

社会情報研究資料センターニュース

第27号 (2017.3月)

目次

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 宮本隆史 中村 覚 1
センター情報 7



Digital Cultural Heritage トップページ

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

宮本隆史 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、デジタルアーカイブ・システムとして、Digital Cultural Heritage (以下 DCH) を構築し公開してきた。このシステムは、複数の資料群を収録し、さまざまな方法で検索を可能にする機能を提供するものとして開発された。提供する検索方法としてトピックマップを使うなど、2010年代初頭のプロジェクトとして野心的なものであった。しかし、DCH は、公開の数年後にはメンテナンスが困難となり、2016年に休止することとなった。現在は再公開に向けた準備が進められている。本稿では、公開休止に至った要因を考察し、その解決策を整理したうえで、再構築に向けた道筋を示す。

2. 『Digital Cultural Heritage』の公開休止に関する考察

Digital Cultural Heritage は、社会情報研究資料センター・高度アーカイブ化事業 (2007-2011年) によって構築された。通常の単語検索だけでなく、地図年表検索、トピックマップ検索など、多様な検索機能を提供することがその特徴であった。また、2016年までに「小野秀雄関係資料」「坪井正五郎関係資料」「森恭三関係資料」「外務省情報部関係資料: 第1次世界大戦期プロパガンダポスター」という4つの資料群を収録するようになり、これらの横断的な検索を可能とした。収録する資料群をさらに増やし、トピックマップによる情報提供を充実させることによって、「より広がりのあるデジタルアーカイブ」¹として成長することが期待されていた。しかし、残念ながらDCHは公開の数年後には休止を余儀なくされてしまうことになった。

DCH は、スクラッチ開発で構築したシステムを、自前のサーバ上で運用したものであった。このことは、開発当時の技術環境と社会情報研究資料センターの運営体制を考慮すれば、妥当な選択であったと考えられる。しかしこれらは、数年後にはシステムの維持を困難にする要因として意識されることになった。開発を進めた大規模プロジェクトが終了したため、プログラムの更新や自前サーバのメンテナンスに大きな費用をかけることは困難となってしまったからである。また、DCHのシステムが完全に独自のものであったことは、担当者の交代に際して学習コストを引き上げる結果になっていたと考えられる。

DCHのような多機能システムの場合、検索機能の一部に不具合が生じただけでも、基本的な部分も含め、全体を停止しなければならない。しかし、データ提供といった基本的機能は技術的にも比較的安定して実現できるものであり、技術が変化しやすい先端的な機能と同じシステム上に同居させておかなければならない理由はない。特に、クラウド環境が普及した現在では、その分離も容易になっている。

またDCHは、トピックマップを導入するなど、ネットワーク性の高いデジタルアーカイブ・システムを指向していた点で、当初から先端的なシステムを目指していた。しかし、その後の技術変化についていくことは残念ながら困難であった。特に、2010年代にLinked Data関連の技術の開発が大きく進んだにもかかわらず対応できずにいた。これは、独自に開発された多機能システムを運用していたために、部分的な修正や開発が難しかったことに一因があるだろう。

以上をまとめると、大規模プロジェクトの終了によって生じたつぎの問題点が浮かび上がってくる。

1. サーバ・マシンの維持の問題
2. システムの継続の開発の問題
3. 多機能システムの維持の問題

1 東京大学大学院附属社会情報研究資料センターのウェブサイト下の「閲覧室・デジタルアーカイブ」<http://www.center.iii.u-tokyo.ac.jp/exhibit> (2017年1月5日閲覧)

4. 技術変化への対応の問題

2.1. 小規模デジタルアーカイブ運用の基本戦略

こうした、DCHの問題群は、社会情報研究資料センターだけが直面したのではなく、デジタルアーカイブ・システムを運用する組織や個人にとって、程度の差はあれ一般的なものであろう。そこで、2017年初頭の技術環境において、小規模な組織や個人がデジタルアーカイブを運用するにあたって、現実的な基本戦略を考察してみたい。

