Unpaywallを使用した日本のOAについての調査

西岡　千文（京都大学附属図書館）

2020年6月8日（月） NII学術情報基盤オープンフォーラム2020

京都大学
KYOTO UNIVERSITY

本日の発表

- 目的・内容
  - 日本全体と世界全体のオープンアクセス（OA）状況を比較して、日本のOA状況を理解する
  - 国際的に頻繁に利用されているUnpaywallのOA種別を利用した調査と、リポジトリに注目したOA状況の調査を実施する
  - （おまけ）機関リポジトリ（IR）のアクセスログを分析することで、「どのようなOA状態の論文がよく利用されているか」調査する
- 参考文献
  - 西岡千文, 佐藤翔. Unpaywallを利用した日本におけるオープンアクセス状況の調査. 2020. http://hdl.handle.net/2433/246424
  - 西岡千文.京都大学におけるオープンアクセス状況の調査. 2020. http://hdl.handle.net/2433/250151

データセット

手法

調査に使用するデータは、UnpaywallとScopusを利用して取得

- **日本の論文（2,000,897件）**
  - Scopusで日本に位置する機関2,611件を取得し、それらの機関の構成員が著者である論文をScopus Affiliation Retrieval APIで取得
  - うち、Unpaywallに収録されているかつ種別がjournal-article（雑誌論文）である論文2,000,897件を対象
- **世界の論文（79,534,697件）**
  - Unpaywallに収録されているかつ種別がjournal-article（雑誌論文）である論文79,534,697件を対象

【参考】Unpaywallは論文の合法的にオープンアクセスとなっている版を提供するWebブラウザの拡張機能であり、拡張機能が利用するデータベースは公開されている。Crossref DOIをもつ全文献（109,905,121件）のメタデータならびにオープンアクセスに関する情報等が収録されている。調査では2019年11月22日に作成されたデータベースのスナップショットを利用している。

論文のOA種別の特定 [1/2]

手法

2種類の調査を実施

- OAの全体像（UnpaywallのOA種別を使用）
- リポジトリによるOA

• OAの全体像

OA 種別	OA 種別が付与される条件
ゴールド（gold）	OA ジャーナルに掲載されている論文
ハイブリッド（hybrid）	購読型雑誌に掲載されているが， CC-BY 等のオープンライセンスが付与されている論文
ブロンズ（bronze）	上記いずれにも該当しないが，出版者で OA となっている論文
グリーン（green）	上記いずれにも該当せず，リポジトリのみで OA となっている論文
クローズド（closed）	OA でない論文
不明	フィールド oa_status が存在しない，または空値である論文

https://support.unpaywall.org/support/solutions/articles/44001777288-what-do-the-types-of-oa-status-green-gold-hybrid-and-bronze-mean-

- Unpaywallでは、各論文に上記のいずれかの値を付与
- 現在までにEU諸国でのOAのモニタリング等、様々な調査で利用

論文のOA種別の特定 [2/2]

手法

- **リポジトリによるOA**
  - 前頁のグリーンは「リポジトリのみでOAである論文」を指し、出版者でOAである論文はリポジトリで公開されたとしても考慮されない
  - 機関リポジトリ（IR）の「世界に対する大学の貢献を形作る新しいチャネル」や「機関の研究成果やその他リソースを発信」という役割に注目すると、出版者でOAである論文もIRに収録されることが望ましい
  - 著者のOAに対する意識を把握するためにも、プレプリントサーバ等その他のリポジトリでのOAについても調査することが求められる

日本の論文のみを対象

OA 種別	OA 種別が付与される条件
リポジトリで OA である論文	フィールド oa_locations に，host_type が repository であるオブジェクトが 1 件以上格納されている論文
機関リポジトリで OA である論文	フィールド oa_locations に，host_type が repository であるかつその url のネットワークロケーションが.jp で終わるオブジェクトが 1 件以上格納されている論文
機関リポジトリ以外のリポジトリで OA である論文	フィールド oa_locations に，host_type が repository であるかつその url のネットワークロケーションが.jp で終わらないオブジェクトが 1 件以上格納されている論文
機関リポジトリとその他リポジトリで OA である論文	上記の「機関リポジトリで OA である論文」・「機関リポジトリ以外のリポジトリで OA である論文」両方に該当する論文
機関リポジトリでのみ OA である論文（出版者でも OA ではない論文）	上記の「機関リポジトリで OA である論文」に該当し，フィールド oa_locations に格納されているオブジェクトが 1 件のみである論文

