論文の被引用数と機関リポジトリにおけるダウンロード数の関係

佐　藤　　翔*，冨　本　壽　子**，逸　村　　裕***

The Relationship between Citations and Number of Downloads in Institutional Repositories

Sho SATO, Hisako TOMIMOTO, Hiroshi ITSUMURA

抄録

大学・研究機関において機関リポジトリの設置が普及するにつれ，リポジトリで公開したコンテンツの利用状況，中でもどのような特徴を持つコンテンツがよく使われるのかに注目が集まっている。本研究では論文の特徴の一つとして被引用数に着目し，機関リポジトリに収録された文献における被引用数とダウンロード数の関係を分析，被引用数からよくダウンロードされる論文を推測することが出来るかどうかを検討する。

分析対象は北海道大学（HUSCAP），京都大学（KURENAI），筑波大学（Tulips-R）の 3 つの機関リポジトリに収録された雑誌掲載論文のうち，Web of Scienceにも収録されている論文である。各論文の被引用数と 2008 年中のダウンロード数をそれぞれ集計し，相関関係の有無を分析した。また，出版年や分野ごとの引用慣行による影響を排除するため，出版年別および物理学分野の論文に限定した場合についても分析を行った。分析の結果，被引用数とダウンロード数の間には相関はなく，被引用数の多寡からは機関リポジトリ収録後によくダウンロードされる論文を推測出来ないことがわかった。

Abstract

As Institutional Repositories become common in universities and research institutions, many have been paying attention to the repository usage data. In this research we focus attention on number of citations as one of features of articles. We reveal the relationship between number of citations and number of downloads of contents in institutional repositories and discuss whether we can predict papers which will be highly-used on the basis of number of citations.

Our samples are journal articles which are included in Web of Science and three institutional repositories; Hokkaido University （HUSCAP）, Kyoto University （KURENAI） and University of Tsukuba （Tulips-R）. We analyzed correlations between number of citations and that of download data in 2008. In addition, to exclude influences of publication year and citation culture of each discipline, we analyzed year-to-year to articles in physics. As a result of analyses, we found that there is no correlation between number of citations and downloads. It has been revealed that we may not predict which papers will get high usage on the basis of number of citations.

* 筑波大学大学院図書館情報メディア研究科博士前期課程
Master Program
Graduate School of Library, Information and Media Studies, University of Tsukuba

** 北海道大学大学院教育学院聴講生
Auditing Student
Graduate School of Education, School of Education, Hokkaido University

*** 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

1．はじめに

1.1　背景と先行研究

1.1.1　機関リポジトリの普及と利用状況への関心の高まり

大学・研究機関等における機関リポジトリの設置と提供コンテンツの拡充が進んでいる。機関リポジトリとは「大学とその構成員が創造したデジタル資料の管理や発信を行うために，大学がそのコミュニティの構成員に提供する一連のサービス」であり[1]，学術情報への障壁のないアクセスを目指すオープンアクセス運動の一翼を担うものとしてあらわれた取り組みである。日本では国立情報学研究所による最先端学術情報基盤(CSI)整備の一環としての学術機関リポジトリ構築連携支援事業等[2]を通してリポジトリを設置する機関数や各機関リポジトリで提供されるコンテンツ数が増大した。2009年 4 月 1 日現在，世界全体で1,301[3]，日本国内でも93の機関リポジトリが存在する[4]。国内総コンテンツ数は58万件以上，本文が閲覧可能なものに限っても39万件を超え[5]，学術雑誌掲載論文をはじめ各機関の紀要論文，学位論文など多くの学術文献をオンラインで，誰でも無料で読むことが出来るようになってきている。

機関リポジトリの設置数，収録コンテンツ数ともに整ってきた状況を受け，近年では機関リポジトリで公開しているコンテンツの利用状況への注目が高まっている。機関リポジトリのように電子的に提供されている文献ではサーバに残るアクセスログを分析することで利用の実態を把握することができ，このログの適切な処理方法の確立と標準化をめぐってサウザンプトン大学によるIRSプロジェクト[6]，英国情報システム合同委員会(JISC)等によるPIRUSプロジェクト[7]，千葉大学等による「機関リポジトリ評価のための基盤構築」プロジェクト[8]など多くのプロジェクトが進行している。

実際にアクセスログから機関リポジトリで公開しているコンテンツの利用状況を見た例として，Nebraska-Lincoln大学のRoysterは同大学の機関リポジトリの月ごとのダウンロード数上位10コンテンツの半数以上が学位論文など同大学の機関リポジトリ以外では読むことが出来ないコンテンツであったとしている[9]。Bonilla-CareloはStrathclyde大学の機関リポジトリに収録された2000〜2005年の物理学分野の文献のアクセスログを分析し，文献のタイプで見ると雑誌掲載論文の 1 コンテンツあたりの平均ダウンロード数が多く，またプレプリントとポストプリントではポストプリントの方が平均ダウンロード数が多いことなどを報告している[10]。佐藤は北海道大学，京都大学の機関リポジトリのアクセスログを分析し，京都大学では紀要論文，北海道大学では紀要論文と教材がダウンロード数上位文献中に多く出現するが，アクセス元ホストの種類(大学，企業，民間プロバイダなど)によって利用に異なる特徴があること，利用者の大部分が大学・研究機関ではなく自宅から機関リポジトリを利用していること，利用の大半はGoogleなどのサーチエンジンの検索結果からリポジトリ掲載コンテンツにアクセスしていることなどを示している[11]。また，佐藤・逸村はアジア経済研究所の機関リポジトリのアクセスログ分析から，同リポジトリではダウンロード数上位文献の多くをワーキングペーパーが占めることを明らかにしている[12]。九州大学の池田らは同大学の業績データベースと機関リポジトリの連携機能から得られるログを用い，業績データベースの各論文から機関リポジトリ収録論文に貼られたリンクのクリック状況と，うちどれだけが本文を閲覧できたか(リポジトリに文献が収録されていない場合も業績データベース側ではリンクが自動で付与される。その場合，クリックしても本文は閲覧できない)を分析している。その結果，リンクがクリックされたが該当文献がリポジトリに収録されておらず利用者が閲覧できなかったケースが 9 割にのぼるとし，コンテンツ整備状況がまだ十分ではないことを示している[13]。

