

DAPCON Newsletter

デジタルアーカイブ
推進コンソーシアム
Newsletter

No. 08

CONTENTS

委員会紹介：

データ活用の持続可能性検討委員会

データがモノを言う時代

中澤敏明 東京大学大学院情報理工学系研究科 特任講師

トピックス



デジタルアーカイブ推進コンソーシアム
DAPCON

委員会紹介：データ活用の持続可能性検討委員会

2018年5月に発足し、月1回のペースで開催している「データ活用の持続可能性検討委員会」では、これまでに計7回の検討委員会を実施しました。本検討委員会では、デジタルコンテンツの保存・活用に関わる課題設定や解決策について検討を進めています。

設立の趣旨

ICT/IoTで発生する膨大なデータの保存活用を核とするデジタルアーカイブ産業の発展にとって、我が国におけるデジタルコンテンツの保存・活用に関わる技術的・制度的共通基盤の整備は不可欠の要件です。そのための課題設定と解決策を探るための検討委員会として2018年5月に発足しました。

検討事項

本検討委員会では下記の4つの項目について、検討を進めています。

- (1) 民産官学すべてのセクターで産み出される大量のデジタルデータの長期間の利活用を前提とした、保存様式・媒体・設備等及びそれらの維持管理のあり方
- (2) 共同データの二次利用促進のための情報空間デザイン・制度整備の方向性
- (3) 共同データ保存活用センターのあり方
- (4) その他国レベルの大量データ保存活用に関わる仕組みづくり

構成メンバー

構成メンバーは下記の通りです。

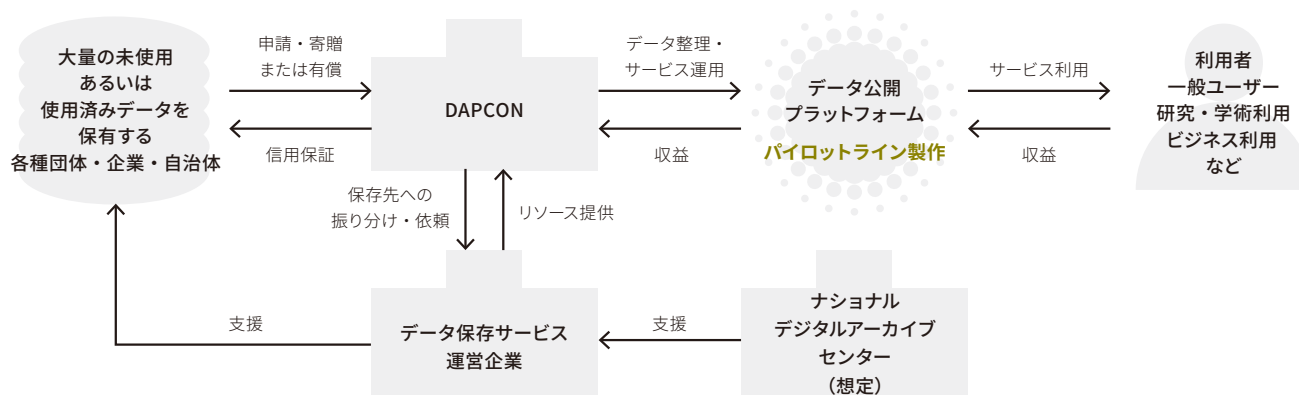
〈研究者所属〉

京都大学東南アジア地域研究研究所
慶應義塾大学理工学部情報工学科
東京大学情報基盤センター
筑波大学図書館情報メディア系

〈DAPCONからの参加企業〉

株式会社IMAGICA Lab.
キヤノン株式会社
ソニーストレージメディアソリューションズ株式会社
寺田倉庫株式会社
日本オラクル株式会社
富士通株式会社
富士フイルム株式会社
株式会社ワンビシアークライブズ

目指すべきデータアーカイブ産業プラットフォームのスキーム



本検討委員会及びPoCの今後の進め方について

これまでの検討結果から、今後は本検討委員会をA,B,Cの3つに分け、検討を進めていくことに決定しました。

A 大量データ保存・活用モデルのためのPoCの実施

- 1 PoCの内容については、「流通フォーマットとしてのデジタルチラシ」の形成と流通をテーマに実施する。
- 2 本検討会の中に企画・実施のための分科会を設置する。
- 3 実施のための体制と予算確保（政府補助金の獲得を主眼とする）を行う。
- 4 補助金獲得先としては、総務省（SCOPE社会展開指向型研究開発プログラム、ICTイノベーション創出チャレンジプログラム）、経済産業省（学びと社会の連携促進事業「未来の教室」、Connected Industries推進のための協調領域データ共有・AIシステム開発促進事業）を対象に考える。但し2019年度・2020年度の応募時期等は未定。