調査の欠点

手法

- Unpaywallは、リポジトリをクローリングしてグリーンである論文を収集している。
- 学術機関リポジトリデータベース（IRDB）に登録されている機関リポジトリ631件のうち、Unpaywallにカバーされているリポジトリは76件である。
- よって、多くの日本のリポジトリが対象となっておらず、結果で示す日本の論文のグリーンの割合やリポジトリによるOAの割合は実際よりも低い。

全体

結果

- OAの全体像
  - 論文のOAの割合は日本のほうが高い
  - 理由として、日本のブロンズの割合の高さが挙げられる
- リポジトリによるOA
  - 日本のほうが出版者でOAである論文をリポジトリによく登録している
  - 多くの論文が、IR以外のリポジトリで公開されている

OA 種別	日本（%）	世界（%）
ゴールド	7.33	7.67
DOAJ 掲載論文・ライセンス付	5.02	5.22
DOAJ 掲載論文・ライセンスなし	0.34	0.66
その他 OA ジャーナル掲載論文・ライセンス付	0.29	0.71
その他 OA ジャーナル掲載論文・ライセンスなし	1.68	1.09
ハイブリッド	3.09	3.37
ブロンズ	24.37	10.45
グリーン	7.04	8.28
クローズド	58.16	70.08
不明	0.01	0.14

OA 種別	日本（%）	世界（%）
リポジトリで OA である論文	19.90	17.57
機関リポジトリで OA である論文	2.36	—
機関リポジトリ以外のリポジトリで OA である論文	18.71	—
機関リポジトリとその他リポジトリで OA である論文	1.17	—
機関リポジトリでのみ OA である論文（出版者でも OA ではない論文）	0.83	—

