

地球科学データと学術論文の対応付け獲得に向けた事例調査

梅原 佳宏[†] 清水 敏之[†] 吉川 正俊[†]

[†] 京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町

E-mail: [†]{y.umehara@db.soc., tshimizu@, yoshikawa@}i.kyoto-u.ac.jp

あらまし 観測機器の発達や情報技術の進歩により、膨大な量の地球科学データが収集、蓄積、解析されており、データの統合利用等による新たな知見の発見が期待されている。しかし、地球科学データは専門性が高く、専門分野の合致しない利用者が適切なデータを検索し、データを理解することは困難である場合が多い。我々は地球科学データに対して、そのデータを言及している学術論文を提示することで、利用者に対するデータの理解、活用支援を行うことが出来ると考えた。DIAS プロジェクトにおけるデータセットとそれを言及している学術論文を対象として、対応付け事例を調査し、自動的に対応付けるための手法の検討を行った。

キーワード 地球科学データ、学術論文

1. はじめに

近年の観測機器の発達や情報技術の進歩により、膨大な量の地球科学データが収集、蓄積、解析されている。それらのデータは、同分野の専門家によって利用される場合もあるが、異なる分野の専門家によって利用される場合もある。異なる分野のデータの統合的な利用によって新たな知見が得られることが期待される。専門分野の合致しないデータ利用者にとっては、データ自体の理解が困難である場合が多いため、メタデータの整備やデータに対するコメントの付与といったデータの理解支援を行う必要がある。さらに、データベース内にどのようなデータがあるのか把握することが困難な場合や、検索要求を適切に表現するのが困難な場合があるため、データを探し出す検索補助のための仕組みが必要である。また、データの提供者が、データがどのように利用され、データを利用した人がそのデータに対してどのような感想を持ったのかというフィードバックを求めている場合もある。

一方、科学者はデータを利用して得られた知見をまとめ、学術的な成果として示すために学術論文を書く。本研究では、地球科学データについて言及している学術論文に着目し、地球科学データと学術論文の対応付けを行うことで、データの理解、活用支援を行うことが出来ると考えた。地球科学データベースにおいて地球科学データが特定された際に、対応付けられた学術論文を提示することで、利用者に対するデータの理解支援を行うことができ、対応付けられた学術論文の情報をを用いることで検索の補助を行うことができると考えられる。また、学術論文の情報により、データ提供者に対してのデータ利用状況に関するフィードバックもできると考えた。

ここで、英国情報システム合同委員会、出版研究コンソーシアム、研究情報ネットワークが出した調査報告書[1]によると、データへのアクセスは重要であるが、アクセスが困難であるという報告がある。一方で、学術論文は、アクセスが比較的容易であるため、地球科学データと学術論文の対応付けを行うことにより、データへのアクセスが容易になると考えられる。また、

データの理解支援として、データに対してコメントを付けることが考えられる。データに対してコメントを付ける場合は、新たな内容の記述が必要となるが、学術論文を利用することにより新たな内容の記述が必要とされないと考えられる。

また、デジタルオブジェクトを識別する ID として、DOI がある[2]。科学データに対して DOI を付与するという試みがある[3]。データセットに DOI が付与され、学術論文でデータを参照する際に DOI を用いることが一般的となれば、地球科学データと学術論文の対応付けの取得が容易になると考えられる。一方で、データセットへの DOI の付与は、十分に浸透していないと考えられ、また、過去に発表された学術論文については、データセットに付与された DOI を用いて、対応付けを行うことはできないため、地球科学データと学術論文を対応付けるための手法が必要であると考えられる。

