CA1894

ドイツ国立図書館（DNB）における オンライン資料を対象にした自動分類

鴇田拓哉（ときたたくや）*

はじめに

ドイツ国立図書館（DNB）では、オンライン資料、具体的にはインターネット上で入手可能な電子書籍や学位論文などの刊行物を対象に、機械的な分類記号の付与（自動分類）が行われている(1)。

本稿で取り上げるDNBの自動分類に比較的近いと思われる、出版物を対象に自動分類を行う事例に、米国議会図書館（LC）で行われているAutoDeweyがあげられる(2)。AutoDeweyとは、小説・戯曲・物語を対象とした目録データ作成の過程でジャンル（小説・戯曲･物語）と時代（例えば1745-1799）を選択すると、既に付与されている「米国議会図書館分類表（LCC）」の分類記号を手がかりにして「デューイ十進分類法（DDC）」の分類記号が自動的に作成されるものである(3)。AutoDeweyの対象である小説・戯曲・物語においては、LCで使用しているLCCとDDCの分類記号の対応づけが、完全一致とはいかないものの、ある程度確実に行えたことがこのような自動的な分類記号付与の実現につながっている(4)。本稿で扱うDNBの自動分類は、類例が少ない点、オンライン資料を対象として行われている点、さらに、後で述べるように目録作業の機械化を積極的に推し進めている点で珍しい事例といえよう。

本稿では、はじめにDNBが自動分類を導入した背景を述べる。続いて、自動分類によって付与される分類記号について具体例をあげて紹介する。そして、分類記号が自動的に付与されるしくみについて述べる。さらに、自動分類によって付与された分類記号の品質に関する調査結果を紹介する。

1．自動分類を導入した背景

DNBが自動分類を導入した背景に、DNBを取り巻く状況をあげることができる。ドイツではドイツ国立図書館法の施行により、2006年からオンライン資料もDNBへの納本対象となった（CA1613参照）。2012年から2015年における印刷資料とオンライン資料の年間受入点数の推移を表1に示す(5)。この4年の間に、印刷資料が年々減少しているのに対し、オンライン資料は年々増加し、2015年には印刷資料の点数を超えていることがわかる。

表1　印刷資料とオンライン資料の年間受入点数
（2012－2015年）

	2012年	2013年	2014年	2015年
印刷資料（点）	61万600	57万8,950	54万6,980	52万5,000
オンライン資料（点）	24万7,660	36万8,510	46万2,290	61万2,000

加えて、各年の印刷資料とオンライン資料を合わせた資料全体の数、つまり目録作業の対象となる資料数が増加し、特にその中に占めるオンライン資料の割合が高くなっている。目録作業を担当する職員の負担の軽減や作業効率の向上を図るために、主題目録作業を人手による作業から機械による作業へと移行することが検討された(6)。

2010年に、オンライン資料については、人手による主題目録作業の停止を決め、機械による主題目録作業に置き換えていくこととなった(7)。以降で説明する自動分類を含む機械による主題目録作業は、2009年から2011年に実施された機械による目録作業の基盤を構築することを目的としたプロジェクトPETRUS（Process-supporting software for the digital German National Library）の成果が土台となっている(8)。

2．付与されている分類記号

自動分類のしくみを述べる前に、自動分類によって付与された分類記号が実際の目録データにどのような形で表れているのかを紹介しておく。DNBでは、電子書籍や学位論文を対象に（1）DDCの3桁の区分を参考にしてDNBが独自に作成した「主題カテゴリ（Subject Categories）」、（2）医学分野の論文については、DDCによる分類記号の桁数を短縮した「DDC短縮記号（DDC Short Numbers）」を自動分類によって付与している。それぞれの概要を表2に示す(9)。

表2　自動分類により付与されている分類記号

区分	主題カテゴリ	DDC短縮記号
付与の対象	オンライン資料（小説を除くすべての分野）	オンライン資料（医学）
付与の開始年	2012年	2015年
カテゴリの数	102	138

主題カテゴリは3桁からなり、「560　古生物学」をはじめ、DDCとほぼ類似した内容となっている。DDC短縮記号は、膨大な数の医学分野の論文に対し、DDCによる分類記号の桁数を短縮した形で表したものである。例えば、本来の分類記号が「618.9298…」となるものを「618.92」と桁数を短縮させている。

主題カテゴリとDDC短縮記号が付与された書誌データの事例を図1に示す(10)。図の下方にある“Sachgruppe(n)”が主題カテゴリ、“DDC-Notation”

*共立女子大学

がDDC短縮記号を表している。その記号の後ろに書かれている“maschinell ermittelte Kurznotation”は、「機械によって付与された」ことを意味している。

