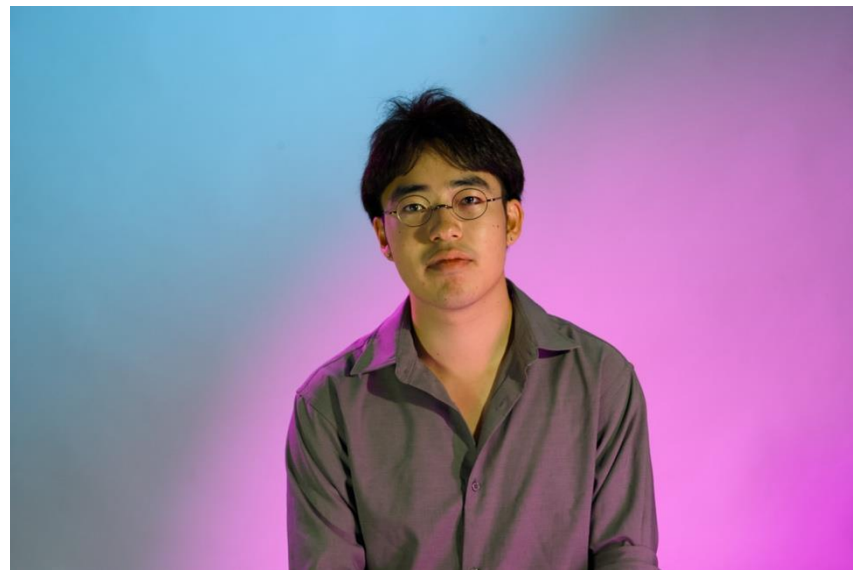


大規模クローリングによる 計量社会学的な分析

2026-07-21 Tatsuhiko Akiyama
@Ultra-Coins総会

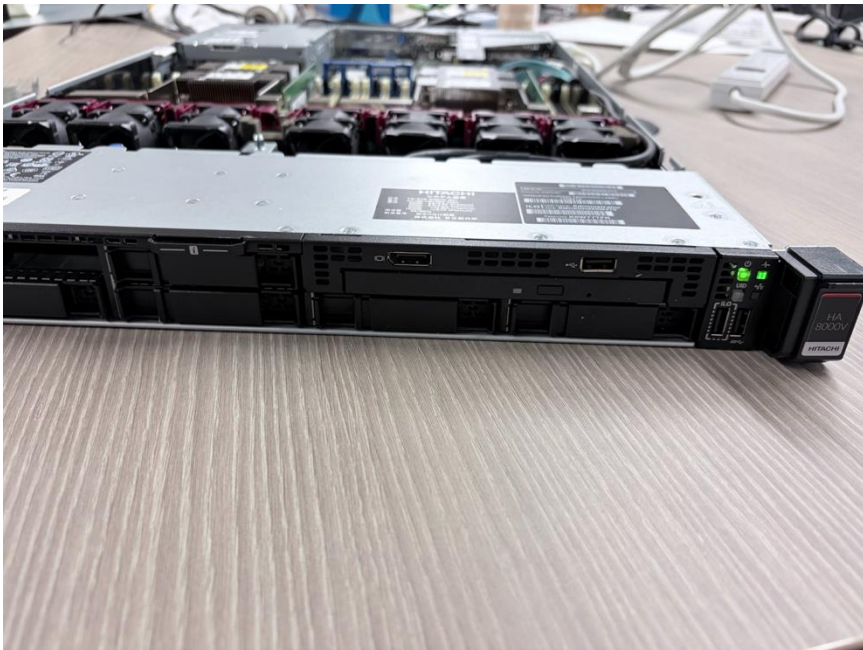
自己紹介

- 秋山 達彦 (あきやま たつひこ)
 - 筑波大学情報学群情報科学類 1年
 - Digital Nature Group (落合研究室)
 - 国立情報学研究所 コンテンツ科学系 技術補佐員
 - Digital Experts ソフトウェアエンジニア
 - 公益財団法人 孫正義育英財団 正財団生
- 研究
 - 大規模言語モデル (LLM)が人を模倣して評価できるか?
 - 人の記憶の変遷や人の創造的な活動を計算機で捉えられるか?



