文学作品におけるオープンデータ化の取り組みとその展望

兼松 篤子*1・浦田 真由*2・遠藤 守*1・安田 孝美*1
Email: kanematsu@is.nagoya-u.ac.jp

*1: 名古屋大学大学院情報科学研究科
*2: 名古屋大学大学院国際開発研究科

◎Key Words 文学作品，オープンデータ，電子書籍制作

1. はじめに

2012 年より政府主導で始まったオープンデータ推進の取り組みは，従来のトップダウン型と市民参加のボトムアップ型との相乗効果により，年々加速している。今年 2016 年は江戸川乱歩や谷崎潤一郎など著名な作家の著作権が切れパブリックドメイン[1]になった年でもある。近年 TPP に関連して話題となっている著作権保護期間についての関心が高まっている。

本稿では，著者の非常勤先である金城学院大学文学部日本語日本文化学科の学生を対象に開講されている「電子書籍制作」における取り組みを例に，文学作品のオープンデータ化とその展望について述べる。昨年度の授業では，『村岡花子童話集たんぽぽの目』を底本[2]に，EPUB3 形式による電子書籍制作を行った。そこでの結果と考察(1)を踏まえ，今年度は実際に著作権が消滅した江戸川乱歩の作品を底本に選び，デジタルデータ化と EPUB3[3]の電子書籍の制作に取り組んでいる。これらのデータは後に，文学作品のオープンデータ活用へと繋がっていくことを期待するものである。

2. オープンデータと文学作品

オープンデータ[4]とは，政府など公共機関や民間事業者，企業などが保有するデータを機械判読に適した形式，かつ誰もが 2 次利用可能な利用ルールのもと公開されているデータである。ウェブや Linked Data の創始者であるティム・バーナーズ＝リー（Tim Berners-Lee）は，機械判読可能なデータ形式について5★オープンデータ[5]を提唱している（表 1）（図 1）。そこでは，オープンデータのための 5 つ星スキームとそれに伴うコストや利益についても提唱されている。データを公開する際オープンラーセンスを基本とし，クリエイティブ・コモンズ・ライセンス[6]（以下，「CC ライセンス」）に基いている。そしてそれらデータは，CC ライセンスによる利用ルールのもと，データのマッシュアップや社会活用などによって，新たな価値を生み出すことが期待されている。

表 1　5★オープンデータ

星の数	データ形式の例	公開の状態
★	PDF，JPG	OL-Open License （計算機により参照できる，可読）
★★	XLS，DOC	RE-Readable(Human&Machine) （コンピュータでデータが編集可）
★★★	XML，CSV	OF-Open Format （アプリケーションに依存しない）
★★★★	RDF，XML	URI-Universal Resource Identifire（リソースのユニーク化，Web リンク）
★★★★★	LOD，RDF スキーマ	LD-Linked Data（データ間の融合情報が規定。検索可）

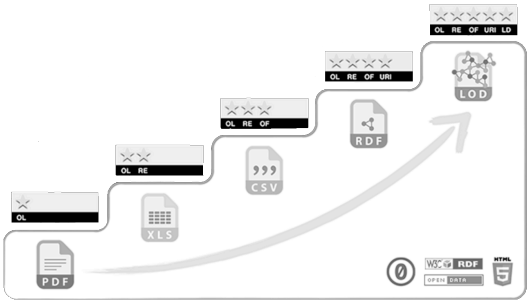

図 1　5★オープンデータ

[1] パブリックドメインとは，著作物など知的創作物において，知的財産権が発生していない，もしくは消滅した状態をいう。CreativeCommonsJapan,
https://creativecommons.jp/licenses/

[2] テキスト入力の際，元となる本のこと。底本を選ぶ,
http://www.aozora.gr.jp/aozora-manual/#3

[3] EPUB3 とは，電子書籍のファイルフォーマット規格の一つである。EPUB2 から EPUB3 へアップデートされたことで，日本語組版などで使われる縦組みやアラビア語などで見られる右から左へ文字が書かれる言語に対応した。日本電子出版協会,
http://www.jepa.or.jp/ebookpedia/201512_2781/

[4] 総務省政府全体の取組,
http://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/opendata02.html/

[5] 5★オープンデータ，http://5stardata.info/ja/

[6] Creative commons JAPAN,
https://creativecommons.jp/licenses/

※ URL は全て 2016 年 6 月時点の閲覧。

表２　機械判読可能なデータ形式
「オープンデータの５つの段階」別の利用可能なデータ形式[8]から一部抜粋

区分	主な データ形式	特徴	1段階 (OL)	2段階 (RE)	3段階 (OF)	4段階 (URI)	5段階 (LD)
文字	.TXT	オープンライセンス			○		
複合 文書	.XHTML	オープンライセンス(Web 標準)			○		
	.XML	オープンライセンス(Web 標準)			○		
	.PDF	現在は，仕様が公開		○			
	.epub	オープンライセンス(Web 標準)		○			

