2009年7月9日（木）

平成20年度　CSI委託事業報告交流会（コンテンツ系）　機関リポジトリの更なる普及と新たな価値創出に向けて

ZSプロジェクト

リポジトリ登録は、次の引用を喚起するか：これまでの成果と今後の課題

参加機関：北海道大学、京都大学、筑波大学、日本動物学会（データ提供）
千葉大学、金沢大学、大阪大学、広島大学
報告者：筑波大学大学院図書館情報メディア研究科
佐藤翔　（min2fly@slis.tsukuba.ac.jp）

それでは、ZSプロジェクトについて発表いたします。

報告は、筑波大学図書館情報メディア研究科の佐藤翔が行わせていただきます。

概要

1. ZSプロジェクトとは？
2. これまでの成果
   - リポジトリの利用状況分析
   - ダウンロード数と被引用数の関係
3. 今後の課題

2

発表の概要はこちらの通りです。
最初に、ZSプロジェクトとは何かについて。

1. ZSプロジェクトとは？

- プロジェクトの目的
  - 機関リポジトリによるオープン・アクセス（OA）の効果を検証する
  - 「リポジトリに登録することが新たな引用を引き起こすか？」
- 参加機関
  - 北海道大学、京都大学、筑波大学、日本動物学会（データ提供）
  - 千葉大学、金沢大学、大阪大学、広島大学

3

ZSプロジェクトは日本動物学会発行の雑誌、”Zoological Science”掲載論文を対象に、「機関リポジトリによるオープンアクセスの効果」を検証すること、具体的には「機関リポジトリに登録することは論文の新たな引用を引き起こすか」を検証することを目的とするプロジェクトです。
北大、京大、筑波大、日本動物学会を中心にご覧の機関に参加いただいています。

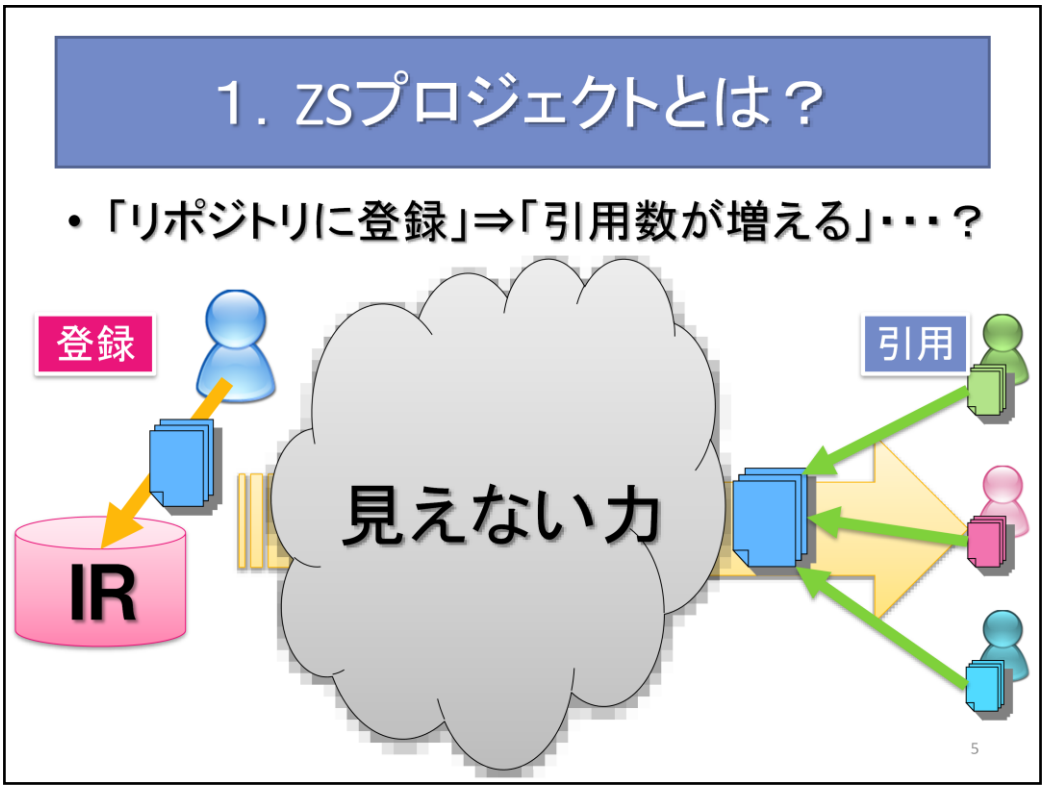
1. ZSプロジェクトとは？

- OAの被引用数増効果に関する研究は多い
  - 分野別リポジトリ（arXiv等）
  - 電子ジャーナルの部分公開
  - OAであれば対象を限定しない
- 機関リポジトリに限定した研究はない
  - 分野別リポジトリや電子ジャーナルでの結果を援用できるようにも考えられない
  - 「リポジトリに登録することの効果」を検証する必要

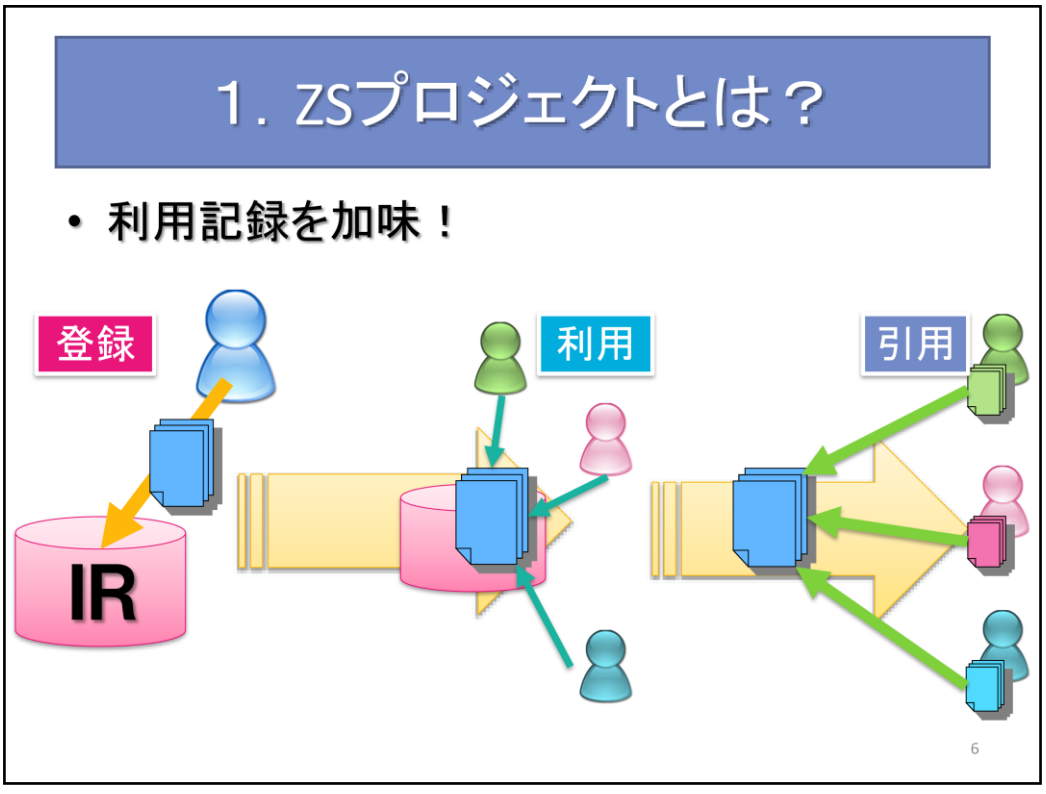
4

これまで「オープンアクセスは被引用数を増加させるか」ということについては、それが研究者にとって論文をオープンアクセスにする動機になると考えられていることもあり、多くの研究で検証されてきました。
しかしそのほとんどはarXivなどの分野別リポジトリ登録文献や電子ジャーナル掲載論文の一部を無料で見られるようにする、あるいはオープンアクセスであれば手段は問わず対象とするもので、機関リポジトリに限定して効果を見た研究はありません。

