Jan. 15, 2021

統合データサイエンスプラットフォームの現状

田中良昌[1,2,3], 梅村宜生[4], 阿部修司[5], 齊藤泰雄[2],
南山泰之[6], 新堀淳樹[4], 中野慎也[7]

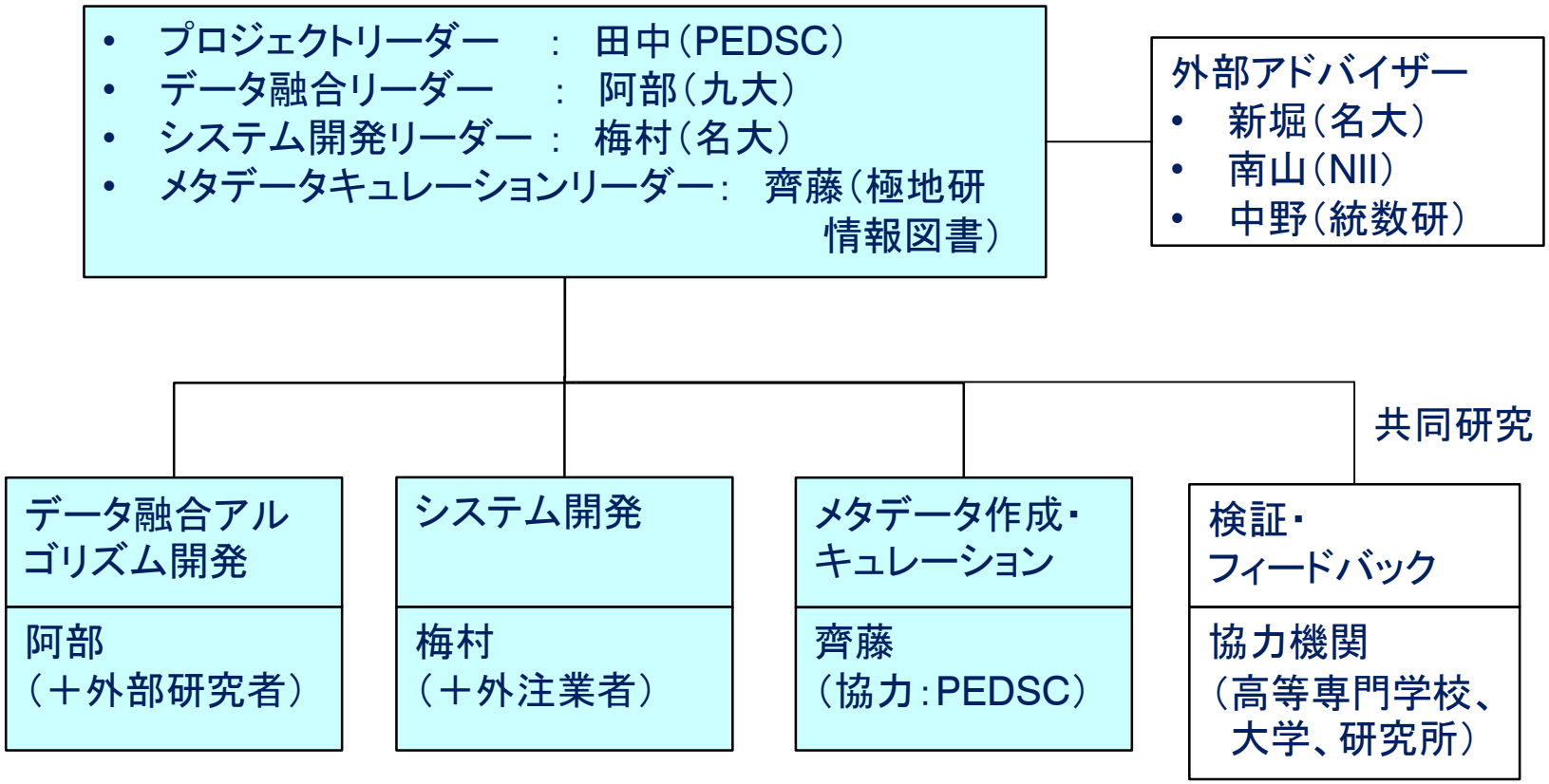
1. 極域環境データサイエンスセンター, 2. 極地研, 3. 総研大
4. 名大ISEE, 5. 九州大ICSWSE, 6. 情報研, 7. 統計数理研究所