地球科学データと学術論文の対応付けについては、一部の地球科学データベースおよび論文データベースにおいて実際に行われている。例えば、地球科学データライブラリ PANGAEA^(注1)のデータセットと、Elsevier^(注2)が出版するフルテキストデータベース Science Direct^(注3)のデータセットに対応する学術論文との相互リンクが進められている。この場合、対応付けを行うために、学術論文の著者は、Elsevier に学術論文とデータ識別子を提出している。我々は、過去に発表された学術論文について、どのように対応付けを行うのかという問題や、データ識別子を提出することが学術論文の著者にとっての負担になるのではないかと考え、地球科学データと学術論文を自動的に対応付けるための手法が必要であると考えた。そこで、本研究では、地球科学データとそれを言及している学術論文を対象として、対応付け事例を調査し、自動的に対応付けるための手法の検討を行った。

ここで、我々の過去の研究[4]では、地球科学データと学術

(注1): <http://pangaea.de/>

(注2): <http://www.elsevier.com/>

(注3): <http://www.sciencedirect.com/>

論文の関連付けを行った．地球科学データと学術論文との間には，地球科学データとそれを言及している学術論文のような直接的な関係と，地球科学データと学術論文が，地球科学に関連する語彙，人・組織，分野に関連しているような間接的な関係があると考えた．そこで，主に，地球科学データと学術論文の関連付けについて，間接的な関係に基づく関連付けの取得に取り組んだ．本研究では，直接的な関係に着目し，地球科学データとそれを言及している学術論文について，対応付け事例を広範囲に調査し，自動的に対応付けるための手法の検討を行った．

2. メタデータ及び対応付けの現状

2.1 地球科学データベースのメタデータ

既存の地球科学データベースとしては，アメリカ航空宇宙局 NASA が管理する地球科学データのメタデータベース GCMD(Global Change Master Directory)^(注4)や，国内では，DIAS(Data Integration and Analysis System)^(注5)[6] が多種多様な地球科学データを管理している．また，FAO GeoNetwork^(注6)などの地球科学データのメタデータポータルでは，空間情報を持つデータのメタデータ編集・検索機能を提供するオープンソースのソフトウェア GeoNetwork^(注7)が利用されている．地球流体データのデータベース，解析，可視化のためのデスクトップツール兼サーバである Gfdnavi(Geophysical fluid data navigator)^(注8)[5] では，Web ベースでデータ解析，可視化できる地球流体データのデータベース構築が可能である．メタデータファイルとして，YAML ファイルや SIGEN ファイルを使用することができ，データベースに登録されるデータに情報を付加することが可能である．

地球科学データベースではデータセットに対してメタデータを付与することが一般的であり[7]，例えば，GCMD では，DIF(Directory Interchange Format) というフォーマットで多様な地球科学データのメタデータを収集，分類し，検索機能を提供している．また，DIAS では，専用のウェブツールを用いてデータセット単位で地球科学分野のメタデータ標準である ISO19115[8] および ISO19139[9] に準拠したメタデータを作成している．我々はこのようなメタデータを介して地球科学データと学術論文の対応付けを行うことにした．

2.2 対応付けの現状

地球科学データと学術論文の対応付けが実際に行われている例としては，地球科学データライブラリ PANGAEA のデータセットと，対応する Science Direct の学術論文との相互リンクが進められている．具体的な対応付け事例としては，図 1 に示す PANGAEA のデータセットと，図 2 に示す Science Direct の学術論文である．

地球科学データに対するメタデータのの一つとして参考文献の情報が格納されていることがあり，その場合は地球科学データ

と学術論文が対応付けられているとみなすことが出来る．しかし，現状，このような手動での作業による対応付けは十分に整備されているとは言い難い．例えば，DIAS では，地球科学データに対するメタデータのの一つとして参考文献があるが，多くのデータセットに関して参考文献情報の記載がない．

また，学術論文中に，データセットに対しての謝辞が記述されている場合があり，そこから地球科学データと学術論文の対応付けを抽出することも考えられるが，我々の調査した範囲では，現状，データセットに対して直接謝辞を記述しているものは少ない．そのため，メタデータ中の参考文献や学術論文中の謝辞を利用するだけでは，対応付けを得るのに不十分であると考えられる．