このように機関リポジトリにおけるコンテンツの利用状況を分析した研究はいくつか存在し，中でもどのようなコンテンツがよく使われるのか，今後どのようなコンテンツを収集すべきかが議論の対象となっている。しかし分析方法はダウンロード数上位コンテンツに限定してコンテンツの傾向を確認する，特定の外部サービスからリポジトリ掲載論文を利用したログのみを分析するといった部分的なものにとどまり，よく使われるコンテンツの特徴は必ずしも明らかにはなっていない。

1.1.2　利用（ダウンロード）数と被引用数の関係

一方，機関リポジトリ以外に目を向けると電子化の進展以降，学術文献の利用に関する研究は数多くなされており，Tenopirによれば1997〜2003年の期間だけで200件以上の文献が発表されている[14]。中でも注目されている主題のひとつは「論文の利用(ダウンロード)数と被引用数の関係」，すなわち「よく読まれている論文はよく引用されているのか／よく引用されている論文はよく読まれているのか」である。

O'LearyはDecision Support System誌のダウンロード数と，Google Scholar，SSCI，Scopusなど複数のソースに基づく被引用数の関係を分析し，ダウンロード数はい

ずれのソースに基づく被引用数との間にも強い正の相関関係（r＝0.7〜0.8以上）にあったことを示している[15]。WatsonはJournal of Vision誌を対象にダウンロード数とScopusにおける被引用数の関係を分析し，両者は強い正の相関関係にあり（r＝0.74），さらに論文出版年ごとに見ると出版から時間がたつほどトータルのダウンロード数と被引用数の相関は強くなり，2007年に出版された論文ではr＝0.4程度であった両者の相関が2003年出版論文ではr＝0.8前後になるとしている。その理由としてWatsonは論文を読んでから（ダウンロードしてから）引用するまでにかかるタイムラグの存在を挙げている[16]。PernegerはBMJ誌のダウンロード数とWeb of Scienceにおける被引用数の関係を分析し，出版後最初の１週間のダウンロード数と出版後５年間の被引用数の間にr＝0.54で有意な正の相関関係があることを示した[17]。また，Nature Neuroscience誌では出版後90日間のダウンロード数とScopusにおける後の被引用数にはr＝0.648と強い正の相関関係があり，ダウンロード数の集計期間を出版後180日間に伸ばすと両者の相関はさらに強くなった（r＝0.724）と報告されている[18]。

以上は特定の雑誌掲載論文について，出版者の持つアクセスログを見た分析であるが，特定雑誌に限らずダウンロード数と被引用数の関係を見た研究としては分野別リポジトリのアクセスログを対象とするものがある。BrodyらはPernegerと同様に初期のダウンロード数から後の被引用数を予測できるかを，物理学を中心に数学，コンピュータサイエンス等の論文を収録する分野別リポジトリであるarXivを対象に分析しており，出版後６カ月間のダウンロード数と２年後の被引用数の間に相関があるとしている（r＝0.4程度）[19]。一方，DavisはBrodyらと同じくarXivを対象に，数学分野の論文についてダウンロード数とMathSciNetにおける被引用数の関係を分析し，両者は有意に正の相関はしているもののr＝0.15と弱い水準にとどまったとしている[20]。

ダウンロード数と被引用数の関係を分析する研究の多くは，「よく読まれている論文はよく引用されているか」，すなわち同一期間中に行われたダウンロード数と被引用数の関係か，「よく読まれた論文は後によく引用されるか」，すなわちある時点でのダウンロード数と後の被引用数の関係に注目している。その中でMoedはTetrahedron Letters誌に掲載された論文のダウンロード数とSCIにおける被引用数を比較し，ある時点での被引用数と後のダウンロード数の関係をある時点でのダウンロード数と後の被引用数の関係と区別して分析している。その結果，被引用数と利用のオブソレッセンス（加齢，時間とともに利用されなくなっていくこと）が関係すること，はじめて論文が引用された直後の３カ月間のダウンロード数は引用されなかった場合に期待される値（当該論文が引用されるまでのダウンロード数と，一度も引用されたことのない論文群の出版後のダウンロード数の推移に基づき計算）より25％増えることなどを示した一方，ダウンロード数全体に見た場合には被引用後に増加する割合は1.7％程度にとどまり，被引用後のダウンロード数増は一時的なもので全体で見ると大きな違いはないとしている[21]。

また，論文単位ではなく雑誌単位での被引用数とダウンロード数の関係を見たものであるが，Bollenらはカリフォルニア州立大学（CSU）における電子ジャーナルのダウンロード数にインパクトファクターと同様の計算式を適用した場合の指標（Usage Impact Factor＝UIF[i]）とJournal Citation Reportsにおけるインパクトファクター（JIF）の関係について分析している[22]。その結果，UIFはJIFとCSU購読誌全体で見ても，分野を区切って見ても相関はないか負の有意な相関があることが多かったとしている。ただし教員・大学院生数が学部生数より多い分野の雑誌ではUIFはJIFと正の相関傾向があるともしており，その理由としてUIF，JIFそれぞれの算出元となるサンプルの違いを挙げている。すなわち被引用数に基づいて算出されるJIFは学術的な文献の著者の，国際的なコミュニティを代表する指標である一方，UIFはCSUの，教員だけでなく学部生やスタッフ，実務者も含むコミュニティ基づいたものであり，両者が相関しないのはこのサンプルの違いによるもので，院生・教員の多い分野でUIFとJIFに正の相関が見られるのはUIFのサンプルがJIFの場合とより近いコミュニティになっているからであるとしている。ここから，被引用数とダウンロード数の関係は，ダウンロード数を算出する際の元となるコミュニティの影響を強く受けることが示唆されている。

1.2　本研究の目的

本研究の目的は機関リポジトリに収録されたコンテンツの，ある時点での被引用数と後のダウンロード数の関係を明らかにすることである。被引用数とダウンロード数の関係については，先行研究で挙げたように，

