

# Web サーバのログを用いた GSJ 地質ニュース記事のアクセス解析 (2023 年度)

大野 哲二<sup>1</sup>

## 1. はじめに

GSJ 地質ニュースは地質調査総合センター (GSJ) の広報誌として位置付けられているが、本誌に限らず一般的な情報発信において、どのような内容のものが読まれているのかを知ることは、編集方針等に関わる重要な要素である。GSJ 地質ニュースも、その発信主体が Web に移行したことにより、アクセスログを解析することで、比較的簡単に読者の興味の対象を知ることができる環境は整えられている。一方でアクセスログは単純にそれを数値化すれば良いというものではなく、配慮すべき点も多々存在する。今回、試行的に 2023 年度のアクセスログを解析した結果について簡単に紹介する。

通常、このような解析は Google Analytics 等、専用のサービスやアプリで行うことが多いが、GSJ のウェブサイトでは導入していない。また詳細に行う初めての解析であることからその傾向を把握する意味もあり、手動で解析を行っている。

## 2. アクセスログとは

GSJ 地質ニュースは、GSJ のウェブサイト (サーバ) の下、<https://www.gsj.jp/publications/gcn/index.html> というアドレス (URL) で閲覧することができる。閲覧にはブラウザ (Chrome, Edge, Safari, Firefox 等のソフトウェア) を用いるが、サーバとブラウザ間で行われていることはかなりシンプルである。プロトコル (手順) としては、

- 1) ブラウザがサーバ (<https://gsj.jp>) に、目的のファイル (例えば `index.html`) が欲しいとのリクエストを出す。
- 2) サーバは、該当するファイルがあればその内容を、なければエラーを返す。

それだけである。

現在では、ブラウザは欲しいファイル名だけではなく付帯情報を送信するし、サーバが返すものもファイルに限るわけではないが、その本質はかわらない。そしてこの手順

の内容を記録したものがアクセスログである。

## 3. アクセスログの内容

アクセスログにはどのような内容が記載されているのであろうか。それはサーバ上で動いているプログラムの種類や設定によって異なるが、多くの場合に共通して記録されている情報もある。以下に、メジャーなサーバプログラムである Apache 2.4 のアクセスログのサンプルを示す (Apache Software Foundation, 2024)。ログは通常、1 アクセスにつき 1 行で記録される。

[例]