3. 調査対象論文の取得

現在，DIAS プロジェクトにおける地球科学データを対象として，対応付け事例の調査を行っている．各データセットに対して 3.2 節から 3.4 節で述べる手法を適用し，論文検索エンジンによって検索を行い，論文本文が取得できた学術論文を用いた．本研究では論文検索エンジンとして Google Scholar^(注9)を用いた．

3.1 対象データセットおよびメタデータ

DIAS では多種多様なデータセットが管理されており，そのメタデータは DIAS 俯瞰・検索システム^(注10)等を通して入手可能である．2013 年 1 月時点で 192 データセットに対するメタデータが公開されている．

DIAS におけるメタデータは以下の項目を含む[7]．

- データセット名，バージョン，データセット名略称
- 問合せ先
- メタデータ作成者，作成日
- データセット作成者，作成日
- データセット概要
- 時間情報
- 地理的範囲
- グリッド
- キーワード
- データセットに関するオンライン情報
- データ周辺情報
- データ配布情報
- データセット作成に関する加工過程や履歴の説明
- 統合データの品質に関する問題点
- データセット利用制約条項
- 参考文献

今回の調査では，データセット名，データセット名略称，データセット作成者，データセット概要，データセット利用制約条項を利用した．今後，本論文では，DIAS におけるデータセットのメタデータをデータセットメタデータと呼ぶこととする．

(注4): <http://gcmd.nasa.gov/>

(注5): <http://www.editoria.u-tokyo.ac.jp/dias/>

(注6): <http://www.fao.org/geonetwork/>

(注7): <http://geonetwork-opensource.org/>

(注8): <http://www.gfd-dennou.org/arch/davis/gfdnavi/>

(注9): <http://scholar.google.co.jp/>

(注10): <http://dias-dss.tkl.iis.u-tokyo.ac.jp/ddc/>

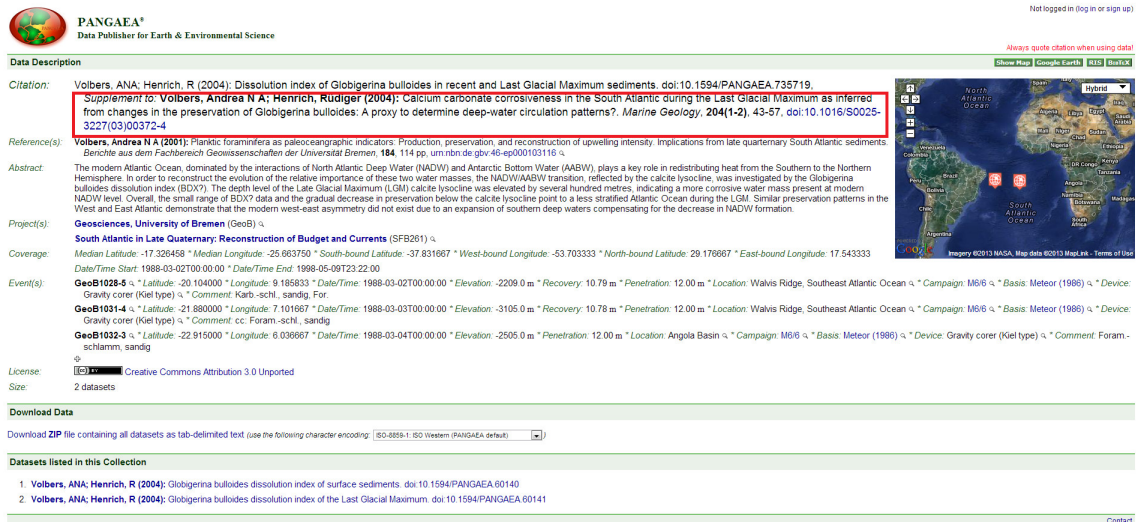


図 1 PANGAEA におけるデータセットの例



図 2 Science Direct における学術論文の例

3.2 データセット名の利用

地球科学データと学術論文の対応付けとして、学術論文中でデータセット名を記述している場合が考えられる。例えば、“西太平洋熱帯域に設置されたトライトンおよび TAO ブイのデータ相互比較”^(注11)という論文では、“トライトンブイデータ”の記述があり、“トライトンブイデータ”というデータセットに対して言及していると推測できる。

