機関リポジトリ収録コンテンツにおける利用数とアクセス元、アクセス方法、コンテンツ属性の関係

佐藤翔　（筑波大学大学院図書館情報メディア研究科）min2fly@slis.tsukuba.ac.jp
逸村裕　（筑波大学大学院図書館情報メディア研究科）hits@slis.tsukuba.ac.jp

1. 研究背景と目的

大学・研究機関による機関リポジトリの設置とコンテンツの整備が進むにつれ、リポジトリに収録されたコンテンツの利用状況に注目が集まっている。中でも「誰がコンテンツを使っているのか」（アクセス元）、「どこから機関リポジトリにアクセスしているのか」（アクセス方法）、「どのようなコンテンツが利用を集めるのか」（コンテンツ属性）の３点はリポジトリ運用の参考になるとともに機関リポジトリが果たしている役割を考える上でも重要である。

このうちアクセス方法について Organ は外部サイトからのアクセスの 96%が Google からであったとしており [1]、また利用の多いコンテンツについて Royster は Nebraska-Lincoln 大学リポジトリのアクセス上位論文の多くがリポジトリ以外で公開されていないコンテンツであったとしている [2]。しかしこれらの調査はリポジトリソフトウェア上の単純な統計やランキングに基づいており、詳細な分析は行っていない。利用の詳細を見た研究としては Bonilla-Carelo による Strathclyde 大学リポジトリの利用分析があり、利用数とアクセス元の国の数の間に相関があること等が示されているが [3]、分析対象は物理学分野の文献に限られておりコンテンツ属性の分析は十分には行われていない。

そこで本研究では機関リポジトリ収録コンテンツの利用数とアクセス元（ユーザドメイン）、アクセス方法、コンテンツの属性（文献タイプ、記述言語、出版年等）の関係を明らかにすることを目的に、４つの機関リポジトリのアクセスログ分析を行った。また、分析結果から合わせて利用数に影響するその他の要因についても明らかにすることを試みる。

2. 調査方法

分析対象はアジア経済研究所（ARRIDE）、北海道大学（HUSCAP）、京都大学（KURENAI）、筑波大学（Tulips-R）の４つの機関リポジトリに収録されたコンテンツ本文に対する、2008 年１年間のアクセスログである（それぞれの収録コンテンツ数等については表 1 参照）。これらのログについて検索ロボットや同一人物による連続アクセス等のノイズを排除するフィルタリングを行った上で分析し、各コンテンツのアクセス元ごと（国内／海外等）、アクセス方法ごと（サーチエンジン／リポジトリ内の別ページ等）および全体でのダウンロード数を集計した。フィルタリングの方法については佐藤義則が提案する電子レコードの統計標準 COUNTER に基づいた方法を採用した [4]。また、コンテンツの属性については各リポジトリ担当者からメタデータ（文献タイプ、記述言語、出版年）の提供を受けた。

表1. 2008 年末時点の各リポジトリ収録コンテンツ数

	ARRIDE	HUSCAP	KURENAI	Tulips-R
総コンテンツ数	640	25,542	28,356	7,899
日本語	398	15,086	16,394	6,922
英語	242	9,871	10,612	923
その他	0	585	1,350	1
雑誌論文	242	2,719	1,413	929
学位論文	0	343	300	6,687
紀要論文	0	22,026	20,750	6
会議予稿	0	132	174	48
発表資料	0	128	52	5
図書	8	13	51	26
技術報告	178	0	1	0
研究報告	1	5	190	133
一般記事	7	65	2,728	2
プレプリント	0	0	0	42
教材	0	24	4	7
データ等	0	0	6	0
ソフトウェア	0	0	2	0
その他	204	87	2,685	0

＊ 総コンテンツ数と各項目の合計の齟齬は、項目データの欠損によるものである。

3. 分析結果

3. 1 リポジトリ全体

表２は各リポジトリへアクセスしてきたユーザのアクセス元の割合を見たものである。

表2. アクセス元の内訳(リポジトリ全体)