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700]
"GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.
example.com/start.html" "Mozilla/4.08 [en] (Win98; I
;Nav)"
```

これら情報の中でアクセスログの解析の際に重要な要素は以下の通りである。

### 1) 127.0.0.1

REMOTE\_ADDR. ブラウザが動いている機械 (例えば PC) の IP address. 機械毎に固有の数値であり、会社名またはプロバイダ名ぐらまでであれば特定することができる。

### 2) [10/Oct/2000:13:55:36 -0700]

アクセスのあった日時。

### 3) "GET /apache\_pb.gif HTTP/1.0"

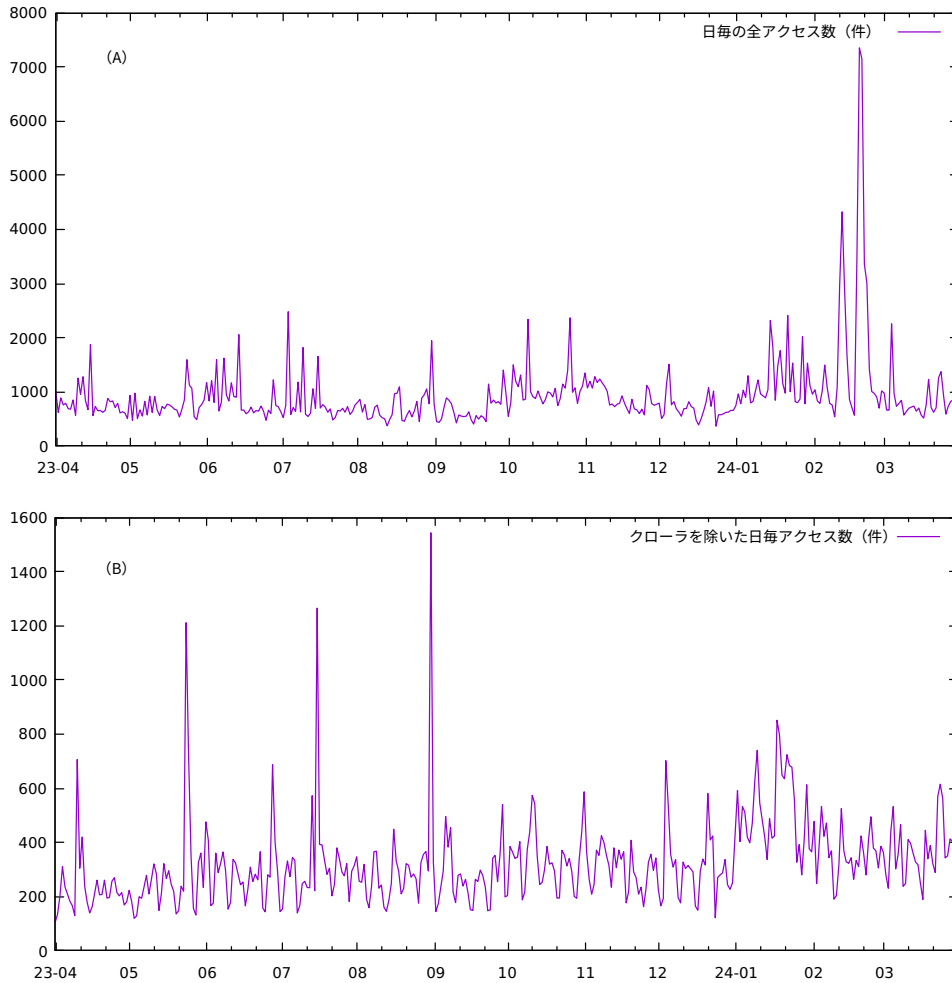
REQUEST\_URI. サーバ中のどこにあるファイルを求めているかの情報。

### 4) "http://www.example.com/start.html"

REFERER. どこからそのアドレスに要求が来たかの情報 (Wikipedia, 2023)。例えば Google 検索で GSJ 地質ニュースの記事を見つけ、リンクをクリックした場合には、ここは "<https://www.google.co.jp/>" といった内容になる。検索エンジンの利用が一般化した現在では、本項目の重要性は低下しており、本解析でも予備的にしか使

<sup>1</sup> 産総研 地質調査総合センター地質情報基盤センター

キーワード: GSJ 地質ニュース, Web サーバ, アクセスログ, ログ解析, クローラ



第1図 GSI地質ニュース記事へのアクセス数の日変化(2023年度)。横軸は月、縦軸はアクセス件数。グラフはそれぞれ、(A)：全アクセス数、(B)：明示的なクローラのアクセスを除いたもの、縦軸のスケールが違っていることに注意。

用していない。

#### 5) "Mozilla/4.08 [en] (Win98; I;Nav)"

USER\_AGENT. ブラウザの種類を示す文字列。この例では単純に「Windows 98 上で動く Netscape Ver.4 (英語)」という程度の意味である(Netscape は Firefox の前身のブラウザ)。

上記の項目の中で、アクセス日時以外の情報はユーザのブラウザから送られてきており、解析において重要であるが、偽装も可能であるので単純に信用することもできない。

#### 4. 解析対象及び解析期間

今回の解析においては、主としてどのような記事が読まれているか(ダウンロードされているか)を調査した。GSJ地質ニュースの記事は全て pdf 形式で公開されているため、調査対象も pdf のみとし、目次等は除外した。対象となった pdf は 2012 年発行の第 1 号(1 月号)から 2024 年

発行の 2・3 月合併号までの 147 か月間分、ファイル数にして 1400 件である。発行月数に比してファイル数が多いのは、記事毎に pdf が分割されているためである。解析対象期間は、2023 年 4 月 1 日から 2024 年 3 月 31 日とした。同期間中のアクセス総数は約 34 万件であった。