目次

1. AMIDERプロジェクトの目的
2. プロジェクトの体制
3. 統合データサイエンスプラットフォームの概要
4. 統合データサイエンスプラットフォームの現状
5. 今後の展望
6. まとめ

- 極地研が所有する多種多様な極域データ（宙空圏、気水圏、地圏、生物圏）のメタデータ、及び、実データを統合的に扱うことができ統合プラットフォームを開発することで、将来的にROISや全国の研究機関や大学の所有する幅広い分野のデータへ展開するための手法・ノウハウを構築し、オープンデータ、オープンサイエンスの促進に貢献する。
- 異種・異分野のデータ間の関連性を計算機により導出し、新たな知見を得る手法を導入し、データ駆動型科学（Data-Driven-Science）の推進に貢献する。

2. プロジェクトの体制

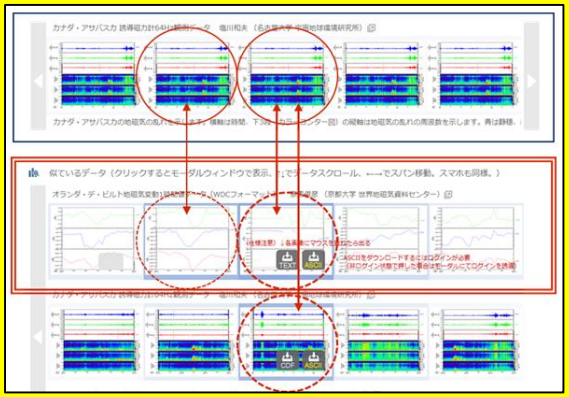

• プロジェクトリーダー ： 田中（PEDSC）
• データ融合リーダー ： 阿部（九大）
• システム開発リーダー ： 梅村（名大）
• メタデータキュレーションリーダー： 齊藤（極地研 情報図書）
外部アドバイザー
• 新堀（名大）
• 南山（NII）
• 中野（統数研）
共同研究
データ融合アルゴリズム開発
阿部
（＋外部研究者）
システム開発
梅村
（＋外注業者）
メタデータ作成・キュレーション
齊藤
（協力：PEDSC）
検証・フィードバック
協力機関
（高等専門学校、大学、研究所）

3. 統合データサイエンスプラットフォームの概要

未知の現象の発見
新しい学術分野の
創出へ

新たな知見
（相関情報）

データ融合計算



異分野データ間の関連性の導出

AMIDER

ユーザ
（研究者、
一般市民）

データ提供者
（NIPRデータPI）

メタデータ
実データ

メタデータ
DB構築

メタデータ作成
実データ標準化
（CDF, NetCDF,...）

メタデータ
実データ

メタデータ
実データ

ファイル変換
→ ASCII
可視化

**データの
相関情報**

実データ
- ASCII
- CDF
- NetCDF

カタログ
- 可視化画像
- メタデータ

カタログ表示

統合データサイエンス
プラットフォーム

ドメインDB
- ADS
- IUGONET
- ….