Link zu diesem Datensatz	http://d-nb.info/110052729X
Titel	A Clinician's Guide to Systemic Effects of Periodontal Diseases
Person(en)	Craig, Ronald G. (Herausgeber) Kramer, Angela R. (Herausgeber)
Ausgabe	1st ed. 2016
Verlag	Berlin, Heidelberg : Springer Berlin Heidelberg
Zeitliche Einordnung	Erscheinungsdatum: 2016
Umfang/Format	Online-Ressource (pdf)
Andere Ausgabe(n)	Erscheint auch als Druck-Ausgabe: A clinician's guide to systemic effects of periodontal diseases
Persistent Identifier	URN: urn:nbn:de:1111-201605188808 DOI: 10.1007/978-3-662-49699-2
URL	http://www.springerlink.com/content/978-3-662-49699-2 (Verlag)
ISBN/Einband/Preis	978-3-662-49699-2
EAN	9783662496992
Anmerkungen	Lizenzpflichtig Langzeitarchivierung gewährleistet
DDC-Notation	617.63 [maschinell ermittelte Kurznotation]
Sachgruppe(n)	610 Medizin, Gesundheit

図１　書誌データの具体例

DDC短縮記号が機械によって自動的に付与されたことが書誌詳細画面上で確認できるのに対し、主題カテゴリが自動的に付与されたものかどうかは、書誌データをMARCXML形式で表示させることで確認できる。

図１の書誌データをMARCXML形式で表示をした内容（一部）を図２に示す[11]。この図の“datafield tag="883"”は、このフィールドが機械により生成されたメタデータの由来に関する情報を記録するフィールド883であることを示している。“subfield code="a"”は、フィールド883のサブフィールドaであることを示す。このサブフィールドaでは、当該メタデータがどのような処理方法によって付与されたかについて記録される。例えば、冒頭で紹介したAutoDeweyもこのサブフィールドの値となりうる。図2での値“maschinell gebildet”は「機械による付与」を意味している[12]。

```
<datafield tag="883" ind1="0" ind2=" ">
  <subfield code="8">1¥p</subfield>
  <subfield code="a">maschinell gebildet</subfield>
  <subfield code="c">0,999</subfield>
  <subfield code="d">20160519</subfield>
  <subfield code="q">DE-101</subfield>
</datafield>
```

図２　書誌データの具体例（MARCXML形式）（一部）

３．自動分類のしくみ

DNBの自動分類[13]では、代表的な機械学習のアルゴリズムであるサポートベクターマシーン（SVM）が採用されている。主題カテゴリの自動分類の対象となるオンライン資料は、ドイツ語か英語で書かれたもので、ファイル形式がPDF（2012年から対象）または電子書籍の規格EPUB（2015年から対象）のもの（小説を除く）である。2012年の開始から2016年３月までの間に44万4,586点の資料に対して自動分類が行われた。

DDC短縮記号の対象となるオンライン資料は、ドイツ語か英語で書かれたもので、ファイル形式がPDFまたはEPUB、かつ主題カテゴリが「610　医学」のものである。2015年10月の開始から2016年３月までの間に8,121点の資料に対して自動分類が行われた。

使用するソフトウェアは、Averbis社（ドイツ）の製品Averbis Extraction Platformをカスタマイズしている。この製品が選ばれたのは、当初からDNBで独自に開発せずに外部のソフトウェアを採用する意図で市場調査を行った結果であること、Averbis社がフライブルク医科大学から派生した会社であり、医学分野の文献に対して目録データを作成する技術に長けていたことがあげられる。

自動分類は、（1）実際に自動分類を行うための準備段階である学習フェーズと、（2）実際に運用していく段階の運用フェーズの二つに分かれる。

（1）学習フェーズ

学習フェーズでは、学習データを選定し、各種パラメータを設定してから、対象データを解析・学習させる（いわゆる「教師あり学習」を行う）。選定される学習データは、ドイツ語または英語で書かれている、オンライン資料のメタデータ・本文データと、紙媒体の資料の目次を電子化した目次データで、件数は45万1,333件（2016年４月時点）となっている。

事前に設定されるパラメータには、対象言語、解析対象とするテキストの長さ、メタデータや本文データの項目間の重みづけなどがある。これらのパラメータは定期的に見直される。なかには、本文の対象を開始から４万文字までとする、本文よりもメタデータのタイトル項目の内容をより重要とみなすなど、実際に運用された経験を踏まえて、原則として見直されず固定化されているパラメータもある[14]。