リポジトリでOAである論文の割合

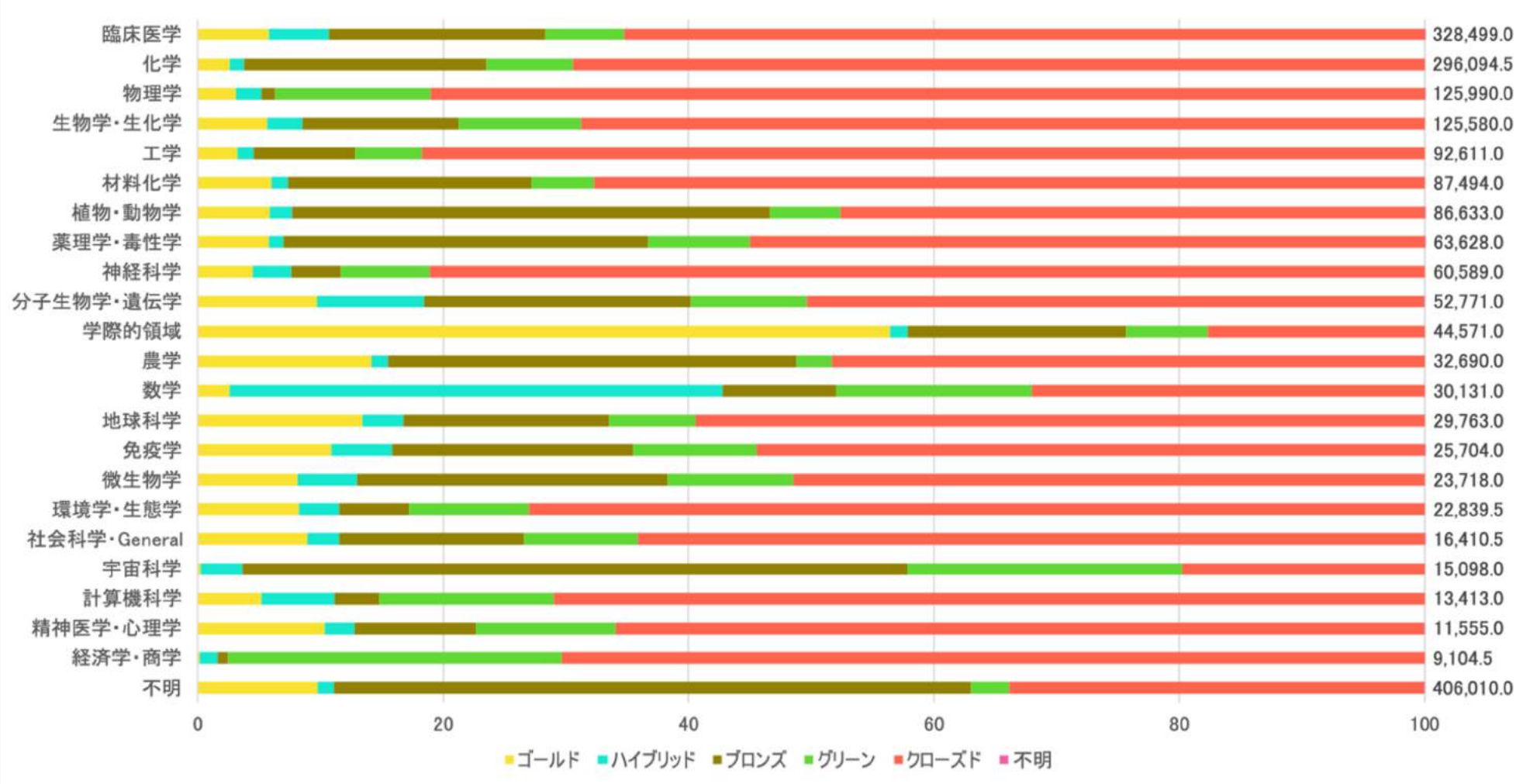
結果

全てのIRがUnpaywallによるクローリング対象となっているわけではない。
よって、実際の機関リポジトリによるOAの割合はもっと高いだろう。
実際にはどのくらいの割合なのだろうか？

- IRDBで「本文あり」「資源タイプがjournal-article」「言語が英語」という3つの条件を満たす論文を検索したところ、106,519件がヒットした（2020年2月26日時点）
- これらの条件を満たす論文にCrossref DOIが付与されていると仮定すると、IRでOAである論文の割合は5.32%である
- この調査でIRでOAである論文の割合は2.36%なので、2.25倍の差異が存在することになる

