

社会情報研究資料センターニュース

第34号 (2024. 3月)

目次

静的サイトジェネレータを用いたデジタルアーカイブシステムの構築： 『Digital Cultural Heritage』への適用 中村 覚 高嶋 朋子	1
センター情報	7

目次



中村芝翫、市川海老蔵、嵐冠十郎（熊谷直実、源義経、平宗清）
[小野秀雄コレクション錦絵]

静的サイトジェネレータを用いたデジタルアーカイブシステムの構築：『Digital Cultural Heritage』への適用

中村 覚 高嶋 朋子

諸言

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（以下、センター）は、2023年度にデジタルアーカイブシステム「Digital Cultural Heritage」の2度目の改修を行った。ここでは当初のDCHをDCH ver.1、初回改修を経たものをDCH ver.2とし、今回の改修版DCH ver.3に至った経緯と今回採用した静的サイトジェネレータを用いたシステム詳細について、以下に述べる。

DCH ver.2 構築の経緯と内容

DCH ver.2の構築は、2007-2011年の大型プロジェクトで開発された独自システムであったDCH ver.1の一時休止が契機となった。DCH ver.1はプロジェクト終了後に人的資源や予算の縮小に直面し、サーバマシンの維持、システムの継続的開発、技術変化への対応および多機能システム維持がそれぞれ困難となったことが一時休止の直接的原因である。筆者は、これら3つの問題点をデジタルアーカイブシステムの維持を目指す小規模な組織が共通して抱える課題ととらえ、レンタルサーバの利用、技術変化の少ない「枯れた」技術を使ったオープンソースソフトウェアの活用、基本機能と応用機能を分離したシステムの確立によって、これらの課題を克服することを提案した¹。DCH ver.2は東京大学情報基盤センターが提供する学内組織向けウェブホスティングサービス「WEB PARK」（実際には、さくらのレンタルサーバ）上でOmeka Sを用いたデジタルアーカイブの基本機能を実現し、Omeka Sが提供するAPIを介してデータ活用を行う仕組みを構築することで、長期運用が可能な安定性と単なる画像提供にとどまらない利便性を両立させている。

DCH ver.2の活用

DCH ver.2の構築について議論し始めた当初は、DCH ver.1で公開していたデータの再公開のみを目的としていたが、センターが所蔵する新聞原紙資料のデジタルアーカイブ化計画が持ち上がったことから、既存のデータを格納するだけでなくコンテンツの新規追加も行う想定へと変更

した経緯がある。センターでは、新聞研究所時代に収集されたものを含めた希少な新聞原紙資料を所蔵しており、クリーニングや補修等、保管のための適切な対策を施してきた。しかしながら、新聞紙の特性を鑑みれば劣化を防ぐことは難しく、原紙を開閉するたびに小さな破片が落ちてくるような状態にある。そこで、利用者に不便をかけず原紙の使用機会を減らすひとつの方策として検討されたのが、新聞原紙資料をデジタルアーカイブ化して提供していくことである。この計画は2017年度から2021年度の「東京大学デジタルアーカイブ構築事業」内の一事業として採択され、少しずつデジタルアーカイブ化を進めて、DCH ver.2で公開してきた。現在においても、デジタルアーカイブ化が完了した新聞原紙資料は未だ所蔵全体量のほんの一部にすぎないが、これらの成果を広く社会に公開することで、学術分野にとどまらない新聞資料の多様な活用を後押ししている。更に、公開した新聞原紙資料の一部については、当該資料の利用価値や背景の理解促進に資するための資料紹介を、社会情報研究資料センターニュースに掲載した。

DCH ver.2が安定的に稼働していることから、センター所蔵資料の管理を担っている情報学環附属図書室とデジタルアーカイブについての相談事項が増え、公開に関して提案を受ける機会も増えた。図書室との協力によって追加した新規公開データは、センター所蔵の貴重資料である小川一写真真帳や情報学環附属図書室所蔵の摺物やポスター類だが、DCH ver.1で一部のみを公開していた「小野秀雄コレクション」については、図書室が管理していた小野秀雄コレクションページ（現在は閉鎖）で提供していた全データを引き継いでの公開とするなど、データの再整備も行うことができた。また、「大井第一小学校寄贈資料」も、図書室からどのように利活用をはかるべきか相談を受けたコンテンツのひとつである。140冊を越える昭和30年代の児童本や少年雑誌の付録コミック本等を中心としたこの寄贈資料群の多くは著作権保護期間を満了していないため、すべてをデジタルアーカイブとして公開することはできない。しかし、手書きによる詳細な目録データがあるにも関わらず公開されていなかった点に着目し、この目録データをデジタル化して各資料の表紙のサムネイル画像とともにDCH ver.2上で公開することで、資料の全容にアクセスしやすい状況を整えることができた。

1 中村覚, 高嶋朋子, 持続性と利活用性を考慮したデジタルアーカイブ構築手法の提案, デジタルアーカイブ学会誌, 5巻, 1号, p. 56-60, 2021.

