ESDC Statistics and LLM Crawlers

Rachana Bhatawdekar[1], Jos de Bruijne[1], Deborah Baines[1], Arnaud Masson[1], Mark Bentley[1], Guido de Marchi[1], Bruno Merin[1], Christophe Arviset[1], Hector Perez[2], and Ignacio Leon[2]

[1]European Space Agency, [2]Starion Group for ESA

IVOA Interop, Gorlitz, Germany, 15/11/2025

ESAC Science Data Centre (ESDC)

ESAC Science Data Centre (ESDC)

European Space Astronomy Center, ESAC
Villanueva de la Canada, Madrid, Spain

ESDC Statistics

Monthly Users (*)	Monthly Downloaded (*)	Archive Total Size
34 778	106.2 TB	1146.2 TB

* Monthly averages in 2024

We are supporting over 30 missions across astronomy, planetary, heliophysics, and human and robotic exploration domains.

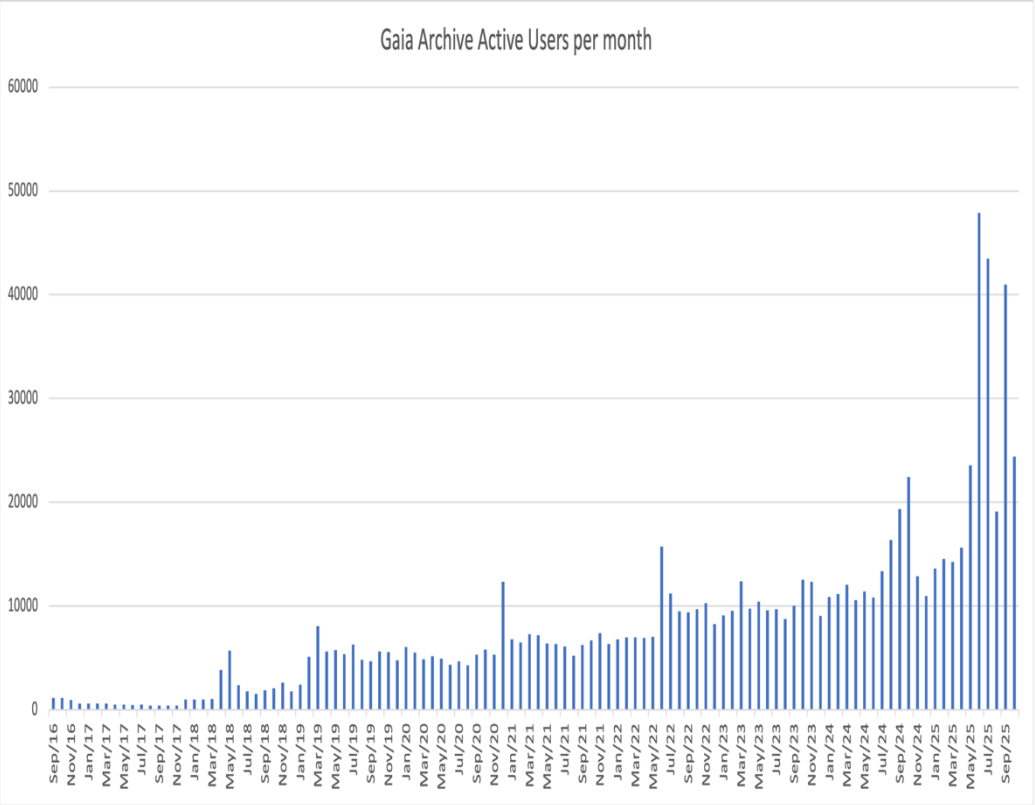
Tracking Archive Usage Statistics at ESDC

- ESDC archive usage has been monitored and reported for many years
- Figures are shared with missions, funding bodies, and the Science Programme Committee
- Defining usage metrics is challenging— missions differ in community size, data volume, user interaction
- ESDC tracks three metrics:
    - Total archive size
    - Number of active users (unique IPs),
    - Data download volume
- Sudden, unexpected increases in usage metrics