文学作品におけるオープンデータ化の例として，青空文庫[7]が挙げられる。青空文庫とは，利用に対価を求めることのないインターネット上にある電子図書館である。1997 年に呼びかけ人の代表である冨田倫生を含めた 4 人によって発足した。現在，1 万点を超える文学作品が集まっており，特に明治期から昭和初期にかけての作品が多くを占める。それらデータの入力や校正などの作業は，青空工作員と呼ばれているボランティアの人たちの地道な努力によって支えられている。

青空文庫では，著作権の消滅した作品と著作権は消滅していないが書き手が公開してもよいとした作品を TXT(テキストファイル)と XHTML（一部 HTML）形式で公開している。高度情報通信ネットワーク社会推進戦略本部の電子行政オープンデータに関する決定等[8]によると，TXT と XHTML は 5 つ星スキームのうち 3 段階目に該当する（表 2)。つまり 2 段階に加え，オープンに利用できるフォーマットでデータの公開がされ，特定のアプリケーションに依存しない形式である。機械判読可能なデータ形式として利用価値のある公開レベルであると考えられる。

3.　関連事例と関連研究

3.1　関連事例

まず教育の場における事例について述べる。本授業のみならず他大学においても関連した内容の授業などがおこなわれている。例えば，愛知大学にて文学部時実ゼミ青空文庫班[9]として著作権の消滅した文学作品をデジタルデータ化し青空文庫に登録する授業が行われていた。京都大学においてはサークル活動にて京都大学電子テクスト研究会として電子化したデータを青空文庫に登録する活動が行われている。

次にデータの活用事例について述べる。青空文庫を手軽にスマートフォンやタブレットなどモバイル端末で読むためのアプリケーションが App Store や Google Play からリリースされている。いくつか例を挙げる。i 読書-青空文庫リーダー[10]は，青空文庫を手軽に楽しむためのアプリケーションである。他にも Android 版の青空文庫ビューア Ad[11]や類似のアプリケーションとして青空読手，読み上げ機能に特化した音声文庫などがある。

また関連サイトとして星空文庫[12]が挙げられる。星空文庫とは，小説を中心とした文芸作品の掲載，閲覧サービスである。プロ・アマ問わず，執筆した小説に CC ライセンスを付け登録することができ，小説の他に随筆・エッセイ，韻文詩，自由詩が登録されている。投稿された小説は PDF の他，TXT,EPUB といったオープンデータとしても活用可能なデータ形式で自由にダウンロードし，CC ライセンスに従って利用することができる。専用のアプリケーションも用意されており，モバイル端末を使って，電子書籍として楽しむことができる。

3.2　関連研究

研究分野において，文学作品とオープンデータに関連した研究が進められている。情報技術開発の分野では，青空文庫に登録されているデータを使い，近代文学作品に見られる難解な語句に自動で注釈を付けるシステム開発の研究（速水，2014）(2)，日本語学習者を対象とし，文学作品に注釈や画像，例文などを表示することで読解支援を行うサイトの開発（久米，2014）(3)，電子書籍の読書中に気になった場所に線を引いたり，メモ書きをするためのモバイルシステム開発（中島，2013）(4)，同義語・類義語抽出ツールの同義語抽出制度やクエリに対する応答性を青空文庫を活用して性能評価を行った研究（吉田，2009）(5)がある。言語推移に関する分野では，「キリスト・キリシタン」の意味と表記の変遷について国語辞書と青空文庫をもとに考察した研究（李，2009）(6)，「イタダク」の意味推移について青空文庫のデータをもとに考察した研究（秋山，2009）(7)がある。

---

[7] 青空文庫，http://www.aozora.gr.jp/
※ URL は全て 2016 年 6 月時点の閲覧。

[8] データ形式・構造，データカタログに関する技術について，http://www.kantei.go.jp/jp/singi/it2/densi/wg/dai1/siryou7.pdf/

[9] 愛知大学文学部時実ゼミ青空文庫班における活動の報告は第 14 回図書館総合展のポスターセッション（パシフィコ横浜展示ホール D／アネックスホール他，2012.11.20-22）にて発表されている。
http://2012.libraryfair.jp/taxonomy/term/1652/

[10] i 読書-青空文庫リーダー，
https://itunes.apple.com/jp/app/i-du-shu-qing-kong-wen-kurida/id534970999?mt=8/

[11] 青空文庫ビューア Ad，
https://play.google.com/store/apps/details?id=jp.dip.sys1.aozora&hl=ja/