分野：日本のOAの全体像

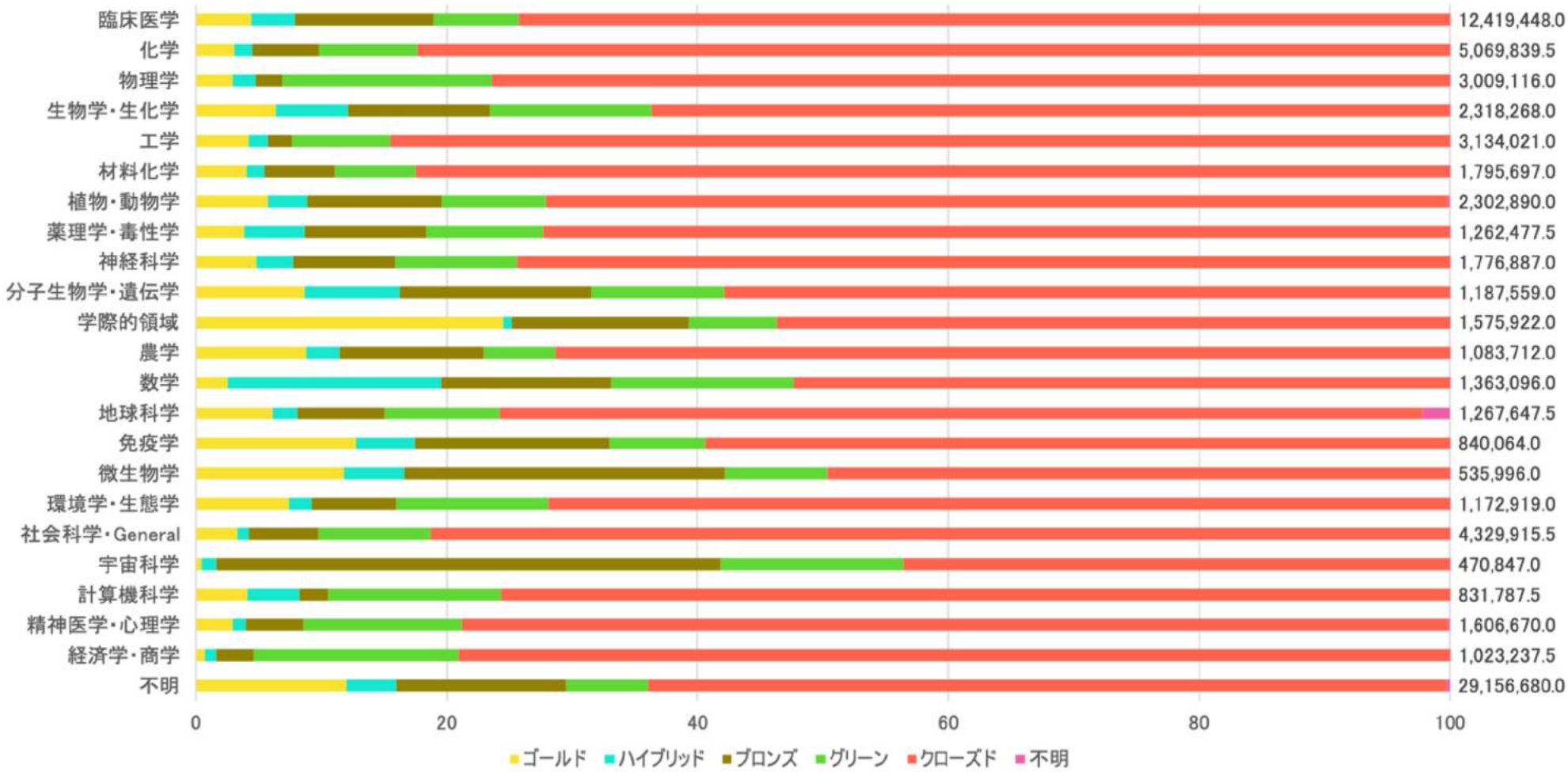
結果



各論文の分野は、ESI Master Journal Listを使用して取得した。ESI Master Journal Listでは、雑誌ごとに分野が与えられている。論文が収録されている雑誌のISSN、EISSNによって紐付けを行い、論文の分野を特定した。図の右側の数値は、各分野の論文件数である。