設置している機器

- Hitachi HA8000V DL360
 - CPU: 48 x Intel® Xeon® Gold 6136 CPU @ 3.00GHz (2 Sockets)
 - Memory: 235.78GiB
 - Storage: 3.47TiB
- NIC (10GbE)およびSSDの一部 (990GB)はUltra-Coinsの備品を借用中



大規模クローリングによる計量社会学的な分析

- クローリングしたデータを用いてある程度社会学的なことは言えるかもしれない
- 収集したデータを用いてこれまで行ったこと
 - Internet Engineering Task Force (IETF)のメーリングリストの分析
 - 政府系のサイト (lg.jpやgo.jpドメイン)のクローリングおよびインデックス
 - 匿名掲示板 (はてな匿名ダイアリー)の投稿層の推定
 - co.jpドメインを用いた日本企業のDXに関する分析
- 本日はクローリングをすることでどのようなことがわかるか?についてお話します

ケース: IETFの分析

- IETFの標準化の議論はメーリングリストによって公開されている
 - これまでの全ての議論がWebサイト上にアーカイブされ、自由にダウンロードしたり検索したりできる
 - このメーリングリストを分析すれば良いのでは?
- 実験
 - rsyncでメーリングリストの全データ (33GB)を取得する
 - DuckDB上に格納する
 - DuckDB + Pythonを使っているいろいろ分析する
- 特筆すべき分析手法について
 - 所属についてはメールアドレスのドメインからUnique
 - 国の判定もドメインで行った
 - Gmailなどの一般的なドメインは排除した

Communications

Mailing list archives

The Mail Archive provides a unique and permanent URL for messages received by public email lists.

- Mail Archive website: <https://mailarchive.ietf.org/>
- rsync: `rsync.ietf.org::mailman-archive`
individual list archives can be found by appending /LISTNAME

Mailing list archives can also be [accessed using IMAP](#).

Chat archives

Members of IETF groups often use chat to communicate during and between meetings.

- Pre-2021 (Jabber) chat archives:
 - <https://get.ietf.org/archive/jabber/logs>
- 2021-current (Zulip) archives are available via Zulip zulip.ietf.org

Liaison statements

[Liaison statements](#) are official communications between the IETF and other organizations.