Large-Scale Non-Human Access

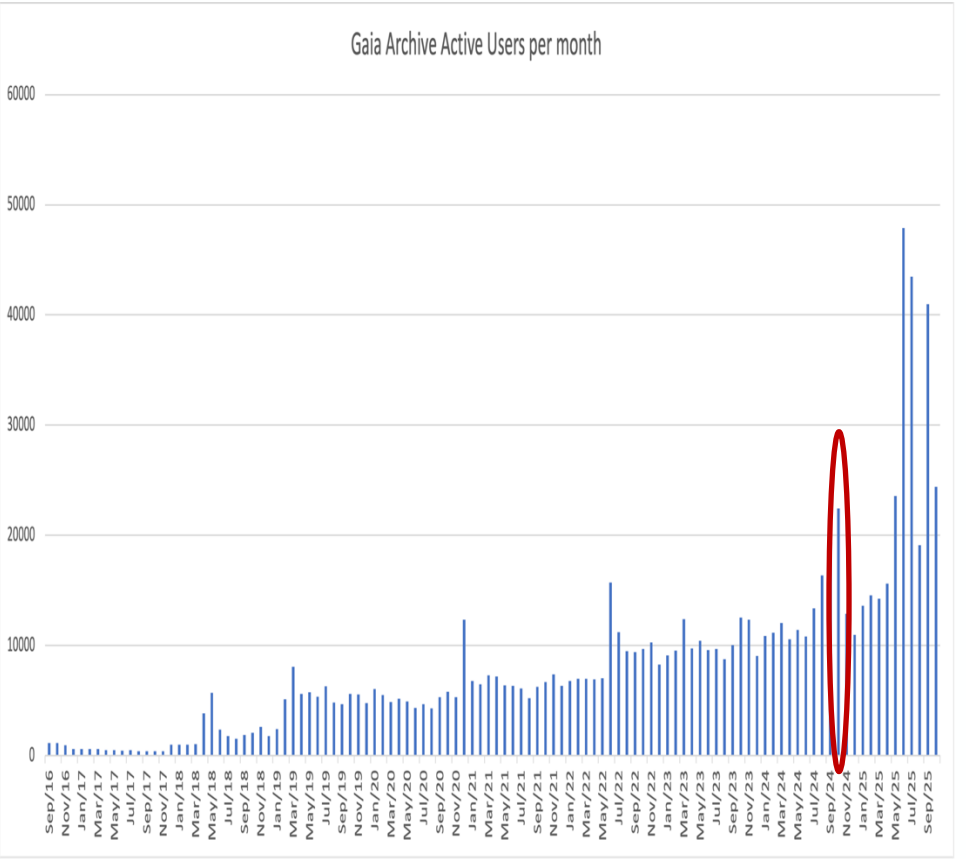
- Growing trend of automated, large-scale access to ESA's public science archives
- Many of these requests appear to come from LLM crawlers — bots that harvest online data to train or improve AI systems
- These activities significantly distort web metrics, inflate usage statistics, and strain infrastructure.
- Other data centres (e.g., PDS, SETI, ISRO) already face this issue and have tried various mitigations (logins, Cloudflare, nightly reboots)
- **New: CDS also reports the same phenomenon — "Abuse of Service Crawling" poster presented this week at ADASS 2025**