実データ

実データ

リポジトリ
（NII）

4.1. 統合データサイエンスプラットフォームの現状（1）

- データサイエンスの推進（データマネジメント）

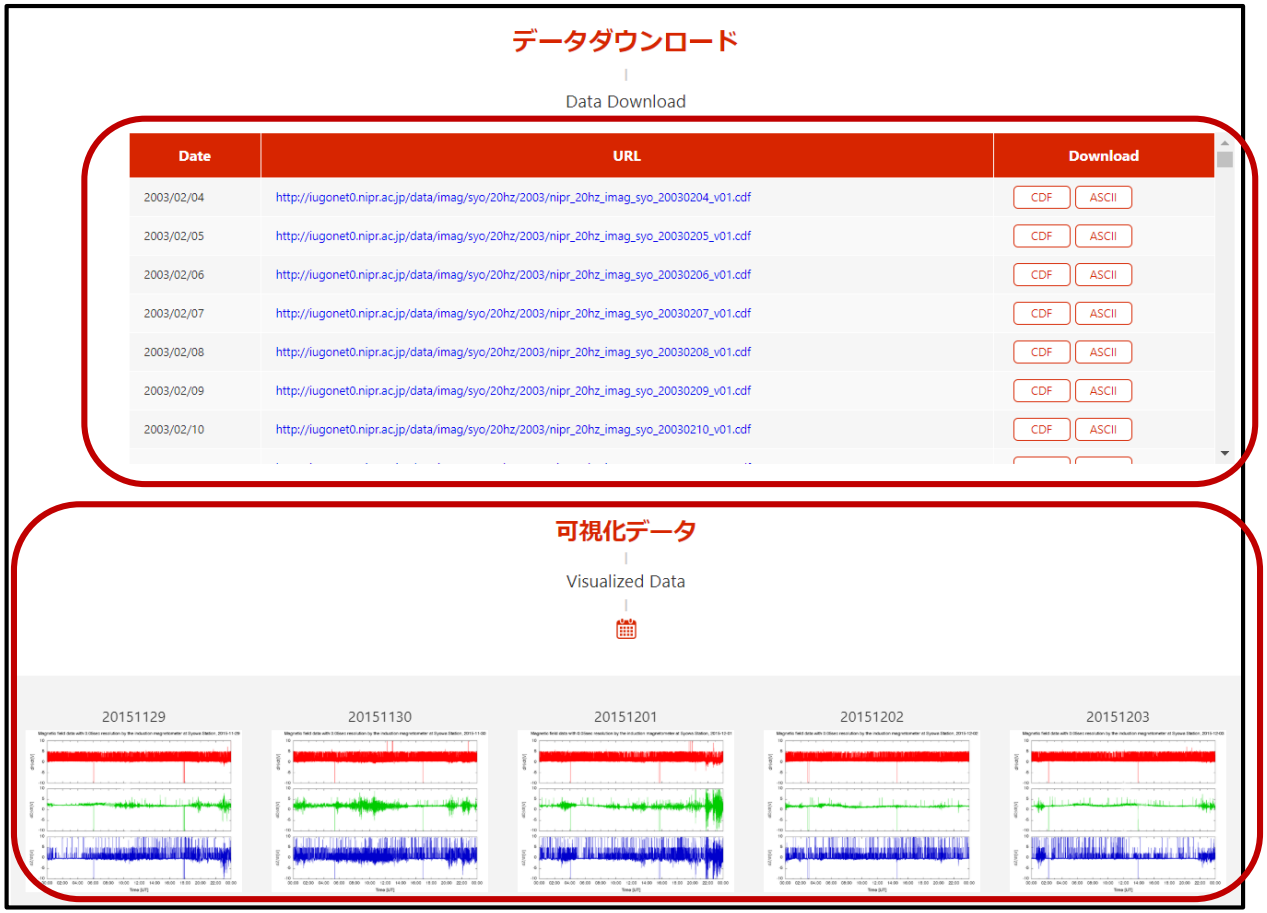
現在、極地研究所が所有する極域科学データを登録し、関係者内で公開中。

異分野データを同一システムで管理。

日・英両言語に対応。（研究者、及び、一般市民を対象）

- 異分野データをまとめて検索、表示することができる統合プラットフォームを構築・公開。
- メタデータフォーマットとしてISO19139、SPASEを採用し、ドメインのサイエンス情報を保持。
- キーワードやカテゴリ検索で異種データを抽出することで、隣接分野データへの接続を促す。
- データの参照数、登録数、利用数を統計・可視化し、データPIに提供（インセンティブの提供）。

- 解析の強化、データ駆動型科学の創出（データ融合）Ⅰ



実データを、CDF、NetCDF、GCMD準拠ASCII等の機械可読のフォーマットで公開。

ASCII形式への変換・ダウンロード機能の実装。

QLプロットの表示機能の実装。

- データの用途（研究、教育等）や得られる知見等の新しいメタ情報を付加し、利用を促進。
- 実データを、分野標準フォーマット and/or 機械可読フォーマット（CDF、NetCDF、GCMD準拠ASCII）で公開し、解析へ誘導。
- 上記実データをASCIIフォーマット変換・ダウンロードする機能を実装。
- QLプロットを自動生成、並べて表示し、現象の発見を支援。