- Web: <https://datatracker.ietf.org/liaison/>

図: 自由にrsyncしろって書いてある <https://www.ietf.org/about/open-records/>

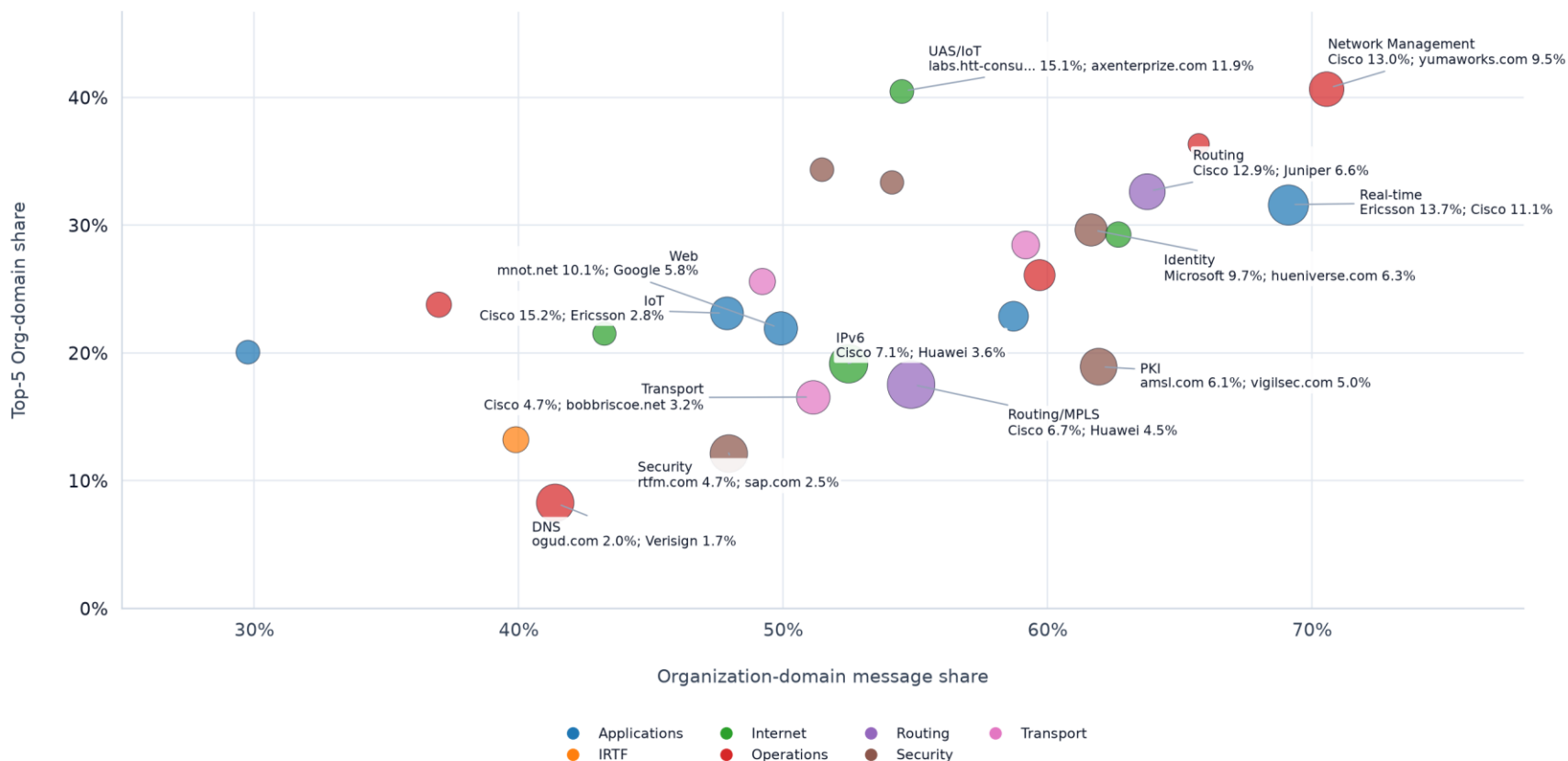
ケース: IETFの分析

- Working Groupに關与している会社の傾向は事業とある程度一致する？

Fig. 2 Organization-domain share by broader IETF field

Unit: one major field from the broader IETF corpus. Unmapped fields and fields under 1,000 messages are excluded.