サーバ・マシンの維持については、できるだけセキュリティ面のメンテナンスを外注することが、小規模なプロジェクトにとっては好ましいだろう。現在、費用の面でも現実的な選択肢のひとつとして、いわゆるウェブホスティング・サービスなど、インフラストラクチャーの管理を任せてしまえるサービスが考えられる。ただし、こうしたサービスでは、利用できる言語やRDBMSの種類に限られるなど、技術的な制約がともなう。典型的には、いわゆるLAMP環境(Linux, Apache, MySQL, PHP)での運用となることが多いだろう。

システムの継続的開発の問題に関しては、スクラッチ開発を極力避け、既存のコンテンツ・マネジメント・システム(CMS)を活用することで、開発にかかる費用をさげることができる。PHPのような軽量のスクリプト言語で書かれたCMSを用いることで、一般的なウェブホスティング・サービスで提供されるサーバに実装できるだけでなく、管理者の学習費用をおさえられるだろう。ただし、既存のCMSを使ってできることに限界はある。

多機能システムの維持が困難であるという問題については、ひとつのシステムで多くの機能を実現しようとせず、機能ごとに切り分けるという解決策が考えられる。このようにすることで、機能のひとつに不具合が生じて停止せざるをえなくなっても、他の機能は維持することができる。とくに、目録情報と画像などのデータだけは最低限提供しつづけられるように、基本機能だけを最小構成のシステムとして切り分けておくことが望ましい。この場合、Linked Data化やデータ可視化のためのアプリケーションなどは、最小構成システムの視野からは外すことになるだろう。

技術変化への対応の問題については、最小構成システムには標準的で今後の技術変化が少ないと思われる「枯れた」技術を使い、長期的な安定的運用を目指すことができる。

ただし、このような選択をする場合、最新の先端的な技術を最小構成システムに応用することはできず、アプリケーションによる活用の段階で行なうことになる。その意味でも、情報を提供するための基本的なデジタルアーカイブ・システムと、情報を活用するためのアプリケーションは切り分けておいたほうがよい。

情報を長期的に低コストで提供しつづけるためには、以上のようにデータ連携やデータ活用のための機能を切り分けて別システムとし、基本的な情報提供機能だけを維持することが現実的だと考えられる。DCHでは、目録データと画像データを提供する機能、異なる資料群の目録情報を連携させる機能、そして多様な検索機能が同じシステム上で動いていた。これらを切り分けることによって、たとえば検索機能のひとつが動かなくなっても情報の提供が継続できるようにすることが望ましい。いちど別個のシステムとしたうえで、それらを連携させることによって、相互の弱点を補完できるような生態系を構築するのである。具体的には、つぎのような3段構えのシステム群を用意することで、より柔軟かつ長期的にメンテナンス可能な環境を構築することを狙いとする。

- 最小構成アーカイブ・システム
- データベース化のためのシステム
- データ活用のためのアプリケーション

2.2. 最小構成アーカイブ・システム

データ提供に特化した最小のシステムを構築する。オペレーティング・システム(OS)やプログラミング言語の更新にあまり影響を受けない、長期的にデータを提供しつづけるための単純なシステムである。このためには、高度なプログラミング技術を必要とせず、LAMP環境で運用できるオープンソースのCMSを使うことが現実的な方法のひとつと考えられる。これによって、たとえば一般的なウェブホスティング・サービスでも、最低限のデータ提供が継続できるようになる。目録には、広く使われるメタデータの語彙を可能なかぎり使い、RDF/XMLやOAI-PMHなどの静的ファイルとして提供することを基本とする。また、画像データは、URI(Uniform Resource Identifier)を与えて提供する。

