『Digital Cultural Heritage』リニューアル公開について

高嶋朋子　中村覚

1. 緒言

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では、2018年11月1日より、デジタルアーカイブ・システム、Digital Cultural Heritage（以下、DCH）のリニューアル公開を開始した。

旧DCHは、社会情報研究資料センターの高度アーカイブ化事業（2007-2011年）で構築されたアーカイブシステムで、地図年表検索、トピックマップ検索などの多様な検索機能を装備した先駆的なものであった。しかし、当該事業終了後、充分な予算と人的資源の確保が難しくなったセンター体制では、サーバの維持やシステムの継続的開発およびその維持が困難となったため、2016年より一時休止を余儀なくされた[1]。この一時休止と同時に、小規模組織が運用するに最適なデジタルアーカイブ・システムのあり方を模索しつつ、再公開のための計画を立ち上げた。

本稿では、リニューアルしたDCHについて、その収録コンテンツとシステム詳細を解説する。

2. コンテンツ

2.1 旧Digital Cultural Heritageに収録されていたコンテンツ

旧DCHでは、2012年から2016年まで「小野秀雄関係資料」「外務省関係資料」「森恭三コレクション」「坪井家関係資料」4資料群の目録データと一部の画像データを公開してきた。いずれもさまざまなメディアや研究者らによって利用されてきた貴重資料であり、今回リニューアルしたDCHでも、旧DCHで公開されてきたデータを継承している。

各資料群の詳細は以下の通りである。

小野秀雄関係資料

東京大学新聞研究所初代所長で日本の新聞研究を牽引した小野秀雄が作成・収集した研究資料群である。小野の研究活動の足跡をたどることができる「小野秀雄研究資料」、かわら版や錦絵などを中心とした「小野秀雄コレクション」あわせて約4,700件の目録データに加え、一部画像データを公開している。

外務省関係資料

外務省情報部が収集したと考えられる資料群で、当センターの前身である新聞研究所の何初彦教授によって収集・受け入れられた。内容はポスターや伝単などの戦時資料であり、DCHでは「第1次世界大戦期プロパガンダポスターコレクション」約660件について、目録データと画像データを併せて公開している。

森恭三コレクション

当センター前身の社会情報研究所に寄贈された元朝日新聞論説主幹森恭三著作物を収録した『森恭三著作集CD-ROM版』の目録データ、約3,500件を公開している。

坪井家関係資料

坪井家より情報学環に移管された資料群である。このうち、明治期の人類学者である坪井正五郎が収集及び作成した資料群である「坪井正五郎関係資料」と、正五郎の祖父と父で蘭方医だった坪井信道・信良が収集及び作成した資料群「坪井信道・信良関係資料」、あわせて約7500件について、目録データと一部画像データを公開している。

2019年3月現在、これらの資料群のデジタルデータは、個人が調査・研究・教育・学習を目的とする非営利の場合に限り、申請不要で印刷やダウンロードを許可するが、二次利用については、規定の申請を義務付けている。また、データ改変は禁じている。

2.2 社会情報研究資料センター所蔵 新聞原紙資料のデジタルコンテンツ

DCHリニューアルに着手した当初は、2.1で説明した旧DCHに収録されていたコンテンツのみを公開対象として、できるだけ早期に再公開できるよう計画をすすめてきた。しかし、社会情報研究資料センターは、2017年度より東

京大学デジタルアーカイブズ構築事業 [2] の支援を受け、新たに貴重新聞原紙のデジタル化事業に着手したため、この成果も DCH の新規コンテンツとして扱うこととなった。

新聞のデジタルアーカイブ化は、世界的に見れば、発行元による商用ベースのアーカイブ整備以外にも公的機関によって推進されるケースが散見され、多くの人が過去の新聞に容易にアクセスできる環境が整えられてきた [3][4]。しかし、日本の現状ではまだ新聞のデジタル化は各発行元が主体という域を出ていない。そして、近代に発行されていた業界紙などの場合は特に、1930 年代末からの新聞統制を経て、既に発行元団体が解散や廃業していることが多い。また、存続していたとしても活発な活動を行っておらず、デジタル化に着手するための資金や人材を調達することが難しいケースもある。

社会情報研究資料センターではこうした現状を踏まえ、他館所蔵がなく、且つ発行母体に直接の後継組織や団体等の存在が確認できず、今後のデジタル化が期待しにくい 1900 年代前半に発行された業界紙を中心として事業を進めてきた。

現在 DCH で公開しているのは、2017 年度の事業成果で『新聞之日本』536 件、『日本毛織物新報』234 件、『日本糸物新報』79 件である。各資料とも画像データと併せて、Dublin Core などの語彙を用いた、タイトル、資料 ID、出版社、出版年月日、号数、サイズ、面数といったメタデータを提示している。各資料の概要については、「社会情報研究資料センター所蔵新聞原紙資料のデジタル化事業 −対象資料の紹介を中心に−」（『社会情報研究資料センターニュース』28 号、2018 年）に詳細をまとめたので参照されたい。なお、2018 年度の事業成果としては『日刊新聞興信所報』『新聞興信所報』のデジタル化を進めており、2019 年 4 月には公開する予定である。