（2）運用フェーズ

日々行われる自動分類の処理は、収集されたオンライン資料のIDリストを手がかりに行われる。リストの各IDに紐付けられるメタデータと本文データが、書誌データベースとコンテンツデータベースから抽出され、圧縮データの解凍、ファイル形式の変換、正規化処理などが行われ、自動分類を行うシステムへ転送される。自動分類が終了すると、その結果が書誌データベースの該当書誌レコードに上書きされる構図となっている（図３参照）。

4．自動分類によって付与された分類記号の品質管理

自動的に付与された分類記号の品質を保つために、全件は無理ではあるものの、何件かについて担当者による内容確認を行っている。既に分類記号が付与されている印刷資料が電子化されたのであれば、自動分類の結果と照合できる。間違いを見つけた場合にはその都度自動分類の結果を修正する。自動分類の最終的な判断を人間が行っている点は、人手による作業と機械による作業の位置づけおよび両作業の役割分担の在り方という観点から考えると非常に興味深い。

ここで、自動分類によって付与された分類記号の品質を検討する材料として、自動的に付与された主題カテゴリとDDC短縮記号に関するDNBによる調査結果をあげておく。2012年から2015年までの間に自動的に付与された主題カテゴリ41万3,363件に対し、人手による確認を行った７万3,509件（全体の18％）の正答率は75％であった。DDC短縮記号は、2015年10月から12月までの間に自動的に付与された4,072件に対し、人手による確認を行った574件（全体の14％）の正答率は74％であった。

5．おわりに

本稿では、DNBで行われている自動分類について述べた。自動分類を導入した背景にDNBを取り巻く状況をあげたが、これらの状況はDNBだけでなく、日本を含む各国の国立図書館においてもあてはまるように思われる。

この事例を考えるにあたっては、先ほどあげた自動分類の正答率75％と74％をどのように捉えるのかがポイントとなろう。DNBにおける自動分類はまだ始まったばかりといえる。今後の動きや成果を見つつ、図書館における自動分類というテーマについて長期的な視点で見ていくことが必要であると思われる。

(1) 後でも述べるように、DNBでは、本稿で扱う分類記号の付与だけでなく、件名の付与も含めた主題に関する記録に伴う作業（主題目録作業）が機械化されている。本稿では、機械的に分類記号を付与する作業を「自動分類」として説明を進める。

(2) “AutoDewey”. Library of Congress. https://www.loc.gov/aba/dewey/practices/autodewey.html, (accessed 2017-01-13).

(3) Beall, Julianne; Saccucci, Caroline. “AutoDewey”. https://www.loc.gov/aba/dewey/practices/autodewey-presentation.pdf, (accessed 2017-01-13).

(4) LCCとDDCは記号の構成が異なっているため、AutoDeweyで対象としている小説・戯曲・物語においても完全に一致した記号の対応づけとまではいかない。しかし、刊行された年代という観点で両者の対応づけを考えると、AutoDeweyで実現している程度の機械的な変換は可能となる。例えば、LCCのPR6051からPR6076は1961年から2000年の間に特定の著者によって書かれた英語の文学作品に対する記号である。6051から6076は、6051が「A」、6052が「B」、6076が「Z」から始まる著者名の頭文字で分かれている（さらに「A」の中で細分される）。一方、DDCでは、英語で書かれた文学作品は「822　戯曲」、「823 小説」が割り当てられており、さらに刊行された年代に応じて細分化されている。例えば、1945年から1999年に刊行されたのであれば「822」や「823」のうしろに「914」をつけて「822./914」「823./914」のように表す。このことから、LCCのPR6051からPR6076は、ジャンルと刊行された年代からDDCの記号に機械的に変換が可能となる。以上の説明は、下にあげる文献を参考にした。
Beall, Julianne; Saccucci, Caroline. “AutoDewey”. https://www.loc.gov/aba/dewey/practices/autodewey-presentation.pdf, (accessed 2017-01-13).

(5) Busse, Frank. “Machine-based issuing of DNB Subject Categories and DDC Short Numbers for Medicine in the German National Library”. http://edug.pansoft.de/tiki-download_file.php?fileId=140, (accessed 2017-01-13).
このほかに、下にあげる文献もDNBにおける自動分類について述べている。
Schöning-Walther, Christa; Mödden, Elisabeth; Uhlmann, Sandro. “Germany（DNB）Automatic classification and indexing”. IFLA Classification & Indexing Section Newsletter. 2013, (47), p. 10. http://www.ifla.org/files/assets/classification-and-