ESDC Archives usage statistics examples



Astronomy Science Archive Statistics: Gaia as example

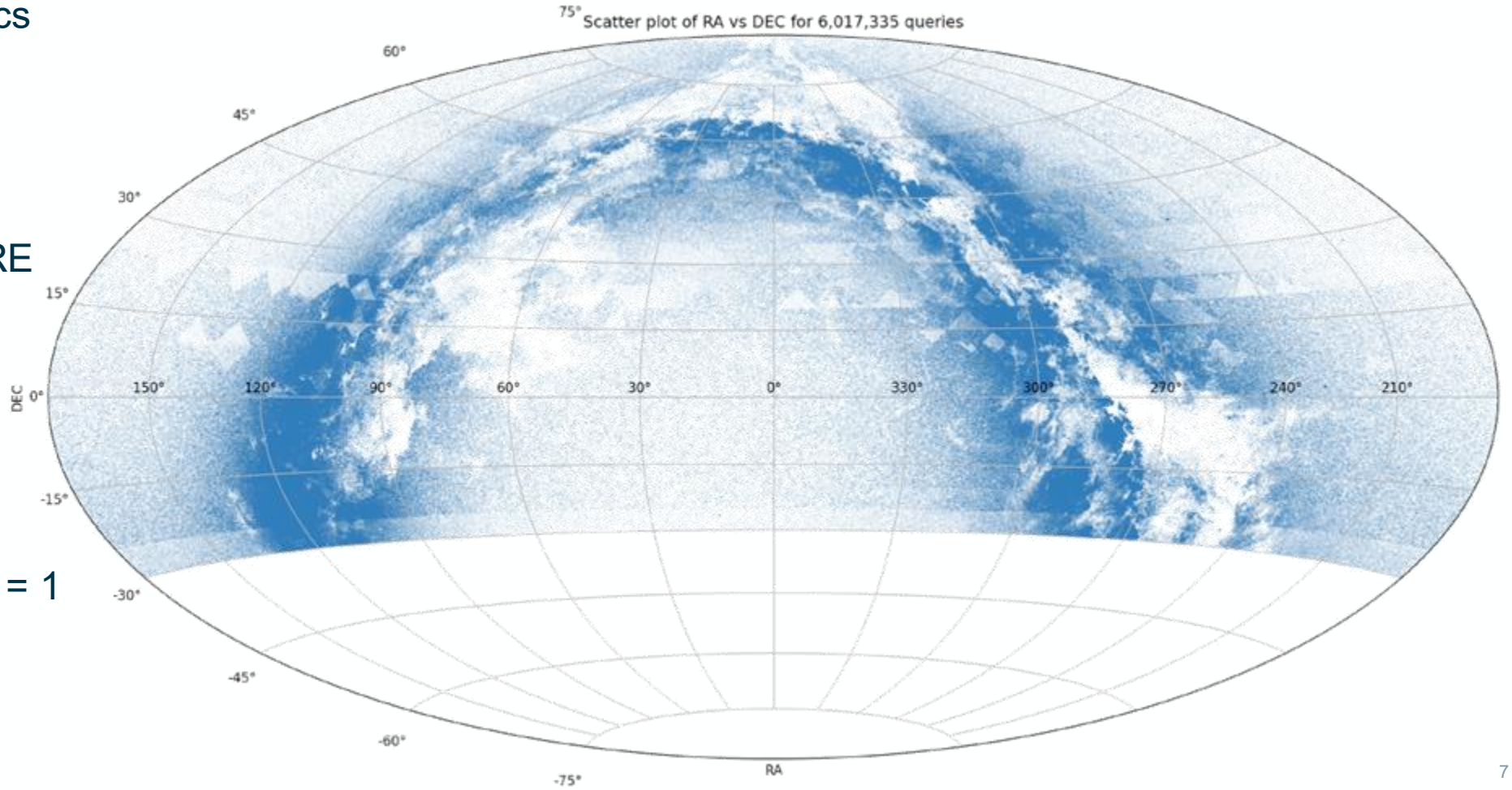
- Gaia ESA Archive usage (= query and download) statistics have always shown reasonable and credible numbers and trends
  - Peaks around data releases
  - Ever-growing interest from one data release to the next data release
  - Strong impact from Python query showers – not a problem since linked to a single IP address
- Starting in summer 2024, the number of monthly unique IP addresses has shown unexplained peaks:
  - Amazon science cloud users – a problem since artificially-multiplied numbers of unique IP addresses
  - Thousands of apparently "normal users" from Brazil – a problem since these are disguised crawlers
    "AI crawlers lie, change their user agent, use residential IP addresses as proxies, and more" [AI crawlers dominate traffic, forcing blocks on entire countries, arstechnica.com]