この最小構成アーカイブ・システムを使うことによって、安価にデータ提供を継続するというのが狙いである。

ここで提供される目録データと画像データは、あくまで静的ファイルであるため、これをそのまま使って高度な検索やビジュアル化などのために利用することは現実的ではない。

2.3. データベース化のためのシステム

最小構成アーカイブによって提供される静的ファイルの情報を横断検索などのアプリケーションに利用しやすくするためには、データベース化する必要がある。このデータベース化のためには、近年開発が著しく進んでいる Linked Data 関連の技術を使うことが有望と考えられる。詳細は次章にゆずるが、そのためには RDF ストアと呼ばれるシステムを実装できるサーバが必要である。これは最小構成アーカイブが想定するようなウェブホスティング・サービスでは実現が困難なことが多く、別個のサーバ環境が必要となるだろう。

このシステムでは、サポートするメタデータの語彙で書かれた目録情報であれば、DCH が収録していた4つの資料群だけでなく、複数のデジタル・コレクションからデータを取り込むことが可能となる。DCH が目指した「より広がりのあるデジタルアーカイブ」という目標は、これによって実現可能となる。いわば、Wikipedia に対する DBPedia のような位置づけにあるシステムであり、これによって提供される RDF データを使うことで、ユーザが直接さわることになる検索等のアプリケーションを構築することが容易となる。

2.4. データ活用のためのアプリケーション

RDF 化されたデータを使うことで、DCH が提供していた複数の検索機能や、データの可視化などのためのアプリケーションを作成することができる。目録データがオープンなライセンスで公開されていれば、基本的には誰でも自分のコンピュータ上にこうしたアプリケーションを作ることができることになる。

これらの3種類のシステムは、同じサーバ・マシン上で運用することも、別々のマシン上で運用することも可能であることが望ましい。従来の DCH のように多機能の単一システムを作り上げるのではなく、単一機能のシステムの集合によって有用なサービスを提供するという方向を目指すのである。そのために有望と考えたのが、前述の

Linked Data 関連の技術を活用することであった。

3. デジタルアーカイブの運用戦略の実装

前章で考察したデジタルアーカイブの運用戦略の実装に向けたシステム構成の全体像を図 1 に示す。本システム構成は欧州 35 ヶ国、3,000 以上の図書館・美術館・博物館・文書館等の文化施設が保有する資源へのアクセスを可能とする横断検索ポータルサイト「Europeana^[1]」を参考としている。

3.1. 最小構成アーカイブ・システム

前章の運用戦略に基づき、「最小構成アーカイブ・システム」を用いた資料公開を行う。具体的には、東京大学の学内組織向けウェブホスティングサービス「WEB PARK 2014」が標準機能として提供するプログラム言語「PHP」やデータベース「MySQL」、CMS「WordPress」等を用いて動作可能なシステムとして構築することにより、サーバのメンテナンス等に要する運用コストを低減する。さらに、設定情報の変更によって異なる組織が所蔵する資料の目録データや画像に対して汎用的に適用可能なパッケージシステムとして構築する。これにより、システム構築に要するコストを低減し、学内外におけるデジタルアーカイブを用いた資料公開を支援する。

新 DCH は、この最小構成アーカイブ・システムを用いたアプリケーションの一つとして位置付けられる。

さらに、RDB (Relational Database) である MySQL に格納されている目録データを RDF データに変換し、RDF ファイルとして公開する仕組みを提供する。これにより、データの保守性を高めると同時に、後述する Linked Data を用いた公開データの利活用を促進する。

3.2. データベース化のシステム

Linked Data とは、Web 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、分野やシステムを超えたデータ共有や再利用が容易となる利点を持つ。複数のデジタルアーカイブ間のデータ連携や、第三者によるアプリケーション開発が容易となり、再利用・再配布による新たな価値の創造が期待されている^[2]。