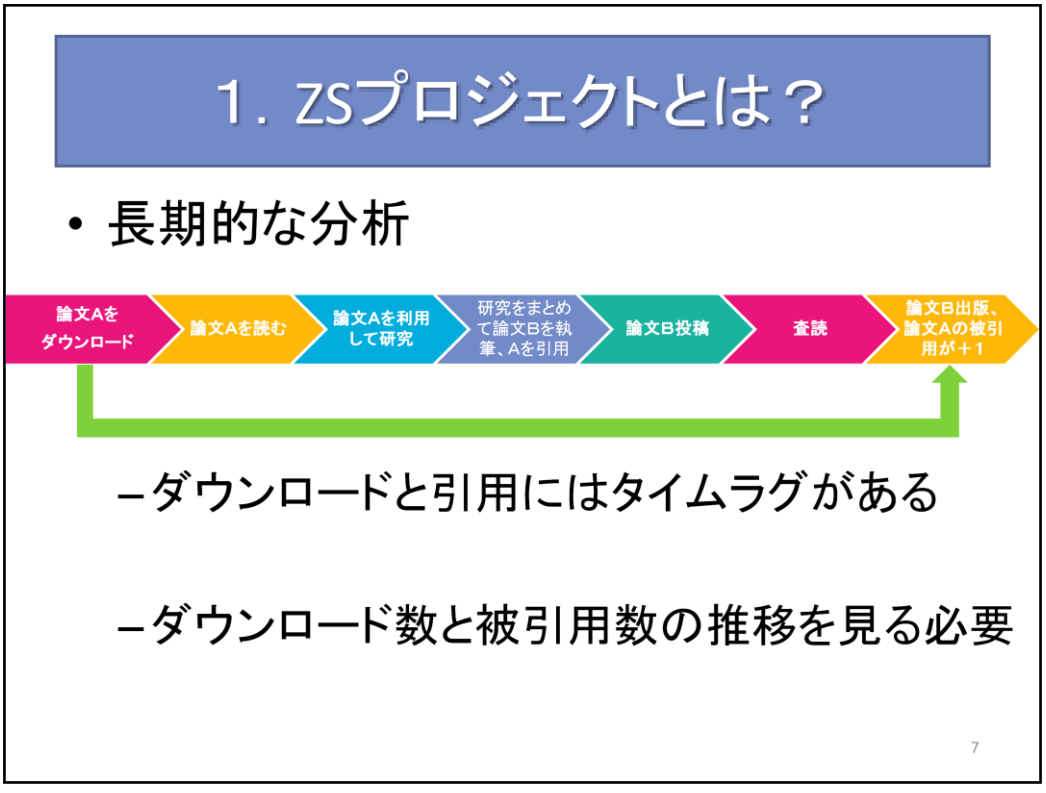
「オープンアクセスの効果がわかれば十分だ」とも考えられるかも知れませんが、もともと研究者が良く使うプラットフォームである分野別リポジトリや電子ジャーナルを使ったオープンアクセスについての研究成果が、機関リポジトリにもあてはまるかは疑問が残ります。
本当に機関リポジトリに登録することがなんらかの効果を持つのか、OA一般ではなく機関リポジトリに限定した検証を行うことがZSプロジェクトの目的です。



ZSプロジェクトのもう一つの特徴は利用記録をあわせて分析することです。
機関リポジトリに登録された論文が実際に使われたかどうかを見ずに、被引用数が増えるかどうかを分析する…というのでは、仮に引用が増えていたとしてもそれがリポジトリ掲載論文が読まれたからなのか、何か違う理由があるのかがわかりません。
論文の登録と引用の間にあるはずのステップがこれだと見えなくなってしまうわけです。



そうではなく、リポジトリに載せたことで、これまでその文献を読めなかった、あるいは読まなかった人も読むようになったので引用が増えたんだ…ということを証明するために、ZSプロジェクトでは被引用数データだけでなくリポジトリのアクセスログも加えた分析も行います。



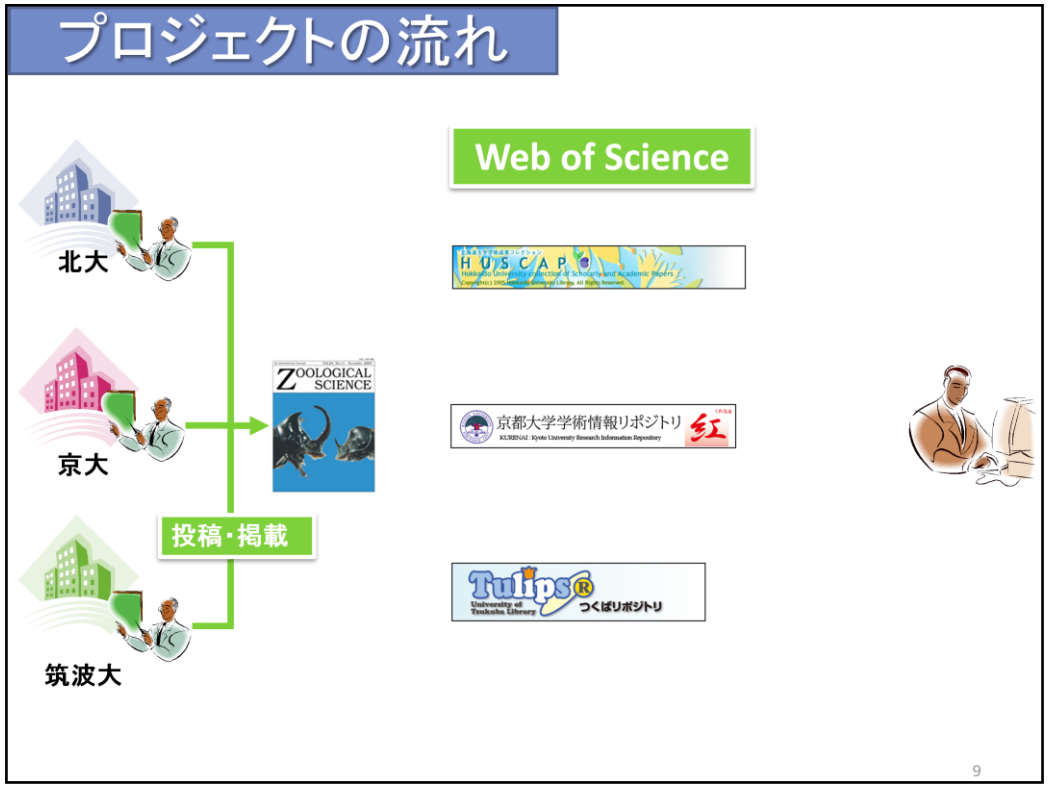
さらに、ZSプロジェクトでは、短期的な効果を見るのではなく長期間、継続して分析を行っていきます。
論文がダウンロードされてから引用されるまでにはここに示したような経緯を経るはずで、相当のタイムラグがあります。
特に今回対象とする動物学分野では、論文が出版されてから引用されるまでのスパンがかなり長い傾向もあり、あまり短期間でリポジトリ掲載の効果が出るようだとかえって不自然と考えられます。
そこでZSプロジェクトではある程度、長期間にわたってダウンロード数と被引用数がどう推移するのかを分析していく計画です。