分野：世界のOAの全体像

結果



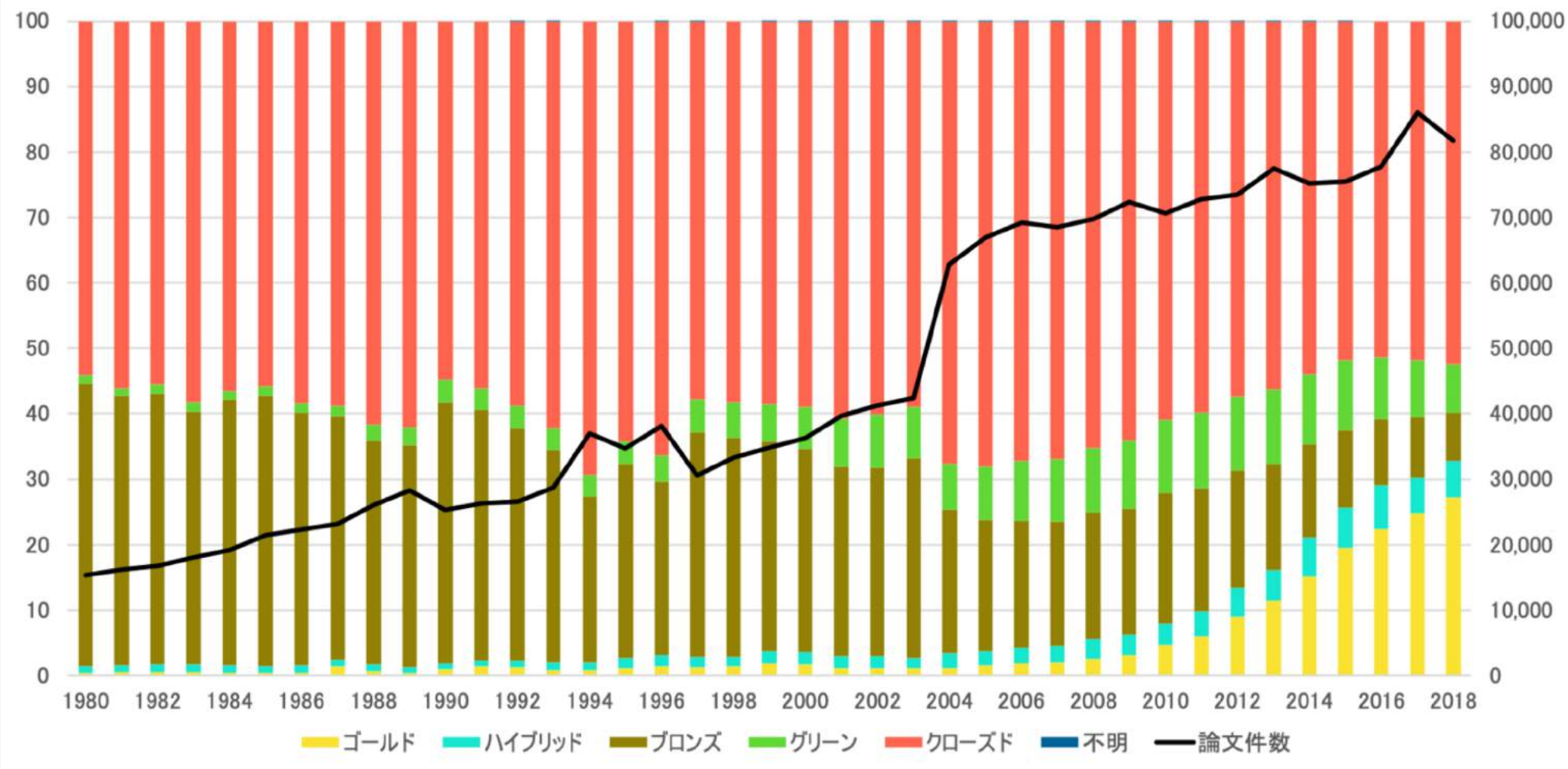
分野

結果

- ほとんどの分野で、日本のOAの割合は世界のそれよりも高い
  - 物理学、生物学・生化学、神経科学、微生物学、環境学・生態学の分野においては、世界の割合のほうが高い
- 日本・世界で共通してOAの割合が高い分野として学際的領域と宇宙科学が挙げられ、日本ではいずれも80%以上である
  - 学際的領域はメガジャーナルの影響を受けてゴールドの割合が高い
  - 宇宙科学は主要な学術雑誌の多くがブロンズである
- リポジトリでの公開が進んでいない分野として、日本・世界共通して、化学、工学、材料工学が挙げられ、1割程度である
- リポジトリでの公開が進む分野として学際的領域が挙げられる
  - オープンなライセンスにより様々なリポジトリでの公開が進んでいる
- IRでの公開が進んでいる分野として、学際的領域と経済学・商学が挙げられる