（1）同一期間中に行われたダウンロード数と被引用

i　雑誌Jの2009年のインパクトファクターは当該雑誌に2007〜2008年に掲載された論文が2009年中に引用された回数を，2007〜2008年の掲載論文数で割った値となる。同じようにある雑誌Jの2009年のUIFは，2007〜2008年の当該雑誌掲載論文が2009年中にCSUでダウンロードされた回数を，2007〜2008年の掲載論文数で割った値である。

数の関係
(2) ある時点でのダウンロード数と後の被引用数の関係
(3) ある時点での被引用数と後のダウンロード数の関係

の 3 つに分けられる。このうち特に多いのは(2)の関係を見る研究である。これは被引用数の多寡が現在の論文・学術雑誌の評価において重視されている一方，論文の出版から引用までには年単位での時間がかかるために，出版後早期に研究を評価しうる指標としてのダウンロード数に着目した研究が多いためであると考えられる。

これに対し本研究では(3)のある時点での被引用と後のダウンロード数の関係に着目する。これは「機関リポジトリでよく使われている文献はなにか，どんな特徴があるのか」を明らかにし，機関リポジトリに重点的に収集すべきコンテンツを選択する際の材料を得る試みの一環として，ある時点での被引用数が後によくダウンロードされる論文を予測する指標足り得るかを検討するためである。機関リポジトリに重点的に収集すべきコンテンツを定める方針はすでに国立情報学研究所の学術機関リポジトリ構築連携支援事業でも取り入れられており，平成20〜21年度委託事業では「学位論文，科学研究費補助金・COE・特色GPなどの助成金による研究成果報告書(付随する研究データ等含む)，テクニカルレポート，紀要論文など」が重点コンテンツとして指定されている[23]。平成22年度以降も同様の取り組みが行われる場合，当然重点コンテンツが見直される可能性があるが，このような重点的に収集すべきコンテンツを定める際にダウンロード数の多寡は判断材料となりうる。もちろん利用の多さのみが機関リポジトリに収集すべきコンテンツを定める際の材料となるわけではないが，利用が多い，すなわち高い潜在需要を持つコンテンツを積極的に収集・提供することの意義についてもまた言を俟たない。コンテンツのダウンロード数そのものは実際にリポジトリに収録して見ないと測ることはできないが，もしある時点で被引用数の多い論文が少ない論文に比べて機関リポジトリ上でよく利用される傾向があるのであれば，被引用数の多さは積極的／優先的に収集すべきコンテンツを選ぶ際の材料の一つとなりうる。一方，もし両者の間に相関がないのであれば，被引用数の多寡からは後の利用の多寡を予測することはできず，被引用数は(利用の面のみから考えれば)重点コンテンツを決定する際の材料にはなり得ないと言える。言い換えれば，「ある時点での被引用数から後に機関リポジトリからよくダウンロードされる論文を推測できるか」を明らかにすることが本稿の目的である[ii]。

なお，被引用数に関するデータ取得の関係上，本稿においては分析対象とするコンテンツを原著論文，それも Web of Science に収録されているもの(大部分は英語論文)に限定している。機関リポジトリ上での利用数と関係する要因としては文献の種別(原著論文なのか教材なのか学位論文なのか)や言語(英語か日本語か)など，他にも多くの可能性が考えられるが，これらの検討については別稿に回すこととしたい。

2．分析対象と方法

2.1　分析対象とするデータ

2.1.1　分析対象とする機関リポジトリ

分析対象は北海道大学の北海道大学学術成果コレクション(HUSCAP)[iii]，京都大学の京都大学学術情報リポジトリ(KURENAI)[iv]，筑波大学のつくばリポジトリ(Tulips-R)[v]の 3 機関のリポジトリである。これらはいずれも同一のソフトウェア(DSpace)によって構築されており，アクセスログのフォーマットも共通であるため比較が容易である。

2.1.2　被引用数

前述の 3 つの機関リポジトリに収録された論文についてトムソン・ロイター社の学術文献データベース Web of Science への収録状況を調査し，収録のあったものにつ

ii　もちろん出版直後の論文を収集する際には，被引用数の多寡からよくダウンロードされる論文が推測できたとしても役には立たない(出版間もない論文はまだ引用されていないか，されていても引用が少ないと考えられるため)。しかし機関に所属する研究者の，出版から一定期間の経った論文を図書館員が(自ら，あるいは著者に依頼して) 遡ってリポジトリに登録することは一般的に行われており，このような場合に被引用数から後によくダウンロードされる論文が推測できるとすれば登録の優先順位を決める材料となり得る (そもそも研究者自身が，出版直後の論文を登録する場合には優先順位等を定める必要自体ない。当該論文を登録するか否か決めるだけである)。

iii　http://eprints.lib.hokudai.ac.jp/dspace/index.jsp

iv　http://repository.kulib.kyoto-u.ac.jp/dspace/

v　https://www.tulips.tsukuba.ac.jp/dspace/index.jsp

いて同データベースから被引用数を取得した。論文の同定は著者名，掲載誌名，掲載巻号等により行った。収録状況・被引用数状況の調査時期はHUSCAP，Tulips-Rは2008年10月，KURENAIは2008年８月であり，機関リポジトリに収録されかつWeb of Scienceにも収録されていた論文の数はそれぞれHUSCAP：2,215件，Tulips-R：624件，KURENAI：1,331件であった。このうち2007年以前に出版された論文への，2007年以前に出版された論文からの被引用数を分析対象とした。

2.1.3　ダウンロード数

ダウンロード数の分析対象期間はHUSCAPは2008年全体の12カ月分，KURENAIとTulips-Rは2008年８～12月の５カ月分とし，各機関リポジトリのアクセスログからこれらの期間のダウンロード数を集計した。分析対象期間がリポジトリによって異なるのは，KURENAIとTulips-Rでは前述の1,331件，624件の論文の多くが2008年中に機関リポジトリに収録されており，2008年全体のログを見てしまうと期間中に新たに収録された論文のダウンロード数が他より低く見積もられてしまうためである。一方で2007年以前に収録された論文に限定すると分析の母数が少なくなってしまうため，両リポジトリではダウンロード数の集計期間を短くすることとした。HUSCAPについては前述の2,215件の論文のうち2007年末時点ですでに2,003件が機関リポジトリに収録されており，これらの論文に限定すれば十分な母数を確保した上で2008年全期間のログを見ることが出来ると考えた。