DCH ver.1 から引き継いだ過去のプロジェクトで構築された貴重なデータに加えて、以上のように、現在のセンター体制のなかでデータの再整備や新たなデジタル化および公開が進んだことは、センターとして非常に大きな収穫であった。しかも DCH ver.2 は、OAI-PMH を用いて「東京大学学術資産等アーカイブズポータル」や「Japan Search」など外部のシステムと連携したことから、こうしたポータルを経由して DCH を利用する層が順調に増加していることが確認されており、より多くの利用者にセンター所蔵の資料を知ってもらう機会を得られたといえるだろう。

DCH ver.2 の課題

上記のように、DCH ver.2 の運用はセンター事業の柱のひとつとなり、極めて安定的な運用を維持してきた。しかしながら、今一度見直すと、いくつかの小さな課題を挙げることができる。

DCH ver.2 で採用した Omeka S は、デジタルコレクションを構築するためのコンテンツ管理システムで、高い柔軟性と拡張性を備えながら簡便に使える点で評価が高い。特徴的な機能として Linked Data への標準対応があり、DCH ver.2 でも OAI-PMH を用いたメタデータ連携によって得られた成果があったことは前章で触れた。IIIF モジュールにも対応しており、Omeka S を採用するメリットは大きかったが、現状では保守運用を請け負う業者が少なく自前でのメンテナンスを要するため、今後の担当者への入れ替わりを想定すると長期運用にあたっての課題となる可能性が高い。また Omeka S は、DCH の運用にあたって必要以上の機能が提供されていた。例えば、ユーザ管理機能、それに伴う動的なコンテンツの表示機能などが挙げられる。

DCH でデータの編集や登録をするのは 2-3 名のみで、データの更新頻度も年に数回程度、コンテンツも数万という小規模なものであり、運用のために大きな予算を確保するのは困難な状況下にある。必要とするのは検索、階層表示、各アイテムの詳細ページというミニマムな機能であることから、よりシンプルでメンテナンスコストが低いシステム構築手法の検討に着手した。従来通り「動的サイト」としてシステムを構築するか、よりシンプルな「静的サイト」として構築するか、という構築手法の根本的な観点から再考を行った。動的サイトおよび静的サイトの違いは以下の通りである。

動的サイト (Dynamic Website) :

- 定義：動的サイトは、訪問者のリクエストに応じてリアルタイムで内容が生成されるウェブページで構成される。

• 特徴 :

- コンテンツの柔軟性：ユーザーのインタラクションやデータベースの情報に基づいて、内容が変わる。
- インタラクティブ：フォーム、フォーラム、E コマース機能など、ユーザーとのインタラクションが可能。
- 複雑な技術：サーバーサイドのプログラミング (PHP、Ruby、Python など) とデータベースの使用が一般的。
- 用途：E コマースサイト、ソーシャルメディア、ニュースサイトなど、頻繁に更新される情報やユーザーインタラクションを必要とするサイトに適している。

静的サイト (Static Website) :

- 定義：静的サイトは、固定された内容のウェブページで構成される。これらのページは、サーバー上にあらかじめ作成され、保存されており、訪問者が見るときにそのまま提供される。
- 特徴 :
 - 変更が少ない：内容は固定されていて、更新するにはファイルを手動で編集する必要がある。
 - 高速：ページは既に作成されているため、読み込みが速い。
 - シンプルな構造：デザインがシンプルで、通常は HTML、CSS、場合によっては JavaScript を使用する。
- 用途：基本的な情報を提供するサイト、ポートフォリオ、小規模なビジネスサイトなどに適している。

上記のような特徴の違い、およびリソース（作業人数、更新頻度、コンテンツ数、利用可能な情報インフラなど）が限られた組織におけるシステムの運用を考慮して、DCH ver.3 では静的サイトを用いた構築手法を採用することとした。