これら資料は、印刷、ダウンロード及び二次利用において申請不要ではあるが、二次利用では所蔵の明示、データ改変についてもその旨の明示を依頼するため、「CC BY 相当」という認識となる。

3. システム

緒言で述べた通り、システムのリニューアルにあたっては、比較的小規模な組織でも持続的に運用可能なシステムを構築することを目的とした。この目的に対するシステムの設計については、姉妹プロジェクトである「東京大学文書館デジタル・アーカイブ」に関する報告 [5] などを参照されたい。本稿では、特にシステムの技術的な側面や工夫点について述べる。

3.1 Omeka S

DCH は、「Omeka S」[6] というソフトウェアを用いて開発している。Omeka とは、デジタルコレクションを構築するためのオープンソースソフトウェア（Open Source Software: OSS）のコンテンツ管理システム [7] であり、ジョージ・メイソン大学の Roy Rosenzweig Center for History and New Media によって開発されている。「テーマ」を用いて UI（ユーザインタフェース）を変更し、「プラグイン」を用いて機能を拡張することができる。2008 年に初版となるパブリック・ベータ版がリリースされ、その後コミュニティによる継続的な開発・改良が進められ、今年 2 月に 10 周年を迎えた。また、2017 年の 11 月には、Omeka S という新しいソフトウェアが正式に公開され、これにより、従来の Omeka の名称が Omeka Classic に変更となった。

Omeka S が提供する機能群のうち、特に特徴的な点は、Linked Data への標準対応である。Linked Data とは、Web 上のデータをつなぐことで、新しい価値を生み出そうとする取り組みであり、データを共有（公開）し、相互につなぐ仕組みを提供する。例えば、Omeka S に登録されたアイテム（資料を管理する単位）やメディア（画像や動画など）には URI（Uniform Resource Identifier）が自動的に与えられ、その URI にアクセスすることにより、当該資源に関する情報（メタデータ）が JSON-LD 形式で出力される。なお、JSON-LD とは JSON（JavaScript Object Notation）という軽量なデータ交換フォーマットを利用して、Linked Data を表現するためのフォーマットである。このデータを介して、アイテム同士の関係やアイテムとメディアとの関係、インターネット上の他のリソースとのつながりを記述することができる。

3.2 DCH の画面例

ここでは、DCH の画面を示しながら、主要な機能群について説明する。詳しい利用方法については、DCH のヘルプページを参照されたい。

検索機能としては、階層検索画面（図 1 左）とキーワード検索画面（図 1 右）を提供する。階層検索では、資

料同士の関係性に基づき、資料を検索することができる。上述したように、Omeka S は Linked Data に標準対応しているため、アイテム間の上下関係を RDF（Resource Description Framework）を用いて記述している。

キーワード検索では、全文検索機能のほか、タイトルや作成者などの特定の項目を指定した検索機能を提供する。検索結果は図 1 右に示すように、サムネイル画像とともに表示される。検索機能については、プラグインの追加により、旧字・異体字の違いを吸収した検索機能を追加している。

図 1　階層検索画面（左）とキーワード検索画面（右）

図 2 に示す閲覧画面では、資料に関する情報（メタデータ）の表示と、画像を持つものについては画像閲覧用のビューアが表示される。

メタデータについては、図 2 左に示すように画面上で表示されるほか、プラグインの追加により、図 2 右下に示すように、JSON-LD や RDF/XML などの形式でのダウンロードを可能としている。メタデータを RDF で記述するにあたっては、語彙として Dublin Core に加え、EAD 標準を Linked Data で記述するための語彙「The LOCAH RDF Vocabulary」[8] 等を利用している。例えば記述レベル（フォンドやシリーズなど）を記述する際には、プロパティとして「http://data.archiveshub.ac.uk/def/level」を使用している。

また、画像については、画像共有のための国際規格である IIIF（International Image Interoperability Framework）に準拠した画像公開を行っている。図 2 右は、IIIF 対応の画像ビューアである Universal Viewer を用いて画像を表示している例を示す。加えて、プラグインを追加することにより、画像の一括ダウンロード機能を追加提供している。