アクセスログについては分析を行う前に，検索ロボットによるクローリングや同一人物による連続したアクセス等を取り除くフィルタリング作業を行った。これは機械的なアクセスを取り除き，実際に人間が利用する目的でのダウンロードに対象を限定するためである。フィルタリングについては電子レコードの統計標準であるCOUNTER[24]に則って行った。COUNTERに基づくフィルタリングは前述の千葉大学等による「機関リポジトリ評価のための基盤構築」プロジェクト[8]でも採用されており，機関リポジトリにおけるフィルタリング基準としては現在最も適切であると考えられる。また，本研究では本文ファイルのダウンロードのみをダウンロード数として集計し，メタデータのみの取得については分析から除外した。

なお各リポジトリのアクセスログ及び論文の機関リポジトリへの収録日データについては各機関リポジトリ担当者から提供を受けた。

2.1.4　論文発表後年数

被引用数，ダウンロード数の双方に関連しうる指標として，Watsonが指摘している論文が発表されてからの経過年数（論文発表後年数）が考えられる[16]。そこで本研究では各論文の発表後年数（出版年－2007）を計算し，分析に加えた。各論文の出版年のデータについては各機関リポジトリ担当者から提供を受けた。

2.2　分析方法

2.2.1　全体での分析

まず論文の被引用数，ダウンロード数についてリポジトリごとの平均値・中央値などの基礎的データを確認した。また，被引用数とダウンロード数，及び両者に関連しうる論文発表後年数の間のスピアマンの順位相関係数を確認した。相関関係の指標としてはピアソンの積率相関係数を見ることも考えられるが，よく知られるように論文の被引用数は少数の被引用数の多い論文と多数の被引用数の少ない論文の非常に偏った分布をしている。また，ダウンロード数も同様に偏りの大きいデータであるため正規分布を前提とするピアソンの積率相関係数を用いることは不適切であると考えスピアマンの順位相関係数を見ることとした。同様にダウンロード数と被引用数の関係をスピアマンの順位相関から見たものとしてはMoedの先行研究等がある[21]。

2.2.2　出版年ごとの分析

Watsonが示しているように[16]論文の被引用数は一般にその出版後の経過年数と相関するため，出版年を分けて見ない分析では各論文の出版年の違いが大きく影響している可能性がある。出版から年数の経っている論文は累積被引用数は多いがダウンロード数は年とともに少なくなっていくことが考えられ，その場合ダウンロード数と被引用数の間の相関関係を見ることは難しくなる。そこで本研究ではWatsonの分析と同様に出版年ごとに論文を区切った場合の被引用数とダウンロード数の関係を分析することとし，全体の場合と同じくスピアマンの順位相関を確認した。なお，出版年別の分析の対象期間は2001～2007年に限定した。これは2000年以前に出版された論文は各リポジトリへの収録数が少なく，出版年別の分析に十分な母数が得られないと考えたためである。

2.2.3　物理学分野での分析

論文の引用慣行は分野によって異なり，分野を区分しない分析では各論文の分野の違いが大きく影響していることが考えられる。例えば一般に物理学分野では多くの

論文を引用するため被引用数が多くなり，数学分野では引用数が少ないため被引用数も伸びないことなどが知られている。そのため被引用数とダウンロード数の関係を正確に捉えるには分野を区切った分析が必要になる。そこで本研究では分析対象とする各リポジトリのいずれにおいても最も収録論文数の多い物理学分野の論文に限定した場合の分析結果についても他と同様に確認することとし，掲載誌名から物理学分野の論文を特定し分析することとした。なお，物理学分野以外については十分な分析の母数を得られないため今回の分析では扱わないこととした。以上をまとめると，各リポジトリの分析対象期間，分析対象論文数等は表 1 の通りとなった。

表 1　分析対象データのまとめ

		HUSCAP	KURENAI	Tulips-R
分析対象とする論文の出版年		2007 年以前	2007 年以前	2007 年以前
Web of Science 調査実施時期		2008.10	2008.08	2008.10
Web of Science に収録されていた論文数		2,215	1,331	624
アクセスログ分析対象期間		2008.01–12	2008.08–12	2008.08–12
分析対象とする論文の機関リポジトリへの収録時期		2007 年以前	2008.07 以前	2008.07 以前
分析対象論文数（全体）		2,003	1,304	599
出版年ごとの分析対象論文数	2007	398	152	144
	2006	599	199	28
	2005	265	178	41
	2004	169	140	39
	2003	133	163	35
	2002	91	153	18
	2001	57	130	42
物理学分野での分析対象論文数	全期間	553	1,005	436
	2007	87	115	48
	2006	127	155	18
	2005	52	146	26
	2004	50	112	21
	2003	45	118	28
	2002	35	129	16
	2001	23	103	39

2.3　本手法における限界

第一に，本研究で明らかに出来るのは被引用数とダウンロード数の「相関関係」であり，両者の因果関係について明らかにするものではない。つまり仮に両者に相関があったとしても，「被引用数が多いから」よく読まれたのか，「利用者がよく引用されていることを知っていたわけではないが，よく引用されるような論文は興味を持つ人も多いので」読まれたのかと言ったことについては明らかに出来ない。相関がない場合も同様である。

第二に，本研究は「ある時点での被引用数と後のダウンロード数の関係（1.2の(3)の関係）を検証するものであり，「ある時点でのダウンロード数と後の被引用数の関係」（1.2の(2)の関係）や「同一期間中に行われたダウンロード数と被引用数の関係」（1.2の(1)の関係）を明らかにすることはできない。現在，ダウンロード数の集計に十分なだけのログが得られているのは2008年分のみ（KURENAI，Tulips-Rについては 8 月以降のみ）であり，「ダウンロード数と後の被引用数」の相関を分析するためにはこれ以降（2009年以降）の被引用数が算出されるのを待つ必要が，「同一期間中のダウンロード数と被引用数の関係」を見るには加えて2009年以降のダウンロード数が得られるのを待つ必要がある[vi]。

vi　2008 年中の被引用数をダウンロード数と後の被引用数の分析に用いることが出来ないのは，Web of Scienceのデータには論文が引用された月日が含まれていない（引用した論文の出版年しかわからない）ため，同じ年のダウンロードと被引用についてはどちらが先に行われたか判別できないためである。同じく月日がないと「2008 年 8 － 12 月に行われた引用」を特定することも出来ないため，KURENAI，Tulips-R については同一期間中のダウンロード数と被引用数の関係の分析を行うには 2009 年以降のダウンロード数が計算できるようになるのを待たねばならない。