“Zoological Science”とは？

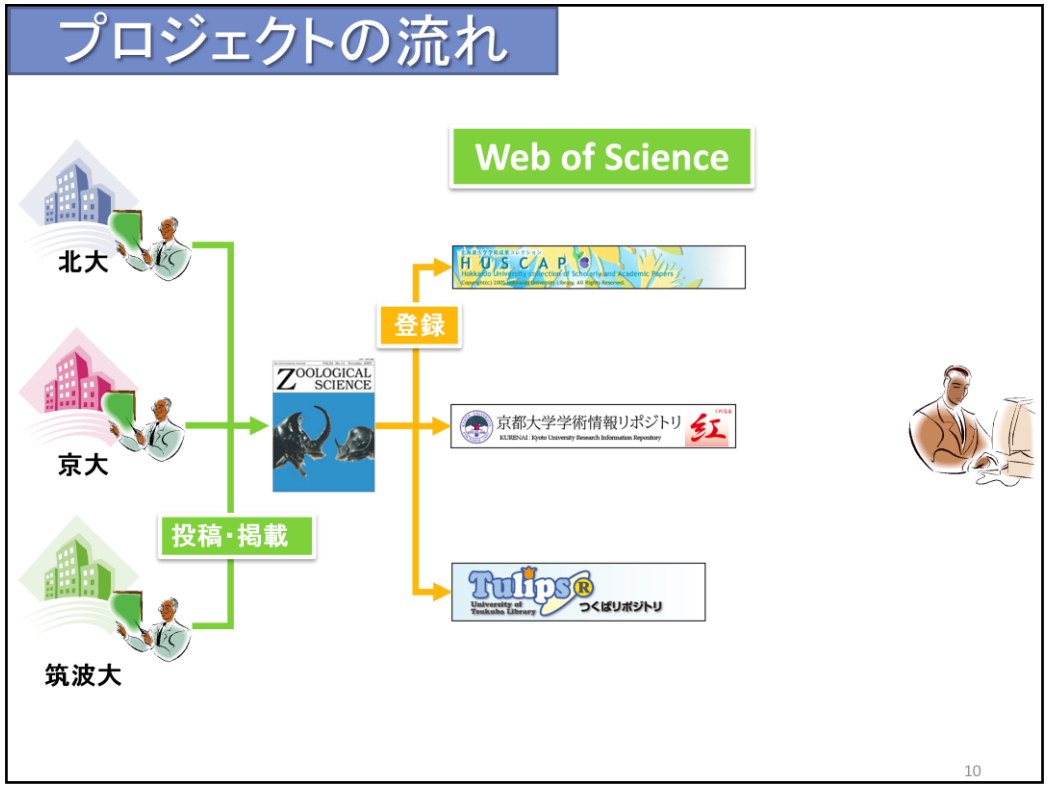
- 日本動物学会発行の月刊誌
- 動物学雑誌（創刊1888年）、動物学彙報（創刊1897年）の統合誌として 1984年に創刊
- 動物学分野の原著論文、レビュー、エッセイ、短報を掲載
- 電子版はBioOne2, UniBio Press等から利用可能
- JCR2008によるデータ
  - 掲載論文数：152
  - IF：1.100, 5年IF：1.227
  - 引用半減期：6.4年

8

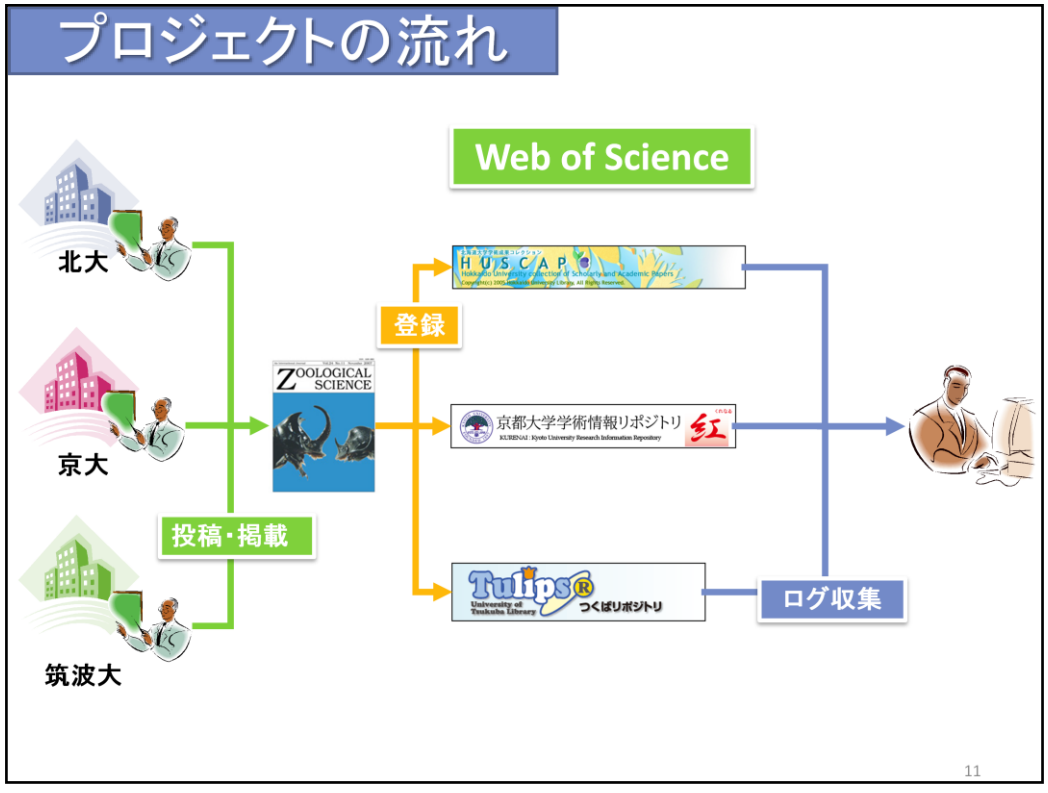
ここで分析対象である雑誌、”Zoological Science”について簡単に紹介します。
Zoological Scienceは日本動物学会発行の英文月刊誌で、1984年に「動物学雑誌」と「動物学彙報」が統合する形で創刊しました。
動物学分野の原著論文、レビュー、短報など毎年150前後の論文が掲載され、冊子の他にBioOne2, UniBio Press等から電子版を利用することが出来ます。
Journal Citation Reports2008によると、2008年のIFは1.1、５年IFは1.227と数年前の論文がよく引用される傾向があるようです。