解析期間における日毎のアクセス数の推移を、第 1 図(A)に示す。縦軸は 1 日あたりのアクセス件数、横軸は時間で、軸ラベルは 2023 年度中の N 月を示す。図中には多くのピークが見られ、特に 2024 年 2 月 20 日及び 2024 年 2 月 21 日のものは 7000 件/日を超えている。しかし、残念ながらこのような急激なピークを示すアクセスは、通常クローラによるものであることが多い。

#### 5. クローラによるアクセス

クローラは Robot または BOT とも呼ばれ「ウェブ上の文書や画像などを周期的に取得し、自動的にデータベース化

するプログラム」のことを言う (Wikipedia, 2022)。従来は Google や Bing, Yahoo など大手検索会社によるアクセスが主であったが、最近では他事業者によるアクセスも増えている。

記事の購読傾向を解析する上で欠かせないのが、このようなイレギュラーなアクセスの排除である。イレギュラーなアクセスには、クローラのほか、同一 IP アドレスからの連続的なアクセスなども含まれる。そのような連続的なアクセスは、目的不明の情報収集や攻撃によるものであることが多く、人間の興味を示すものではないため、今回のような解析においてはノイズとなる。

クローラによるアクセスの判別には一般に、前述の USER\_AGENT の情報が使われる。いわゆる「お行儀のよい」クローラであれば、USER\_AGENT 中に、自分がクローラであること、また詳細を示した URL などを記載している。

例えば Google のクローラであれば、USER\_AGENT に "(compatible; Googlebot/2.1; http://www.google.com/bot.html)" といった文字列が含まれており、容易に判別が可能である。

これを手掛かりに調べた結果、Googlebot (Alphabet, Inc.), bingbot (Microsoft Corp.), Y!J-WSC 及び Y!J-ASR (ヤフージャパン), Amazonbot (Amazon.com, Inc.) などの定番のクローラによるアクセスが見付かった。また、ClaudeBot (Anthropic PBC), GPTBot (OpenAI, Inc.) などの AI 開発系の会社や、NDL (国立国会図書館) からのアクセスも多かった。AI 開発系の会社は LLM (大規模言語モデル) の学習データを収集するため、国立国会図書館は 2002 年よりインターネット資料収集保存事業 (WARP) を実施しているため、それに関係したアクセスと思われる (国立国会図書館, 2013)。

おもしろいところでは、少数ではあるが米国でインターネットベースの論文類似性検出サービスを提供している

Turnitin, LLC. によるアクセスなどもあり、検索以外の分野におけるクローラの必要性について垣間見ることが可能である。

クローラによるアクセスを削除した後の、日毎のアクセス数のグラフを第 1 図 (B) に示す (縦軸の最大値が変更になっていることに注意)。クローラによる影響を除いた後にも、2023 年 5 月 24 日、2023 年 7 月 16 日及び 2023 年 8 月 31 日には大きなピークが残っている。そこで同日について、今度は IP アドレス等を元にした調査を行ったところ、2023 年 5 月 24 日はアンソロピック (anthropic-ai) から、2023 年 7 月 16 日は中国の通信事業者 (bytedance.com) から、2023 年 8 月 31 日はアリババグループ (Alibaba Cloud LLM) からの集中的なアクセスがあったことがわかった。