Gaia example of a Python query shower from one user

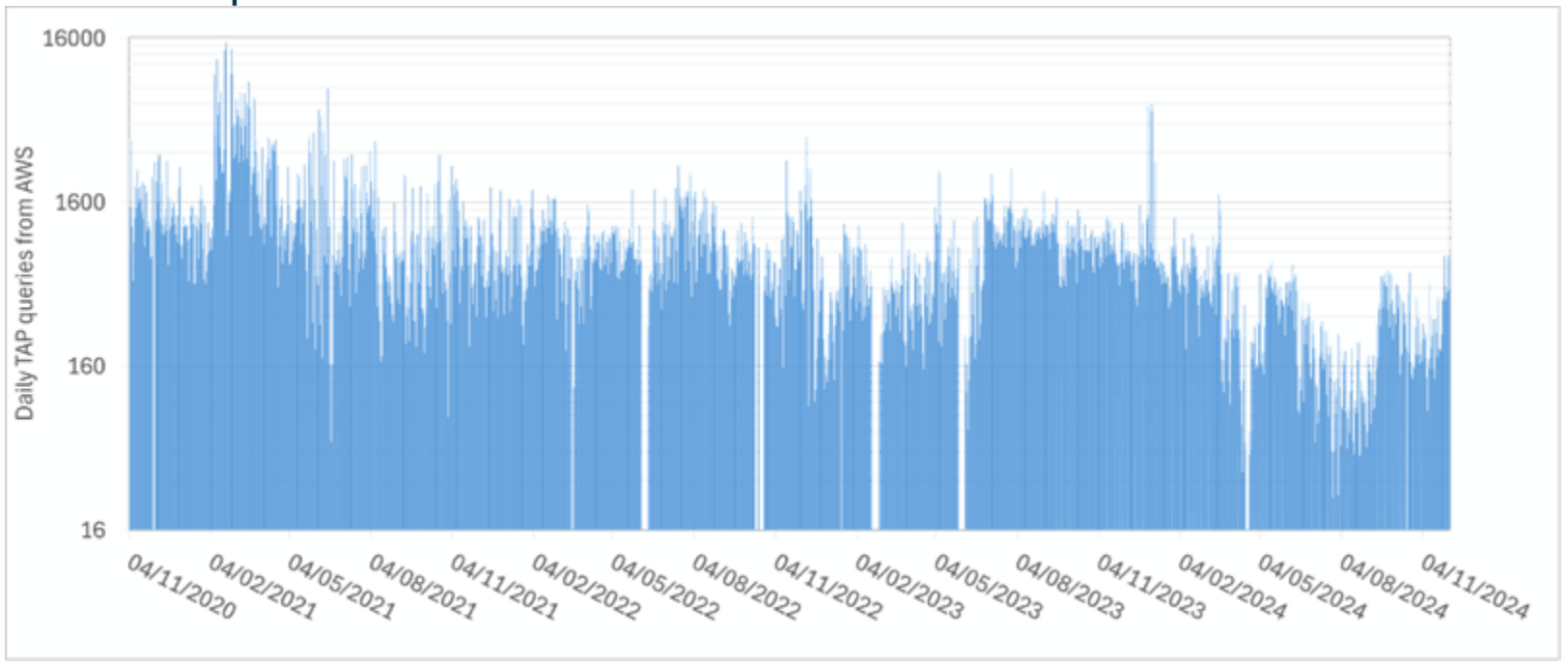
- One user, **one IP address**, 6,017,335 million ADQL cone-search queries launched over 9 days
- Not a problem for statistics

```
SELECT COUNT(*) FROM
gaiadr2.gaia_source WHERE
CONTAINS(POINT('icrs',
gaiadr2.gaia_source.ra,
gaiadr2.gaia_source.dec),
CIRCLE('icrs',
276.2617977807966, -
23.575354635673694,
0.00055555555555556)) = 1
```



Gaia example of an Amazon science cloud user

- One user, **8,022 IP addresses** from Amazon Data Services Singapore (47.128.*), “only” 21,467 ADQL queries
- A problem for statistics:
  - July 2024: 27% of all unique IP addresses come from AWS and (presumably) a single user
  - August 2024: 38%
  - September 2024: 41%