データセットメタデータ中のデータセットタイトルの完全一致検索により調査対象論文を取得した。

3.3 謝辞の利用

地球科学データと学術論文の対応付けとして、学術論文中でデータセットに対する謝辞を記述している場合が考えられる。例えば、DIAS プロジェクトにおける“東北大学全球高解像度客観解析海面水温データ”のデータセットメタデータでは、利用した際に記述する謝辞の指定をしており、“高解像度データを用いた北西太平洋における海上風と海面水温場の関係に関する研究”^(注12)という論文で参照されている。

(注11): 黒田芳史, 曽野和彦, 安藤健太郎, ボールフライタウグ, マイクマックファードン, “西太平洋熱帯域に設置されたトライトンおよび TAO ブイのデータ相互比較”, 海洋科学技術センター試験研究報告, Vol.43, 2001 年。

(注12): 森本直樹, 豊田邦夫, 笠原実, “高解像度データを用いた北西太平洋に

データセットメタデータでは、データセット利用制約条項の中で、学術論文でデータセットを利用する際に用いるべき謝辞を指定している場合がある。謝辞が指定されている場合、その文面を用いて検索を行い、調査対象論文を取得した。

3.4 キーワードの利用

長いデータセット名では、必要以上に絞り込まれ、データセットについて言及している学術論文を取得出来ない場合が考えられる。そこで、データセットメタデータにおける、データセット名略称、データ作成者、概要から有用なキーワードを選択し検索に利用することを考えた。

今回の調査では、データセットメタデータ中の上記の項目から筆者が主観により有用だと判断して選択したキーワードを用いて検索を行い、調査対象論文を取得した。

例えば、DIAS プロジェクトにおける“DIAS 長期再解析プロジェクト (JRA-25)”というデータセットと“CMIP3 マルチ気候モデルにおける日本の気候の再現性比較”^(注13)という論文

における海上風と海面水温場の関係に関する研究”, Journal of The School of Marine Science and Technology, Vol.3, No.2, pp.39-50, 2005 年。

(注13): 道広有理, 佐藤嘉展, 鈴木靖, “CMIP3 マルチ気候モデルにおける日本の気候の再現性比較”, 京都大学防災研究所年報, 第 53 号 B, 2010 年。

に関して、データセット名による検索では取得できなかったが、“JRA-25”と、データ作成組織である“気象庁”と“電力中央研究所”の三つのキーワードを用いて検索することで取得することが出来た。

4. 対応付け事例の調査

3.2 節から 3.4 節の各手法において論文検索エンジンを用いて取得された論文に対して人手により実際にデータセットに対する言及があるかの調査を行った。

一部の DIAS データセットに対して、実際にデータセットに対する言及があったものの数を表 1 に示す。

データセットタイトルを用いて取得された論文では、データセットについて言及している場合が多かった。一方で、特に、長いデータセットタイトルの場合、学術論文中で表記ゆれしたデータセットタイトルを記述して、データセットタイトルと完全一致している場合は少なかつたため、データセットタイトルとの完全一致による対応付けの取得は難しいと考えられる。

謝辞を用いて取得された論文では、データセットについて言及している場合が多かった。一方で、データセットの作成者が指定している謝辞とは異なる表記をしている場合や、データセットに対する謝辞を記述していない場合があり、また、データセットの作成者によっては、謝辞を指定していない場合もあった。

今回調査対象に含めた“DIAS 長期再解析プロジェクト (JRA-25)”では、そのデータセットの作成に関わる論文^(注14)があり、このように、データセットを作成した文献が存在している場合、その文献を参考文献としている学術論文は、そのデータセットについて言及していると考えられる。