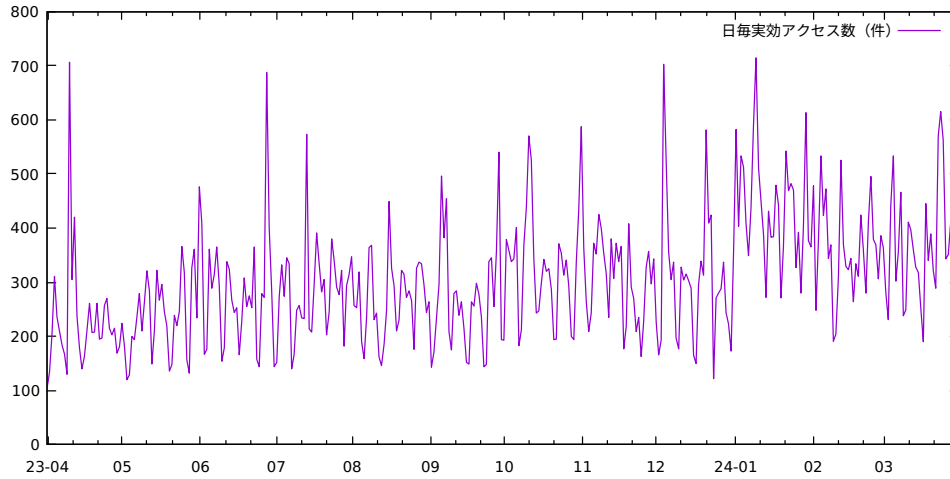
今回除外したクローラの種類と概要、全アクセス数に対する割合を第 1 表に示す。なお、これらクローラ類のアクセスは全体の 67.3 % であった。アクセス中のクローラの占める割合については公式と言える情報はあまり存在しないが、例えば山本 (2019) によれば、J-STAGE において、クローラ対策をする前の全アクセス数に対するクローラの割合は 40-50 % であるとされており、67 % という数字も極端に過大なものではないと考える。

## 6. 解析期間全体を通じたアクセス解析

クローラについての話が長々と続いたが、なにぶん全体の 6 割超を占めるものであったのでご勘弁願いたい。クローラ類のアクセスを除いたアクセス、いわば実効アクセス数は、期間全体で約 12.8 万件であった。これは、1 日あたり 300 件程度になる。日毎の実効アクセス数のグラフを第 2 図に示す。このようにまだピークはあるものの、かなり平準化されていることがわかる。グラフを一瞥したところアクセスには周期性があるように感じられるが、ウイン

第 1 表 GSJ 地質ニュースへのアクセスに占めるクローラの種類とその割合。

No.	クローラ名称等	会社名又はサービス概要	割合 (%)	No.	クローラ名称等	会社名又はサービス概要	割合 (%)
1	Googlebot	Google	24.51	14	PetalBot	(アイルランドのデジタルサービス)	0.46
2	bingbot	Microsoft	12.52	15	Bytespider	百度 (中国系検索サービス)	0.45
3	ClaudeBot	Anthropic (LLM)	9.77	16	naver.me	(韓国系ポータルサイト)	0.43
4	GPTBot	OpenAI (LLM)	6.05	17	BDBot	(OSS系のクローラ)	0.41
5	Amazonbot	Amazon	1.70	18	Alibaba	Alibaba (LLM)	0.40
6	Y!J-WSC, Y!J-ASR	ヤフージャパン	1.67	19	anthropic-ai	Anthropic (LLM)	0.40
7	NDL	国会図書館	1.59	20	scrapy.org	(OSS系のクローラ)	0.29
8	Facebook	Facebook	0.94	21	PhxBot	(OSS系のクローラ)	0.28
9	Linespider	LINE	0.82	22	Baiduspider	百度 (中国系検索サービス)	0.14
10	JPNIC	(インターネット管理団体)	0.77	23	Applebot	Apple	0.13
11	MicrosoftPreview	Microsoft	0.62	24	Turnitin	(論文類似度検索サービス)	0.12
12	Sogou.com	(中国系検索サービス)	0.61	25	その他	接続エラー、攻撃など	1.26
13	DotBot	(OSS系のクローラ)	0.57		合計		66.87



第2図 GSI地質ニュース記事への実効アクセス数(2023年度)。横軸は月、縦軸はアクセス件数。

第2表 GSI地質ニュース記事の年間アクセス数トップ30。

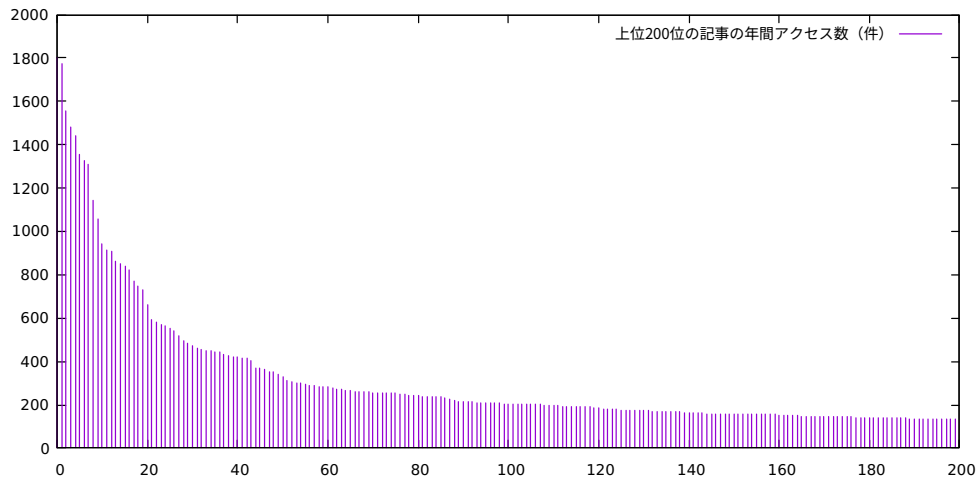
巻号、ページ	アクセス数	記事タイトル	著者
vol2.no7 212-214	1802	南海トラフ三連動型地震・M9はあり得るか？	瀬野徹三
vol11.no2 p49-55	1771	節理と片理	佐脇貴幸
