特別企画 第413回 ＯＵＧライフサイエンス分科会

～第21回情報プロフェッショナルシンポジウムINFOPRO2024ver.～

文献データベースにおけるデータ収録状況

- MEDLINE，Embase，Scopus，Web of Science，CAplus，Dimensions，JSTPlusを対象にLANCETを例に検証する -

2024年7月5日（金）14:25~14:40

INFOSTA　OUG　ライフサイエンス分科会

日本オンライン情報検索ユーザー会　Online Users Group(OUG)

OUG ライフサイエンス分科会

https://www.infosta.or.jp/research/oug-life/

- 医薬分野から広くライフサイエンス全般を対象に検索演習、勉強会、見学会等を開催
- 開催日：毎月第3木曜午後、年間９回程度開催
- 開催方法：主にWeb
- 本会はメンバー全員によるグループ（輪番）制で会の企画・運営を行っております。
- OUG入会手続きは入会申込（OUG/SIG)のページをご覧ください。
  OUG会員の方で初めてライフサイエンス分科会に出席される場合は、予め事務局経由にて主査までご連絡ください。
- OUG非会員はビジター参加も可能です。

背景・目的

■背景

- 文献データベース（以下、文献DB）には収録対象などの特徴があり、調査の際には目的に合わせて文献DBを選択する必要がある
- 特徴の1つに収録対象誌があるが、実際に収録対象誌がどの程度収録されているかを確認したことがない

■目的

文献DBの収録対象誌のうち特定のジャーナルについて調査・検証して今後の活用に繋げる

調査内容

■ **調査対象の文献DB　（（　）は提供システム）**
MEDLINE、**Embase**、**Lancet Titles**、**New England Journal of Medicine**（以上Dialog）
Scopus、　**Web of Science**、　**Dimensions**、
CAplus（STN）、　**JSTPlus**（JDreamⅢ）、　**J-GLOBAL**

■ **調査したジャーナル**
Lancet、　**New England Journal of Medicine**、　**JAMA**

■ **検索内容**

① **予備調査**（2024年2月～5月にかけて実施）
出版年を2000~2024年とし可能な場合はpISSNとeISSNで検索

② **検証のための調査**（2024年5月20日、27日、6月19日に実施）
出版年を2022年、ジャーナルをLancetに限定して各DBで検索

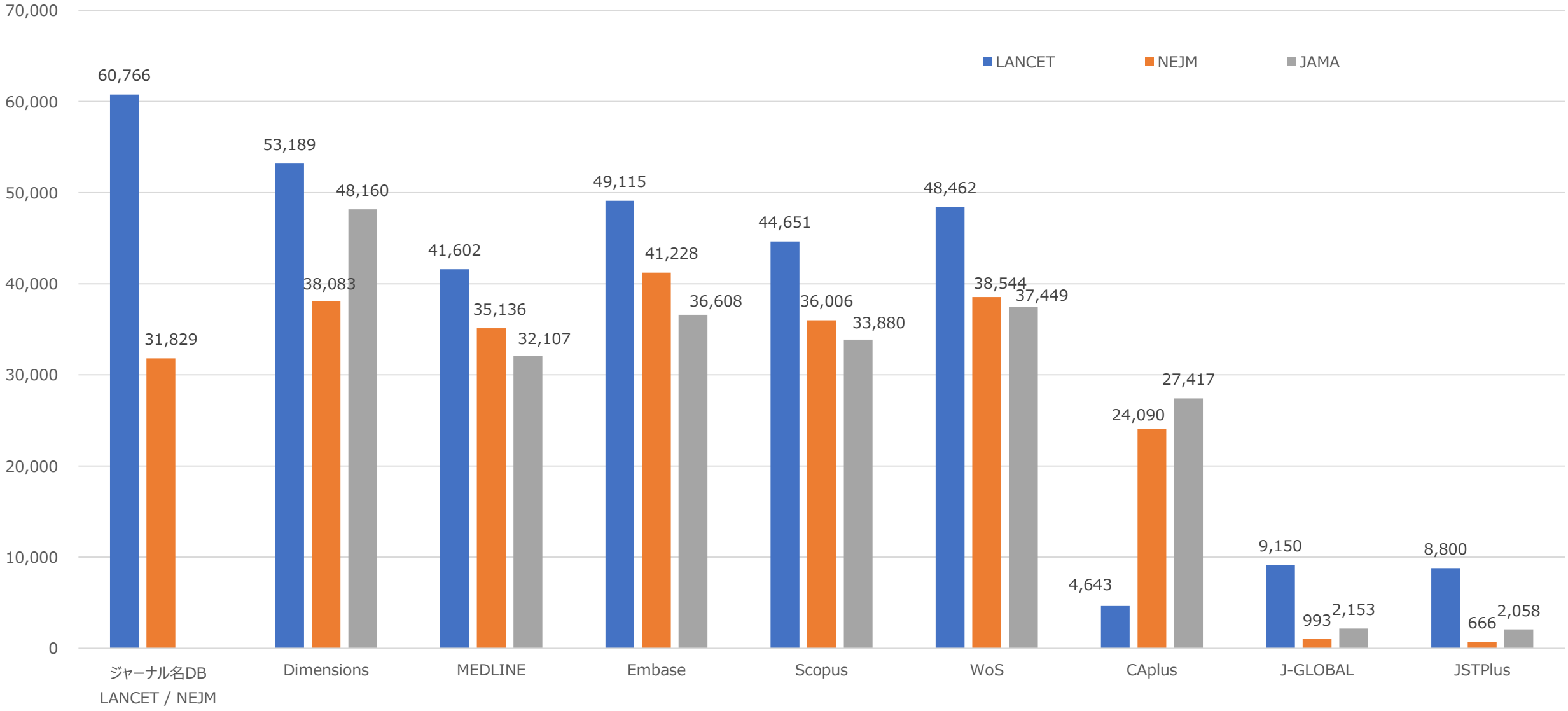
■ **検証内容**
①にて各DBにおける３ジャーナルの収録状況を確認する
②の検索結果について重複状況やデータの中身を調査し検証する