「最小構成アーカイブ・システム」は運用・構築コストを低減しつつ資料公開を支援することを目的としたシステムであるため、静的なデータ公開に限定され、第三者によ

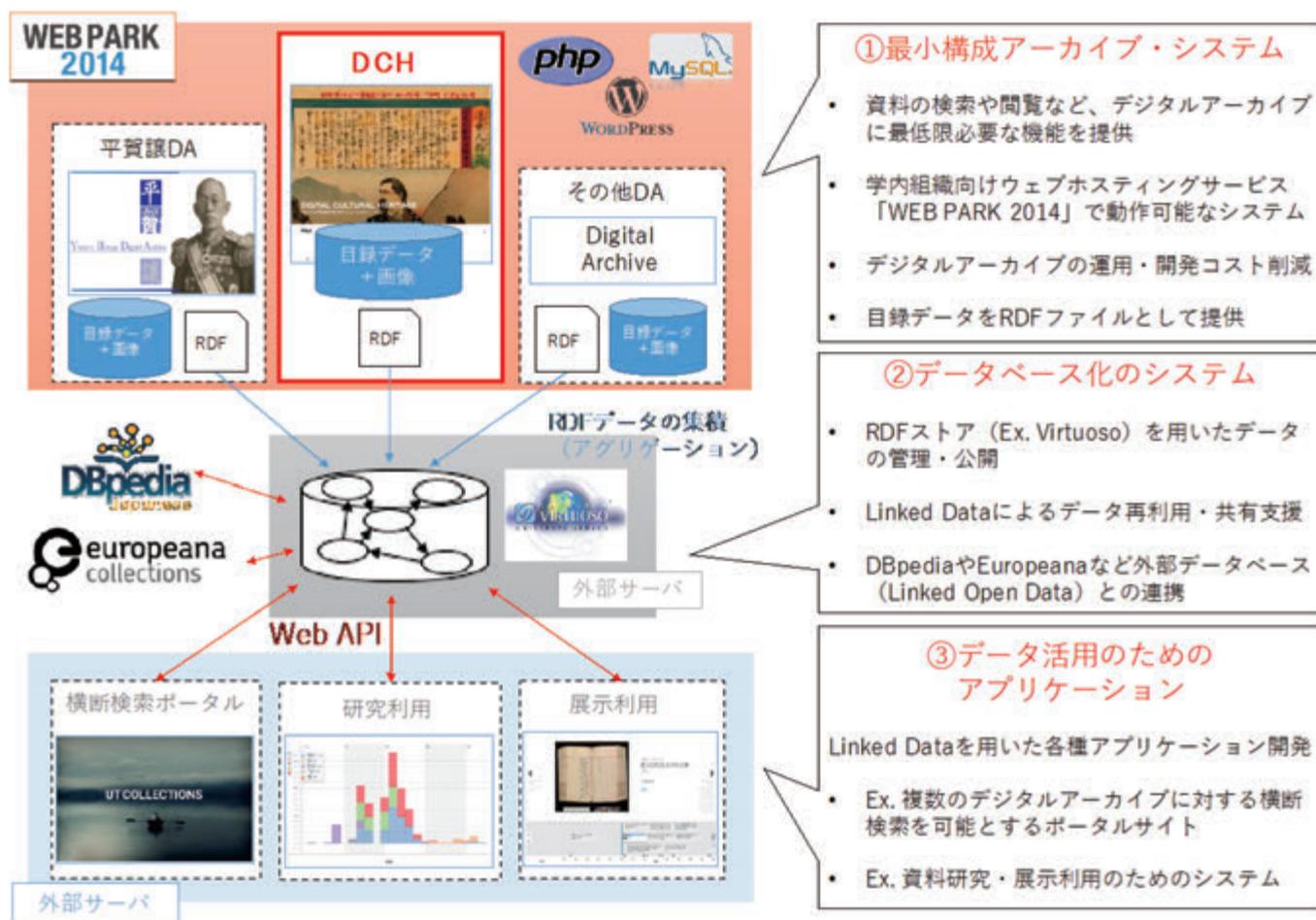


図1 システム構成の全体像

るデータ活用は考慮されていない。したがって、Linked Dataを実現するための記述言語であるRDF (Resource Description Framework) を用い、公開する目録データを「Virtuoso」などのRDFストアに集積し、インターネット上で公開する。これにより、Web APIを用いた外部アプリケーションからのデータ利用や、EuropeanaやDBpedia^[3]などの外部データベース (Linked Open Data) との連携を支援する。