B 大量データ保存活用社会に関わる諸課題に関する勉強会の開催

- 1 学術・実務の両面から標記に関する新しい方向性・アイデアを出すことを目的として、月1回フリーディスカッションを旨とする勉強会を杉本委員長主宰で開催する。
- 2 各回は、研究者1名に発題をお願いし、最新の研究動向や新しいアイデアについて提案していただき、その後討論を行う。
- 3 討論の結果、具体的な検討が有効と思われる場合は、新しい分科会の設置も考慮する。
- 4 とりあえず1年間の継続を想定。

C 本検討会の継続

- 1 年数回開催し、PoC分科会及び勉強会の進捗状況及び今後の取組等について協議する。
- 2 本検討会、PoC分科会、最新動向勉強会については、新規の参加を含め、改めてメンバー構成を考慮する。

会社紹介 DAPCONのメンバー企業をご紹介します。

寺田倉庫

寺田倉庫の取り組み

当社は創業以来、貴重図書や機密書類、美術品やワイン、フィルムや映像・音楽テープ類といった、専門性の高い商品の保存・保管においては、業界内で高い評価を得ており、保存・保管の技術とスペースを活用するノウハウを駆使した、新たな価値を創出するサービス提供の加速と、東京湾のウォーターフロントである天王洲アイルを拠点とした芸術・文化の発信を行なっています。

デジタルアーカイブ関連サービス

多種多様なドキュメントや破損しやすい記録メディアの徹底した温湿度コントロールによる品質保持と、防災・セキュリティ対策が施されたセンターでの保管に加えて、デジタルアーカイブ関連事業にも取り組んでおり、物理保

管の先にあるコンテンツの利活用までを見据えたワンストップのサービス提供を行っています。保管施設内での安全なリフォーマット・デジタル化、電子データに対して信頼のおける確定日時を付与して改ざん検知を可能とするタイムスタンプの提供、データ容量が大きく安定的な長期保存に膨大なコストがかかる映像・音楽ファイルを、ネットワークを介した攻撃を回避して安全に保存できるオフラインストレージ、全編プレビュー／検索項目のカスタマイズ／メタデータの追加や編集／過去作品の共有／本番データの取り寄せ依頼等、コンテンツにまつわる管理を一元的に行うことができる管理システムもご用意し、お客様の大切な文化資産の継承と利活用をサポートしています。



TRC-ADEAC 株式会社

地域社会への貢献

弊社は図書館の総合支援企業である株式会社図書館流通センター（TRC）の関連会社です。TRCは図書館を「人類の英知を未来へと生かす知識と情報の宝庫」と考え、その支援を通して地域社会に貢献することを使命としています。弊社もまたデジタルアーカイブシステムADEAC（アデアック）を運営することにより、TRCの理念と同様に主に史資料の高精度デジタルデータの提供を行うことを主眼としています。

ADEACという名前はA System of Digitalization and Exhibition for Archive Collectionsの略です。命名およびデジタル化・システム化の方式は、2010～2012年度における東京大学史料編纂所社会連携研究部門「図書館等所蔵資料の調査・整備研究」の成果に基づいています。

ADEACの特長と今後

ADEACは地域の文化資源をデジタル化・公開するためのクラウド型プラットフォームです。

- ・資料の高精細画像・Dデータの閲覧
- ・本文データのフルテキスト検索
- ・各搭載データ間の横断検索
- ・安価な運用コスト

等が特長です。コンテンツの搭載については既存データの利用はもちろんのこと、資料の新規デジタル化のご相談から公開コンテンツの作成まで対応し、データ利活用の提案を含むトータルサービスを提供させて頂いています。

2019年1月時点で採用機関が96になりました。NDLサーチ連携やIIIFへの対応が済み、昨年10月には古地図をベースに大容量の情報を高速に提供する「WEB版デジタル伊能図」という新サービスも始まりました。現在、外部システムであるジャパンサーチやEBSCOディスカバリーサービスとの連携といった話も進みつつあります。システム改修は日々行われており、時代の変化や要請に対応した機能追加・改良を続けています。引き続き地域社会への貢献という使命を大切にしながら、デジタルアーカイブ推進に務めて参ります。今後ともよろしくお願いいたします。