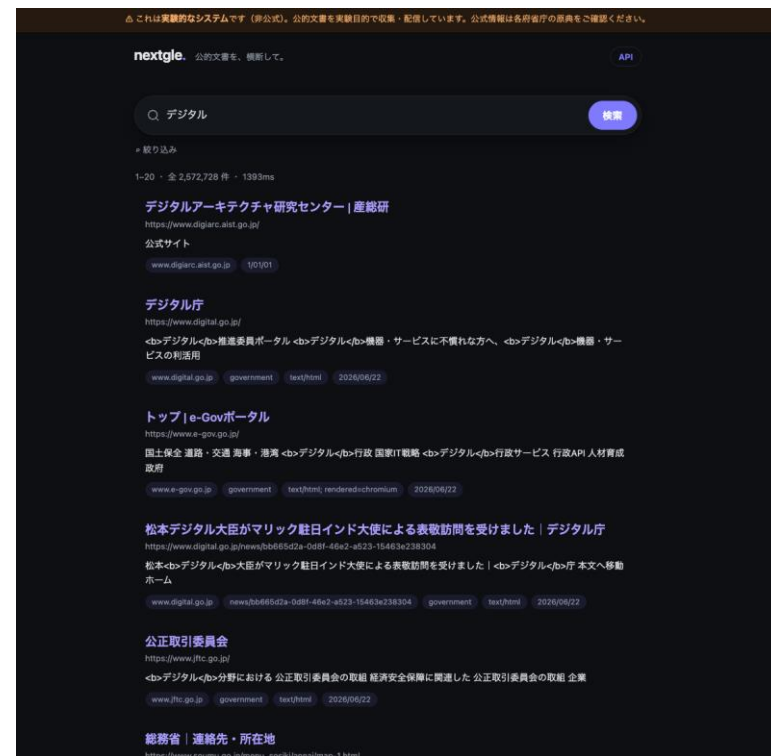
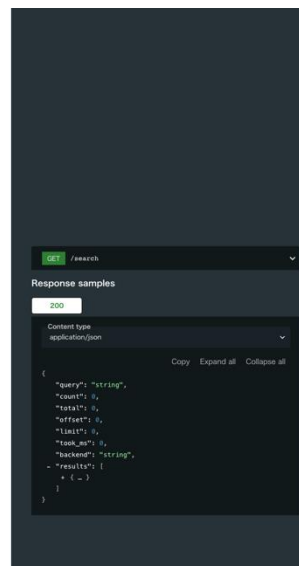
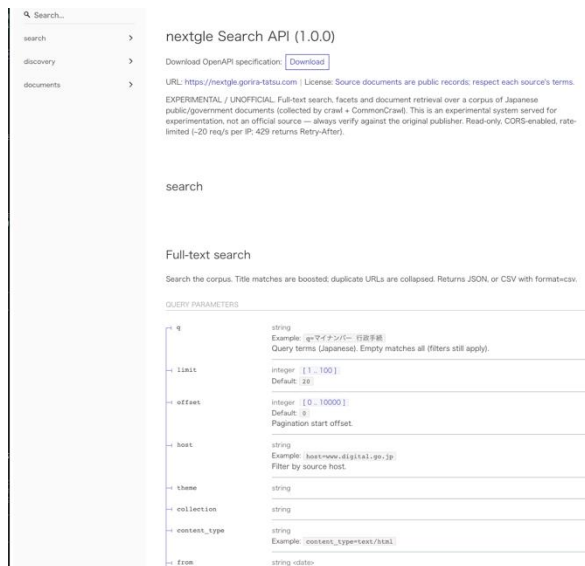
X = Org-domain messages / all field messages. Y = top-5 Org-domain messages / all field messages. Labels show top Org domains.



ケース：政府系のクローリングおよびインデックス

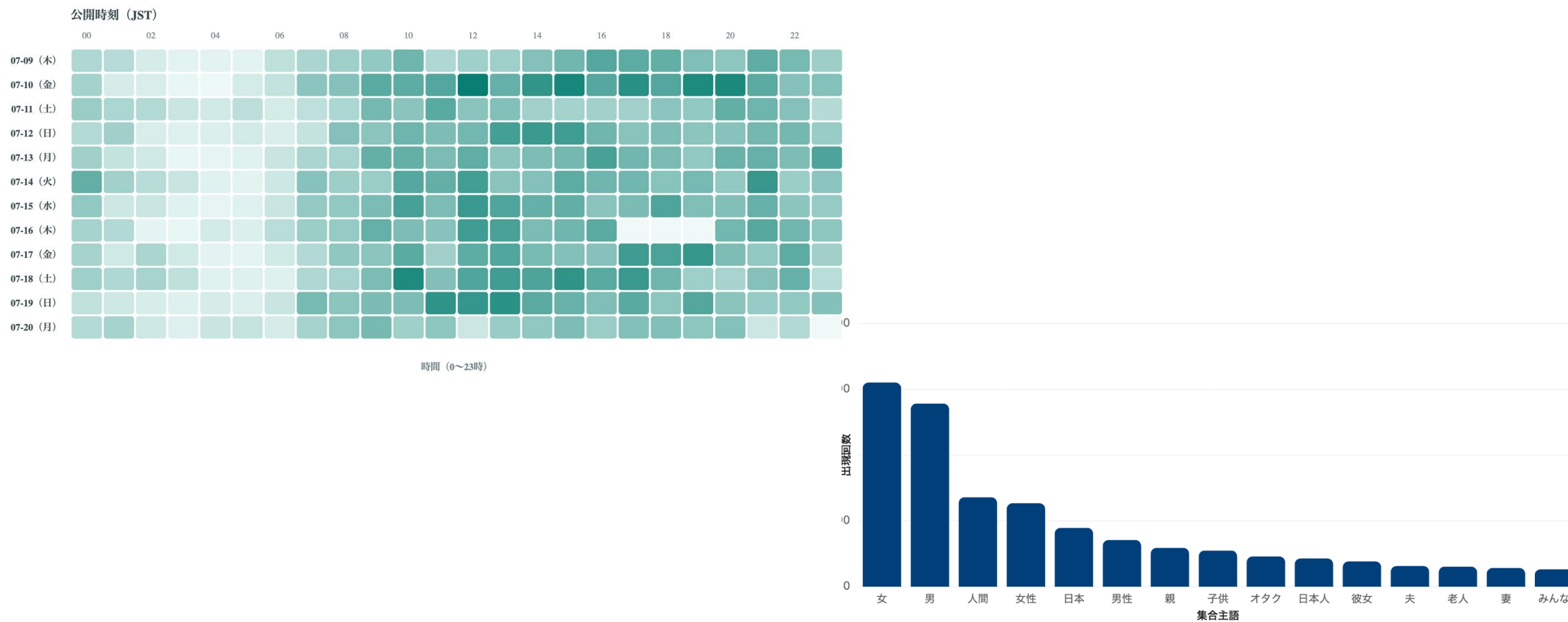
- 政府系のサイトを対象に28321679件のクローリングを行った
 - 政府系サイトへ第三者からの保全、アーカイブを目的としておこなった
 - 本来robot.txtなどには従うことが倫理的観点から重要とされているが、本研究ではアーカイブをする必要がある & 民主主義の尊重の観点 (のちの検証可能性)から従っていない
 - APIは全て外部に公開している
 - レート制限もとくに設けていない
- nextgle.gorira-tatsu.com