3.3. データ活用のためのアプリケーション

Linked Dataとして公開されているデータを用いた各種アプリケーション開発を行う。具体的には、複数のデジタルアーカイブに対する横断検索ポータルサイトの構築や、公開データの研究、展示活動への利用などが挙げられる。本稿ではこれらの詳細な説明は割愛するが、筆者 (中村) の博士論文^[4]や東京大学文書館紀要論文^[5]の中で具体的な活用事例を述べているので参照されたい。

4. 新DCHのインタフェース例

ここでは、現在再構築中の新DCHのインタフェース例を紹介する。基本的には、DCHが提供していた機能をベースとする。異なる点としては、上述したように運用コストの軽減を目的とした「①最小構成アーカイブ・システム」の要件に基づき、学内組織向けウェブホスティング・サービス「WEB PARK 2014」で動作可能なシステムとして実装している点である。トピックマップや地図年表検索などについては、「③データ活用のためのアプリケーション」として今後実装し、新DCHが提供する機能からは除外している。

図2に資料検索インタフェースの例を示す。本インタフェースはユーザが入力した検索条件に基づき、検索結果として得られた資料の一覧を表示する。また、資料の階層情報に基づく検索インタフェースや、図3に示す各資料のメタデータの詳細や画像を閲覧可能なインタフェースの提供を予定している。

DIGITAL CULTURAL HERITAGE
東京大学大学院情報学環 デジタルカルチュラルヘリテージ

資料検索 概要解説

検索条件

タイトル

作成日

検索結果

20件の資料が見つかりました。

No	資料名
1	船火事をのがれ鰐に呑まれるのB-1-4-3-17
2	京の火事物語のG-7-5
3	桐生町出火事件 一ツトせぶしのG-10-3-1-23
4	「正徳三巳年三月廿日火事図」のG-10-5-2-18
5	日本橋本舟町火事 (仮) の017

図2 検索インターフェース例

総合アーカイブ > 総合アーカイブ > 小野秀雄関係資料 > 小野秀雄コレクション > 小野秀雄コレクションから版 >

東京の火事 (仮) の105

資料情報

項目	内容
来歴-所有者	東京大学大学院情報学環図書室/附属社会情報学環
来歴-現物資料の来歴	1970年代後半小野秀雄部より旧新聞研究部にて情報学環本館7F展示室に保存され、情報研究資料センター貴重資料保存スペースへ移転。
利用制限-現物資料利用条件	「東京大学大学院情報学環附属社会情報学環図書室蔵書利用規定」第10条第1項参照
利用制限-デジタル公開データ資料利用条件	「東京大学大学院情報学環附属社会情報学環図書室蔵書利用規定」第10条第1項参照
利用制限-デジタルマスターデータ資料利用条件	「東京大学大学院情報学環附属社会情報学環図書室蔵書利用規定」第10条第1項参照

図3 資料閲覧インターフェース例

5. 結論

DCHの再公開に向けて、目録情報と電子画像データを基本的には静的ファイルとして公開し、最低限の簡潔な検索機能を提供するというかたちが、長期的公開にむけてわたしたちがたどりついた現実的な解決策である。大きな費用をかけることなく利用できるウェブホスティング・サービス上に実装できる、軽量かつシンプルなシステムを構築することで、安定的に運用することをねらいとしている。また、検索機能は広く使われているオープンソースのCMSをベースとして開発している。社会情報研究資料センターだけではなく、デジタル資産をかかえるどのようなプロジェクトや個人も、将来的に利用できるよう配布することを目指す。そのために、最小構成アーカイブ・システムのパッケージ化とドキュメント作成を進めているところである。