3．分析結果

3.1　全体の分析結果

3.1.1　各機関リポジトリの被引用数とダウンロード数の概要

表２は各リポジトリの分析対象論文全体について，被引用数とダウンロード数の平均値，中央値，最頻値，標準偏差を示したものである。

HUSCAPは他の２リポジトリに比べ分析対象論文の被引用数の平均値・中央値ともに低い。HUSCAPは表１の通り，物理学分野の論文の占める割合が少ない。物理学は引用数の多い分野であり，その占める割合が少ないHUSCAPは多い他の２リポジトリよりも被引用数が少なくなっていると考えられる。一方Tulips-Rは掲載論文の平均被引用数は高いが，被引用数の標準偏差も大きく非常によく引用される論文とそれほどでもない論文の二極化傾向が他より強いと考えられる。そのため中央値で見た場合には他との差は縮まる。

ダウンロード数について，HUSCAPで平均値・中央値が高いのは分析対象期間の違いによる。分析対象期間の長さの比率(HUSCAP 12カ月：他２つ ５カ月)以上の差があるが，機関リポジトリにおいては６，７月など大学の学期期間中にアクセスが増える傾向があり，HUSCAPはこれらの期間を含んでいるがKURENAIとTulips-Rの分析期間には含まれていない。そのため単純にここでHUSCAPと他のリポジトリとのダウンロード数の差を見ることは意味がないと言える。KURENAIとTulips-Rでは同一の分析対象期間でダウンロード数(平均値・中央値とも)に差があり，ややKURENAI収録論文の方がダウンロード数が多い傾向がある。ただし差はわずかである。

表２　分析対象論文の被引用数・ダウンロード数の概要

	HUSCAP (*N*=2,003)		KURENAI (*N*=1,304)		Tulips-R (*N*=599)	
	被引用数	ダウンロード数	被引用数	ダウンロード数	被引用数	ダウンロード数
平均値	7.19	47.10	10.51	12.76	18.96	11.12
中央値	2.00	33.00	5.00	7.00	6.00	5.00
最頻値	0.00	20.00	1.00	0.00	0.00	0.00
標準偏差	16.409	54.971	17.053	31.082	43.895	15.895

3.1.2　被引用数とダウンロード数の相関関係

表３～５は各リポジトリにおける被引用数，ダウンロード数と論文発表後年数の間のスピアマンの順位相関係数を示したものである。また，図１は各リポジトリにおける被引用数とダウンロード数を図示したものである(それぞれ１を加えた上で対数化して表示)。表より，いずれのリポジトリでも論文発表後年数と被引用数は有意な正の相関，論文発表後年数とダウンロード数は有意な負の相関関係にある。被引用数と論文発表後年数はどのリポジトリでもρ＝0.5～0.7の範囲内でやや強い相関となっている。当然ではあるが，論文が発表されてから年数がたつほど被引用数の合計は増えると言える。一方，ダウンロード数と論文発表後年数についてはリポジトリにより差があり，KURENAIではρ＝－0.087と非常に弱い負の相関にとどまるが，Tulips-Rではρ＝－0.431とやや強い負の相関を示す。いずれのリポジトリでも古い論文ほどダウンロードされなくなる傾向があり，特にTulips-Rで顕著であると言える。

被引用数とダウンロード数についてもいずれのリポジトリでも有意な相関関係があるが，HUSCAPとTulips-Rでは弱い負，KURENAIでは非常に弱い正と傾向に違いがある。母数が多いため相関関係が示されてはいるが，

表３　HUSCAP における各指標間のスピアマンの順位相関係数（*N*＝2,003）

		被引用数	ダウンロード数	論文発表後年数
被引用数	相関係数	1.000	−0.125(**)	0.610(**)
	有意確率（両側）	.	0.000	0.000
ダウンロード数	相関係数	−0.125(**)	1.000	−0.255(**)
	有意確率（両側）	0.000	.	0.000
論文発表後年数	相関係数	0.610(**)	−0.255(**)	1.000
	有意確率（両側）	0.000	0.000	.

** 相関は、1％ 水準で有意（両側）。

表 4　KURENAI における各指標間のスピアマンの順位相関係数（*N* ＝ 1,304）

		被引用数	ダウンロード数	論文発表後年数
被引用数	相関係数	1.000	0.085(**)	0.539(**)
	有意確率（両側）	.	0.002	0.000
ダウンロード数	相関係数	0.085(**)	1.000	−0.087(**)
	有意確率（両側）	0.002	.	0.000
論文発表後年数	相関係数	0.539(**)	−0.087(**)	1.000
	有意確率（両側）	0.000	0.000	.

** 相関は、1 % 水準で有意（両側）。

表 5　Tulips-R における各指標間のスピアマンの順位相関係数（*N* ＝ 599）

		被引用数	ダウンロード数	論文発表後年数
被引用数	相関係数	1.000	−0.324(**)	0.661(**)
	有意確率（両側）	.	0.000	0.000
ダウンロード数	相関係数	−0.324(**)	1.000	−0.431(**)
	有意確率（両側）	0.000	.	0.000
論文発表後年数	相関係数	0.661(**)	−0.431(**)	1.000
	有意確率（両側）	0.000	0.000	.

