

http access log を解析してみた

瀬川朋紀

計測・制御技術支援室 計測機器技術グループ

概要

名古屋大学宇宙地球環境研究所が運用する 2 つの Web サイトを対象に 2021 年 7 月から 2025 年 12 月までの http access log を解析し、学外からのアクセスの特徴を整理した。本報告書では、初期的な調査ではあるものの、両サイトの利用傾向はほぼ独立していることが確認された。一方で、2025 年 6 月には両サイトにアクセスの増加が認められた。今後は、目的に応じた分析を進めていきたい。

1 はじめに

近年、Web サイトは組織活動に不可欠な情報基盤として定着し、Web サイトを保有しない組織はほぼ存在しないと見えるほど広く普及している。Web サイトの運営においては、その利用実態を把握するための Web アクセスログが重要な情報源であり、配信効率の最適化、ユーザ行動分析、マーケティング戦略の高度化、サイバーセキュリティ対策など多様な場面でログ解析の有効性が指摘されている [1]。また、教員が提出する外部評価資料には、Web サイトに関するアクセス数などの情報も含まれるようになってきた。

本報告書では、名古屋大学宇宙地球環境研究所で運営する Web サイトのうち 2 つのサイトを対象として、2021 年 7 月から 2025 年 12 月までの http access log を基礎的に解析し、その傾向を整理することを目的とする。

2 基本情報

解析対象とした Web サイトは次の 2 つである。

- 宇宙地球環境研究所の公式 Web サイト
- あらせ衛星サイエンスセンターの Web サイト

国際共同利用・共同研究拠点である宇宙地球環境研究所の公式 Web サイト^{*1}には、同研究所の研究紹介や成果などが公開されている。あらせ衛星サイエンスセンターの Web サイト^{*2}は、宇宙航空研究開発機構宇宙科学研究所と名古屋大学宇宙地球環境研究所が共同で運用するあらせサイエンスセンターによって運営されるあらせ衛星の観測データ配信サイトであり、数百万件を超える ERG プロジェクトの科学データファイルをアーカイブ・公開、国内外の研究者に広く利用されている。

いずれのサイトも、OS は Red Hat Enterprise Linux 8 系、Web サーバソフトウェアは Apache HTTP Server 2.4 で運用されている。総当たり攻撃 (brute force attack) や不正アクセス対応として、fail2ban も運用している。Fail2ban の設定は、主に

- http status code 403, 404 : bantime=600, findtime=5, maxretry=10
- phpmyadmin : bantime=600, findtime=5, maxretry=1

などである。

解析期間はいずれも日本時間の 2021 年 7 月 1 日から 2025 年 12 月 31 日までとした。

3 データ

^{*1} <https://www.isee.nagoya-u.ac.jp/>

^{*2} <https://ergsc.isee.nagoya-u.ac.jp/>

両サイトの **http access log** は、**Apache Combined Log Format** に基づき、リモート IP アドレス、リクエスト行、ステータスコード、レスポンスサイズ、Referer、User-Agent など 9 項目から構成される (表 1)。

本報告書で対象とした解析期間におけるログ件数は、ISEE では 109,536,422 行 (約 1.10 億行)、ERGSC では 1,630,995,230 行 (約 16.31 億行) である。ERGSC は ISEE の約 15 倍の規模であり、データ量が非常に大きいことから、従来の shell ベース処理では性能面の制約が大きいことも確認された。なお、shell ベースでの処理に要する時間の詳細は、宇宙地球環境研究所 ISEE 技術部報告集^{*3}にて別途報告予定である。

また、リモート IP アドレスがプライベートアドレス、マルチキャストアドレス、および 133.6.0.0/16・133.47.0.0/16 以外であるアクセスを「学外アクセス」と定義した。この定義に基づく、学外アクセスの総件数は ISEE で 102,461,412 件、ERGSC で 1,581,409,533 件となる。ISEE では約 93.5%、ERGSC では約 97.0% が学外からのアクセス記録である。

両サイトとも対象期間中、法定点検に伴う停電による Web サイトの停止期間を除き、通算 1639 日稼働している。一日平均のアクセス件数は、ISEE で約 66,831.3 件/日、ERGSC で約 995,116.1 件/日であり、学外アクセスに限定すると、ISEE で約 62,514.6 件/日、ERGSC で約 964,862.4 件/日である。