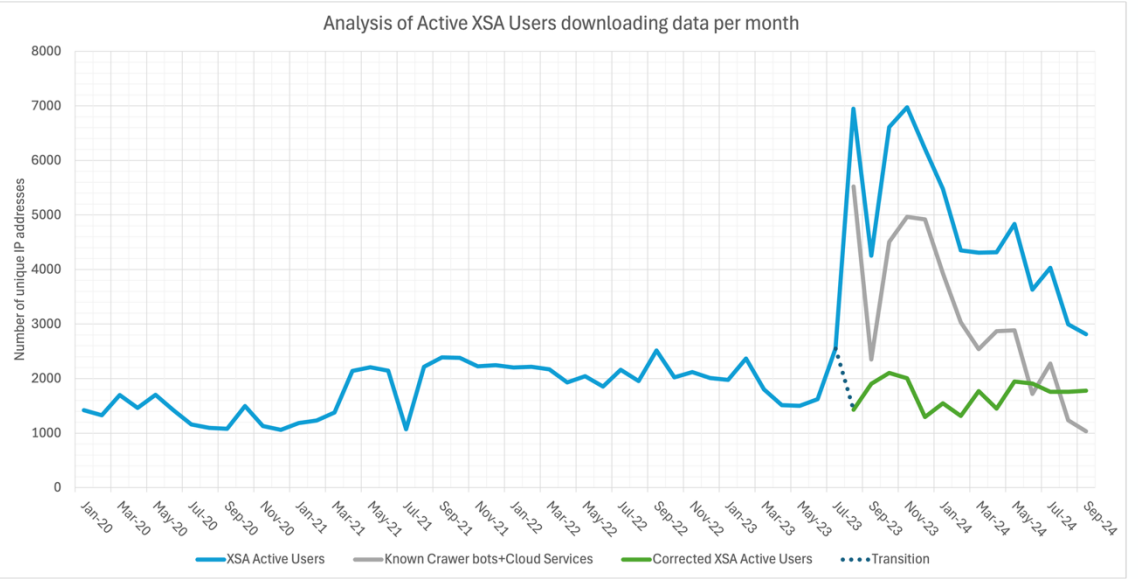
- Daily total number of TAP queries from AWS during 4 years
- No suspicious or excessive patterns → legitimate scientific use
- AWS, however, explodes the number of unique IP addresses from a single user

XMM-Newton Science Archive (XSA) example

- Before 2023, XMM-Newton Science Archive (XSA) statistics showed credible numbers and trends:
  - Consistent, gradual increase in active users over the years (by 2023 average users were ~1500/month)
- From August 2023:
  - Most web crawlers stopped reporting DNS names, providing only IP addresses, indicating a shift in access behaviour
  - Cloud providers start accessing the archive in large numbers.