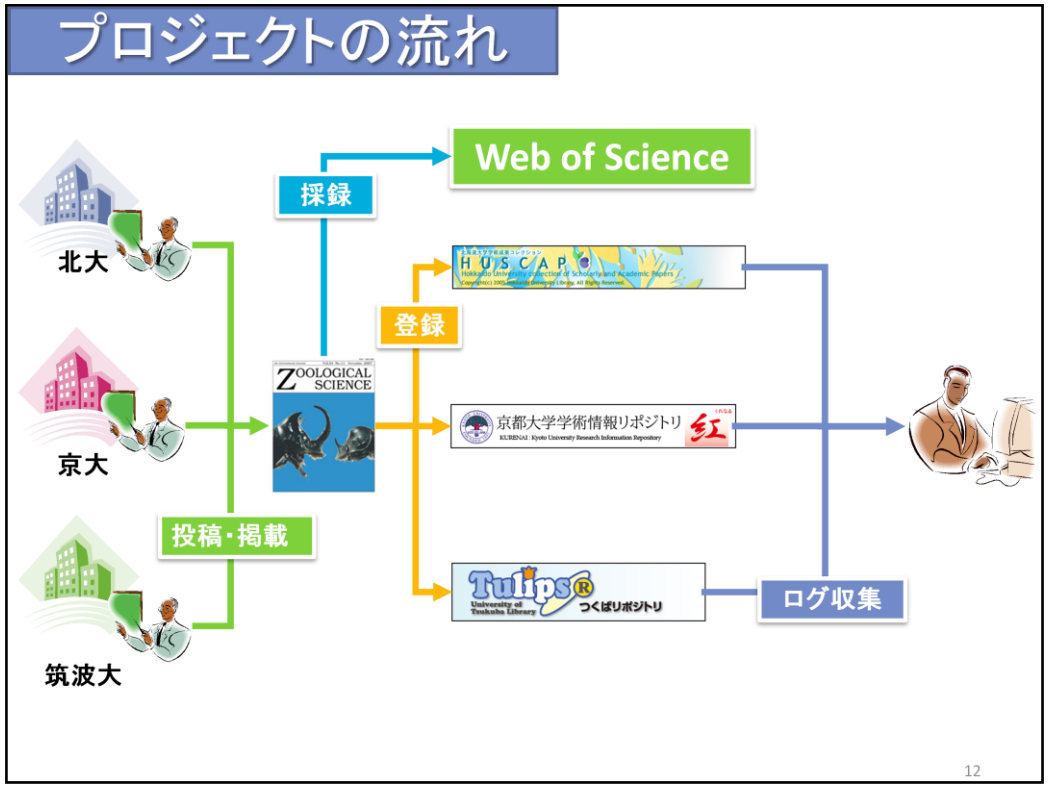
次に実際のプロジェクトの流れですが、まず北大、京大、筑波大の先生方が”Zoological Science”に投稿した論文について、しかるべき期間を置いた後に



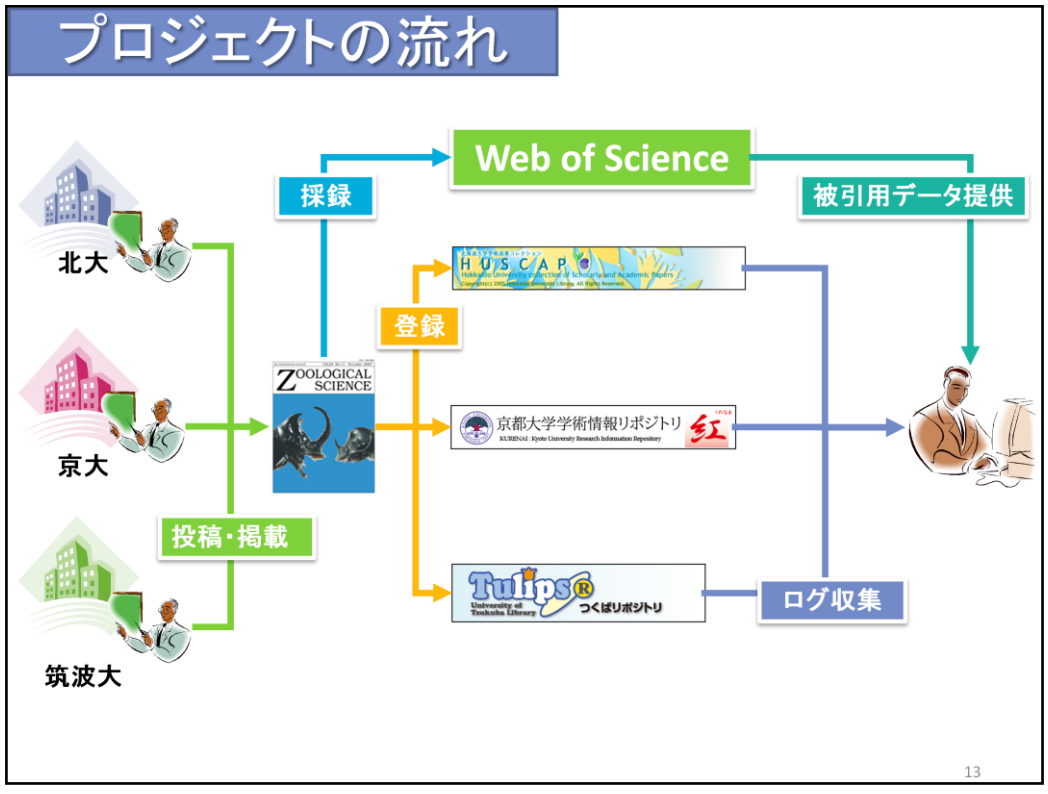
それぞれの大学のリポジトリに登録します。



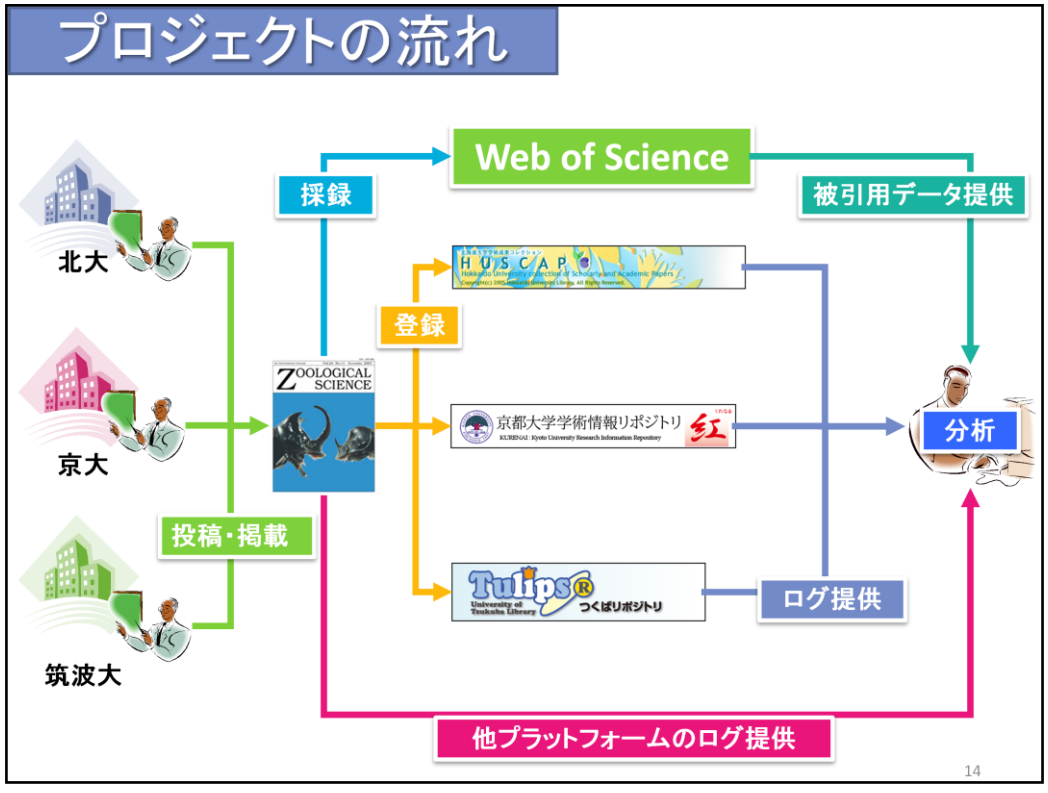
登録後、一定期間ごとに各リポジトリのアクセスログを集め、リポジトリ上でZS論文がどれだけ使われていたかを取りまとめます



