DataCiteを利用した日本の大学における研究データの公開状況についての分析

西岡千文 （京都大学附属図書館）

1. はじめに

近年、政府機関や研究機関によって研究データの公開がより一層推進されている。日本における研究データの公開状況については、アンケート等で調査が行われてきた。本発表では、研究データにDOI の割り当てを行っているDataCiteの全レコードのメタデータを対象とした調査を行うことで、日本の大学における研究データ公開状況を観察する。

2. 手法

データセットと処理.

データセットのDL
- DataCiteの全レコードのメタデータをDL

↓ 19,606,708件

著者の所属機関の抽出・同定
- DataCiteのレコードには、著者のリストがフィールドcreatorsに、各著者の所属機関のリストがフィールドaffiliationに格納
- affiliationの各要素には文字列で所属機関が記載されており、表記揺れが存在することから、NISTEP機関同定プログラム公開版と機関名辞書を使用し、文字列から機関を同定
- 同定された機関のセクターが公立大学、国立高専、国立大学、国立短大、私立大学、私立短大のいずれかである機関に所属する著者が含まれているレコードを対象

↓ 64,953件

対象とするリソースの種別
- リソースの種別がCollection、Dataset、Softwareのうちいずれかであるレコードを研究データとして捉え、対象

↓

分析の対象となるレコード39,128件

研究データ公開件数のカウント方法. 本発表では以下の2要因に基づき、日本の大学における研究データ公開件数を3×2 = 6通りの方法で算出する。

A) 「整数カウント」or「分数カウント」or「第一著者」
- A) 整数カウント：日本の大学に所属する著者がレコードに含まれていれば、1件としてカウント
- B) 分数カウント：レコード1件を著者数、著者が所属する機関数、同定される機関数で按分
- C) 第一著者：研究データの公開は第一著者が主導していると想定し、第一著者のみをカウントした研究データ公開件数を算出

B) HEPDataで公開されているレコードを「含む」or「含まない」

抽出されたレコードの半数以上がHEPDataで公開されているものであった。HEPDataではエネルギー物理学の論文に関連する研究データが公開されている。HEPDataで公開されているレコードを含めると高エネルギー物理学に偏るため、これらを除いた分析も行う。

3. 結果

全体（表1）. 論文ではカウント方法によって国や大学ごとの論文件数が大きく異なることが報告されているが、研究データにも同様のことがいえる。高エネルギー物理学では著者が1,000名を超える論文が存在しており、それらの論文の研究データがHEPDataで公開されていることから、HEPDataを含めた結果では整数カウントと分数カウントの差が特に大きい。DataCiteで公開されている研究データのうち、日本の著者が携わっている研究データ（HEPDataを含めた整数カウント）は0.48%であり、ごく僅かである。

出版年（図1）. 2018年に突起があるものの、2017年以降は概ね増加傾向にある。

他リソースとの紐付け. DataCiteの各レコードはフィールドrelatedIdentifiersにより関連する他リソースの識別子と紐付けられており、他リソースの識別子や関係の種別が記録されている。日本の大学の研究データで最も使用されている関係の種別としてIsSupplementToがある。IsSupplementToは、研究データが補足している論文と紐付ける際によく利用される。世界の研究データと比較すると、日本の大学の研究データは他リソースの補足として公開されている割合が高いことがわかった。

表1：研究データ公開件数

研究データ公開件数カウント方法		研究データ公開件数
HEPData 含	整数カウント	39,128.00
	分数カウント	3,901.81
	第一著者	3,706.00
HEPData 除	整数カウント	5,010.00
	分数カウント	2,566.34
	第一著者	2,968.00
世界		8,175,217.00

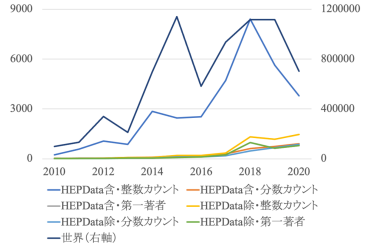

図1：出版年ごとの研究データ公開件数

4. まとめ

分析で明らかになったこと.
- カウント方法によって研究データ公開件数は大きく異なる。
- 近年研究データ公開件数は増加傾向にある。
- 日本の大学の研究データは他リソースの補足として公開されている割合が高い。

分析の課題.
- DataCiteのレコードの著者は延べ200,011,615人存在するが、そのうち101,440,687名に所属機関が記載されていない。よって、実際の日本の大学の研究データ公開件数はさらに高いと思われる。
- 機関の同定に際して、NISTEP機関同定プログラムと機関名辞書を使用した。おおよそ正しく機関同定を行っていたが、一部誤りも観察された。研究データ公開件数のより正確な把握には、機関同定プログラムと機関名辞書の評価も必要となる。