また、データセットの URL が存在している場合、その URL を記述している学術論文は、そのデータセットについて言及していると考えられる。例えば、“反射法地震探査データを用いた海洋物理研究”^(注15)という論文では、DIAS プロジェクトにおける“AVISO 海面高度アノマリ”の URL を参照している。

5. 対応付け手法の検討

対応付けを行う際に有用なキーワードとしては、いくつかの種類があり、例えば、データセットを作成したプロジェクト名や衛星名である。一方で、プロジェクトや衛星は、複数のデータ、領域を扱う場合があるため、さらにキーワードを追加する必要があると考えられる。

また、データセットを作成した期間や場所についても検討の余地があり、例えば、DIAS プロジェクトにおける“マルハナバチモニタリングデータ”というデータセットに対して、キーワードを用いて取得された論文では、期間や場所がデータセットメタデータと異なる学術論文が多くあった。

上記の点に加えて、以下の点に関して、対応付け手法の検討を行っている。

- データセットメタデータにおけるデータセットタイトルと、データセット作成者により指定された謝辞を利用する。謝辞については、学術論文の謝辞に関する節をマッチングの対象とし、表記ゆれをどの程度許容するのか検討する。

- 学術論文中のデータ記述部を発見し、データに関する節で記述されている、または、作成した、利用した、用いた等の記述が続く場合には、データセットについて言及していると考ええる。

6. おわりに

我々は地球科学データに対して、そのデータを言及している学術論文を提示することで、利用者に対するデータの理解、活用支援を行うことが出来ると考えた。データセットメタデータを手掛かりに取得してきた学術論文を対象として、対応付け事例の調査を行った。

対応付け事例の調査より、データセット名を利用すると、学術論文中で表記ゆれが生じている場合や、長いデータセット名の場合に必要以上に絞り込まれることにより、対応付けの取得が困難な場合があることが分かった。また、謝辞の利用に関しては、データセットの作成者が謝辞を指定していない場合、学術論文の著者がデータセットに対する謝辞を記述していない場合、データセットの作成者が指定した謝辞とは異なる表記をしている場合があり、対応付けの取得が困難な場合があった。キーワードを利用すると、データセット名の表記ゆれや、長いデータセット名の場合に必要以上に絞り込まれること等による、対応付けの取得が困難な場合にも対処できると考えられる。一方で、キーワードの選択によっては、データセットについて言及していない学術論文を対応付けてしまう場合がある。そのため、地球科学データと学術論文の自動的な対応付けに向けて、データセット名、謝辞、文献・URL を利用する場合に加えて、地球科学データのメタデータの内、データセット名、データセット名略称、データ作成者、概要等を利用して、必要十分なキーワードを選択する手法について検討していきたい。また、DIAS ではプロジェクトの成果論文を管理しており、それらを対象に含めた対応付け事例の調査についても行っていきたい。

文 献

- [1] Joint Information Systems Committee, publishing research consortium, research information network, “Access to scholarly content gaps and barriers”, 2011.
- [2] 長谷川豊祐, “DOI(デジタルオブジェクト識別子) システムの概要”, 情報の科学と技術, Vol.49, No.1, pp.28-33, 1999 年.
- [3] Natasha Simons, “Implementing DOIs for Research Data”, D-lib Magazine, Vol.18, 2012.
- [4] 梅原佳宏, 清水敏之, 吉川正敏, “地球科学データと学術論文

(注14): 大野木和敏ら, “JRA-25 長期再解析データの概要”, 大会講演予稿集, Vol.89, p.147, 2006 年.

(注15): 黒田芳史, 曾野和彦, 安藤健太郎, ボールフライターグ, マイクマックファーデン, “西太平洋熱帯域に設置されたトライトンおよび TAO ブイのデータ相互比較”, 海洋科学技術センター試験研究報告, Vol.43, 2001 年.