今後の展開としては、外部のデジタルアーカイブとのデータ連携や第三者によるアプリケーション開発を促進するために、RDFストアを構築・公開する必要がある。しかし、たとえば本論で紹介したVirtuosoを実装するためには、簡易的なウェブホスティング・サービスが提供するサーバ・マシンでは不十分である。大規模なRDFストアを構築するためには、相応のリソースが必要となるため、安定的な予算を持つ機関がこの役割を担うことがのぞましい。目下のところ、研究・開発のためにサーバ・マシンを借りて運用している。

さらに、研究やウェブ展示のためには、それぞれのプロジェクトの要件に即したアプリケーションの開発が必要になる。これには、これまでいわゆるデジタル人文学などに関わってはこなかったような、人文・社会科学の研究者、各地の郷土史家、芸術家などの知見が非常に重要となる。これまでのデジタルアーカイブに関する議論は、情報の供給サイドに重心をおく傾向があったが、今後は情報の需要サイドについての考察と働きかけがさらに必要となる。どのようなデータどうしが連携することに意味があるのか、あるいはどのようなデータ活用の方法に意味があるのかといったことについての議論を詰めることが重要だろう^[6]。特に、地域に根ざしたアーカイブ連携とその活用を、今後の展開のひとつとして重視している。デジタルアーカイブを結節点とした、コミュニティの形成が今後の課題となる。

参考文献

- [1] Europeana, Available at: <<http://www.europeana.eu/portal/en>>. Accessed on: Feb 28th 2017.
- [2] トム・ヒース, クリスチャン・バイツァー著, 武田英明監訳, Linked Data: Webをグローバルなデータ空間にする仕組み, 近代科学社, p.7, 2013.
- [3] DBpedia Japanese, Available at: <<http://ja.dbpedia.org/>>. Accessed on: Feb 28th 2017.
- [4] 中村覚. デジタルアーカイブと Linked Data を用いた歴史学研究支援に関する研究, 東京大学新領域創成科学研究科博士論文, 2017.3 (予定).
- [5] 中村覚, 稗方和夫, 満行泰河, 加藤諭, 宮本隆史, 高嶋朋子, 『文部省往復』を中心としたデジタルアーカイブの構築とその活用, 東京大学文書館紀要, 第35号, 2017. (掲載予定)
- [6] 近代史研究からの多少の考察としてつぎを発表している。宮本隆史, デジタル・ヒストリーと制度, 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, 第26号, pp. 7-12, 2016; 宮本隆史, デジタル・アーカイブズと歴史の諸カテゴリー, 人文情報学月報, 61号, 2016 <<http://www.dhii.jp/DHM/dhm61-1>>. Accessed on: Feb 28th 2017.

(宮本隆史 東京大学文書館

中村覚 東京大学大学院新領域創成科学研究科)

センター情報

■社会情報研究資料センター長

平成 28 年度 石 崎 雅 人 (情報学環)

■社会情報研究資料センター運営委員会委員

平成 28 年度の委員の方々です。

石 川 徹 (委員長 情報学環)

板 倉 聖 哲 (情報学環)

鶴 田 啓 (情報学環)

河 井 大 介 (情報学環)

酒 井 麻千子 (情報学環)

■ペレストロイカ・ソ連終焉期の新聞コレクション (寄贈資料) の保管について

和田春樹東京大学名誉教授 (社会科学研究所) から寄贈された当該コレクション 28
タイトルを専用の中性紙保存箱に収納して書庫 M2 階に保管しました。

■非接触型スキャナーの導入

傷んでいたり製本が開ききらない資料を上向きのまま資料に無理をかけずに複写できる、非接触型オーバーヘッドスキャナ (FUJITSU ScanSnap SV600) を導入しました。

■空調機の更新

書庫 1 階マイクロフィルム保存庫の空調機更新を行いました。

■高機能空気清浄機の設置

書庫 1 階マイクロフィルム保存庫に、保存環境整備のため、高機能空気清浄機 (進和
テック フレッシュオドコップ MF) を設置しました。

■除湿機増設

書庫に保存環境整備のため除湿機を増設しました。

■復刻版週刊東日の刊行について

当センターが一部原資料を提供した「復刻版週刊東日」が不二出版より刊行されています。