データがモノを言う時代

中澤 敏明 東京大学大学院情報理工学系研究科 特任講師

コンピュータに何かを自動で行う能力を持たせるには、一般的にコンピュータに「学習」させることが必要であり、この学習のための材料となるのが「データ」である。ここ数年の深層学習を基礎としたAI技術の進歩には、学習データの量が非常に大きな役割を果たしている。深層学習によってコンピュータが自動で何かを行う能力の幅も精度も著しく向上したが、だからといって「AI ってなんでもできるんでしょ?」というのは大きな間違いだ。コンピュータに何かをさせなければ、その目的に合った学習データを大量に用意しなければならない。私は機械翻訳の研究をしているが、機械に翻訳をさせるのにもデータは非常に重要である。ちなみに機械翻訳の学習に必要なデータは、対訳コーパスと呼ばれる2言語以上で同じ内容が書かれた文の集合である。

ここでよく問題になるのが、どの程度の量のデータを用意すればいいのか?ということである。一般的に学習データは多ければ多いほどコンピュータの能力は向上するが、データ量と能力との間に明確な関係があるわけではない。問題の複雑さと、用意したデータが入力のパリエーションをどれくらいカバーしているかによる。極端な例として、入力の種類が2つしかなければ、学習データも2つ用意すれば十分である。しかし現実の問題では入力は多種多様であるため、学習データもそれに合わせて様々なものを用意しておく必要がある。現在の機械翻訳システムでは数十万から数百万、多い時には数億といった対訳文を学習に使うが、それでも全ての入力に対する学習データを用意することは不可能であるため、限られた学習データでいかに未知の入力に対応できるかが、学習手法の腕の見せ所ということになる。

さて、何か新しいことをAIを使ってやりたい場合、上述のように学習データが必要なのであるが、一から新しくデータを整備するにはかなりの時間とお金がかかってしまう。例えば100万文からなる対訳コーパスを一から翻訳して作るとなると最低でも数億円必要になる。しかし、実はこのような学習に使えるデータは毎年大量に作られているのである。このページ(<https://journal.jtf.jp/eventreport/id=721>)によると2017年度における翻訳の市場規模は推計2561億円だそうだ。つまり毎年何億文もの翻訳が生成されており、元の文と合わせることで対訳コーパスとして利用可能なデータが日々生み出されているのである。しかし残念ながらこれらのデータの多くはきちんと管理されておらず、その場限りの利用で終わってしまっている。某新聞社でも日本語記事を元に英語記事を作っているのだが、記事の対応情報が保存されておらず、後で別の目的に再利用することが難しい状態なのだそうだ。

データは後から利用できてこそ初めて価値があるものであり、他のデータと結びつけることで新たな価値も生まれるであろう。これには、1. データの所有者がその価値を正しく認識すること、2. データを利用しやすい形式で保存すること、3. どこで何が役に立つかわからないので、可能な限り多種多様なデータを保存することなどが大切だと思う。さらにデータの共有というのも大事な点である。これは各社の利益やプライバシー保護など色々と言葉があると思うが、共有することでより大きなメリットがあるということを実例を挙げて説明できれば、次第に共有の大切さが伝わるはずであり、これは私のような研究者やDAPCONの課題であろう。

トピックス

- ▶12/18, 1/23, 2/25 DAPCON: テキストデータ再活用推進検討会
- ▶1/10, 2/19 DAPCON: デジタルアーカイブ産業賞選考委員会作業部会
- ▶2/6 DAPCON: データ活用持続可能性検討委員会
- ▶2/11 JSDA: 研究会「デジタルアーカイブIN岐阜」開催

主催はデジタルアーカイブ学会人材養成部会・コミュニティアーカイブ部会ほか。(岐阜女子大学にて)

- ▶2/13 DNP 学術電子コンテンツ研究寄付講座: 3周年記念シンポジウム開催

DNP 学術電子コンテンツ研究寄付講座の3年間の活動成果およびレビューと、「我が国のデジタルアーカイブ: これからの課題」をテーマにパネルディスカッションをおこなった。

- ▶3/12 JSDA: 研究会「コミュニティ映画・映像のアーカイブを考える」開催

主催はデジタルアーカイブ学会コミュニティアーカイブ部会。ヤフー本社 LODGE にて開催。

- ▶3/15, 16 JSDA: 第3回研究大会(京都大学にて)

DAPCON: デジタルアーカイブ推進コンソーシアム
DARA: デジタルアーカイブ研究機関連絡会
JSDA: デジタルアーカイブ学会