限定的な収録方針のDBについて

- JSTPlus、 J-GLOBALは内容が同一で、海外文献の収録対象を記事の種類（原著など）で選定した後に著者抄録のあるもののみが収録されるため収録が限定される

- CAplusは化学を中心に収録され、臨床医学文献は収録されない特徴がある

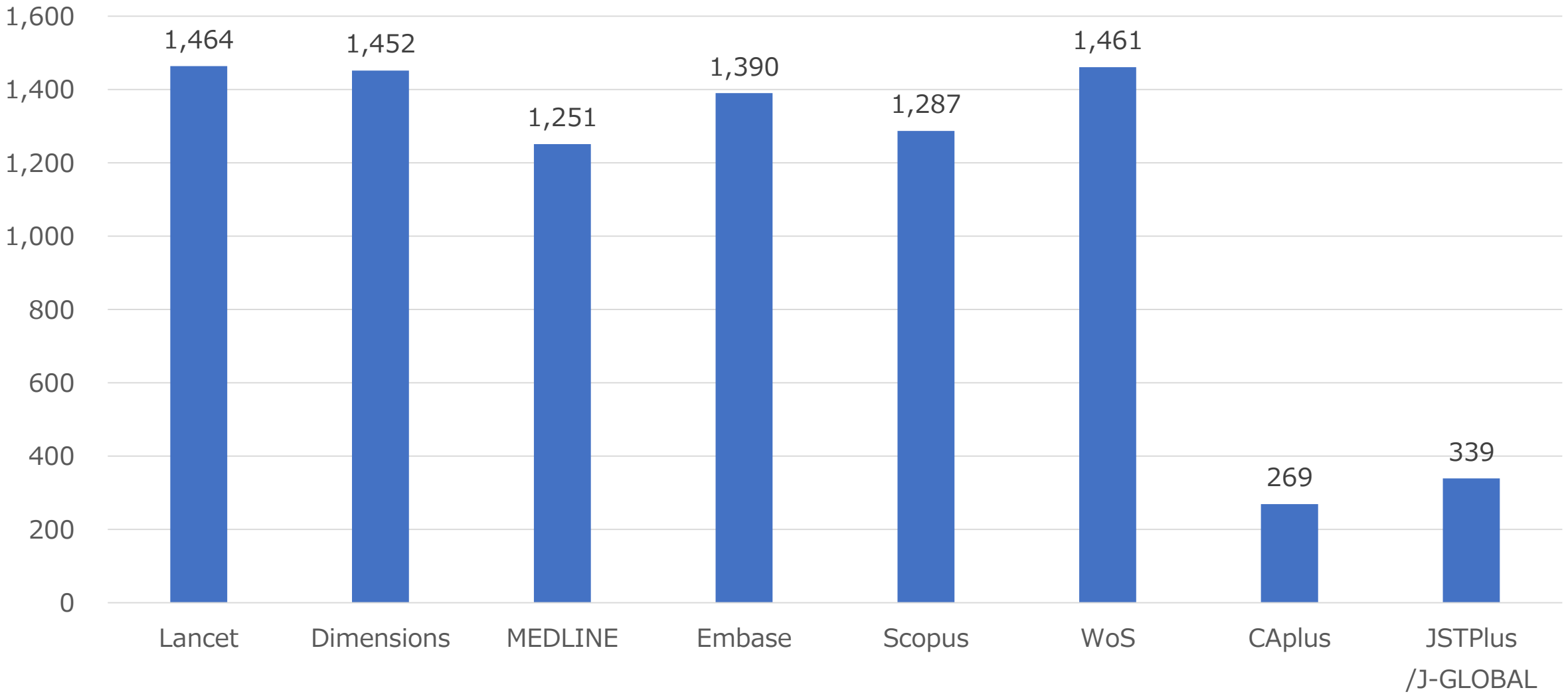
①各DBのジャーナル収録状況（2000-2024）



結果①　ジャーナル収録状況　2000～2024年

- LANCETはジャーナル名のDBの件数が最多だったが、NEJMはジャーナル名のDBが最多ではなかった。NEJMには除外記事があり収録方針によるものと思われた。
- CAplus、J-GLOBAL、JSTPlusは収録方針通り結果件数は少なかった。
- CAplusは医学雑誌3誌のうちJAMAを多く収録していた。
- 全体的にDimensions、Embase、WoS、Scopus、MEDLINEの順に収録データが多い傾向が見られた。
- LANCETに関してはLANCET Titles（以下、LANCET）の件数が最も多く、ジャーナル記事を網羅的に収録していると考えられたことからLANCETを取り上げて詳細を検証することとした。

②LANCET　出版年2022年の検索結果



②LANCET　2022年の検索結果の内訳

DB	ISSN	件数	実行日	計	備考
LancetTitles	pISSN	1,464	2024/06/19	1,464	（pISSNのみヒットしeISSNはヒットしない）
	eISSN	-	-		
Dimensions	-	1,452	2024/05/16	1,452	
MEDLINE	pISSN	1,251	2024/06/19	1,251	pISSNとeISSNは同じ結果
	eISSN	1,251	2024/06/19		
Embase	pISSN	1,247	2024/06/19	1,390	eISSNのみ 143件（eの1,439件中49件が重複、pにもあり）
	eISSN	1,439	2024/06/19		
Scopus	pISSN	1,119	2024/05/20	1,287	eISSNのみ 168件（eの1,293件中6件が重複、pにもあり）