vol10.no7 p148-152	1557	東京低地の沖積層	小松原純子
vol6.no4 p113-120	1483	東西日本の地質学的境界【第六話】日本海の拡大	高橋雅紀
vol5.no8 244-250	1440	東西日本の地質学的境界【第二話】見えない不連続	高橋雅紀
vol4.no12 337-345	1355	磁化率(magnetic susceptibility)を巡る雑感	森尻理恵・中川 充
vol10.no10 p235-241	1326	斑れい岩類：その種類・成因と特徴	山崎 徹
vol2.no12 357-360	1311	浦安市における液状化被害・復旧状況と不動産取引における地質情報の活用策	本間 勝
vol4.no10 297-305	1141	つくば市谷田部の地名「福田坪」と「要害」の由来と地形・地質瞥見	杉山雄一
vol6.no5 p149-157	1055	東西日本の地質学的境界【第七話】火山フロントのずれ	高橋雅紀
vol6.no8 p251	943	東西日本の地質学的境界【第九話】幻の利根川構造線	高橋雅紀
vol4.no11 315-331	914	大阪湾岸の東西性正断層「高石断層」と深部流体の貫入モデル	杉山雄一・今西和俊
vol8.no11 p301-307	909	鹿沼土の話①一探掘から製品まで	徐 維那ほか
vol6.no10 p315-331	860	東西日本の地質学的境界【第十話】待ち構えていた難問	高橋雅紀
vol3.no12 357-365	852	環境中のベリリウムとその地球化学	金井 豊
vol3.no8 238-244	837	温故知新：宮沢賢治と地震	加藤碩一
vol5.no10 311-319	820	東西日本の地質学的境界【第四話】関東平野下の地帯配列	高橋雅紀
vol2.no10 289-292	770	アイスランドの地質	山崎 徹・庄山紀久子
vol10.no7 p153-158	749	東京都区部の台地を構成する地層の層序ー東京層と下総層群ー	納谷友規・中澤 努
vol4.no3 69-74	728	伊勢神宮式年遷宮「お白石持」行事における白石の起源	内野隆之
vol12.no8 p248-250	660	書籍紹介「人類の起源 古代DNAが語るホモ・サビエンスの『大いなる旅』」	七山 太
vol5.no9 279-286	595	東西日本の地質学的境界【第三話】銚子の帰属	高橋雅紀
vol2.no3 67-68	580	誕生石の鉱物科学ー3月 アクアマリンー	奥山康子
vol8.no3 p81-82	571	新刊紹介「地球46億年 気候大変動」	七山 太
vol3.no3 73-78	567	地質図とは何かー地質図幅からシームレス地質図へー	斎藤 眞
vol3.no1 31-32	554	誕生石の鉱物科学ー1月 ガーネットー	奥山康子
vol5.no7 218-225	544	東西日本の地質学的境界【第一話】事の発端	高橋雅紀
vol2.no5 129-130	517	大阪の地史/地質情報展2012	おおさか事務局
vol8.no2 p31-40	497	5万分の1地質図幅「身延」の紹介	尾崎正紀
vol9.no7 p.195-200	483	鹿沼土の話②一鹿沼土を観察してみる	地下まゆみほか

ドウなどを考えず簡易にフーリエ変換してみたところ 60 日前後の周期が強いとの結果になり、わかりやすい理由が推測できるものではなかった。

全期間を通じてのアクセス数上位 30 件の記事を第 2 表に示す。人気のある記事の傾向を読み取ることは難しいが、時事的な理由から地震や地盤に関連するもの(後述する)、そして連載ものや解説が読まれているように感じられる。特に「東西日本の地質学的境界」の連載は上位 30 件中 8 件がはいる人気の記事であるが、これは著者である高橋雅紀氏の知名度も影響しているかもしれない。