[12] 星空文庫，http://slib.net/
※ URL は全て 2016 年 6 月時点の閲覧。

また，文学作品や記録資料データのデジタル・アーカイブに関する研究や人文学分野におけるオープンデータ化に関する研究も行われている。デジタル・アーカイブの構築と情報共有に関する研究(藤村, 2016)[8]，デジタル時代の人文学のあり方を追究とテキストデータベースのオープンデータ化についての考察（永崎, 2015）[9]，人文学分野におけるオープンデータの普及に関する研究（橋本，2015）[10]が例として挙げられる。

4. 授業の目的と底本の選定

本授業の目的は，受講している学生たちが著作権の消滅した書籍を底本に電子書籍の制作を体験することで，電子書籍がどのようにできているのかその仕組みについて学ぶことである。電子化したテキストデータは青空文庫に登録し，社会貢献につながっていくことを意識している。併せて，文学作品のオープンデータ化やそれらテキストデータの活用（2 次利用）についてもオジリナルの電子書籍を作り体験することで，理解を深めることができるよう工夫している。

今年度は，2016 年元旦に著作権が消滅し，パブリックドメインになったばかりの江戸川乱歩の作品を選んだ。本授業は日本語日本文化学科の 3 年生から 4 年生を対象に開講している。学生はそれぞれゼミに所属し，江戸時代以前の古典文学から近現代までそれぞれ研究テーマとしている時代は異なるが，ファンが多く人気のある作家であること，著作権が消滅したばかりでオープンデータ化に取り組んでいるケースがほとんどなかったこと，学科内に江戸川乱歩を専門とし研究を進めている教員がいることや，江戸川乱歩が愛知県の出身であることなどの理由から興味をもって授業に取り組めるのではないかと思い，底本の作家として選んだ。

5. 授業の環境とその手順

本授業では，MacPC（OS，X）と複数のフリーソフトウェアを使用している。特にソフトウェアに関しては，無料で配布されているものを使うことで，学生が授業内容に興味を持ち授業終了後も個人的に続けることを希望した際，すぐに自宅でも同じ環境を作ることができるよう考慮している。使用したソフトウェアの詳細については以下の通りである。

・ mi（2.1.12r5）

mi（旧名：ミミカキエディット）[13]はMac 用のテキストエディタである。特徴の一つとして自由度の高いカスタマイズが挙げられる。他のユーザによって開発，配布されているモード機能を必要に応じて追加することで，個々のユーザが使いやすいようカスタマイズすることができる。

― 電書協 EPUB 用 XHTML モード

電書協 EPUB 用 XHTML モード[14]は，mi のモード機能に追加し使用する。電書協 EPUB3 制作ガイド[15]で規定されている全てのプロパティをメニューツールのドロップダウンリストから選択するだけで，簡単にタグ付けをすることができる。プログラミング言語に不慣れな学生のタグ付け作業に対するハードルを下げることを目的とし使用している（図 2）。

図 2　電書協 EPUB 用 XHTML モードのツール

― 青空文庫モード

青空文庫モード[16]は，mi のモード機能に追加し使用する。青空文庫の注記一覧で規定されているもののうち，頻度の高いプロパティを中心に採用され，メニューのドロップダウンリストから選択することで，前述の電書協 EPUB 用 XHTML モードと同じく，タグ付けの作業が容易になる（図 3）。

図 3　青空文庫モードのツール

[13] mi, http://www.mimikaki.net/

[14] 電書協 EPUB 用 XHTML モード, http://densyodamasii.com/?p=1973/

[15] 電書協 EPUB3 制作ガイド, http://fairfield.minibird.jp/other_resources/mi-%E7%94%A8-%E9%9D%92%E7%A9%BA%E6%96%87%E5%BA%AB%E3%83%A2%E3%83%BC%E3%83%89/

[16] 青空文庫モード, http://fairfield.minibird.jp/other_resources/mi-%E7%94%A8-%E9%9D%92%E7%A9%BA%E6%96%87%E5%BA%AB%E3%83%A2%E3%83%BC%E3%83%89/

※ URL は全て 2016 年 6 月時点の閲覧。

・ ePubZip/Unzip (2.0.1)

ePubZip/Unzip は，圧縮ソフトウェアの中で epub に特化したアプリケーションである。電書協 EPUB3 制作ガイドのテンプレートを元に，原稿や画像を入れたものを，この ePubZip/Unzip で圧縮することで，「.epub」に生成することができる。

・ Readium

Google Chrome のアプリケーションである Readium[17] は，ePubZip/Unzip で生成した「.epub」のファイルを読むために使用する。ローカルファイルのデータを読み込む他，パッケージ化されていないフォルダを読み込む事も出来るため，制作途中の検証用としても使うことができる。