	eISSN	1,293	2024/05/20		
WoS	pISSN	1,463	2024/05/20	1,461	pISSNで2件重複、eISSNで2件重複 doi無し1,030件、doiのPrefixの誤り：5件
	eISSN	1,463	2024/05/20		
CAplus	pISSN	270	2024/05/15	269	化学を重点的に収録される
	eISSN	0	2024/05/15		
JSTPlus/ JGLOBAL	pISSN	339	2024/05/13	339	著者抄録があるもののみ収録される
	eISSN	-	-		

②LANCET　2022年の収録状況の確認

収録状況	件数	収録DB と *論文タイプ*　（（　）はデータの件数　）
①1つのDBで収録	**45**	**Dimensions(45)**： *401巻(43)402巻(2)で出版年は2023年*
②2つのDBで収録	**3**	**Dimensions/MEDLINE(3)**：　*Department of Error(2) Letter(1)*
③3つのDBで収録	**179**	**Lancet/Dimensions(179)**　Embase(104)　WoS(74)： *Conference Abstract(104)はEmbaseで全件収録。　論文タイプなし(73)*
④4つのDBで収録	**22**	**Lancet(22)** Embase/Scopus(20) Dimensions/MEDLINE(12) WoS(2): *Note(16) Comment(7) Journal Article(6) など*
⑤5つのDBで収録	**132**	**Embase/Scopus(132)　LANCET(129)** MEDLINE(101) WoS(85) Dimensions(79)：　*Journal Article(87) Note(67) など*
⑥全DBで収録	**1,134**	*Journal Article(712)　Note(362)　Letter(255)　Editorial(80)　など*
全レコード	**1,515**	