出版年：日本のOAの全体像

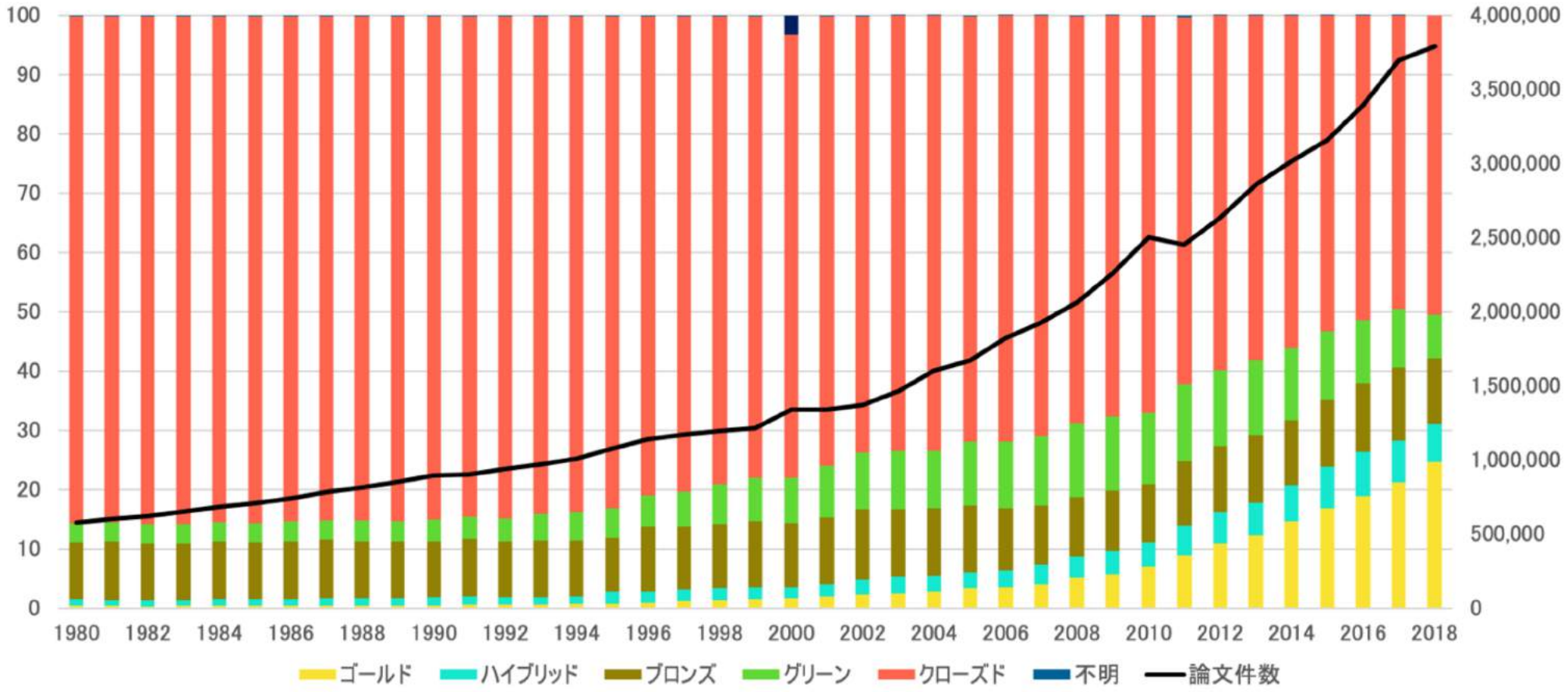
結果



左の軸は各OA種別の割合、右の軸は論文の件数を表している。
2003年から2004年にかけての論文の件数の急増は、JSTリンクセンターの始動（2002年9月）によって、論文へのCrossref DOIの付与が増加したことに起因する（？）。