28321679



ケース:匿名掲示板の投稿層の推定

- 匿名掲示板であるはてな匿名ダイアリーをクロールし、投稿内容や時間を分析した



ケース:co.jpドメインを用いたDXに関する分析

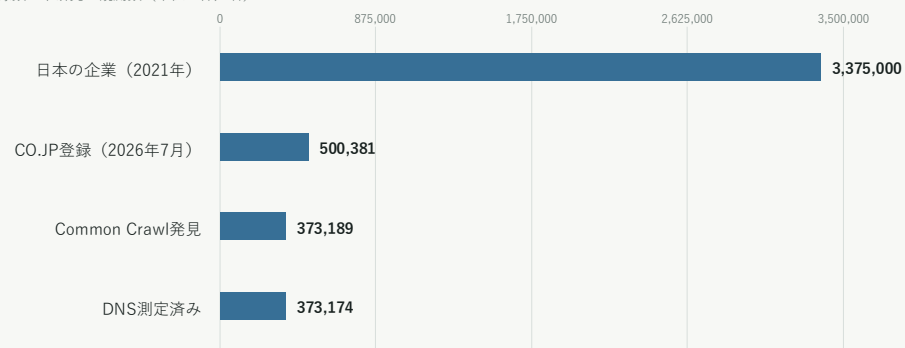
- JRPSが管理しているco.jpドメインは取得に厳格な規則が存在する
 - 日本の企業限定 (かつ登記が必須)
 - 1組織1登録制
- ドメインから辿ることのできるデータ
 - WHOIS情報
 - DNSレコード
 - A
 - AAAA
 - MX
 - TXT
 - NS
 - DNSレコードに紐づくサーバー
- co.jpから日本企業 (おもに中小企業)の分析ができるかも？

データソース

- Common Crawl (CC)と呼ばれるWebのクロールを2008年から行っている財団がデータセットを公開している
 - 少なくとも1000言語のカバー、毎月約100TBほどのクロールをするなど巨大なデータセット
- CCはクロールしたドメインのリストを公開しており、ここからco.jpのドメインのみを抽出する
 - 2008年のsnapshotから遡り373,189件の集合に
 - JRPSの保有しているデータによれば500,381なので、約75%のカバレッジ
- WHOISの情報はJRPSから取得
- DNSはunboundから権威DNSへ