照合はdoiにより行った。なおWoSでDOI無し／誤りのレコードは、巻号ページ情報を使用して照合した。
※JSTPlusとCAplusは収録方針によりデータが限定されるため本確認では除いた。

②LANCET　2022年の各DB収録状況

LancetTitles	Dimensions	MEDLINE	Embase	Scopus	WoS	件数
	①					45
	②	②				3
③	③				③	74
③	③			③		1
③	③		③			104
④	④	④			④	2
④		④	④	④		10
④	④		④	④		10
★1	⑤	⑤	⑤	⑤	⑤	3
⑤	⑤	★3	⑤	⑤	⑤	30
⑤	⑤	⑤	⑤	⑤	★4	46
⑤	★2	⑤	⑤	⑤	⑤	53
⑥	⑥	⑥	⑥	⑥	⑥	1,134

①1つのDBで収録	45
②2つのDBで収録	3
③3つのDBで収録	179
④4つのDBで収録	22
⑤5つのDBで収録	132
⑥全DBで収録	1,134
合計	1,515

LancetTitles	Dimensions	MEDLINE	Embase	Scopus	WoS	件数
★1	⑤	⑤	⑤	⑤	⑤	3
⑤	⑤	★3	⑤	⑤	⑤	30
⑤	⑤	⑤	⑤	⑤	★4	46
⑤	★2	⑤	⑤	⑤	⑤	53

★1　LANCETに無かった3件：書誌に間違いはなかったが収録されておらず理由は不明だった。

★2　Dimensionsに無かった53件：すべて399巻。LANCET399巻の出版年は2022年だがDimensionsでは早期公開の年を出版年としている（Crossrefをデータソースとしている）ためヒットしなかった。

★3　MEDLINEに無かった30件：大半が記事タイプNoteだった。他のDB上でPMIDがあるが、MEDLINE上に収録が無いレコードが1件あった。（PubMed上にもなし）。

★4　WoSに無かった46件：記事タイプに傾向はみられず、PMIDは全件で付与されていた。

②の検証結果

- pISSNとeISSNの検索結果はDBごとに重複状況が異なった
- WoSはdoiの無いデータが多く、また、doiの誤りもあった
- LANCETに無いデータが他DBに収録されている場合のあることを確認した
- Dimensionsは出版年データが他と異なるデータがあった
- DB収録に特徴のある記事タイプが存在した
  - ➢Conference Abstractは　LANCET、Dimensions、Embaseに収録、MEDLINEとScopusは限定的な収録、WoSは0件だった
  - ➢MEDLINEのみに収録の無いデータはNoteが多かった
- PMIDがあるがMEDLINEで検索されないデータを確認した

考察

- 収録誌とされているジャーナルの各DB収録の違いを確認した
- ジャーナルDBのLANCETで必ずしも全データが収録されているものではないことを確認した
- 論文タイプでの収録のバラつき、出版年と巻号データの不一致、PMID付与データのMEDLINE収録なし、などの状況を確認した
- 調査時には収録誌の確認とともに、その収録方針にも留意が必要である
- 検索結果は検索条件とともに、使用したDBとシステム名および検索日時を記録し保管することが望ましい

今後に向けて

- DBの収録データは収録対象誌から予測できるが想定外の未収録の可能性もあることに留意したい
- そのため複数DBを参照して検索結果を確認することが望ましい
- 今回の結果には示さなかったが、重複除去に関して、DBシステムによってはタイトルや資料名をキーに除去されることを確認している
  （例：タイトル“Department of Error”がほぼ除去される、など）
  重複除去はdoiなどのIDを活用して必要以上の除去は無くすべきものと考える
- 収録データと出力結果の関係性について今回の検証では検索条件と出力結果とを照合して精査できたが、今後、AIベースの検索では検証も困難と思われ、検索プロセスの開示が望ましいと考える

最後に

今回の検証にあたり、データベース作成・提供機関のJST様、ジー・サーチ様、DigitalScience様など関係の皆様のご協力をいただきましたことに感謝申し上げます。

ご清聴をありがとうございました。