表 1 収集している情報

No.	Format String	Description
1	%a	リモート IP アドレス
2	%l	リモートログ名
3	%u	リモートユーザ (認証による)
4	%t	リクエストを受付けた時刻
5	%r	リクエストの最初の行
6	%>s	Status
7	%>b	レスポンスのバイト数
8	%{Referer}i	サーバに送られたリクエストの Referer ヘッダの内容
9	%{User-Agent}i	サーバに送られたリクエストの User-Agent ヘッダの内容

4 解析

本章では、学外アクセスに限定して解析していく。

4.1 全期間

両サイトのステータスコード別の分布を比較する。本報告書で扱うステータスコードの意味は、表 3 に示す。

表 2 に示すとおり、ISEE では 200 が 88,620,143 件 (約 86.5%) と最も多く、次いで 304 の 7,581,641 件 (約 7.4%) が続く。200 番台および 300 番台の合計は 97,221,772 件 (約 94.9%) と大部分を占めている。一方、400 番台は 5,169,079 件 (約 5.0%) で、その多くが 404 (4.6%) であった。

ERGSC では、304 が 836,123,241 件 (約 52.9%) で最も多く、次いで 200 の 662,450,590 件 (約 41.9%) が続く。200 番台および 300 番台の合計は約 1,515,221,499 件 (約 95.8%) と ISEE と同様に高い割合を示す。ERGSC の 400 番台は 61,965,424 件 (約 3.9%) であり、ISEE の 400 番台の 5.0% と比較して、相対的に少ないことがわかった。なお、500 番台は両サイトとも 0.1% 程度であった。

これらより、ISEE は 200 の応答が支配的で、404 の割合が比較的高い、ERGSC は 304 と 200 の応答が大半を占め、キャッシュ利用が多いことが特徴的といった傾向が確認された。

^{*3} <https://tech.isee.nagoya-u.ac.jp/>

4.2 月別

月別アクセス数を表 4 に示す。ISEE と ERGSC の月別の学外アクセス数の Pearson 相関係数は -0.0692 であり、ほぼ無関係とわかった。しかし、ISEE、ERGSC とともに 2025 年 6 月にアクセス数の大きな増加があった。

ERGSC では、2025 年 6 月にステータスコード 403、404 の件数が他の月より多いことが確認された (図 2)。しかし、ISEE では同時期に特筆すべき増加はみられなかった (図 1)。

これらの増加要因としては、特定期間に外部から集中的なアクセスが発生した、あるいはクローラや自動化ツールの増加など複数の可能性が考えられるが、現段階では特定に至っていない。

また、ISEE では、2025 年 6 月を除くと、年毎のアクセス数の多い月が 2021 年 12 月に 2,243,488 件、2022 年 10 月に 2,223,357 件、2023 年 12 月に 2,511,318 件、2024 年 12 月に 2,902,915 件、2025 年 12 月に 2,287,868 件と分かった。毎年 12 月がピークになりやすい、と推測される。これらの要因に関しても、詳細にログを分析する必要がある、現段階では特定に至っていない。

表 2 学外アクセスにおけるステータスコード別の件数。
ステータスコードの内容は表 3 に記載した。

Status	ISEE		ERGSC	
	count	%	count	%
200	88620143	86.5	662450590	41.9
204	0	-	38	0.0
206	609817	0.6	14444315	0.9
301	313324	0.3	1019809	0.1
302	96847	0.1	1183506	0.1
304	7581641	7.4	836123241	52.9
400	28430	0.0	58071	0.0
401	69614	0.1	38219985	2.4
403	333504	0.3	5210103	0.3
404	4735273	4.6	17765601	1.1
405	128	0.0	3660	0.0
406	0	-	330	0.0
408	10	0.0	419206	0.0
409	286	0.0	0	-
412	2	0.0	0	-
414	0	-	3	0.0
416	1741	0.0	288290	0.0
417	3	0.0	0	-
421	88	0.0	175	0.0
500	61468	0.1	2581503	0.2
501	63	0.0	25	0.0
502	148	0.0	76816	0.0
503	8825	0.0	1563898	0.1
504	57	0.0	367	0.0

表 3 http status code.

