては、中村らが公開している日本国内の IIIF 準拠画像に対する横断検索システム「IIIF Discovery in Japan」を利用している[16]。この横断検索システムでは、「漢籍・碑帖拓本資料」に加え、共用サーバ上で公開している他の学術資産、および本事業でデジタル化を実施した東京大学文書館、東京大学大学院情報学環附属社会情報研究資料センターが公開する IIIF 画像も取り込まれており、日本国内の他の機関が公開する IIIF 画像と合わせて検索することが可能となっている。

他の活用事例としては、『捃拾帖』の内容検索を可能とするシステム「電子展示『捃拾帖』」の開発が挙げられる[17]。『捃拾帖』とは、明治時代の博物学者である田中芳男が収集した、幕末から大正時代にかけてのパンフレットや商品ラベルなどを貼り込んだ膨大なスクラップブックである。東京大学総合図書館はこれらの画像を冊単位で公開しているが、貼り込まれた資料単位での検索が望まれていた。この課題に対して、本研究では IIIF のアノテーション機能を利用し、各頁の貼り込み資料単位で画像を切り出し、検索可能なシステムを開発した。また、東京大学史料編纂所の「摺物データベース」が提供する、貼り込み資料単位のメタデータと組み合わせることで、内容情報に基づく検索を可能としている。本研究はその他、複数の機関が提供する各種リソース（IIIF・オープンデータ）を組み合わせて利用している点に特徴があり、デジタルアーカイブの利活用を検討する上での一事例を示すことを目的としている。さらに、本システム開発において人手で抽出した貼り込み資料に関するデータを、機械学習における教習データとして利用し、深層学習（ディープラーニング）を用いた貼り込み資料の自動検出、および類似画像検索システムの開発などにも取り組んでいる[18]。

4. まとめ

人文学分野における研究のデジタル化とオープン化の進展に合わせて、東京大学デジタルアーカイブズ構築事業では、本報告で述べた通り、デジタル化とオープンデータ化による学術資産の利活用を支援する基盤整備を進めている。また、2018 年 11 月に開催した第 2 回東京大学学術資産アーカイブ化推進室主催セミナー「かわいい子には旅をさせよ　―デジタルアーカイブとオープンデーター」など、オープンデータの推進に関する各種セミナーの実施も行なっている[19]。これらの取り組みが学内外におけるデジタルアーカイブの構築と活用、およびオープンデータ化の推進に寄与することができれば幸いである。

参考文献

1. 中村覚. 東京大学デジタルアーカイブズ構築事業におけるオープンデータに関する取り組み, シンポジウム 「オープンデータと大学」| 九州大学附属図書館, 2019.1, http://hdl.handle.net/2324/2197530, (参照 2019-2-13).
2. 北本朝展. 人文学研究のデジタル化とオープン化, Japan Open Science Summit 2018, http://doi.org/10.20676/00000336, (参照 2019-2-13).
3. Open Knowledge International, Open Data Handbook, http://opendatahandbook.org/guide/ja/what-is-open-data/, (参照 2019-2-13).
4. 5-star Open Data, http://5stardata.info/, (参照 2019-2-13).
5. Tim Berners-Lee. Linked Data, Design Issues, 2009.
6. LOUD: Linked Open Usable Data, https://linked.art/loud/ , (参照 2019-2-13).

7. 西岡千文. IIIF を利用した富士川文庫資料の再統合の試み, じんもんこん 2018 論文集, Vol. 2018, pp. 291-296, 2018.
8. 永崎研宣／下田正弘. オープン化が拓くデジタルアーカイブの高度利活用：IIIF Manifests for Buddhist Studies の運用を通じて, じんもんこん 2018 論文集, Vol. 2018, pp. 389-394, 2018.
9. 北本朝展／本間淳, Tarek Saier. IIIF Curation Platform：利用者主導の画像共有を支援するオープンな次世代 IIIF 基盤, じんもんこん 2018 論文集, Vol. 2018, pp. 327-334, 2018.
10. 鈴木親彦／髙岸輝／北本朝展. 顔貌コレクション（顔コレ）：精読と遠読を併用した美術史の様式研究に向けて, じんもんこん 2018 論文集, Vol. 2018, pp. 249-256, 2018.
11. 山本和明. 人文学の検証可能性とオープンデータ, Japan Open Science Summit 2018, http://doi.org/10.20676/00000335, (参照 2019-2-13).
12. 橋本雄太. 歴史地震研究における異分野連携とシチズンサイエンス, Japan Open Science Summit 2018, http://doi.org/10.20676/00000337, (参照 2019-2-13).
13. 京都大学図書館機構. 【図書館機構】学術情報リポジトリ「KURENAI」が世界ランキングで第 5 位になりました, https://www.kulib.kyoto-u.ac.jp/bulletin/1380077, (参照 2019-2-13).
14. 東京大学学術資産等アーカイブズ共用サーバ-データセット, https://archdataset.dl.itc.u-tokyo.ac.jp/collections/, (参照 2019-2-13).
15. 中村覚, 成田健太郎, 永井正勝. Linked Data 化した典拠データと IIIF を用いた法帖の異版比較支援システムの開発, じんもんこん 2018 論文集, Vol. 2018, pp. 297-302, 2018.
16. 中村覚, 永崎研宣. 日本国内の IIIF 準拠画像に対する横断検索システムの構築, 研究報告人文科学とコンピュータ（CH）, Vol. 2018-CH-118, No. 8, pp. 1-6, 2018.8.
17. 電子展示『捃拾帖』, https://kunshujo.dl.itc.u-tokyo.ac.jp/, (参照 2019-2-13).
18. 貼り込み資料画像検索プロトタイプ, http://kunshujo-i.dl.itc.u-tokyo.ac.jp, (参照 2019-2-13).
19. 東京大学附属図書館. 【11/22 開催】学術資産アーカイブ化推進室セミナー「かわいい子には旅をさせよ」, https://www.lib.u-tokyo.ac.jp/ja/library/contents/event/20181012, (参照 2019-2-13).