システム

本章では、DCH ver.3 のシステム的な側面について説明する。まず採用したシステムの構築手法について述べ、次にシステムの使用方法等について紹介する。

構築手法

上述した通り、リソースが限られた組織におけるシステムの運用に関する特徴を踏まえて、デジタルアーカイブシステムを静的サイトとして構築する手法を採用した。関連す

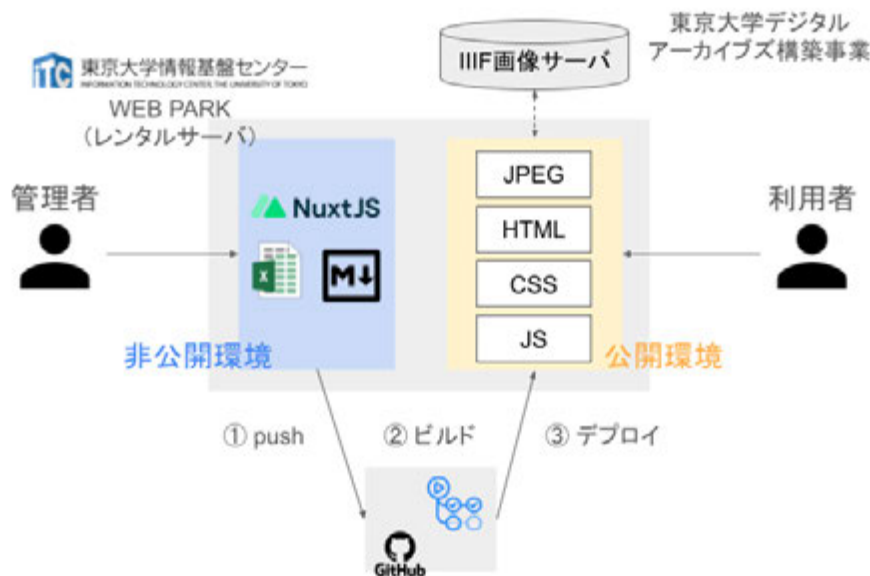


図1 構築手法の概要

る取り組みとして、国内の事例としては、江草²、阿達³らの取り組みが挙げられる。国外の事例としては、Wax⁴やCollection Builder⁵などのツールが開発されている。

具体的には、ウェブアプリケーションを作成するためのフレームワークの一つであるNuxt⁶を用いて、本フレームワークが提供する静的サイトジェネレータ（SSG: Static Site Generator）を使用した。これにより、DCH ver.3で提供されるページはサーバーサイドのプログラミング言語を使用せず、HTMLやCSS、JavaScriptファイルなどの静的ファイルのみで構成される。

構築手法の概要を図1に示す。インフラとして、WEB PARK、GitHub、および東京大学デジタルアーカイブズ構築事業が提供するIIIIF画像サーバの3つから構成される。次節において、これらのインフラを使った運用方法について説明する。

2 江草由佳:「小規模図書館における“必要十分 な” デジタルアーカイブの構築について / Development of Digital Archives in small libraries,2020」. <https://speakerdeck.com/yegusa/development-of-digital-archives-in-small-libraries-2020> (2024年1月31日参照)。

3 阿達 藍留, 山田 俊幸, 大向 一輝, DAKit: 低コストなデータ共有のための静的デジタルアーカイブジェネレータの提案, 情報知識学会誌, 2022, 32巻, 4号, p. 406-409, 公開日 2023/01/27, Online ISSN 1881-7661, Print ISSN 0917-1436, https://doi.org/10.2964/jsik_2022_035, https://www.jstage.jst.go.jp/article/jsik/32/4/32_2022_035/_article/-char/ja

4 Wax, <https://minicomp.github.io/wax/>, Accessed on 2024-01-26.

5 CollectionBuilder, <https://collectionbuilder.github.io/>, Accessed on 2024-01-26.

6 Nuxt: The Intuitive Vue Framework, <https://nuxt.com/>, Accessed on 2024-01-26.

運用方法

コンテンツ管理

DCHが管理するコンテンツの種類は、画像、メタデータ、および「お知らせ」や「利用条件」といったテキストデータに大別される。これらのコンテンツはWEBPARK上で共有し、コンテンツの編集にあたっては、WEB PARKが提供するサーバ内のファイルをウェブブラウザ上で管理する機能「ファイルマネージャー」を用いる。コンテンツ管理にあたっては、TinaCMS⁷などのGitベースのCMSや、DrupalなどのCMSを採用する方法も検討したが、使用するツール群をシンプルにするという観点からこれらの採用は見送った。