** 相関は、1 % 水準で有意（両側）。

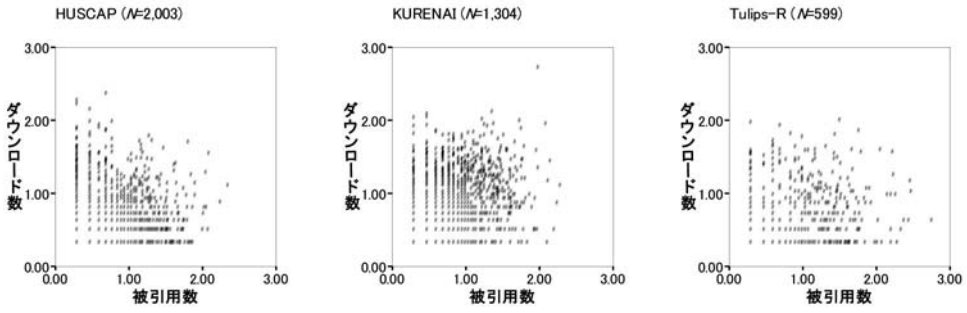

図 1　各リポジトリにおける論文の被引用数とアクセス数の関係（散布図）

実際にはほとんど被引用数とダウンロード数の間には関係がないものと考えられる。図 1 からも被引用数の多寡にかかわらずダウンロード数はおおむね均一に分布していると言え，両者の間に関係性を読み取ることは出来ない。

3.2　出版年ごとの分析結果

表 6 は論文出版年別の，各リポジトリにおけるダウンロード数と被引用数のスピアマンの順位相関係数を示したものである。HUSCAPの2001年出版論文，KURENAIの2007年および2005年出版論文を除き，ほとんどのリポジトリで出版年を区切ると被引用数とダウンロード数の間に有意な相関関係は見られない。Tulips-Rについては分析の母数が十分でないことも考えられるが，HUSCAP・KURENAIについては出版年ごとの論文数も相当数あり，出版年を分けてみた場合には被引用数とダウンロード数の間に有意な相関関係はないものと考える方が妥当である。

3.3　物理学分野での分析結果

表 7 は各リポジトリに収録された物理学分野の論文について，被引用数とダウンロード数の平均値・中央値等を見たものである。全体で見た場合(表 2)と異なり，HUSCAPの方がKURENAIより被引用数の平均値が高い(中央値はKURENAIの方が高い)。全体で見た場合のHUSCAPの被引用数の少なさが物理分野の占める割合の少なさによるものであったことがここから裏付けられる。

ダウンロード数についてはいずれのリポジトリでも全体で見た値よりも平均値が低く，KURENAIを除くと中央値も低い。物理学分野の論文はリポジトリに収録された雑誌掲載論文の中ではダウンロードされる回数が少な

表６　出版年別の各リポジトリにおけるダウンロード数と被引用数のスピアマンの順位相関係数

HUSCAP		KURENAI		Tulips-R	
2007（*N*=398）	−0.002	2007（*N*=152）	0.213(**)	2007（*N*=144）	−0.117
2006（*N*=599）	0.002	2006（*N*=199）	0.120	2006（*N*=28）	−0.095
2005（*N*=265）	0.087	2005（*N*=178）	0.270(**)	2005（*N*=41）	−0.199
2004（*N*=169）	0.095	2004（*N*=140）	0.135	2004（*N*=39）	−0.293
2003（*N*=133）	0.010	2003（*N*=163）	0.147	2003（*N*=35）	0.120
2002（*N*=91）	0.027	2002（*N*=153）	−0.038	2002（*N*=18）	−0.146
2001（*N*=57）	0.281(*)	2001（*N*=130）	0.125	2001（*N*=42）	0.083

** 相関は、1％ 水準で有意（両側）。
* 相関は、5% 水準で有意（両側）。

表７　物理学分野の論文の平均被引用数・ダウンロード数

	HUSCAP（*N*=553）		KURENAI（*N*=1,005）		Tulips-R（*N*=436）	
	被引用数	ダウンロード数	被引用数	ダウンロード数	被引用数	ダウンロード数
平均値	12.13	33.01	11.76	11.97	25.36	7.46
中央値	4.00	26.00	7.00	8.00	10.00	3.00
最頻値	0.00	6.00/19.00	1.00	3.00	0.00	0.00
標準偏差	23.650	31.890	17.638	19.579	49.918	12.148

い傾向があると言える。特にHUSCAP，Tulips-Rでは差が顕著であり，両リポジトリにおいて物理学分野の論文は被引用数は他より多いがダウンロード回数は他より少なくなっていたと言える。このことが両リポジトリで被引用数とダウンロード数の間に負の相関関係があったことの一因と考えられる。

また，表８～10は各リポジトリにおける物理学分野の論文の被引用数，ダウンロード数と論文発表後年数の間のスピアマンの順位相関係数を示したもの，図２は各リポジトリにおける物理学分野の論文の被引用数とダウンロード数を図示したものである（それぞれ１を加えた上で対数化して表示）。物理学分野に限定した場合もいずれのリポジトリでも論文発表後年数と被引用数の間には有意な正の相関が，論文発表後年数とダウンロード数の間には有意な負の相関があり，発表から時間がたっている論文ほど合計の被引用数は大きくなるが，ダウンロードされる回数は少なくなっていく傾向がある。

一方，被引用数とダウンロード数の関係についてはKURENAIでは有意な相関がなく，HUSCAPとTulips-Rでは非常に弱い有意な負の相関関係があった。Tulips-Rについては全体で見た場合よりも相関が弱くなっており（全体で $\rho = -0.324$，物理学に限定すると $\rho = -0.102$），全体で見た場合に中程度の負の相関が見られるのは分野の異なる論文が混じっていることの影響があると考えられる。ただしHUSCAPではダウンロード数と被引用数の負の相関は物理学に限定した方がわずかに強くなっている。また，いずれのリポジトリでも全体での分析と同様に図からは特に傾向を読み取ることはできない。さらに出版年別の，物理学分野における被引用数とダウンロード数の関係を示したものが表11である。分野

表８　HUSCAP・物理学分野における各指標間のスピアマンの順位相関係数（*N*＝553）

		被引用数	ダウンロード数	論文発表後年数
被引用数	相関係数	1.000	−0.133(**)	0.670(**)
	有意確率（両側）	.	0.002	0.000
ダウンロード数	相関係数	−0.133(**)	1.000	−0.306(**)
	有意確率（両側）	0.002	.	0.000
論文発表後年数	相関係数	0.670(**)	−0.306(**)	1.000
	有意確率（両側）	0.000	0.000	.