	ARRIDE	HUSCAP	KURENAI	Tulips-R
国内(jpドメイン)	31.3%	60.1%	64.1%	78.0%
海外(非jpドメイン)	68.7%	39.9%	35.9%	22.0%
民間・個人(ne, net)	40.8%	43.5%	50.2%	44.2%
民間・団体(or, org)	3.2%	4.8%	5.2%	5.3%
大学等(ac, edu)	17.8%	20.3%	16.1%	24.0%
企業(co, com)	18.1%	12.5%	12.2%	11.7%
政府(go, gov)	1.1%	1.7%	1.2%	2.2%
その他	19.1%	17.2%	15.2%	12.7%

国内・海外別ではARRIDEのみ海外からのアクセスが多いが、他では国内が６～８割を占める。また、所属機関ではいずれのリポジトリでも民間・個人（自宅等）が４～５割を占め、次いで大学または企業からのアクセスが多い。

また、表３は各リポジトリへのアクセス方法（本文にアクセスする前に見ていたwebページ）の内訳を示したものである。

表3. アクセス方法の内訳(リポジトリ全体)

	ARRIDE	HUSCAP	KURENAI	Tulips-R
直接アクセス	17.8%	15.5%	15.8%	14.4%
リポジトリ内部	17.6%	33.0%	22.7%	4.2%
サーチエンジン	17.8%	48.7%	55.8%	79.8%
その他	46.7%	2.8%	5.8%	1.7%

ARRIDEでは「その他」からのアクセスが最も多く、ほとんどが経済学分野の分野別リポジトリRePEcからのアクセスである。他ではサーチエンジンからのアクセスが最も多く全体の半数前後を占めるが、そのほとんどがGoogleの検索結果からのアクセスである。

３．２ 文献タイプ

表４は文献タイプごとの平均ダウンロード数を示したものである。

表4. 文献タイプごとの平均ダウンロード数

	ARRIDE	HUSCAP	KURENAI	Tulips-R	合計
雑誌論文	18.4	45.2	26.8	25.9	35.7
学位論文	-	31.9	90.4	15.4	19.3
紀要論文	-	9.9	21.0	19.5	15.3
会議予稿	-	39.5	54.1	83.5	52.7
発表資料	-	41.5	29.7	20.4	37.6
図書	63.1	77.0	244.1	4.1	143.5
技術報告	60.9	-	20.0	-	60.7
研究報告	8.0	22.4	74.5	7.3	46.4
一般記事	10.0	32.5	11.8	20.5	12.2
プレプリント	-	-	-	2.4	2.4
教材	-	1304.4	41.8	286.7	956.5
データ等	-	-	36.2	-	36.2
ソフトウェア	-	-	35.5	-	35.5
その他	13.0	18.5	10.8	-	11.2
合計	28.9	15.6	21.2	17.1	18.4

全体で見ると最もよく利用されるのは教材、次いで図書、最も利用が少ないのはプレプリントとなっている。ただしこれらのコンテンツは表１に示したように登録数自体が多くはない。全体で5,000件以上の登録がある３つのタイプ（雑誌論文、学位論文、紀要論文）に絞って見ると、最も利用が多いのは雑誌論文である。学位論文はKURENAIでは平均90回以上使われるコンテンツであるがTulips-Rで利用が伸びず、紀要はHUSCAPで最も利用の少ないコンテンツになっている等、リポジトリによっても差がある。

３．３ 記述言語

表５は記述言語（英語・日本語）ごとの平均ダウンロード数を示したものである（その他の言語によるコンテンツ数は限られるため分析から除いた。国内からのアクセスと海外からのアクセスの合計が全体の値に一致しないのは、アクセス元ドメインが判別できないケースが全体の集計には含まれているためである）。

表5. 言語ごとの平均ダウンロード数(全体・国内／海外別)

	全体		国内から		海外から	
	日本語	英語	日本語	英語	日本語	英語
ARRIDE	12.2	56.4	7.0	3.2	1.9	29.3
HUSCAP	14.6	17.5	10.3	2.1	1.5	9.3
KURENAI	20.5	20.1	14.4	1.9	2.4	10.9
Tulips-R	15.8	24.2	14.4	2.4	1.5	13.1
全体	17.3	19.5	14.4	2.0	1.9	10.5

全体で見ると英語と日本語ではダウンロード数の差はほとんどないか、英語の方が多い。国内からのアクセスと海外からのアクセスを分けて見ると国内からは日本語、海外からは英語コンテンツの利用が多い。逆（国内から英語、海外から日本語）の利用はいずれも少ない。