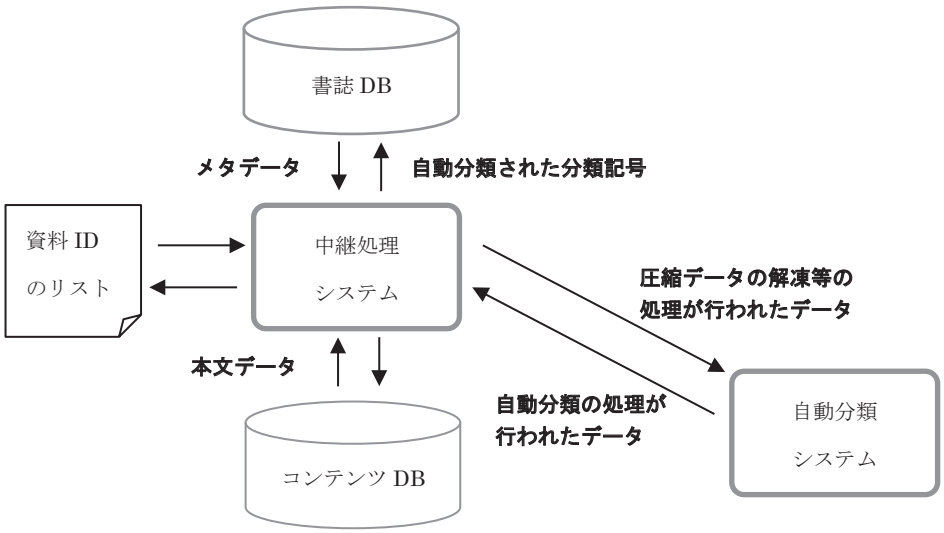

図３　運用フェーズの行程

indexing/newsletters/newsletter_june_13.pdf#page=10,（accessed 2017-01-13）.

(6) 検討自体は以前から行われていたと思われるが、実際に動き出したのは、本文の次の段落で述べているPETRUSが最初であると思われる。

(7) Busse, Frank. “Machine-based issuing of DNB Subject Categories and DDC Short Numbers for Medicine in the German National Library”. http://edug.pansoft.de/tiki-download_file.php?fileId=140,（accessed 2017-01-13）.

(8) “PETRUS - Process-supporting software for the digital German National Library”. Deutsche Nationalbibliothek. http://www.dnb.de/EN/Wir/Projekte/Archiv/petrus.html,（accessed 2017-01-13）.

(9) 表内の主題カテゴリの付与開始年である2012年は、主題カテゴリが自動的に付与されるようになった年を示している。主題カテゴリは自動的に付与される前（2004年）から付与されていた。DDC短縮記号の付与開始年の2015年も、自動的に付与されるようになった年を示している。DDC短縮記号が開発されたのは2006年から2007年にかけてである。DDC短縮記号の「カテゴリの数」の「138」は次の文献を参照した。
“DDC-Notationen für medizinische Dissertationen Untergliederung der DDC-Hauptklasse 610”. Deutsche Nationalbibliothek. http://www.dnb.de/Subsites/ddcdeutsch/SharedDocs/Downloads/DE/anwendung/ddcGliederungMedizin.pdf?__blob=publicationFile,（accessed 2017-02-15）.

(10) “Ergebnis der Suche nach: idn=110052729X”. Deutsche Nationalbibliothek. http://d-nb.info/110052729X,（accessed 2017-01-13）.

(11) “MARC21-XML-Repräsentation dieses Datensatzes”. Deutsche Nationalbibliothek. http://d-nb.info/110052729X/about/marcxml,（accessed 2017-01-13）.

(12) 主題カテゴリの値は機械によって付与されたことがMARCXML形式で表示させたときに確認できる（別の言い方をすれば、通常の書誌詳細画面上では確認できない）。これに対し、DDC短縮記号はMARCXML形式で表示させても確認できず、通常の書誌詳細画面上でのみ確認できる。以上の内容と本文の記述は、国立国会図書館の塩崎亮氏がDNBのFrank Busse氏から聞き取った内容を整理したものである。

(13) 本稿の3章および4章の説明は、下にあげる文献をもとにしている。
Busse, Frank. “Machine-based issuing of DNB Subject Categories and DDC Short Numbers for Medicine in the German National Library”. http://edug.pansoft.de/tiki-download_file.php?fileId=140,（accessed 2017-01-13）.

(14) この段落における、パラメータの見直しや固定化されているパラメータなどについての説明は、国立国会図書館の塩崎亮氏がDNBのFrank Busse氏から聞き取った内容をもとにしている。

［受理：2017-02-16］

Tokita Takuya

Automatic Classification of Online Resources in the German National Library