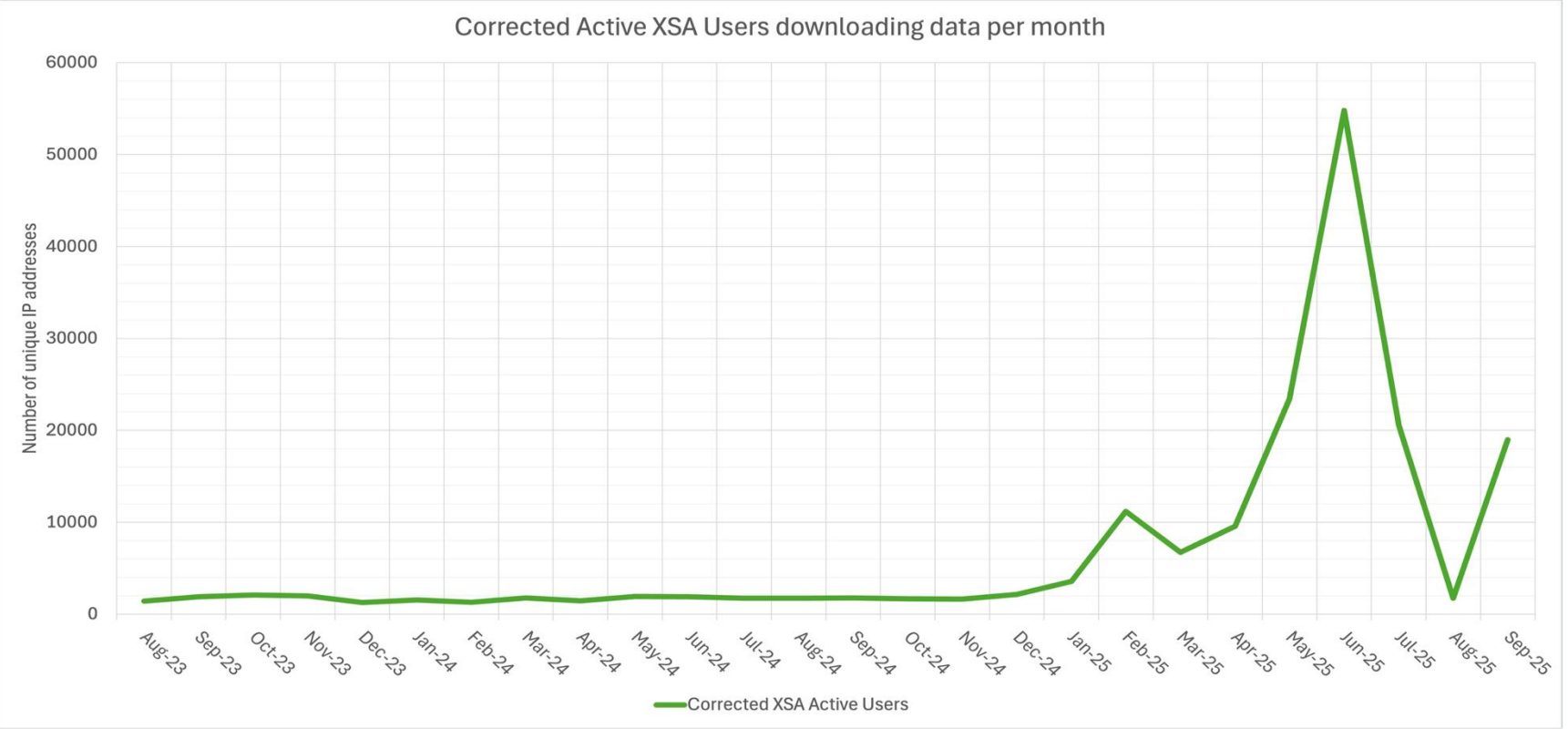
- After removing the known web crawlers and cloud-based accesses, users statistics returned to pre-August 2023 levels.

- Problem solved?...