** 相関は、1％ 水準で有意（両側）。

表 9　KURENAI・物理学分野における各指標間のスピアマンの順位相関係数（*N* ＝ 1,005）

		被引用数	ダウンロード数	論文発表後年数
被引用数	相関係数	1.000	−0.008	0.584(**)
	有意確率（両側）	.	0.791	0.000
ダウンロード数	相関係数	−0.008	1.000	−0.142(**)
	有意確率（両側）	0.791	.	0.000
論文発表後年数	相関係数	0.584(**)	−0.142(**)	1.000
	有意確率（両側）	0.000	0.000	.

**　相関は、1 % 水準で有意（両側）。

表 10　Tulips-R・物理学分野における各指標間のスピアマンの順位相関係数（*N* ＝ 436）

		被引用数	ダウンロード数	論文発表後年数
被引用数	相関係数	1.000	−0.102(*)	0.465(**)
	有意確率（両側）	.	0.033	0.000
ダウンロード数	相関係数	−0.102(*)	1.000	−0.238(**)
	有意確率（両側）	0.033	.	0.000
論文発表後年数	相関係数	0.465(**)	−0.238(**)	1.000
	有意確率（両側）	0.000	0.000	.

**　相関は、1 % 水準で有意（両側）。
*　相関は、5 % 水準で有意（両側）。

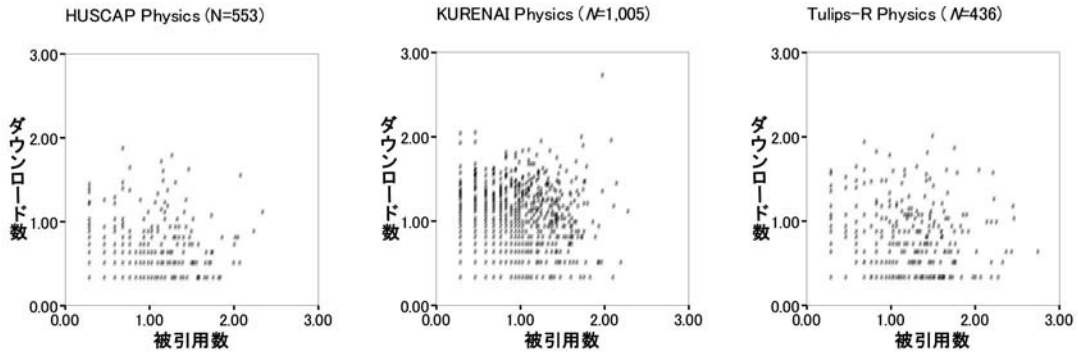

図 2　各リポジトリにおける物理学分野論文の被引用数とアクセス数の関係（散布図）

表 11　出版年別の各リポジトリにおけるダウンロード数と被引用数のスピアマンの順位相関係数
（物理学分野の論文に限定）

HUSCAP		KURENAI		Tulips-R	
2007（*N*=87）	0.050	2007（*N*=115）	0.134	2007（*N*=48）	−0.164
2006（*N*=127）	0.169	2006（*N*=155）	0.010	2006（*N*=18）	0.103
2005（*N*=52）	0.081	2005（*N*=146）	0.195(*)	2005（*N*=26）	−0.130
2004（*N*=50）	0.174	2004（*N*=112）	−0.015	2004（*N*=21）	−0.319
2003（*N*=45）	0.064	2003（*N*=118）	0.082	2003（*N*=28）	0.010
2002（*N*=35）	0.156	2002（*N*=129）	−0.102	2002（*N*=16）	−0.015
2001（*N*=23）	0.000	2001（*N*=103）	0.155	2001（*N*=39）	0.175

*　相関は、5 % 水準で有意（両側）。

と出版年を区切ってみた場合にはKURENAIの2005年に出版された物理学分野の論文を除き，ほとんど有意な相関関係は存在しない。Tulips-RおよびHUSCAPの2001・2002年については分析の母数の少なさも要因であると考えられるが，KURENAIでは十分な母数があるにもかかわらず2005年を除き有意な関係がなく，また非有意の場合の相関係数の値についても正負が必ずしも定まっていない。ここから，分野・出版年を揃えて分析した場合，被引用数とダウンロード数の間にはほとんど相関がないと言うことができる。

4．結果の考察と今後の課題

3章の分析結果より，分野と出版年を揃えて分析した場合，機関リポジトリに収録された論文のある時点での被引用数と後のダウンロード数との間にはほとんど相関関係がない。このように被引用数とダウンロード数の間に相関が見られない理由として，先行研究等から以下の2点が考えられる。

(1) 機関リポジトリに限らず，ある時点での被引用数と後のダウンロード数は相関しない
(2) 被引用数と機関リポジトリにおけるダウンロード数の算出元となるサンプルの違い

(1)についてはMoed[21]の先行研究から，引用された直後のダウンロード数増は一時的なもので全体で見ると大きな差はないことが指摘されている。Moedの分析対象はTetrahedron Letters誌の出版者プラットフォームであるが，機関リポジトリにおいても同様の傾向があてはまるために被引用数とダウンロード数が相関しなかったことが考えられる。

(2)についてはBollenらがUIFとJIFの比較分析の中で両者を算出する際のサンプルの違いに言及しており[22]，UIFがカリフォルニア州立大学の，教員だけでなく学部生やスタッフ，実務者も含む，すなわちローカルかつ研究者以外も含むサンプルに基づくものであるのに対し，JIFは学術論文の著者から成るグローバルなサンプルに基づく指標であり，そのサンプルごとの興味・関心の違いがダウンロード数に基づくUIFと引用数に基づくJIFの違いの一因となっていることを指摘している。機関リポジトリの利用者は教育・研究機関以外のドメインからの来訪者を含むこと，一方で(日本からのアクセスが中心ではあるものの)世界中から利用があることが佐藤に指摘されている[11]。この「グローバルかつ研究者以外も含む」サンプルの関心が，被引用数のサンプルである学術論文の著者らの関心と異なるために，機関リポジトリにおけるダウンロード数と論文の被引用数の間には相関関係が見られないのではないかと考えられる(これに対し多くの特定雑誌における被引用数とダウンロード数の間に高い相関が見られるのは，論文出版直後に出版者のプラットフォームからダウンロードする利用者と論文を引用する著者がいずれも似たようなコミュニティから成っているためである，と考えられる)。