図1 観測対象の範囲

企業数・CO.JP登録数・本研究の観測数（単位：件／者）

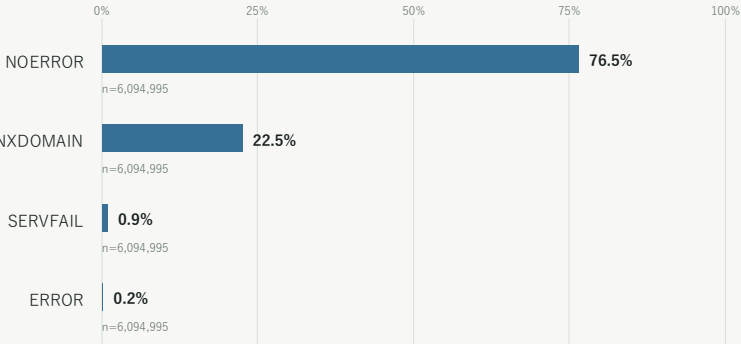


DNSに登録されているデータの分析

- 37万件の取得したドメインからDNSへ問い合わせ

図11 DNS応答状態

最新問い合わせに占める割合

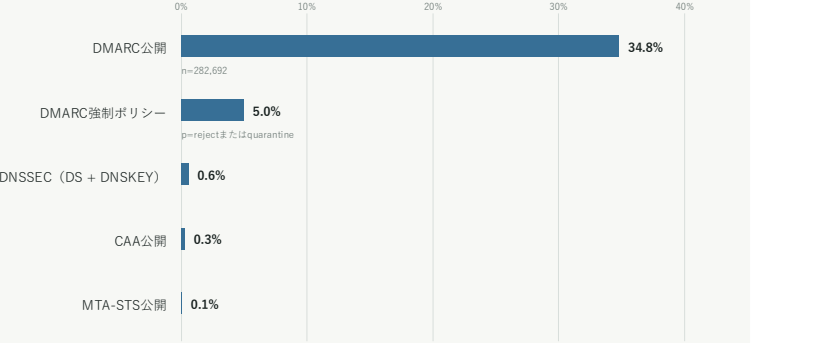


出典：本研究DNS測定。問い合わせ単位。

生成: 2026-07-20 13:32 UTC

図10 公開DNSで確認されたセキュリティ設定

詳細測定済みドメインに占める割合

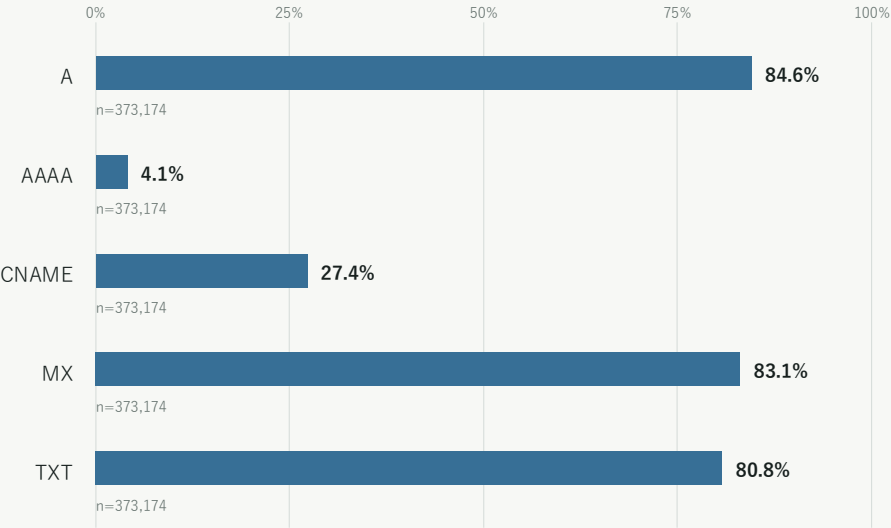


出典：本研究。組織内の情報セキュリティ対策や方針の有無を示すものではない。

生成: 2026-07-20 13:32 UTC