表 1 データセット名及数

データセット名	キーワード	データセット名の利用	謝辞の利用	キーワードの利用
DIAS 長期再解析プロジェクト (JRA-25)	DIAS 長期再解析プロジェクト JRA-25	0.00(0/0)	0.00(0/0)	0.00(0/0)
	気象庁 電力中央研究所 JRA-25			0.84(21/25)
気象庁提供の GPV(格子点データ) アーカイブ	気象庁 提供 GPV 格子点データ アーカイブ	1.00(1/1)	0.00(0/0)	1.00(1/1)
	GPV GSM 気象庁			0.70(7/10)
	GPV 全球数値予報モデル			0.75(3/4)
RR2002 人・自然・地球 共生プロジェクト 第一課題	RR2002 人 自然 地球 共生プロジェクト 第一課題	0.00(0/0)	指定なし	0.00(0/0)
	RR2002 地球シミュレータ			0.50(1/2)
	RR2002 共生			0.67(8/12)
マルハナバチモニタリングデータ	マルハナバチ モニタリングデータ	0.00(0/0)	指定なし	0.00(0/15)
	マルハナバチ モニタリング			0.00(0/20)
CEOP 衛星データセット (TRMM > PR > 2PRECI)	CEOP 衛星 データセット TRMM TMI 2PRECI	0.00(0/0)	0.00(0/0)	0.00(0/0)
	CEOP TRMM TMI 衛星			0.50(1/2)
	TRMM TMI 降水			0.71(22/31)
MAHASRI Pathfinder アジア域降水量格子点データ	MAHASRI Pathfinder アジア域降水量格子点データ	0.00(0/0)	0.00(0/0)	0.00(0/0)
	MAHASRI 降水量			0.50(1/2)
	MAHASRI JAMSTEC			0.27(3/11)
みらい CTD データセット	みらい CTD データセット	0.00(0/0)	0.00(0/0)	0.42(8/19)
	JAMSTEC みらい CTD			0.50(9/18)
	みらい CTD 水温			0.42(8/19)
トライトンバイデータ	トライトンバイ データ	0.75(3/4)	指定なし	0.70(23/33)
	TRITON バイ			0.73(19/26)
東北大学全球高解像度客観解析海面水温データ	東北大学 全球高解像度 客観解析 海面水温 データ	0.00(0/0)	0.00(0/0)	0.40(2/5)
	東北大学 海面水温			0.10(3/33)
	大気海洋陸域結合データ 海面水温			1.00(2/2)
AVISO 海面高度アノマリ	AVISO 海面高度 アノマリ	0.00(0/0)	1.00(38/38)	1.00(2/2)
	AVISO アノマリ			1.00(2/2)

の関連付け”，第 4 回データ工学と情報マネジメントに関するフォーラム，2011 年．

- [5] Takeshi Horinouchi, Seiya Nishizawa, Chiemi Watanabe, Akinori Tomobayashi, Shigenori Otsuka, Tsuyoshi Koshiro, Yoshiyuki Hayashi, GFD Dennou Club, “Gfdnavi, Web-Based Data and Knowledge Server Software for Geophysical Fluid Sciences, Part I: Rationales, Stand-Alone Features, and Supporting Knowledge Documentation Linked to Data”, Lecture notes in computer science, Vol.6193, pp. 93-104, 2010.
- [6] 絹谷弘子, 清水敏之, 吉川正俊, 喜連川優, “DIAS における多分野研究者連携による地球科学データ公開に向けた協働”, 電子情報通信学会技術研究報告, Vol.110, No.328, DE2010-33, pp.45-50, 2010 年.
- [7] 清水敏之, 絹谷弘子, 吉川正俊, “多様な地球科学データに対する俯瞰・検索システムの開発”, 電子情報通信学会技術研究報告, Vol.110, No.328, DE2010-32, pp.39-44, 2010 年.
- [8] International Organization for Standardization: “ISO19115: 2003, Geographic information Metadata”.
- [9] International Organization for Standardization: “ISO/TS 19139:2007 Geographic information Metadata XML schema implementation”.