出版年：世界のOAの全体像

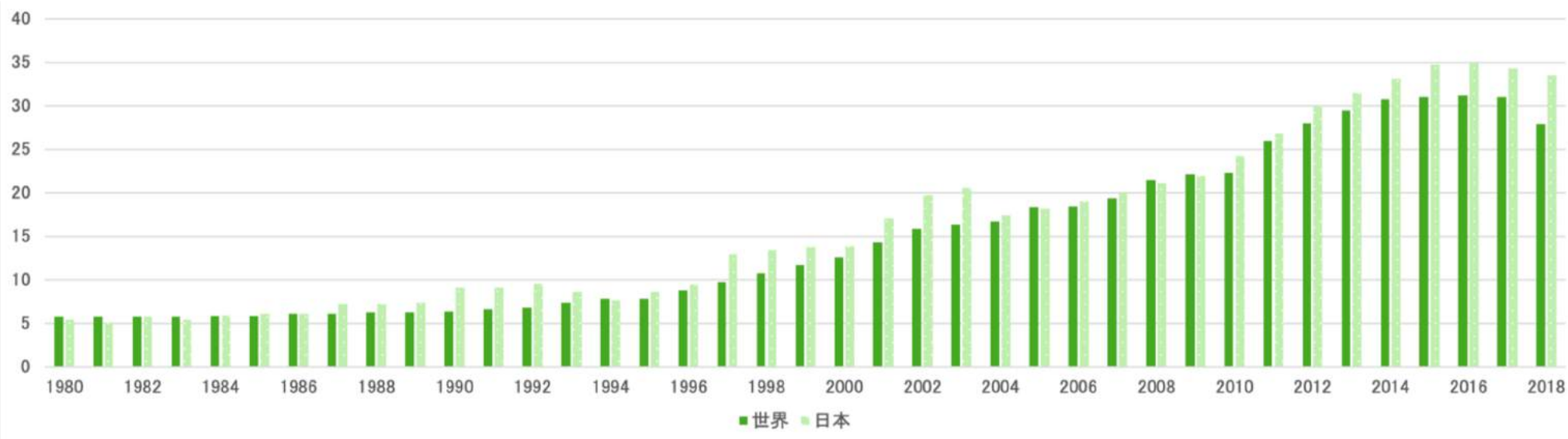
結果



- 過去10年間では日本・世界ともに約40%がOAである
- 近年、ゴールドの割合が伸びている
- 日本の過去の論文でのブロンズの割合の圧倒的に高い

出版年：リポジトリによるOA [1/2]

結果

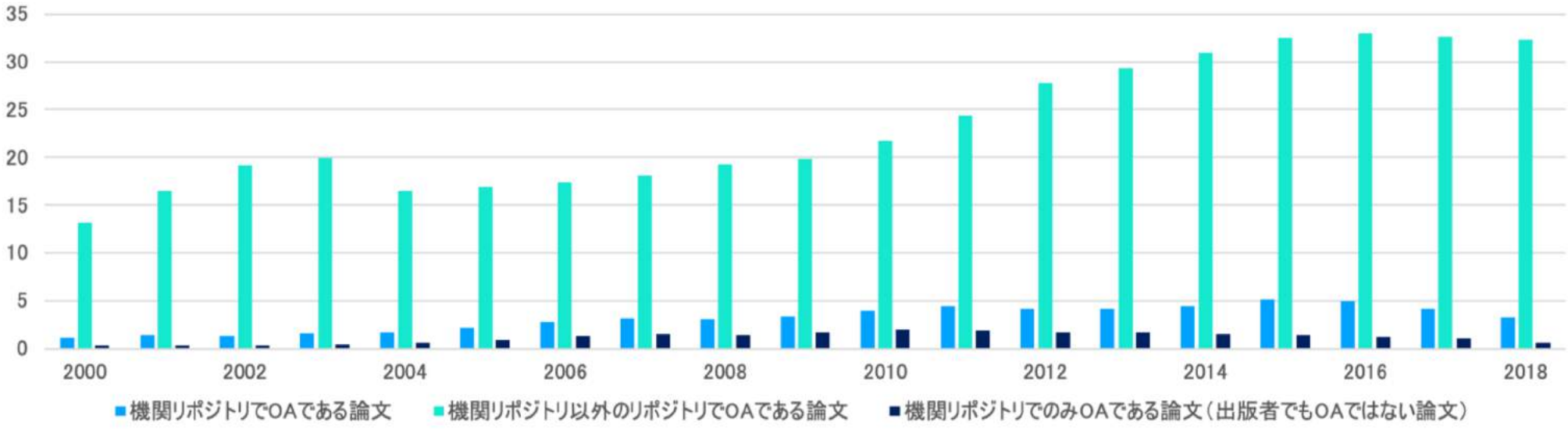


出版年ごとの日本と世界のリポジトリによるOA状況

- 日本と世界は概ね同様の傾向であり、2000年以降リポジトリで公開される論文は増加している
- 近年では日本のほうがリポジトリで公開される論文の割合が若干高い

出版年：リポジトリによるOA [2/2]

結果



日本における出版年ごとのリポジトリによるOA状況の詳細

- IR以外のリポジトリで公開されている論文の割合が高い
  - これらのリポジトリとしてSemantic Scholar、PubMed等が挙げられる
- ゴールドによって再利用（公開）可能な論文が増加したことで、リポジトリで公開される論文も増加している
  - 実際に「リポジトリのみでのOA」を指すグリーンは伸びていない
- IRのみでOAである論文の割合は減少傾向にある

OAに使用されているリポジトリ

結果

- 1～4位は日本・世界で共通である
- 1～3位は著者によるアップロードを主要としないリポジトリである
- 著者のアップロードによる公開が主要であるIR以外で日本の論文の公開に使用されているリポジトリとしては，**arXiv**（4位）、**figshare**（7位）、**CERN Document Server**（8位）が挙げられる
- 日本で10位以内に入っていない世界で使用されているリポジトリとしては、**Zenodo**（5位）、**Hyper Articles en Ligne** (HAL)（6位）が挙げられる