図13 DNSレコード取得範囲

DNS測定済みドメインに占める回答確認率



出典：本研究DNS測定。レコードの存在はサービス到達性を保証しない。

生成: 2026-07-20 13:32 UTC

メールについて

- MXレコードに記載されているアドレスから利用しているサービスを推定することが可能
- Gmailらが手動したポリシーによりDMARC対応が課題

図14 推定メール基盤

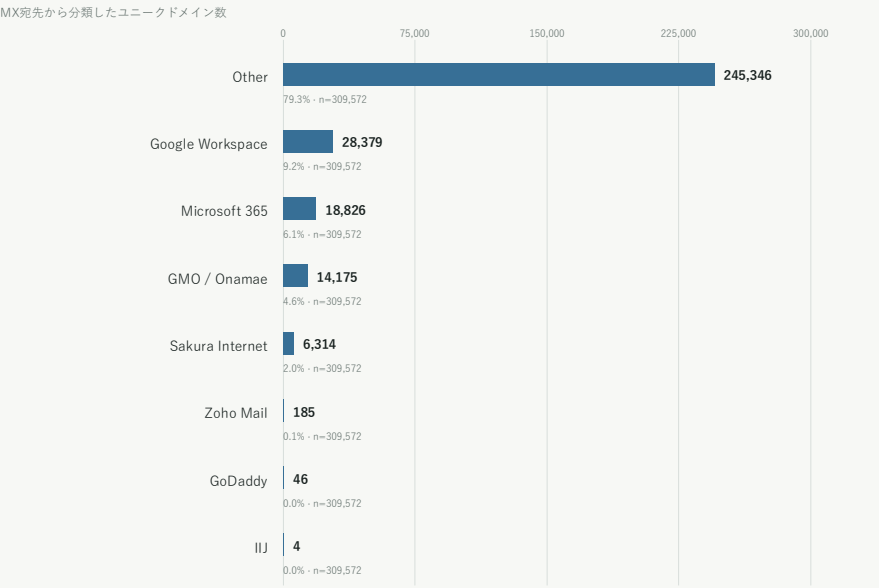


図15 上位MX宛先

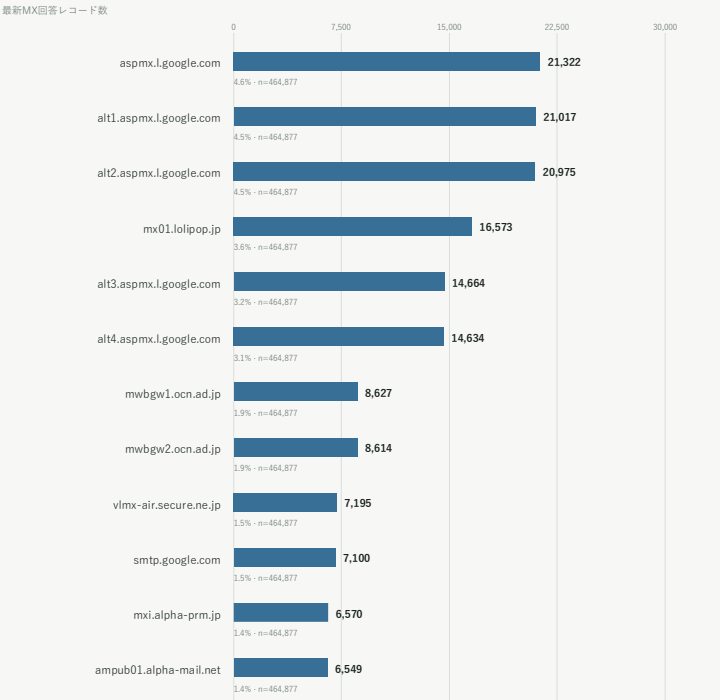


図17 メール保護設定の組合せ

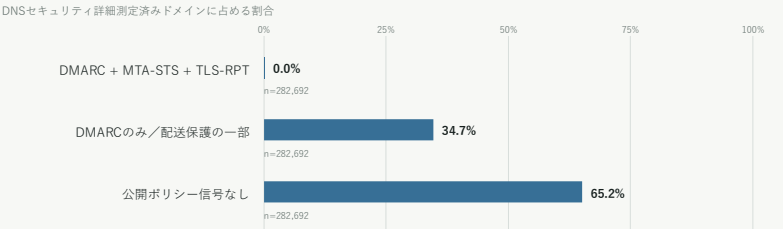


図20 推定メール基盤別DMARC公開率