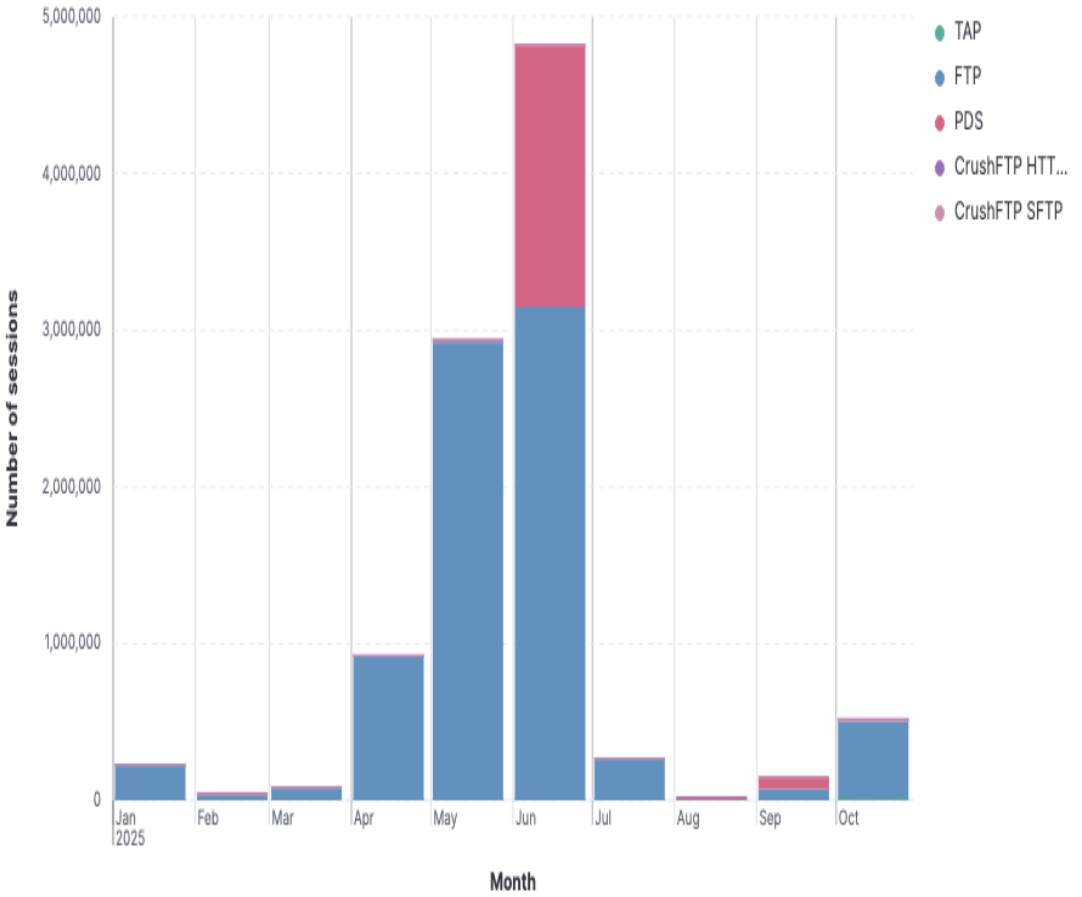
XMM-Newton Science Archive (XSA) example

- Problem solved? Not so fast, enter 2025:
  - The same issue observed in the Gaia archive appeared in the XSA: thousands of apparently “normal users”, mainly from Brazil – a problem since these are disguised crawlers.
  - Removing known web crawlers and cloud-based accesses is no longer solving the issue:



Planetary Science Archive Statistics

- Usage of the user interface and underling TAP service have mostly normal statistics
  - we see, for example, large spikes driven by (in particular) social media campaigns
- PSA also offers a simple web-directory of public data (for curl/wget type retrieval)
  - due to the very "web scrapable" nature of the interface, this is hit hard by non-human accesses
  - e.g. June 2025 recorded over 4 million unique IPs
- Analysis of the IPs, user agents, etc. lead us to suspect LLM crawlers masquerading as desktop browsers
  - clustering accesses to datasets in temporal bursts from many separate IPs
- So far system performance and reliability has not been dramatically reduced
  - but we continue to monitor, and expect the situation to only get worse



Heliophysics Science Archive Statistics

- Since 2021, sharp rise in IP connections for some (but not all) heliophysics archives
- New actors identified:
  - Likely LLM crawlers or download bots (often cloud-based)
  - Swarming access from many IPs in short bursts
  - Many disguise as normal browsers; others self-identify
- Possible link to publication of DOIs with direct data links
- Download volumes increased slightly — connection numbers exploded
- Challenges: hard to detect, classify, and filter bots
- Impacts: distorted usage statistics and reduced service performance (DoS-type effects)

Possible Solutions?

- Apply some kind of rate-limiting
  - but the IPs used by the swarms are often located in similar geography or IP space
- Block IPs known to be problematic
  - but there are too many to make this manageable
- Use an external service to blacklist IPs (or at least remove from stats)
  - this works for low numbers of requests, but would be prohibitively expensive for e.g. https
- Fight fire with fire, and develop some kind of ML technique to identify and filter out these “users”
- Restrict access to registered users
  - across all interface (UI, SFTP, APIs etc.)
  - probably the easiest mid-term solution
    - but a change to the ethos of data access in the space science archives…

IVOA experiences?

- Have other members experienced something similar?
- What solutions have you tried?
- How will you handle this going forward?