一方でZoological ScienceはWeb of Scienceの収録対象誌でもあり、各論文がどれだけ引用されたかはWeb of Science側に記録が残っています



この被引用データについても提供いただき、リポジトリのログと合わせて分析します



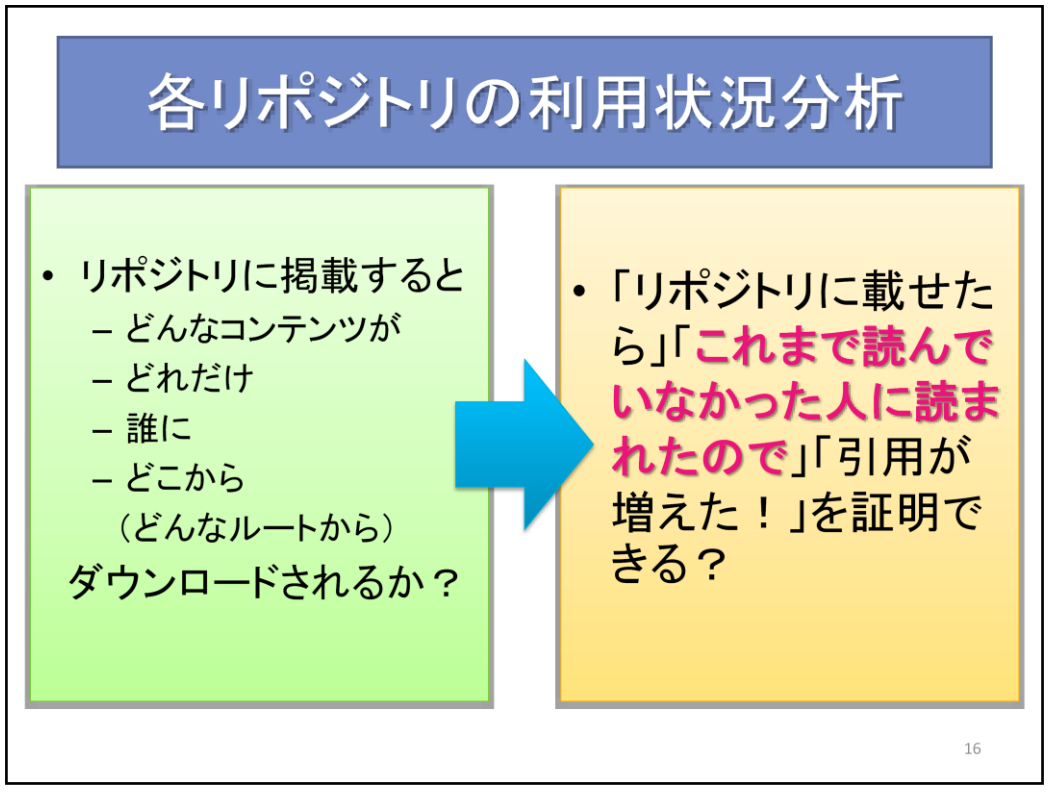
これに加えてBioOneなどの電子ジャーナル側での利用ログについても提供いただくことで、リポジトリに限らず各文献が全体としてどのように利用され、引用されているかを追いかけることが出来るようになります。
これが本プロジェクトの概要となります。

2. これまでの成果

- 各リポジトリの利用状況分析
- ダウンロード数と被引用数の関係

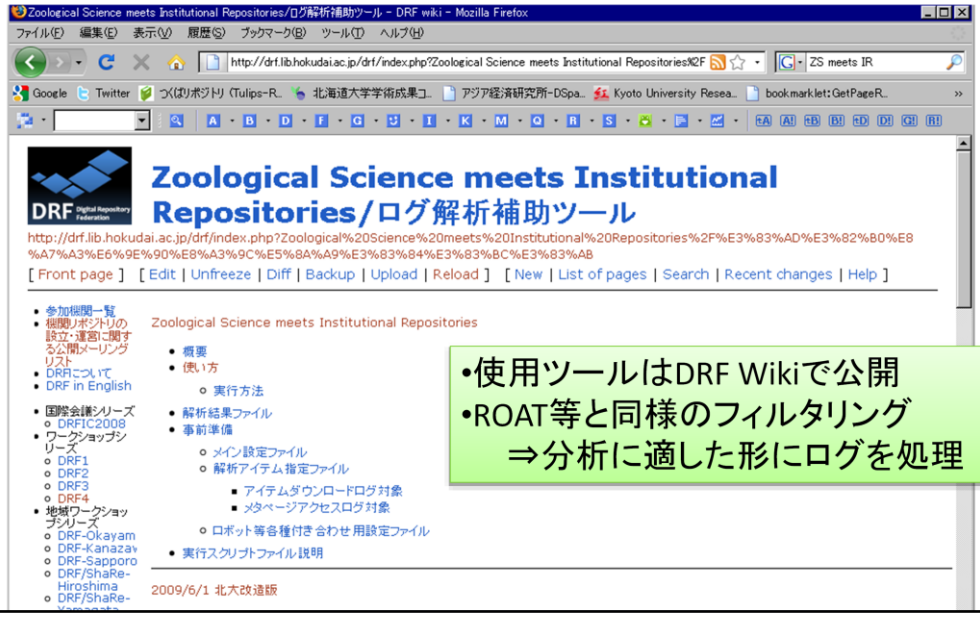
15

次に現在までの成果についてですが、現在までに各リポジトリのログ分析の手法を確立し、リポジトリ登録直後段階でのダウンロード数と被引用数の関係について分析をはじめています。



先に言ったように本プロジェクトではリポジトリに論文を収録したことで「これまで読んでいなかった人が読むようになった」か否かが焦点の一つになります。
それを知るためには、リポジトリに掲載した論文のうちどんなコンテンツが、誰に、どうやって使われているのか・・・例えば購読機関の人なのかそれ以外の人なのかとか、あるいはサーチエンジンからなのかデータベースからの利用なのかと言った利用状況の詳細を見ていく必要があります。

各リポジトリの利用状況分析



- 使用ツールはDRF Wikiで公開
- ROAT等と同様のフィルタリング
  ⇒分析に適した形にログを処理

この分析に使うためのツールは昨年度末に北大の野中さんたちが開発し、現在DRF Wikiにて公開されています。
これはROAT等と同じようにアクセスログのフィルタリングを実施した上で、分析に適した形にログを整形処理してくれるというものです。