また、表６は同じく記述言語ごとのダウンロード数をアクセス方法別に示したものである（こちらは平均値と合わせて中央値も示している）。サーチエンジンからの利用は平均・中央値いずれも英語コンテンツの方が多いのに対し、リポジトリ内の別ページからのアクセスは平均、中央値いずれも日本語の方が多い。日本語コンテンツについてはサーチエンジンからの利用に劣らずリポジトリ内からの利用もあるのに対し、英語はサーチエンジンからの利用に偏っていると言える。

表6. 言語ごとの平均ダウンロード数(アクセス方法別)

		直接 DL	リポジトリ内	サーチエンジン	その他
日本語	平均	2.0	5.1	8.9	0.2
	中央	0.0	2.0	1.0	0.0
英語	平均	3.9	3.0	11.3	0.6
	中央	1.0	1.0	2.0	0.0
全体	平均	2.7	4.4	9.7	0.4
	中央	0.0	2.0	1.0	0.0

3. 4 出版年

図1はリポジトリ別に、コンテンツの出版年（5年区切り）ごとの平均ダウンロード数の推移を示したものである。

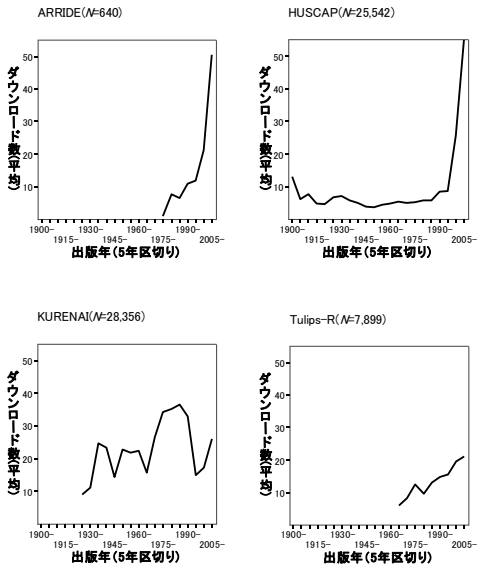

図1. 出版年ごとの平均ダウンロード数（リポジトリ別）

ARRIDE、Tulips-R では新しい文献の方がよく利用される傾向があるが、いずれも収録期間が短い。過去に遡って多くのコンテンツが登録されているのは HUSCAP と KURENAI であるが、HUSCAP では顕著に最新の論文が良く使われ、1990 年代以前の論文の利用がほとんど限定的であるのに対し、KURENAI ではグラフの凹凸が激しく必ずしも新しい論文が良く使われ古い論文は使われなくなる、と言った傾向は見られない。そこで KURENAI についてさらに詳細に分析するために、登録コンテンツの大部分を占める紀要論文について、記述言語とアクセス方法別の出版年ごとのダウンロード数を示したものが図2である。

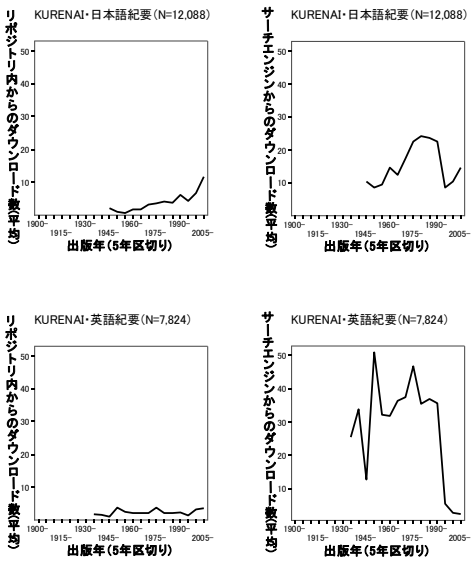

図2. 出版年ごとの平均ダウンロード数
（KURENAI 紀要論文・記述言語×アクセス方法）

日本語・英語とも最近の論文のサーチエンジンからの利用が減るのは、1995 年以降分から収録されている『数理解析研究所講究録』（収録数 10,688 件）のサーチエンジンからの利用が極端に少ないためである。この影響を除くと日本語論文は新しいものの方がよく使われる傾向があるが、英語論文へのサーチエンジンからのアクセスはやはり古い論文までよく利用している。