画像

画像ファイルの大部分については、東京大学デジタルアーカイブズ構築事業が提供するIIIIF画像サーバを利用している。本サーバにアップロードされた画像は、Cantaloupe⁸というソフトウェアを通じて、IIIIF画像が配信される。一方、いくつかのコレクションについては、WEB PARK上で画像を公開している。

メタデータ

メタデータはMicrosoft Excelによって管理し、コレクション毎にExcelファイルを作成している。フォーマットとしては、DCH ver.2で使用していたOmeka Sのモジュール

7 The Markdown CMS | Tina, <https://tina.io/>, Accessed on 2024-01-26.

8 Cantaloupe Image Server, <https://cantaloupe-project.github.io/>, Accessed on 2024-01-26.



図2 階層検索

ルの一つである「BulkImport」の記述⁹を参考している。例えば、非公開情報として扱いたい列については、列名の末尾に「§ private」という文字列を追記するようにしている。また、URLとして扱いたい列については、「^^uri」を追記する、といったルールを定めている。

テキストデータ

テキストデータはMarkdown（マークダウン）ファイルで管理する。マークダウンファイルを採用した理由として、シンプルさ、今回採用したNuxtをはじめとするウェブアプリケーションフレームワークとの親和性の高さ、などが挙げられる。なお、マークダウンファイルを直接編集することが難しいと感じる作業者に対しては、オンラインのマークダウンエディタ¹⁰を併用することで、直感的なテキスト編集を支援する。

静的サイトの公開

静的サイト（SSG）という仕組みを導入するうえで、上述したExcelファイルやMarkdownファイルからHTML

ファイルを作成するプロセス（以下「ビルド」）が必要となる。

本サイトでは、Nuxtというフレームワークを採用しているため、ビルドにあたってはNode.jsというサーバサイドの言語が必要になる。しかし、2024年1月時点において、WEB PARKにはNode.js環境が標準で用意されていない。そのため、GitHubが提供するGitHub Actionsという仕組みを使ってビルドを行っている。このビルドによって作成されたHTMLファイルなどをWEB PARKにコピー（デプロイ）することで、さくらレンタルサーバ上での静的サイトの公開を実現する。

画面紹介

DCH ver.3はDCH ver.2の機能を基本的な機能を引き継ぎ、以下の3つの機能を提供する。

- － 階層検索
- － キーワード検索
- － アイテム詳細ページ

階層検索

階層検索の画面例を図2に示す。従来のシステムでも提供されている機能であるが、DCHに含まれるコレクションの階層関係を表示するとともに、各コレクションをクリックすることにより、当該コレクションに含まれるアイテムに絞り込むことができる。

9 Daniel-KM/Omeka-S-module-BulkImport: Module for Omeka S that allows to create extensible importers and to process bulk import of any source (spreadsheet, another Omeka S...), <https://github.com/Daniel-KM/Omeka-S-module-BulkImport?tab=readme-ov-file#spreadsheet>, Accessed on 2024-01-26.

10 Free Collaborative Markdown Editor, <https://onlinemarkdowneditor.dev/>, Accessed on 2024-01-26.



図3 キーワード検索

キーワード検索

キーワード検索の画面例を図3に示す。画面上部の検索フォームから、入力したキーワードに基づく検索が可能のほか、「メディア（画像の有無）」「コレクション」「種別」「分類」といった項目（ファセット）を用いて、アイテムの絞り込みを行うことができる。

各アイテムをクリックすると、アイテムの詳細画面に移る。

アイテム詳細ページ

アイテムの詳細ページの画面例を図4に示す。

本画面では、IIIF 対応ビューアの一つである Mirador を用いて画像を閲覧できるほか、JSON:API¹¹ に準拠した JSON データのダウンロード機能、および IIIF マニフェストファイルのダウンロード機能などを提供する。これらの JSON データを活用することにより、公開コンテンツに対する機械的な処理も支援する。

考察

今回採用した静的サイトとしてデジタルアーカイブを構築することによって、DCH ver.2 が抱えていた課題を解決

することができた。

ただしデジタルアーカイブシステムの長期的な安定運用には、定期的な更新やシステム移行など、環境の継続的な見直しが求められる。今後、次にあげるような点を引き続き考慮する必要がある。