アクセス数トップの記事は 1802 件/年、30 位の記事は

483 件/年であるが、この後はどのような傾向を示すのであろうか。上位 200 位までの記事についての年間アクセス数を第 3 図に示す。特に示さないが、アクセス数はこの後も漸減を続け、いわゆる「ロングテール型」の傾向となっている。また一般に「80 対 20 の法則」(この場合であれば、上位 20% の記事がアクセスの 80% を占める)と言われるものがあるが、GSJ 地質ニュースの場合、累計アクセス数の 80% の時点の記事数は上位約 1/3 の地点にあり、一般に言われるほどの集中度ではない、つまり比較的平均的に読まれていることがわかる。



第3図 アクセス数上位 200 件の記事について、年間のアクセス数をグラフ化したもの。横軸は順位、縦軸はアクセス件数。

## 7. 個別記事の年間アクセス傾向

2023 年度のトップアクセス記事は「南海トラフ三連動型地震・M9 はあり得るか？」(瀬野, 2013)であったが、同記事の年間のアクセス傾向には大きな特徴があった。同記事の日毎のアクセス数を第4図(A)に示す。また比較として、アクセス数2位の記事である「節理と片理」(佐脇, 2022)のアクセス数を第4図(B)に示す。後者は年間を通して同程度のアクセス数であるが、前者は2024年1月1日以降に急激な増加を見せている。同日は令和6年能登半島地震が起きた日であり、大地震に興味を持った読者が多くいたこと、またその興味が少なくとも3か月に渡って続いていることが見て取れる。

地震後に同様の傾向を示した記事はこれだけではない。第5図に年間アクセス数上位20位までの記事について、日毎のアクセス数を累積した結果を示す。2024年1月1日を境にアクセス数が急増した記事については黒線、それ以外のものについては赤破線で示したが、4つの記事において急増が見られることがわかる。ちなみに急増した記事は、前述の1位の記事と4位「東西日本の地質学的境界【第六話】日本海の拡大」(高橋, 2017)、8位「浦安市における液化化被害・復旧状況と不動産取引における地質情報の活用策」(本間, 2013)、12位「大阪湾岸の東西性正断層『高石断層』と深部流体の貫入モデル」(杉山・今西, 2015)である。4位の記事については関連性は判然としませんが、他は地震や地震災害に関連する記事であることがわかる。

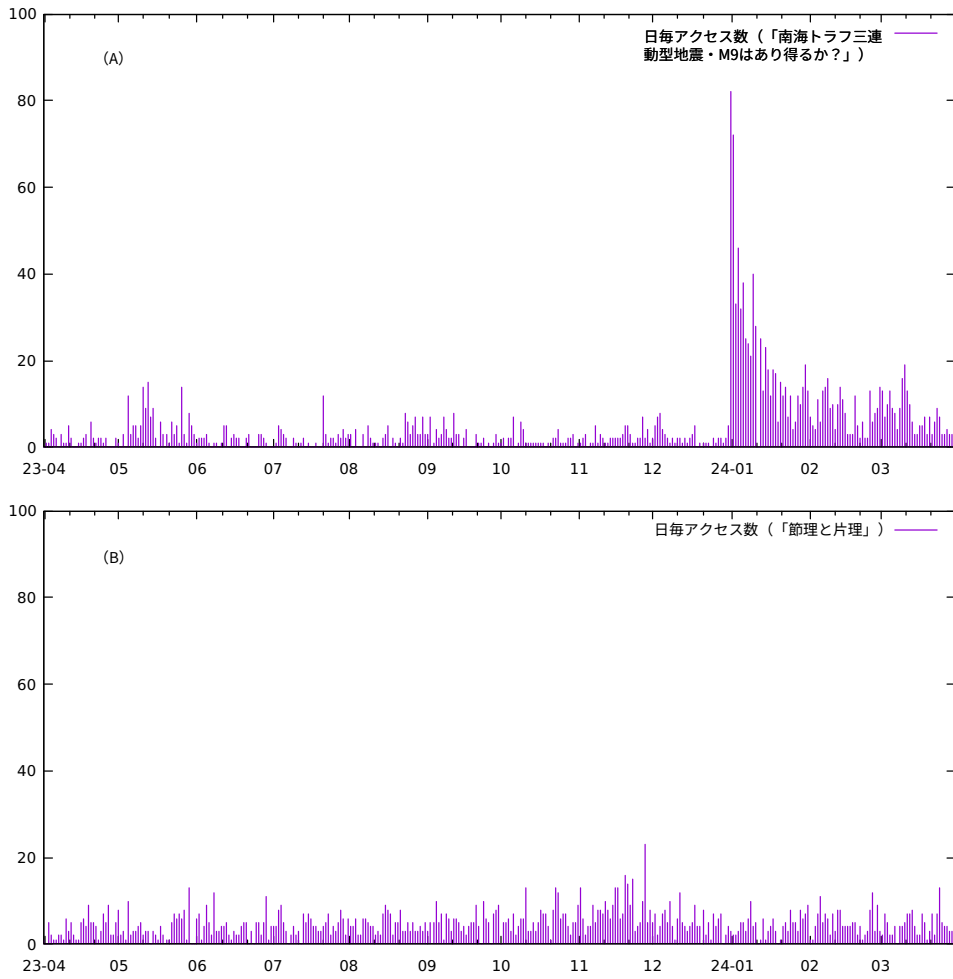
さらに詳細に第5図を見ると、20記事中の少なくとも半分程度の記事は2024年1月1日以降にアクセス数を伸

ばしているように思える。全体のアクセス数の平均を見ると、2023年12月31日までが278.3件/日であるのに対し、2024年1月1日以降は388.8件/日となっており、大きく増加している。この増加の幾分かは地震による影響ではなく、そのような記事に触れて新規読者となった方々のアクセスであると考えるのは期待しすぎであろうか。

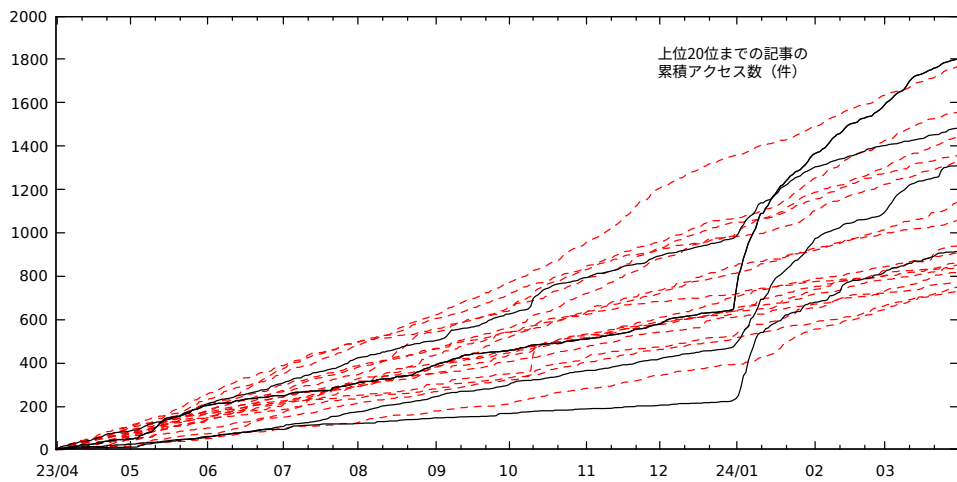
## 8. 2023 年度出版の記事のアクセス傾向

一方で2023年度に新たに出版された記事のアクセス傾向はどうなっているであろうか。第2表に示した上位30位までの記事のうちでは、2023年度に出版されたものは21位の「書籍紹介「人類の起源 古代DNAが語るホモ・サピエンスの『大なる旅』」」(七山, 2023)1件のみであった。