3. 5 テキスト化

HUSCAP で古い論文の利用が顕著に少ない理由を調べたところ、コンテンツの大半を占める紀要の過去分の利用が少ないためであることがわかった。北海道大学附属図書館の担当者への聞き取りから、HUSCAP 収録紀要の過去分は図書館がスキャンした PDF ファイルを登録しており、その際に OCR 等によりテキストデータを付与していない場合が多いとのことであった。そこでテキスト化の有無の影響について調べるために、HUSCAP の紀要論文本文をダウンロードし、テキスト抽出によりテキスト化の有無を判定した結果と利用数の関係について

分析した（表７）。

表7. テキスト化の有無と平均ダウンロード数

		リポジトリ内	サーチエンジン	全体
全体	テキストなし（N=20,597）	3.4	2.0	6.4
	テキストあり（N=1,396）	17.2	34.7	60.7
2005 年以降発行分	テキストなし（N=746）	14.3	2.5	19.5
	テキストあり（N=911）	22.9	43.6	77.7
“Eurasian Journal of Forest Research” 2006 年発行分	テキストなし（N=5）	3.8	2.4	7.8
	テキストあり（N=5）	10.4	101.0	136.6
「北海道大学文学研究科紀要」2006 年発行分	テキストなし（N=9）	5.3	0.9	8.0
	テキストあり（N=13）	18.2	18.8	44.8

全体で見るとテキスト化されているコンテンツの平均ダウンロード数はされていないコンテンツの 10 倍近く、特にサーチエンジンからの利用に顕著な差がある。テキスト化されていない中に古いコンテンツが多いせいもあるが、2005 年以降出版されたコンテンツに絞った場合、あるいは同一タイトル・同一年に掲載された論文に絞った場合でもいずれもテキストがある方がない場合の数倍～十数倍利用されており、特にサーチエンジンからの利用の差が大きいという結果になった。

4. まとめ

４つの機関リポジトリの分析結果からわかったことは以下の通りである。

（1）自宅等からのアクセスが多い。

（2）国内からの利用が多いが、これは日本語コンテンツが多いためであり実際には英語論文には海外から、日本語論文には国内からそれぞれ同程度の利用が集まっている。

（3）機関リポジトリのアクセス方法としてもっとも多いのはサーチエンジンであるが、日本語論文はサーチエンジン以外からの利用も相当数あるのに対し英語論文は ARRIDE を除くとサーチエンジンからの利用に偏っている。

（4）日本語論文への利用は新しいものが多いのに対し英語論文は古いものまでよく利用されている。

（5）テキスト化の有無と利用数に顕著に関係があり、テキスト化されている方が多い。

機関リポジトリに論文を収録することでサーチエンジン等から、これまで学術論文を読む機会の限られていた人々がコンテンツを利用しやすい環境が実現し、実際に利用されていると言える。特に英文紀要については従来限られた流通の中にあったものが、誰もが利用できるようになったことで世界的に、かつ過去の論文まで多くの利用を集めるようになっていると考えられるが、一方で海外からのアクセス方法はほとんどサーチエンジンに限られており、テキストデータが付与されていないなどサーチエンジンから発見しにくいコンテンツはほとんど利用されないことも今回の分析で明らかになった。図書館がスキャンしたファイル等を登録する場合のテキストデータの付与を徹底すると同時に、国内における CiNii のようなサーチエンジン以外の機関リポジトリコンテンツへの誘導方法を模索することが必要であると考えられる。

謝辞

本報告は「科学研究費補助金（基盤研究（C）機関リポジトリへの登録が学術文献流通に及ぼす効果についての定量的分析」の支援により行われたものです。データをご提供いただいた各機関に感謝いたします。

引用文献

1) Organ, Michael. Download statistics: What do they tell us?. D-lib magazine. 2006, vol.12, no.11, http://www.dlib.org/dlib/november06/organ/11organ.html, (accessed 2009-09-06).

2) Royster, Paul. Publishing original content in an institutional repository. Serials Review. 2008, vol.34, no.1, p.27-30.

3) Bonilla-Calero, A. I. Scientometric analysis of a sample of physics-related research output held in the institutional repository Strathprints (2000-2005). Library Review. 2008, vol.57, no.9, p.700-721.

4) 佐藤義則. 動向レビュー：機関リポジトリの利用統計のゆくえ. カレントアウェアネス. 2008, vol.296, p.12-16.