<https://www.iana.org/assignments/http-status-codes/http-status-codes.xhtml> より本報告書で取り扱うステータスコードのみ抜粋した。

200	OK	正常に処理された.
204	No Content	要求は成功したが返す本文なし.
206	Partial Content	クライアントが Range 指定した部分的レスポンス.
301	Moved Permanently	リソースが恒久的に別の場所に移動.
302	Found	リソースが一時的に別の場所にある.
304	Not Modified	キャッシュが最新なので再送不要.
400	Bad Request	リクエストが不正.
401	Unauthorized	認証が必要／認証情報が無効.
403	Forbidden	アクセス権限がないため、閲覧禁止.
404	Not Found	指定されたリソースが存在しない.
405	Method Not Allowed	リソースがその HTTP メソッドを許可していない.
406	Not Acceptable	要求ヘッダ（ Accept 等）の条件に合う内容を返せない.
408	Request Timeout	クライアント側が時間内にリクエストを完了しなかった.
409	Conflict	リクエストが現在のサーバーの状態と競合.
412	Precondition Failed	前提条件付きリクエストが満たせなかった.
414	URI Too Long	URI が長すぎて処理不可.
416	Range Not Satisfiable	クライアントがファイルの部分的な取得 (Range リクエスト) を求めた際、その指定範囲が実際のデータサイズを超えている、または無効である場合にサーバーが返すエラー.
417	Expectation Failed	Expect リクエストヘッダーで指定された期待がサーバー側と適合しない.
421	Misdirected Request	違うサーバに向けられた誤ったリクエスト.
500	Internal Server Error	サーバ内部エラー.
501	Not Implemented	サーバがその機能／メソッドを実装していない.
502	Bad Gateway	中継サーバが上流から不正応答を受け取った.
503	Service Unavailable	過負荷・メンテなどで一時的に利用不可.
504	Gateway Timeout	上流サーバからの応答がタイムアウト.

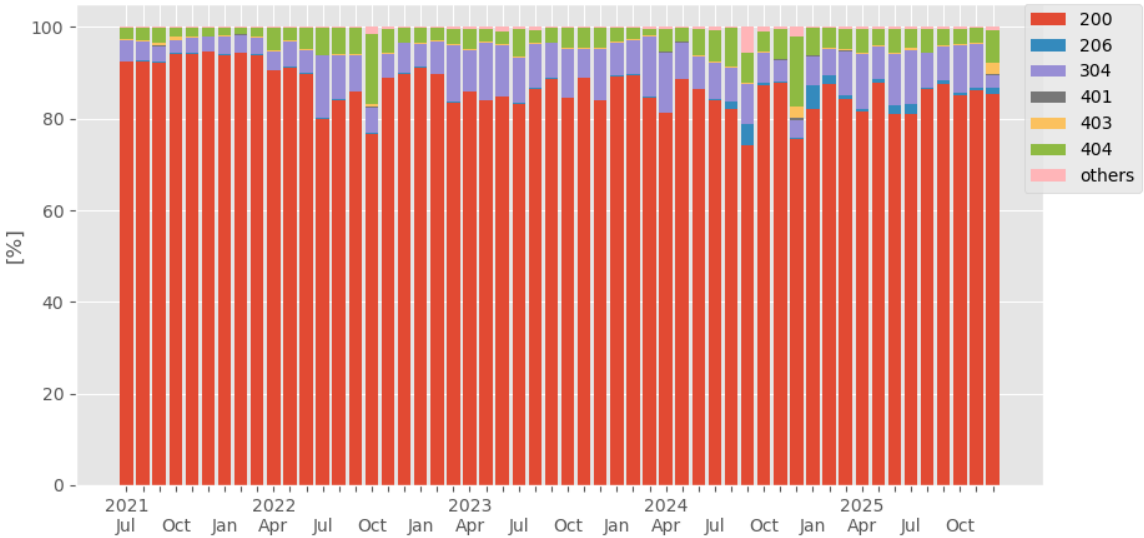
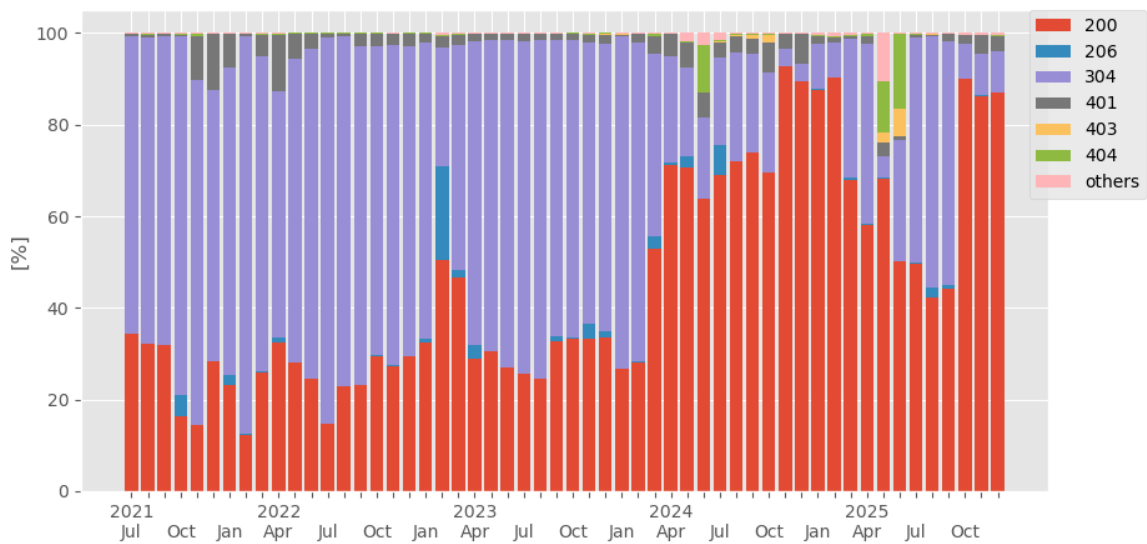