フレームワークへの依存

本手法では、Nuxt というフレームワークを採用したが、このフレームワークの将来性を考慮する必要がある。今回は静的サイトとしてシステムを公開するため、このフレームワークに依存するセキュリティリスクは、動的サイトと比較して低いものと考えられるが、定期的な更新や将来的な他のフレームワークへの移行は避けられない。一方、このような課題はどのようなフレームワークおよびシステムを採用しても発生しうるものである。

これに対して、本手法では Excel ファイルや Markdown ファイルを使ってデータを管理する方式をとっているため、システム移行に伴うコストは比較的小さいものになると考えられる。

GitHub サービスへの依存

本手法では、アプリケーションの公開、特にビルドにおいて、GitHub サービスを利用している。GitHub が今後すぐにサービスの変更や終了を行うことは考えにくいですが、代替手段を検討しておく必要がある。

11 SON:API — A specification for building APIs in JSON, <https://jsonapi.org/>, Accessed on 2024-01-26.



図4 アイテム詳細画面

具体的には、Node.js が動作するサーバが必要となる。選択肢としては、さくらのクラウドや AWS (Amazon Web Services) といった商用サービスの利用が考えられる。または、大学・研究機関で共創する産学官連携のためのデータプラットフォームである mdx¹² などの利用が挙げられる。これらはビルド時にのみサーバを起動する形をとることで、コストを抑える他、セキュリティリスクの軽減をはかる。

結言

本稿では、「静的サイトジェネレータ」を用いた DCH の改修について述べた。リソースが限られた組織におけるシステムの運用を考慮して、静的サイトによるデジタルアーカイブシステムの構築手法について説明した。

最後に、DCH ver.3 へ移行するタイミングで、全てのコンテンツの利用条件を CC-BY 相当に変更したことについて触れておきたい。

DCH ver.2 では、コンテンツによって利用条件が異なっていた。例えば、利用頻度の高い小野秀雄コレクションや第一次世界大戦プロパガンダポスターを二次利用する場合は、資料原本の出陳の場合と同様に、定形フォーム

での利用申請を義務付けていた。ネット上で画像を容易に確認できても、二次利用にあたっては事務手続きをしなければならない不便さを利用者に強いるルールではあるが、センター側はこの申請を受けることによってデジタルアーカイブ化した所蔵資料の利用状況を明確に把握できるという利点があった。予算の獲得や担当者の配備といった基盤を整えるところからさまざまな工夫をしながら運営しているデジタルアーカイブは、利用状況の把握がその維持に大きく影響することはいうまでもない。デジタルコンテンツの自由な利用とその利用状況の把握を両立させるため、どのような利用条件を課すかについては、他機関の対応も例として参考にしながら議論を重ねた。

結論として、DCH ver.3 のコンテンツは全て、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と、それらコンテンツを利用した成果物の提供または利用実績の報告を依頼する「CC BY 相当」での利用条件と決定した。明示や提供の類はすべて「依頼」ではあるが、利用者自身がデジタルアーカイブ事業を推進する側のひとりでもあることを理解してもらい、積極的にデジタルコンテンツを利用して、この「依頼」を受けていただきたい。

(なかむらさとる 東京大学史料編纂所 助教)
(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)

¹² mdx, <https://mdx.jp/>, Accessed on 2024-01-26.

センター情報

■社会情報研究資料センター長

令和5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

令和5年度委員

石崎雅人(委員長 情報学環)

丹羽美之(副委員長 情報学環)

■Digital Cultural Heritage (DCH) の利用条件変更

DCH ver.3への改修を機に、画像利用条件を「CC BY」(クリエイティブ・コモンズ 表示 4.0 国際ライセンス)相当に統一した。

■Digital Cultural Heritage (DCH) 公開資料追加

- ・バジル・ホール・チェンバレン旧蔵小川一写真真帖
- ・情報学環・学際情報学府図書室所蔵摺物(合戦絵と大火のかわら版)
- ・日中戦争・太平洋戦争期ポスターコレクション(支那事変国債に関する広告など)
- ・新聞「日本」

■小野秀雄コレクション新聞錦絵「艶琴誌」のデジタル化

未整理資料であった「是ヨリ勸善懲悪風雅街道 艶琴誌 壹号」をデジタル化した。今後はデジタルデータをDCHで提供すべく作業を進める。

■令和5年度の資料利用サービス状況

新型コロナウイルス感染症対策として利用者同士の接触を避けるため、事前予約制により資料利用サービスを提供している。学外者に対しての来館利用(事前予約制)も6月1日から再開している。