図 2　資料の詳細画面（左）と画像の閲覧画面（右）

3.3 その他の機能：API・データのダウンロード

ここでは、3.2 で説明した画面とは別に、API 等を用いた公開データへのアクセスを可能とする機能群について述べる。

まず、Omeka S が標準機能として提供する REST API を用いることで、DCH で公開されているデータを機械的に取得することができる。個々の資料のメタデータは上述したプラグインを用いてダウンロードすることができるが、各種検索条件に応じた資料の一覧を一括で取得する際などには、API によるアクセスが有効である。API のエンドポイントは「http://dch.iii.u-tokyo.ac.jp/api」であり、利用方法については公式ドキュメント [9] を参照されたい。本機能の想定する利用例の一つとして、他のシステムとの連携時におけるデータ作成が挙げられる。例えば、東京大学デジタルアーカイブズ構築事業では、東京大学内の学術資産を横断的に検索可能なポータルシステム「東京大学学術資産等アーカイブズポータル」（以下、ポータル）の公開を予定している。このポータルに提供するデータを作成する際、API を用いて DCH で公開中のデータを一括取得し、データ加工を行うといった利用を想定している。

API を用いることにより DCH で公開されている各種データをダウンロードすることが可能な一方、この方法では機械可読なデータを人間可読に変換するためのプログラムや処理が必要となる。そこで、Excel 形式などに事前に変換済みのデータをダウンロード可能なデータセットを合わせて提供している。Excel 形式のデータは DCH の「公開コンテンツ概要」のページからダウンロードできる。また、RDF や JSON-LD などの機械可読な形式のファイルを「GitHub（ファイルのバージョン管理サービス）」上で公開している [10]。このように、公開システムである DCH とは別にデータ単体を公開することにより、システムとデ

ータを分離し、持続的なデータ提供を可能とすることを目的としている。

さらに、この GitHub で公開するデータの一つに、DCH で公開している画像に関する情報（マニフェストファイル）の一覧を格納したファイル（IIIF コレクション）がある。本ファイルを用いることで、DCH 上で公開されている画像の一覧を容易に確認することができ、これらのデータを二次利用する第三者や計算機には有益な情報となる。実際、このファイルを活用することで、中村が別途開発・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」[11] に対して、DCH 上で公開されている画像を機械的に登録している。このように、公開システムである DCH とは異なるソフトウェアやアプリケーションにおいても、画像をシームレスに利用することができる点が、画像の相互運用性を高める IIIF の利点の一つである。

4. 結言（今後の展望）

今回リニューアル公開した DCH は、過去のシステムの公開休止の経験を踏まえ、比較的小規模な組織でも長期的に運用可能なシステムを設計・提示しつつ、単なる画像公開に留まらず、API やデータセットの提供といった第三者や計算機による二次利用を支援する機能提供を行っている点に特徴がある。

コンテンツ整備の今後の見通しとしては、貴重新聞原紙のデジタル化を、センターの基幹事業として継続していく。来年度についてはすでに旧外地で発行された日本語新聞のデジタル化が計画されている。他にも社会情報研究センターに寄贈された資料でリストが公開されていないものを優先的に、整理、デジタル化に着手する予定である。

技術的な今後の取り組みとしては、IIIF の機能を利用したコンテンツの利活用を検討している。具体的には、紙面の部分領域の切り出し（キュレーション）による新たな情報提示手法を開発したい。また、現在試験的な機能として提供している IIIF Search API と TEI（Text Encoding Initiative：人文科学のテキストの符号化・交換のための標準規格）を用いた画像内テキスト検索 [12] の実用化などを行う。

参考文献

[1] 宮本隆史 , 中村覚 . 社会情報研究資料センター『Digital Cultural Heritage』の公開休止に関する考察と再構築 . 東京大学大学院情報学環社会情報研究資料センターニュース , No.27, pp.274-279, 2017.3.

[2] 東京大学デジタルアーカイブズ構築事業 , https://www.lib.u-tokyo.ac.jp/ja/library/contents/archives-top,（参照：2019-2-15）

[3] 時実象一 , 欧州における新聞デジタル・アーカイブ Europeana Newspapers, 情報の科学と技術 67 巻 1 号 , pp.34-37, 2017.1

[4] 時実象一 , 欧米の新聞デジタル・アーカイブ　民間の新聞デジタル・アーカイブ , 情報の科学と技術 67 巻 7 号 , pp.383-388, 2017.7

[5] 宮本隆史 . Omeka S を活用した東京大学文書館デジタル・アーカイブの公開 , カレントアウェアネス -E, No.361, 2019.1.

[6] Omeka S, https://omeka.org/s/,（参照：2019-2-11）

[7] Omeka, https://omeka.org,（参照：2019-2-11）

[8] The LOCAH RDF Vocabulary, https://lov.linkeddata.es/dataset/lov/vocabs/locah,（参照：2019-2-11）

[9] REST API - Omeka S Developer Documentation, https://omeka.org/s/docs/developer/key_concepts/api/,（参照：2019-2-11）

[10] GitHub, https://github.com/iii-dch,（参照：2019-2-15）

[11] 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.

[12] Digital Cultural Heritage · 画像内テキストの検索機能について , http://dch.iii.u-tokyo.ac.jp/s/dch/page/textmanual,（参照：2019-2-15）

（たかしまともこ　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　特任助教）
（なかむらさとる　東京大学情報基盤センター　助教）