- 解析の強化、データ駆動型科学の創出（データ融合）Ⅱ

データ融合計算結果の例：

時間連続 vs 時間連続

	A	B	C	D	E	F
1	Base Data					
2	DATASETID: 13453, DATASET_NAME: 10-minute averaged meteorological data (Solar Radiation) obtained at Kizahashi Hama, Skarvsnes on Soya Coast, East Antarctica					
3	PI: Kudoh Sakae (NIPR), FIELD: Natural sciences, FIELD_DETAIL: Earth sciences. Geology, LAT/LON (N, S, E, W): -69.473611, -69.473611, 39.611944, 39.611944					
4						
5	Correlations * The same field data is displayed at the top rows. If you want to revert to the default sort, use column 'A'.					
6	SE	DATA	DATASET_NAME	CORR_20141219	CORR_20141220	CORR_20141221
7	1	13456	10-minute averaged meteorological data (UV) obtained at Skallen on Soya Coast, East Antarctica	0.979567	0.962917	0.940498
8	2	13453	10-minute averaged meteorological data (UV) obtained at Kizahashi Hama, Skarvsnes on Soya Coast, East Antarctica	0.959774	0.965356	0.949691
9	3	13456	10-minute averaged meteorological data (PAR) obtained at Skallen on Soya Coast, East Antarctica	0.973125	0.963216	0.914490
10	4	13456	10-minute averaged meteorological data (Solar Radiation) obtained at Skallen on Soya Coast, East Antarctica	0.970996	0.972709	0.914850
11	5	13450	10-minute averaged meteorological data (UV) obtained at Yukidori zawa, Langhovde on Soya Coast, East Antarctica	0.970750	0.945310	0.943949

	A	B	C	D	E
1	Base Data				
2	DATASETID: 620, DATASET_NAME: Meteorite found around the Asuka Station, Antarctica in the 1986-1987 season (Asuka-8602)				
3	PI: Yamaguchi Akira (NIPR), FIELD: Natural sciences, FIELD_DETAIL: Earth sciences. Geology, LAT/LON (N, S, E, W): -71.7200, -73.00, 28.1700, -21.6700				
4					
5	Correlations * The same field data is displayed at the top rows. If you want to revert to the default sort, use column 'A'.				
6	SE	DATA	DATASET_NAME	CORR	URL
7	1	7679	Meteorite found around the Yamato Mountains, Antarctica in the 1986-1987 season (Yamato-86036)	0.999359	https://amider.jp/data/7679
8	2	2655	Meteorite found around the Meteorite Hills, Antarctica in the 1978-1979 season (Meteorite Hills-78028)	0.999134	https://amider.jp/data/2655
9	3	732	Meteorite found around the Asuka Station, Antarctica in the 1987-1988 season (Asuka-87196)	0.998206	https://amider.jp/data/732
10	4	757	Meteorite found around the Asuka Station, Antarctica in the 1987-1988 season (Asuka-87222)	0.997874	https://amider.jp/data/757
11	5	633	Meteorite found around the Asuka Station, Antarctica in the 1987-1988 season (Asuka-87029)	0.997783	https://amider.jp/data/633

異なる分野のデータを比較し、分野を跨いだ新しい発見を支援する。

成分vs成分

進捗状況（2020年12月現在）：

データ種	時間連続	空間分布	成分
時間連続	直接算出	共通因子抽出・比較	共通因子抽出・比較
空間分布	共通因子抽出・比較	直接算出	共通因子抽出・比較
成分	共通因子抽出・比較	共通因子抽出・比較	直接算出

- **直接算出**： 同種のデータ同士の関連性は、算出可能であることを確認。
- **共通因子抽出・比較**： 実データおよび紐づけられたメタデータ（時間情報・空間情報）から、共通する因子を用いて群を作成し、時間連続・空間分布データと比較する方法。

4.4. 統合データサイエンスプラットフォームの現状(4): データ掘り起し(PEDSCとの連携)

実データ、メタデータ作成数:

分野	データ種	実データ数	フォーマット	メタデータ数	フォーマット
宙空	オーロラ	33	CDF, JPEG	33	SPASE
	地磁気	24	CDF	24	SPASE
	レーダー等	8	CDF	8	SPASE
気水	昭和基地気水モニタリングデータ	4	NetCDF	4	ISO19139
	スバールバル気水モニタリングデータ	4(*)	NetCDF	4	ISO19139
	氷床コア	18(*)	未定	18	ISO19139
	その他	9(*)	未定	9	ISO19139
地圏	地震データ	6(変換中)	CDF	6	SPASE
	インフラサウンド	16(変換中)	CDF	16	SPASE
	磁場・重力	19(*)	CDF	19	ISO19139
	隕石バルク成分	1168	ASCII	1168	ISO19139
	隕石標本			10563	ISO19139
生物	宗谷海岸露岩域気象データ	27	NetCDF	27	ISO19139
	海洋観測データ	(*)	NetCDF	4	ISO19139
	スバールバル環境データ	(*)	NetCDF	8	ISO19139
	生物標本			2678	ISO19139
合計		1,309		14,575	