実際には(1)，(2)双方が理由となり，ある時点での論文の被引用数と機関リポジトリにおける後のダウンロード数は相関しないものと考えられる。「ある時点での被引用数から後に機関リポジトリからよくダウンロードされる論文を推測すること」は出来ず，利用の観点から見た場合，被引用数は機関リポジトリに重点的に収集すべきコンテンツを定める際の材料には出来ないと言える。機関リポジトリ収録後，よくダウンロードされる論文を推測するには被引用数以外の要因を探すことが賢明である。

今後の課題としては被引用数以外の機関リポジトリにおけるダウンロード数に関係しうる要因(書かれた言語，文献種別など)について検討する必要がある。また，本稿では「ある時点でのダウンロード数と後の被引用数の相関」については扱わないこととしたが，今後はデータの整備を待って「同一期間中のダウンロード数と被引用数の関係」，「ある時点でのダウンロード数と被引用数の関係」についても同様に分析することが考えられる。

謝辞

本研究は「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」および国立情報学研究所次世代学術コンテンツ基盤共同構築事業委託事業(領域2)「機関リポジトリへの登録が学術文献流通に対して及ぼす効果についての定量的解析のための文献蓄積及びデータ整理」による支援により行われたものである。

引用文献

1) Lynch, Clifford A. “Institutional repositories : essential infrastructure for scholarship in the digital age”. ARL Bimonthly Report. 2003, 226, http://www.arl.org/resources/pubs/br/br226/

br226ir.shtml,（参照 2009-04-20）.
2） 国立情報学研究所.“学術機関リポジトリ構築連携支援事業”. http://www.nii.ac.jp/irp/,（参照 2009-04-20）.
3） Brody, Tim.“Registry of Open Access Repositories（ROAR）”. http://roar.eprints.org/,（参照 2009-04-20）.
4） 国立情報学研究所.“機関リポジトリ一覧”. 学術機関リポジトリ構築連携支援事業. http://www.nii.ac.jp/irp/list/,（参照 2009-04-20）.
5） 国立情報学研究所.“IRDB コンテンツ分析システム”. http://irdb.nii.ac.jp/analysis/index.php,（参照 2009-04-20）.
6） Brody, Tim.“About the project”. Interoperable Repository Statistics. http://irs.eprints.org/about.html,（参照 2009-04-20）.
7） PIRUS : Publisher and Institutional Repository Usage Statistics. "Developing a global standard to enable the recording, reporting and consolidation of online usage statistics for individual journal articles hosted by institutional repositories, publishers and other entities（Publisher Metadata and Interoperability Project 3）Final Report". 2009, 20p., http://www.jisc.ac.uk/media/documents/programmes/pals3/pirus_finalreport.pdf,（参照 2009-04-20）.
8） 千葉大学附属図書館.“機関リポジトリ評価のための基盤構築”. http://www.ll.chiba-u.ac.jp/~joho/CSI/standardization.html,（参照 2009-04-20）.
9） Royster, Paul. Publishing original content in an Institutional repository. Serials Review. 2008, vol.34, no.1, p.27-30.
10） 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.
11） 佐藤翔.“誰が，何を読んでいるのか：アクセスログに基づく機関リポジトリの利用実態”. SPARC-Japanセミナー2008「日本における最適なオープンアクセスとは何か？」. 東京，2008-10-14，SPARC-Japan, 2008, http://www.nii.ac.jp/sparc/event/2008/20081014.html ,（参照 2009-04-20）.
12） 佐藤翔, 逸村裕. アクセスログから見るARRIDEの利用状況. アジ研ワールド・トレンド. 2009, no.162, p.10-12.
13） 池田大輔, 星子奈美, 井上創造. 外部連携サービスによる機関リポジトリの潜在需要の解析. ディジタル図書館. 2009，no.36，p.62-68.
14） Tenopir, Carol.“Use and users of electronic library resources: an overview and analysis of recent research studies”. Report for the Council on Library and Information Resources. 2003, http://www.clir.org/pubs/reports/pub120/pub120.pdf,（参照 2009-04-20）.
15） O'Leary, Daniel E. The relationship between citations and number of downloads in Decision Support Systems. Decision Support Systems. 2008, vol.45, p.972-980.
16） Watson, A. B. Comparing citations and downloads for individual articles. Journal of Vision. 2009, vol.9, no.4, p.1-4.
17） Perneger, Thomas V. Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ. BMJ. 2004, vol.329, p.546-547.
18） Deciphering citation statistics. Nature Neuroscience. 2008, vol.11, no.6, p.619.
19） Brody, Tim; Harnad, Stevan; Carr, Leslle. Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology. 2006, vol.57, no.8, p.1060-1072.
20） Davis, Philip M.; Fromerth, Michael J. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles?. Scientometrics. 2007, vol.71, no.2, p.203-215.
21） Moed, Henk F. Statistical relationships between downloads and citations at the level of individual documents within a single journal. Journal of the American Society for Information Science and Technology. 2005, vol.56, no.10, p.1088-1097.
22） Bollen, Johan; Van de Sompel, Herbert. Usage Impact Factor: the effects of sample characteristics on usage-based impact metrics. Journal of the American Society for Information Science and Technology. 2008, vol.59, no.1, p.136-149.
23） 国立情報学研究所.“次世代学術コンテンツ基盤共同構築事業 学術機関リポジトリ構築連携支援事業 平成 20-21 年度委託事業公募要項”. http://www.nii.ac.jp/irp/rfp/2008/kobo_yoko2008-2009.pdf,（参照 2009-4-20）.
24） Counting Online Usage of Networked Electronic Re-

sources. “The COUNTER Code of Practice. Journals and Databases. Release 3”. 2008, 38p., http://www.projectcounter.org/r3/Release3D9.pdf,（2009-04-20 入手）.

（平成21年 4 月30日受付）
（平成21年 8 月27日採録）