図 1 ISEE における Status Code 別の割合.

表 4 月別学外アクセス数.

year	month	ISEE	ERGSC
2021	07	1605765	26804625
2021	08	1781195	28812428
2021	09	1587721	29612022
2021	10	2022592	13602620
2021	11	2149202	16866899
2021	12	2243488	17726193
2022	01	1918957	15462438
2022	02	1828286	16016847
2022	03	1561583	14521038
2022	04	1349205	18403844
2022	05	1301557	28528489
2022	06	1396237	29429946
2022	07	1267614	34764061
2022	08	1415839	49795614
2022	09	1590966	37442742
2022	10	2223357	35179737
2022	11	2062692	31364302
2022	12	1841291	36093019
2023	01	2019918	39227102
2023	02	1886075	21070585
2023	03	1622221	27011485
2023	04	1245611	44029908
2023	05	1351836	43016245
2023	06	1084679	42677476
2023	07	1174059	44247413
2023	08	1326171	40507691
2023	09	1587089	43973046
2023	10	2157531	42046601
2023	11	2351894	39497574
2023	12	2511318	42690708
2024	01	2318729	40229171
2024	02	2240763	40277146
2024	03	2232784	25446993
2024	04	2303684	17067887
2024	05	2672851	21723137
2024	06	1935987	19807456
2024	07	1757073	23489557
2024	08	2108629	18030872
2024	09	1725237	19271422
2024	10	2152477	13246131
2024	11	2084130	26762431
2024	12	2902915	20559811
2025	01	2177806	13560559
2025	02	2015202	12495063
2025	03	1704560	31765514
2025	04	1739776	36570375
2025	05	2220735	35791302
2025	06	3509367	58534834
2025	07	1924387	35633989
2025	08	1797630	33071520
2025	09	1560795	24706421
2025	10	1751278	18698141
2025	11	1870800	23652239
2025	12	2287868	20594864



5 まとめ

本報告書では、ISEE および ERGSC の 2021 年 7 月から 2025 年 12 月までの http access log を対象として、学外アクセスを中心に基礎的な解析を行った。主な結果を以下にまとめる。

1. アクセス規模の違い.

ERGSC は ISEE の約 15 倍の学外アクセスがあり、両サイトの利用目的の違いが確認された。

2. ステータスコードの特徴.

ISEE では 200 応答が 86.5% と支配的であり、404 の割合が 4.6% と比較的高かった。一方 ERGSC では 304 が 52.9% と最多で、キャッシュを利用したアクセスが非常に多いことが特徴的であった。

3. 2xx/3xx 応答が全体の約 95% を占める.

両サイトとも正常系（200）とキャッシュ系（304）を中心に安定したアクセス傾向を示していた。

4. 月別の変動と相関.

ISEE と ERGSC の月別アクセス数の相関係数は -0.0692 であり、両サイトの増減はほぼ連動しないことがわかった。また、両サイトとも 2025 年 6 月にアクセス数が増加したが、増加要因は本データのみでは特定できなかった。

5. 注意すべきアクセスの存在.

両サイトともに 400 番台の応答は 3～5% 程度含まれ、特に ERGSC の 403/404 の増加が一部の期間で確認された。継続的な監視と詳細分析の対象となる。

今後は、アクセス増加要因の特定や 400 番台の応答の詳細調査などが課題となる。

参考文献

- [1] 大塚真吾, 喜連川優, Web アクセスログとその利活用. 人工知能学会誌, Vol. 21, No. 4, pp. 410-415, 2006., <https://doi.org/10.11517/jjsai.21.4.410>