(*): 2021年2月以降に作成予定。

- PIに個別にインタビューを行い、適切なフォーマット、公開方法を検討。
- 実データは、分野標準のCDF、NetCDF、及び、GCMD準拠ASCII等の機械可読フォーマットで公開。
- ASCII→CDF・NetCDF変換ツールを開発し、PIに提供。
- メタデータは、キュレーションチームがフォーマットを提案・品質管理。

4.5. 統合データサイエンスプラットフォームの現状(5): メタデータの品質管理 (図書館連携)

メタデータ記述ポリシーの策定:

- メタデータ作成ガイドライン:

1. データ基本情報 (観点)　　朱書き: 必須事項
*1 必須欄にNが付与されているものは、複数個記入してもよい。

No.	項目	必須/任意*1	説明	語句例	補足	文例
1	データセット名	必須	(以下の項目に対する内容を、文章で構成する。)			
1.1	データの種類	必須	取得したデータの種類、物理量、試料名等を記入する。	オーロラ画像		A-1
1.2	取得場所	必須	データを取得した場所の具体名称を記入する。	南極昭和基地	隕石・岩石試料、生物試料等の場合で、場所が不明な場合は省略可とする。	
1.3	取得期間	任意	データを取得した日時を、西暦の世界標準時で記入する。他のデータセットと区別するために必要な場合は、必須とする。			
1.4	取得手段名	任意	観測装置、採取方法、データ処理方法、計算方法など、取得手段の名称を記入する。他のデータセットと区別するために必要な場合は、必須とする。	全天カラーデジタル一眼レフカメラ		
1.5	識別子	任意	1.6「分野固有の識別子」を参照する。他のデータセットと区別するために必要な場合は、必須とする。		隕石・岩石試料、生物標本は必須とする。それ以外の分野では任意とする。	
2	メタデータ作成日	必須	メタデータを作成した日を、西暦の世界標準時で記入する。書式は、YYYY-MM-DD、または、YYYY-MM-DDThh:mm:ssとする。(YYYYは年、MMは月、DDは日、hhは時、mmは分、ssは秒を表す。)	2018-11-29T00:00:00	ISO 8601の日付、時刻表記の規格に基づいて記入する。	
3	概要	必須	(以下の項目に対する内容を、文章で構成する。文章は、「です・ます調」で記入する。)			
3.1	データの種類	必須	1.1.1「データの種類」を参照する。	オーロラ画像		B-1
3.2	取得場所	必須	1.1.2「取得場所」を参照する。	南極昭和基地	隕石・岩石試料、生物試料等の場合で、場所が不明な場合は省略可とする。	
3.3	取得期間	必須	1.1.3「取得期間」を参照する。	2005年9月から観測を開始	隕石・岩石試料、生物試料等の場合で、日時が不明な場合は省略可とする。	

メタデータ格納ポリシーの策定:

- メタデータスキーマ(ISO, SPASE): 統合プラットフォームに組み込み
- 日本語用スキーマ用語リスト: 統合プラットフォームに組み込み
- SPASE/ISO19139マッピング表:

その他(マニュアル等):

- キュレーションマニュアル(作成中)
- PI用メタデータ作成マニュアル(作成中)

異分野データへの適用

- AMIDERで構築したシステムやデータ融合の手法を、ROIS-DSの各センターや、大学、研究機関、学会等が所有する異分野の多様なデータに応用してみる。

データ融合の手法を用いて、分野を跨いだ新しい研究成果、学術分野の創出へ

例：

- 文献におけるオーロラ、隕石の記述と、オーロラ、隕石データ等（文理融合）
- 南極の隕石と「はやぶさ」探査機が持ち帰ったサンプルの比較

来年度に、ROIS内公開を経て、2022年度に一般公開する。

オープンサイエンスの推進

- NIIのオープンサイエンス基盤研究センターと協力し、大学の研究活動の支援やオープンサイエンス推進に貢献する。
- 2020年度から「次期JAIRO Cloud実証実験」に参加し、極地研のいくつかのデータを次期JAIRO Cloudで公開する試みを開始している。

- AMIDERプロジェクトは、データ駆動型科学や異分野融合の推進を目的として、2018年度以降、統合データサイエンスプラットフォームの開発を行ってきた。
- これまでに、PEDSCが公開する多様な極域データ（宙空圏、気水圏、生物圏、地圏）に適用し、ある程度利用可能であることを確認した。
- 今後、共同研究ベースでROISや大学、研究機関のデータに応用し、その効果を調べる。
- 2021年度のROIS内公開を経て、2022年度に一般公開する。