＊補助ツールの主な機能

- 本文（bitstream）、メタデータ（handle）へのアクセスのみ抽出
- フィルタリングの実施
- IPアドレスからのドメイン解決
- 参照元種別の記録（限定的）
- 検索語のエンコード
- 分析用のログの切り離し

18

具体的な機能としては生のアクセスログから本文あるいはメタデータへのアクセスを抽出した上で検索ロボットのアクセス等をフィルタリングするとともに、
IPアドレスから相手のドメイン名を取得する、
参照元がサーチエンジンや他のサービスなのか等を判別するなど色々あります。
詳細についてはWikiをご確認いただければ幸いです。

リポジトリ全体

- HUSCAP, KURENAI, Tulips-R
  ＋ ARRIDE（アジア経済研究所）のログを分析
  - アクセス元ドメイン（所属機関種別、地域）
  - アクセス方法（参照元、検索キーワード）
  - アクセス先の詳細情報

19

このツールを使って、現在ZSプロジェクト参加３機関に加え、ご協力いただけることになったアジア経済研究所のリポジトリ、ARRIDEを含めた4つのリポジトリのログの分析を進めています。
主な分析内容はアクセス元の所属機関や地域等のドメイン、参照元などのアクセス方法、そしてアクセス先の文献種別や出版年などとアクセス数の関係です。

表．　文献タイプごとの平均ダウンロード数（全体およびドメイン別）

文献種別	搭載文献数	全体	ac,edu	co,com	ne,net
Journal Article	5,303	35.69	4.63	3.83	8.26
Thesis or Dissertation	7,330	19.27	3.46	1.88	6.65
Departmental Bulletin Paper	42,782	15.26	2.04	1.32	5.83
Conference Paper	354	52.67	5.18	7.08	15.83
Presentation	185	37.59	6.93	3.05	11.94
Book	98	143.48	25.09	10.22	56.69
Technical Report	179	60.65	5.79	7.59	11.77
Research Paper	329	46.36	6.36	4.86	15.23
Article	2,802	12.23	1.33	1.06	5.24
Preprint	42	2.38	0.33	0.33	0.60
Learning Material	35	956.54	167.63	64.74	439.40
Data or Dataset	6	36.17	6.00	2.17	13.33
Software	2	35.50	4.00	3.50	11.50
Others	2,976	11.16	1.25	1.10	4.00
合計	62,423	18.43	2.55	1.70	6.48

20

これは実際に4つのリポジトリのログから、junii2における文献種別ごとの2008年のダウンロード数の平均を、全体およびドメイン別に示したものです。

表.　文献タイプごとの平均ダウンロード数（全体およびドメイン別）

文献種別	搭載文献数	全体	ac,edu	co,com	ne,net
Journal Article	5,303	35.69	4.63	3.83	8.26
Thesis or Dissertation	7,330	19.27	3.46	1.88	6.65
Departmental Bulletin Paper	42,782	15.26	2.04	1.32	5.83
Conference Paper	354	52.67	5.18	7.08	15.83
Presentation	185	37.59	6.93	3.05	11.94
Book	98	143.48	25.09	10.22	56.69
Technical Report	179	60.65	5.79	7.59	11.77
Research Paper	329	46.36	6.36	4.86	15.23
Article	2,802	12.23	1.33	1.06	5.24
Preprint	42	2.38	0.33	0.33	0.60
Learning Material	35	956.54	167.63	64.74	439.40
Data or Dataset	6	36.17	6.00	2.17	13.33
Software	2	35.50	4.00	3.50	11.50
Others	2,976	11.16	1.25	1.10	4.00
合計	62,423	18.43	2.55	1.70	6.48

昨年10月のSPARC-Japanセミナー等では「アクセス上位には雑誌掲載論文は少ない」とお話ししてきましたが、上位文献に限らず登録文献全体を見ると雑誌掲載論文は学位論文や紀要論文以上に利用されています。

雑誌論文が平均的にどれも一定以上のアクセスを得ているのに対し、紀要論文はよく利用されるものも多い一方で全く利用されないものも相当数あるため、平均して見ると利用が少なくなるようです。

また、博士論文は京大では平均90回以上利用されているのですが、筑波や北大では利用が少ないため全体で見るとアクセスが少なくなっています。

表.　文献タイプごとの平均ダウンロード数（全体およびドメイン別）

文献種別	搭載文献数	全体	ac,edu	co,com	ne,net
Journal Article	5,303	35.69	4.63	3.83	8.26
Thesis or Dissertation	7,330	19.27	3.46	1.88	6.65
Departmental Bulletin Paper	42,782	15.26	2.04	1.32	5.83
Conference Paper	354	52.67	5.18	**7.08**	15.83
Presentation	185	37.59	6.93	3.05	11.94
Book	98	143.48	25.09	10.22	56.69
Technical Report	179	60.65	5.79	**7.59**	11.77
Research Paper	329	46.36	6.36	4.86	15.23
Article	2,802	12.23	1.33	1.06	5.24
Preprint	42	2.38	0.33	0.33	0.60
Learning Material	35	956.54	167.63	**64.74**	439.40
Data or Dataset	6	36.17	6.00	2.17	13.33
Software	2	35.50	4.00	3.50	11.50
Others	2,976	11.16	1.25	1.10	4.00
合計	62,423	18.43	2.55	1.70	6.48

アクセス元のドメイン別でもよく利用される論文の傾向は異なり、例えばco,comドメイン、すなわち企業内からの利用は全体で見ると大学や個人の利用よりも少ないのですが、会議発表論文やテクニカルレポートへのアクセスは大学からより多くなっています。一方で教材やに対するアクセスは他のドメインでは特別多いのに対し、企業ドメインからは多いことは多いのですがそれほどでもない…と言ったかたちです。

ZS誌の利用状況

表. ZS誌・ドメインごとのDL統計（2008年）

	全体	ac,edu	co,com	ne,net
コンテンツ数	171	171	171	171
総ダウンロード数	5047	558	572	1121
平均	**29.51**	3.26	**3.35**	**6.56**
中央	20	2	2	4
最小	0	0	0	0
最大	**210**	22	47	51

表. ZS誌・参照元ごとのDL統計（2008年）

	直接ＤＬ	リポジトリ内部	サーチエンジン	その他
総ダウンロード数	1112	449	3361	122
平均	6.51	2.63	**19.65**	0.71
中央	5	2	13	0
最小	0	0	0	0
最大	48	17	153	47

Zoological Science掲載論文に絞ってみると、こちらはTulips-Rは含んでいない、HUSCAPとKURENAI掲載分についての2008年の分析結果を示したものですが、平均すると1本あたり30回近く、多いものでは年に200回以上ダウンロードされています。
ドメイン別では大学よりも企業、個人からのアクセスが多く、アクセス手法別ではサーチエンジンからの利用がメインです。
現在はまだ購読機関からのアクセスか否かの分析は行っていませんが、企業や個人ドメインからのアクセスが多いことを考えるとおそらくは大部分の利用はZoological Scienceを購読していない、これまでは読めなかった利用者が、サーチエンジンの検索結果からアクセスしてきているのではないかと考えられます。