作業手順については以下のとおりである。

1）底本の原稿を分担
2）Google drive の機能 Google ドキュメントを使い，原稿を打ち込み。もしくは原稿をスキャンした後，OCR にかけたデータを原稿と読み比べ間違いを直す。（学生は自分にあった方を選択することができる。）
3）mi の電書協 EPUB 用 XHTML モードを使い，2）で打ち込んだ原稿にルビや改行などのタグを付ける。
4）電書協 EPUB3 制作ガイドのテンプレートに 3）で作成したテキストデータを挿入し，表紙部分に好きなイラストを入れるなどの形成作業。
5）ePubZip/Unzip で圧縮し，「.epub」に書き出す。
6）EPUB３のオリジナル電子書籍の完成。
7）Readium を使い，ランダムで選んだ別の学生による原稿の第１校正を行う。
8）第１校正とは別の学生をランダムに選び，第２校正を行う。
9）mi の青空文庫モードを使って，2）で打ち込んだ原稿にルビや改行などのタグ付けを行う。ここでは，電書協 EPUB 用 XHTML モードとの違いを学ぶ。

6. 考察とまとめ

今年で２年目となる「電子書籍制作」の取り組みは，現在３分の２が終了し，残すところあと５回となった。今年度はオープンデータ化を意識し，作業手順を見直すとともに，去年行った「電子書籍制作」の手順に青空文庫記法を新しく加えた。作業の手順については，普段パソコンを使う機会が少なく，ソフトウェア環境に慣れていない学生もタグ付けなど特に困ることなく順調に作業が進んでいる。また，デジタル化したデータが電子書籍という目に見える形になることで，オープンデータ化された文学作品のデータの活用についてのイメージがしやすく，学生にとって理解を深める助けになっているようである。

しかし，その他の点において再考が必要な課題が出てきている。一つ目は，底本の原稿の打ち込み作業である。昨年選んだ『村岡花子童話集たんぽぽの目』は，童話でひらがなが多く打ちやすかったため，2 時間から 3 時間程度で履修している全ての学生が打ち終えることができた。しかし，今年度の江戸川乱歩の作品は漢字が多く，中には普段使わないような漢字も含まれている。非常に時間がかかり 5 時間から 7 時間を費やした。

二つ目は，底本の選定である。本来は学生が自分で底本を選ぶところから始めたいと思っているが，青空文庫への入力申請の手続きの関係で今年も筆者が予め選び用意した。江戸川乱歩は推理小説を得意としており，作風が独特で個人の好みが分かれるところでもあるため，一部の学生からは不評であった。

最終的な成果とまとめについては，ポスターセッションにおいて述べるが，今後の授業では，実際にオープンデータ化され，公開されている文学作品のデータを活用した取り組みも取り入れ，更に深みあるものへとしていきたい。

謝辞

本研究の一部は，JSPS 科研費 25280131，15K00448，15K16097 の助成を受けたものです。

参考文献

(1) 兼松篤子，遠藤守，安田孝美：“EPUB3 による電子書籍制作の取り組み”，コンピュータ利用教育学会，2015PC Conference，pcc060，(2015)
(2) 速水秀平，井上潮：“注釈の自動生成による青空文庫の読書支援”，情報処理学会，第 76 回全国大会講演論文集，IPSJ-Z76-5P-8，(2006).
(3) 久米朋子，江見圭司：“日本語学習者を対象とした日本文学作品の読解支援さいと『JL 文庫』の作成～『インターネット図書館青空文庫』を題材として”，情報処理学会，研究報告コンピュータと教育（CE），IPSJ-CE14123009，(2014)
(4) 中島一，白石修二：“Markable book system の開発”情報処理学会，第 75 回全国大会講演論文集，IPSJ-Z75-6D-2，(2013)
(5) 吉田和弘，吉田稔，中川裕志：“文字列検索に基づく同義語・類義語抽出ツールとその性能評価”，情報処理学会，研究報告音声言語情報処理（SLP），IPSJ-SLP09076019，(2009)
(6) 李明心：“「キリスト・キリシタン」の意味と表記の変遷：国語辞書と青空文庫を中心に”，明海大学紀要明海日本語 14，77-83，(2009)
(7) 秋山智美：“「イタダク」の意味推移-文学作品における用例から”，東京交通短期大学研究紀要（15），129-138，(2009)
(8) 藤村涼子：“アーカイブズ情報共有のあり方を考える-機関リポジトリによるデジタル・アーカイブ構築の実践を通して”，国文学研究資料館紀要 12，57-73，(2016)
(9) 橋本雄太：“人文学資料オープンデータの可能性と現状（<特集>オープンデータ）”，情報の科学と技術 65（12），525-530，(2015)
(10) 永崎研宣：“SAT 大蔵経テキストデータベース人文学におけるオープンデータの活用に向けて”，情報管理 58（6），422-437，(2015)

[17] Readium,
https://chrome.google.com/webstore/detail/readium/fepbnnnkkadjhjahcafoaglimekefifl?hl=ja/
※ URL は全て 2016 年 6 月時点の閲覧。