日本

	リポジトリ	論文件数
1	pdfs.semanticscholar.org	272,902
2	www.ncbi.nlm.nih.gov	154,672
3	europepmc.org	115,888
4	arxiv.org	28,875
5	repository.kulib.kyoto-u.ac.jp	11,937
6	eprints.lib.hokudai.ac.jp	9,322
7	figshare.com	5,038
8	cds.cern.ch	4,944
9	tsukuba.repo.nii.ac.jp	4,494
10	kanazawa-u.repo.nii.ac.jp	3,555

世界

	リポジトリ	論文件数
1	pdfs.semanticscholar.org	8,714,986
2	www.ncbi.nlm.nih.gov	4,549,458
3	europepmc.org	3,830,596
4	arxiv.org	895,370
5	zenodo.org	339,266
6	hal.archives-ouvertes.fr	285,491
7	cds.cern.ch	162,320
8	hdl.handle.net	124,104
9	dergipark.org.tr	103,632
10	babel.hathitrust.org	101,256

IRでの論文の利用状況

結果

方法：京都大学のIRであるKURENAIのアクセスログ（2017/2/27～2019/9/30）を分析して、IRでの異なるOA種別のあいだのアクセス数の差異を観察

結果：特に機関リポジトリのみでOAである論文のアクセス数が多い

- 論文のOA版の提供を著者へ依頼する際には、各論文のOA状況を把握しOA版が存在しないものを優先して依頼することが望ましい
- このことは利用者からの需要に応えること、さらには機関リポジトリのプレゼンスの向上につながる

表 8: OA 種別ごとの機関リポジトリでのアクセス数の平均と標準偏差.

OA 種別	論文件数	アクセス数	
全体	8,939	195.12	(155.65)
ゴールド	3,603	157.74	(103.64)
ハイブリッド	588	166.23	(166.41)
ブロンズ	545	194.08	(162.49)
グリーン	4,000	235.09	(179.28)
クローズド	203	157.59	(179.26)
不明	0	—	(—)
DOAJ 掲載論文・ライセンス付	3,226	159.11	(102.67)
DOAJ 掲載論文・ライセンスなし	52	195.33	(162.25)
その他 OA ジャーナル掲載論文・ライセンス付	89	142.67	(110.41)
その他 OA ジャーナル掲載論文・ライセンスなし	236	136.35	(93.89)
リポジトリで OA である論文	8,634	196.51	(154.62)
機関リポジトリで OA である論文	8,351	198.01	(154.14)
機関リポジトリ以外のリポジトリで OA である論文	5,418	178.33	(119.68)
機関リポジトリとその他リポジトリで OA である論文	5,135	179.77	(116.74)
機関リポジトリでのみ OA である論文（出版者でも OA ではない論文）	2,379	247.75	(205.74)

まとめ

- 日本のOAの割合は世界よりも高い。理由として、過去の論文でのブロンズの多さが挙げられる。
- 2000年以降、リポジトリで公開される論文は増加している。これらの多くはIR以外のリポジトリで公開されている。
- IRでOAである論文の割合やIRのみでOAである論文の割合は少ない。
- ただし、Unpaywallにカバーされていない日本のIRが多く存在するため、IRでのOAに関する数値は実際にはさらに高いと考えられる。
- IR以外で日本の著者が公開に利用しているリポジトリとしては、arXiv、figshare、CERN Document Serverが挙げられる。
- （京都大学のIRの利用状況の解析結果）IRでは「IRのみでOAである論文」のアクセス数が多い。
- 論文のOA版の提供を著者へ依頼する際には、各論文のOA状況を把握しOA版が存在しないものを優先して依頼することが望ましい。このことは利用者からの需要に応えること、さらにはIRのプレゼンスの向上につながる。さらに、学術情報空間全体でのOAの割合を増加させる。
- 上記を可能にするには、各機関がOA状況を継続的に把握できる仕組み（例：ドイツにOpen Access Monitor、デンマークにおけるOpen Access Indicator）が必要である。

謝辞

- 本研究は、オープンアクセスリポジトリ推進協会（JPCOAR）コンテンツ流通促進作業部会でのプロジェクトの一つとして進められた
- KURENAIを使用した分析は、京都大学重点戦略アクションプランオープンアクセス推進事業の一部として取り組んだ