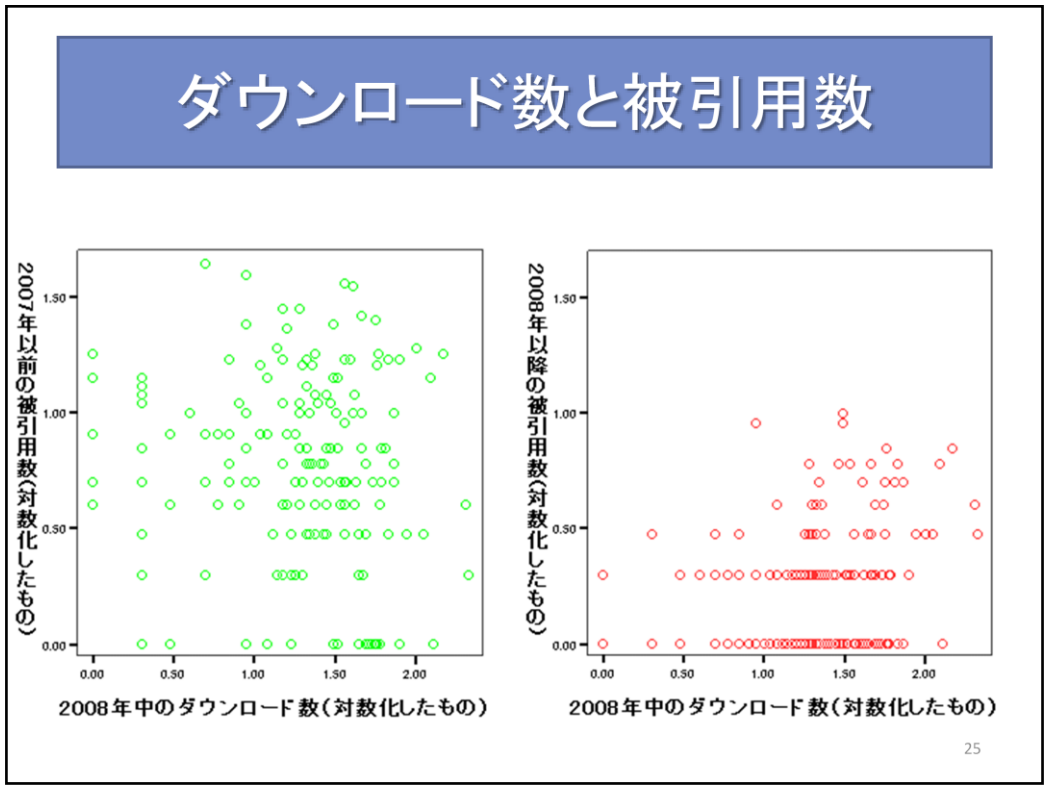
ダウンロード数と被引用数

- HUSCAP, KURENAI収録のZS誌論文について、
  - 2008年のダウンロード数
  - 2007年以前の被引用数
  - 2008年以降の被引用数
  - 論文の出版年

の関係を見てみた

24

さらにこれらの文献について、Web of Scienceにおける被引用数および出版年との関係を見てみると、



こちらは左が2008年のダウンロード数と2007年以前の被引用数、右が2008年のダウンロード数と2008年以降の被引用数、それぞれに1を加えて対数をとったものを散布図に示したものです。左がリポジトリに載せる前の被引用数とリポジトリ掲載後のダウンロード数の関係を示したもの、右がリポジトリに載せた後の被引用数とダウンロード数の関係を示しているということになります。左はご覧のように完全に分散しているのに対して、右は被引用数がまだ多くないために下に偏っていますが、やや右上がりの傾向があります。

ダウンロード数と被引用数

表. ダウンロード数・被引用数・出版年間のスピアマンの順位相関係数

	2008年中のダウンロード数	2008年以降の被引用数	2007年以前の被引用数	出版年
2008年中のダウンロード数		0.37(**)	−0.05	0.55(**)
2008年以降の被引用数			0.28(**)	0.32(**)
2007年以前の被引用数				−0.39(**)
出版年				

** 印は$p<0.01$で有意な相関あり

26

そこで両者の相関の有無を確認して見ると、登録後の被引用数と登録後のダウンロード数には0.37と弱いながらも有意な相関があるのに対して、登録前の被引用数とダウンロード数には有意な相関はありません。

つまりリポジトリに登録する前にどれだけ引用されていたかと、登録してからどれだけ使われるかは無関係で、「良く引用されていた論文」は必ずしもリポジトリで「よく利用される論文」ではない、と言えます。

一方でリポジトリ登録後は一見ポジティブな関係があるわけですが、最初の方で言ったようにダウンロードされてから引用されるまでにはタイムラグがあるはずなので、これが即「リポジトリ上でよく使われると被引用数が増える」ことを意味するとは考えられません。

ダウンロード数と被引用数

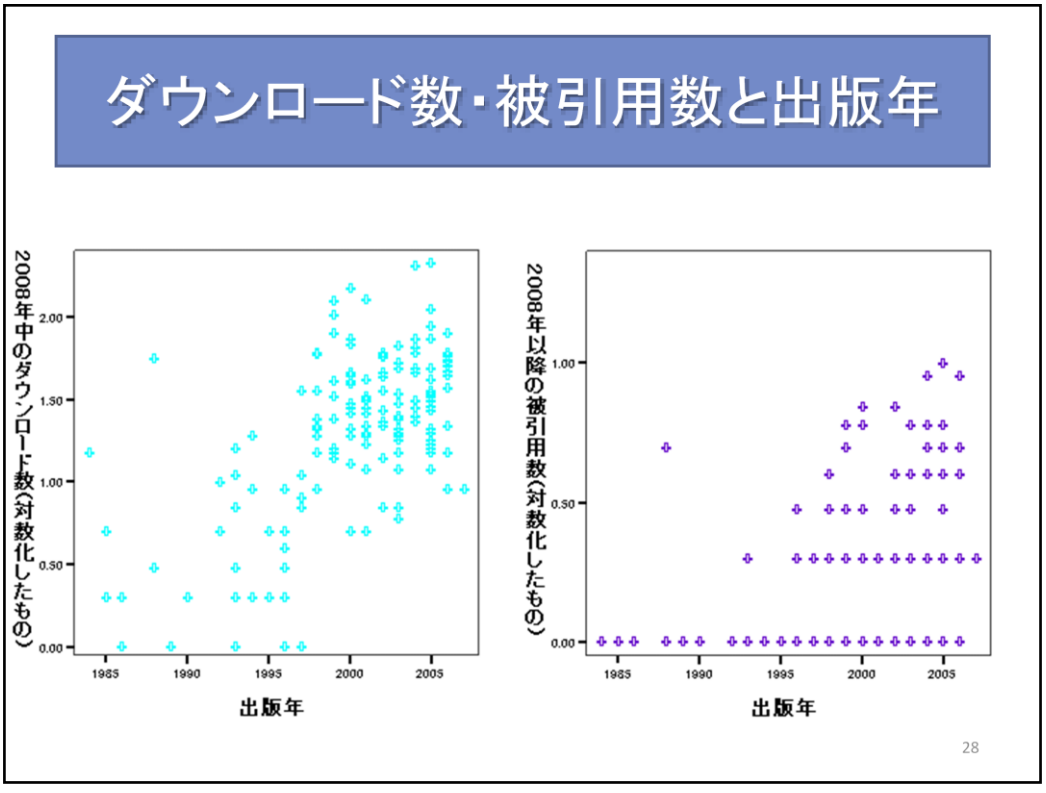
表. ダウンロード数・被引用数・出版年間のスピアマンの順位相関係数

	2008年中のダウンロード数	2008年以降の被引用数	2007年以前の被引用数	出版年
2008年中のダウンロード数		0.37(**)	−0.05	0.55(**)
2008年以降の被引用数			0.28(**)	0.32(**)
2007年以前の被引用数				−0.39(**)
出版年				