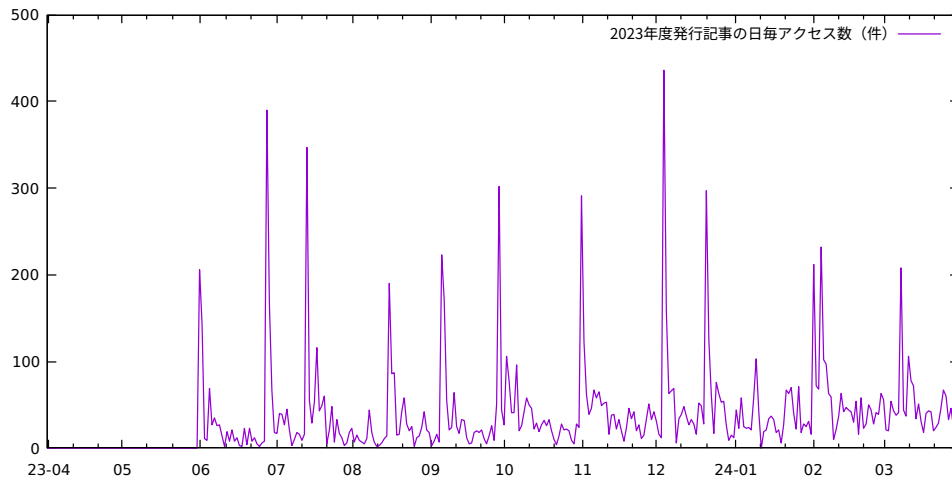
ひとつの試みとして、2023年度に出版された記事のみを抽出して日毎のアクセス傾向を調べてみた。その結果を第6図に示す。複数の明白なピークがあるが、これは新しい号が公開された日にほぼ一致しており、各月の号が公開されると短期間のうちに購読されることがわかる。これは非常に意外な結果であった。というのは、最新号の公開は地質調査総合センター内部でのみアナウンスを行っていたからである。しかし、ログを確認してみた結果、ピークを構成するアクセスのうち内部のもの割合は限定的であり、多くは外部からのアクセスであった。外部の購読者がどのようにして公開日の情報を得ているかは興味深いところである。



第4図 特徴的なアクセスを示した2記事について、アクセス数の日変化を示したもの。横軸は月、縦軸はアクセス件数。(A)は年間アクセス数1位の記事であるが、2024年1月1日に発生した能登半島地震の後に急速にアクセス数が伸びている。対して(B)の記事は、年間を通じて同程度のアクセスがある。



第5図 年間アクセス数上位20位までの記事について、その累積アクセス数を示したもの。横軸は月、縦軸はアクセス件数。1月1日以降に急激な伸びを示した記事が複数あることがわかる。



第 6 図 2023 年度に発行した記事 (vol. 12, no. 4 から vol. 13, nos. 2-3 まで) のみのアクセス数の日変化を示したもの。横軸は月、縦軸はアクセス件数。公開日が不定期であるためわかりにくいだが、200 件/日を越えるピークを示している日が、ほぼ公開日に一致する (1 月上旬の後ろのピークを除く)。

## 9. X (旧 Twitter) による広報の効果について

GSJ 地質ニュースを外部にアピールするための試みの 1 つとして、2024 年 1 月号ならびに 2・3 月合併号については、産総研公式 X (@AIST\_JP) によるポストを行っている。1 月号については 2 月 2 日に公開し、2 月 5 日にポスト ([https://twitter.com/AIST\\_JP/status/1754452343557337221](https://twitter.com/AIST_JP/status/1754452343557337221), 閲覧日: 2024 年 5 月 14 日)、2・3 月合併号については 3 月 8 日に公開し、3 月 19 日にポスト ([https://twitter.com/AIST\\_JP/status/1769858867360338066](https://twitter.com/AIST_JP/status/1769858867360338066), 閲覧日: 2024 年 5 月 14 日) を行った。

第 6 図を見ると、1 月号については公開日とポストした日の両方にアクセスのピークがあり、ポストに一定の効果があったであろうことがわかる。しかし 2・3 月合併号についてはこれといった影響は見られないようである。とはいえ、1 月号の 2 回のピークについて詳細に見ると、ポストにて写真を上げて宣伝した「地質標本館企画展『生痕化石—大地に刻まれた生命の痕跡』開催報告」(清家ほか, 2024) のアクセス割合が 9.0% から 29.3% と伸びており、一定の効果があったことが伺える。

## 10. おわりに

GSJ 地質ニュースには多くの記事が寄稿されるが、どのような記事が読まれているのかアクセス状況が知りたい、という要望は常にあり、最低限のクローラの除外を含む簡易的な解析結果については内部で公開していた。しかし、