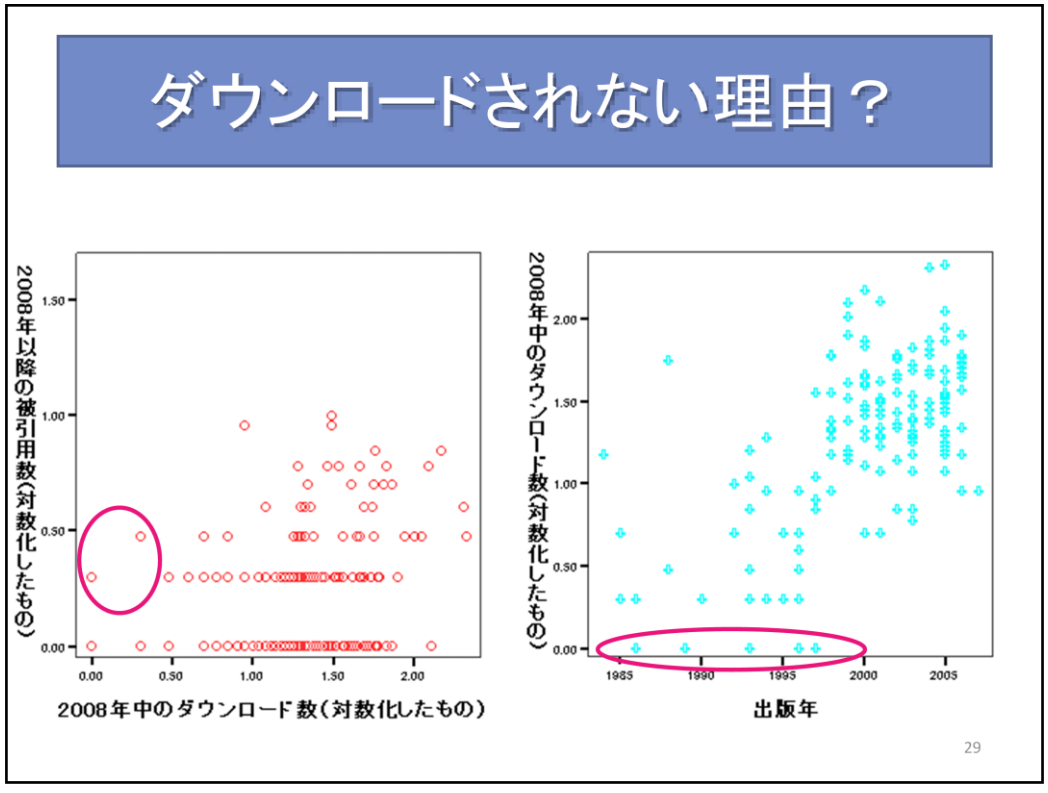
** 印は$p<0.01$で有意な相関あり

27

今回の場合、2008年中のダウンロード数と被引用数はどちらも出版年と相関関係にあり、これがダウンロード数と被引用数の相関に影響していると考えられます。



こちらは左が出版年とダウンロード数、右が出版年と2008年以降の被引用数の関係を示したものですが、ご覧のようにどちらも右上がりの傾向があり、特にダウンロード数については顕著です。
今のところはリポジトリでよく読まれたから引用数が増えたというよりも、リポジトリ上での利用も、Web of Science上での引用もどちらも新しいものの方が多いだけ…と考えた方が妥当でしょう。
今後はもっと長いスパンで、ある時点でダウンロードが増え出した論文の被引用数がその後どうなるか、その時ほかのプラットフォームでの利用はどうだったか…といったことを見ていく必要があります。



ところで被引用数の話からはやや外れるのですが、Zoological Science掲載論文の中で一年間全くアクセスされていないコンテンツも実は相当数あります。
特にHUSCAPで顕著で、2008年中に登録して2009年５月まで１回もアクセスされていないという論文も複数あります。
古くなって誰も興味を持たなくなった結果なのか…とも思いましたが、一方で全くアクセスがないものの中にも引用はされ続けている論文もあり、どうも論文の中身以外の理由がありそうです。

ダウンロードされない理由？

表5. ダウンロード数上位10位のテキスト付与状況

	全体	HUSCAP	KURENAI
テキスト付与コンテンツ数	10	10	10
テキスト付与率	100.0%	100.0%	100.0%

表6. ダウンロード数下位10位のテキスト付与状況

	全体	HUSCAP	KURENAI
テキスト付与コンテンツ数	0	0	1
テキスト付与率	0.0%	0.0%	10.0%

＊ファイル形式はすべてPDF

30

そこでアクセス下位10件と上位10件について、全体およびHUSCAP、KURENAI別で調べてみたところ、アクセスが多いコンテンツは１件の例外もなくテキスト付PDFであるのに対し、アクセスが少ない方は全体ではすべてイメージのみのPDF、リポジトリ別にみてもKURENAIの１件を除きすべてテキストがついていないPDFでした。
アクセスの大部分はサーチエンジンの検索から来ており、テキスト化しないことはリポジトリにとって致命的である…という話は昨年のSPARC-Japanセミナーでもさせていただきましたが、今回あらためて強調させていただきたいと思います。
HUSCAPの方はカバーページをつける等して一応、タイトル中に含まれる言葉からの検索はできるようになっているのですが、それでも全文が検索できる場合に比べるとアクセス数は天地の差で、テキストデータを付与しなかったりコピー不許可のPDFを登録することは論文が利用される機会のほとんどを失ってしまうことになっています。
また、こういう論文の内容とは関係のない理由でアクセス数が大きく変わってしまう点については分析時に注意を要するとも言えます。

3. 今後の課題

- アクセスログ側の詳細分析
  - 雑誌購読の有無
  - 地域の詳細（途上国or先進国？）
- 長期的な変動の観察
  - 動物学は引用のスパンが長い分野

31

最後に今後の課題ですが、アクセスログ側については先にも言ったようにZoological Scienceを購読している人としていない人のアクセスを分けて分析していきます。
また、被引用数とダウンロード数の関係については長期的にみてどう変動していくのかを観察していくことが重要になります。
動物学は引用のスパンが長い分野…とのことですし、節目ごとに結果をまとめつつ気長にやっていく必要がありそうだと考えています。

3. 今後の課題

- アクセスログ側の詳細分析
  - 雑誌購読の有無
  - 地域の詳細（途上国or先進国？）
- 長期的な変動の観察
  - 動物学は引用のスパンが長い分野
- まあ**気長に**やっていきましょう

32

僕も博士後期課程に進む予定ですので、
これからもしばらくはこのプロジェクトに
関わらせていただきたいと思っています。

データ提供をいただいた
各機関に感謝いたします

発表は以上です。
なお、今回の発表ではZSプロジェクト参加機関以外にアジア経済研究所からもデータをご提供いただいています。
この場を借りてお礼申し上げたいと思います。

参加機関名一覧

- CSI委託事業として
  - 北海道大学、京都大学、筑波大学、日本動物学会（データ提供）
  - 千葉大学、金沢大学、大阪大学、広島大学
- アジア経済研究所
- トムソン・ロイター（被引用数データ）