時折不可能な結果が表れることがあり、その理由について悩むことがあった。しかし、今回アクセスログを精査したことにより、かなり実態に近い結果が得られたのではないかと考える。クローラによるアクセスについていえば、AI (LLM) の開発はまだ競争の途中にあることから、例えば、Open Source な LLM の訓練を目的とした小規模集団の情報収集のためのアクセスなど、増加、多様化するのではないかと予想される。

今回特に記載しなかったが、2023 年度の解析結果は、地震の影響が大きいように見えて、実は 2024 年 1 月 1 日の前後でアクセス上位記事の入れ替わったものは 2 件だけであった。これは、長期に読まれる記事は固定化しているということであり、ある意味アクセス解析の必要性を否定しかねないものである。今後は、より短期的なアクセス傾向を把握するための手法を検討する必要があると考える。

## 文 献

- Apache Software Foundation (2024) ログファイル - Apache HTTP サーバ バージョン 2.4. <https://httpd.apache.org/docs/2.4/logs.html> (閲覧日: 2024 年 4 月 30 日)。
- 国立国会図書館 (2013) 国立国会図書館インターネット資料収集保存事業. <https://warp.ndl.go.jp/> (閲覧日: 2024 年 4 月 30 日)。
- 本間 勝 (2013) 浦安市における液状化被害・復旧状況と不動産取引における地質情報の活用策. GSJ 地質

- ニュース, 2, 357-360.
- 七山 太 (2023) 書籍紹介「人類の起源 古代DNA が語るホモ・サピエンスの『大いなる旅』」. GSJ 地質ニュース, 12, 248-250.
- 佐脇貴幸 (2022) 節理と片理. GSJ 地質ニュース, 11, 49-55.
- 清家弘治・森田澄人・瀬戸口 希・都井美穂 (2024) 地質標本館企画展「生痕化石—大地に刻まれた生命の痕跡」開催報告. GSJ 地質ニュース, 13, 14-18.
- 瀬野徹三 (2013) 南海トラフ三連動型地震・M9 はあり得るか? GSJ 地質ニュース, 2, 212-214.
- 杉山雄一・今西和俊 (2015) 大阪湾岸の東西性正断層「高石断層」と深部流体の貫入モデル. GSJ 地質ニュース, 4, 315-331.
- 高橋雅紀 (2017) 東西日本の地質学的境界【第六話】日本海の拡大. GSJ 地質ニュース, 6, 113-120.
- Wikipedia (2022) クローラ. <https://ja.wikipedia.org/wiki/%E3%82%AF%E3%83%AD%E3%83%BC%E3%83%A9> (2022/10/11 版).
- Wikipedia (2023) HTTP リファラ. <https://ja.wikipedia.org/wiki/HTTP%E3%83%AA%E3%83%95%E3%82%A1%E3%83%A9> (2023/11/9 版).
- 山本浩万 (2019) J-STAGE アクセス統計とクローラについて. 日本リモートセンシング学会誌, 39, 156-160.
- 
- OHNO Tetsuji (2024) Access analysis of articles on GSJ Chishitsu News for 2023FY using web server logs.
- 

(受付: 2024年5月17日)