Ops IG Session

Saturday November 15 2025 11:00-12:30, Room: Wichernhaus "nadir"
Notetaker: Tamara Civera

Improving the data access services in the VO ecosystem (Euro-VO Weather Report)

Speaker: Henrik Norman

- Review of the “weather report” for the Euro-VO registry.
- Since last interop they have a slightly different approach in that we are not just presenting the status but also doing a deep dive into issues faced by data providers.
- They have contacted data providers actively to encourage fixes.
- They have services where providers can register their resources or just validate their own resources using any of their validators. There’s a new validator now.
- Twice a month, they run a full validation cycle over the VO ecosystem to all data access services. And they store the validation results.
- Since the start of the year, they have reached out to providers about problems.
- Since the beginning of the year they started to contact people trying to persuade them to fix their services. And they have analyze: TAP, SIAP1, SIAP2, SSA and SCS and started sending out emails in the beginning of October.
- So many data providers have not had a lot of time to fix their issues and many already have fixed their issues anyway.
- High response rate: 53 out of 54 data providers responded. This shows that data providers care about their services and are willing to improve.
- They analyzed:
  - 88 TAP services
  - 26 SSA services
  - 81 SIAP1 and 5 SIAP2 services
  - 1546 SCS services (about 1.400 are Vizier/IRSA)
- They have analized and contacted:
  - 100% faulty SIAP1, SIAP2 and SSA service owners
  - 98% faulty SCS (25 not analyzed)
  - 84% faulty TAP (17 not analyzed)
  - Where “faulty service” = a service failing a mandatory requirement of the standard (“must” in IVOA spec). Warnings / recommendations (“should”) are not considered in this iteration.
  - In the last two weeks a lot of responses have been coming in and people at the data providers are really committed to fixing almost all the issues that they have pointed out. Only one data provider has not responded yet and they are affected by the US government shutdown.
-Classification of issues:

  - Catastrophic failures: service doesn’t respond, crashes, sends errors; e.g., TAP servers that break when receiving ADQL. Some were “zombie” services (abandoned) → removed from registry. 45 SCS, 29 SIAP1, 3SSA and 9 TAP.
  - Less severe issues: missing UCDs, missing mandatory columns, wrong format (e.g., returning JSON instead of a VOTable), validator-service communication problems (e.g., SSL certificate mismatches).
- Many of the issues found in first iteration (SSL, firewall) have been resolved.
- Common kind of response from providers: gratitude (“thanks for reporting”), promise to fix.
- Some providers claim service is no longer active → they continue follow-up to decide whether to remove their resource from the registry.
- Others want to fix but need help → community help: Mark Taylor, Marcus Demleitner, etc., have helped with error messages, upgrading services, schema issues.
- Validators used: they have 4 own validators (SCS, SIA v1 and v2 and SSA) + TAP (STILTS) + MOC (CDS).
- Every service is classified twice a month: red (no compliant), green (compliant).
- By service:
- SCS:
  ~100 more services are compliant compared to a year ago.
  - People are getting emails and you can see a small increase there in the number of greens.
  - Some providers are still catching up (recently emailed, a month ago).
  - Main common issue: missing “main ID” field with proper UCD.
    - There's a discussion about allowing cone-search queries with a zero radius (i.e., radius = 0). It's kind of not super clear in the standard itself. The consensus right now seems to be that it should be allowed to return data.
    - Vizier: 30,000 services (most of them green). Almost 1,000 got fixed last month. Only a few hundred left, less than 300 now.
- SIAP v1:
  - Steady. Managed to remove a lot of the red ones.
  - Slow improvement until emails were sent.
  - Most common issues: UCD missing crucial info, missing test query declaration.
  - Recommendation: migrate to version 2.
- SIAP v2:
  - All issues solved (or shortly).
- SSA:
  - Steady improvement. Most common issue: missing Utypes.
- TAP:
  - TAP is the most complex protocol.
  - Fewer red services than a year ago. Most common issues: missing UCD, TAP schema mismatches, VOSI info mismatches, tables declared but missing.
  - High turnover: green services from a year ago not the same. Compliance can drop when new data is published, if metadata is not updated correctly.
  - Requirement failures dropped from ~4,000 to <1,000 in a year.
- Conclusions:
  ~30% of contacted providers already fixed their issues.
  ~50% say they will fix soon.
  ~10% plan to fix in 2026.
  ~10% say “when we have time.”
  - Very positive responses overall.
- Call to action: if you received emails, address the issues. Help available for compliance, registration, or maintenance problems.

QUESTIONS:

Q1 (room): Thank you for the service. It's very important and useful, the input I received from you.

IRSA VO accomplishments in 2025 - and what's next?

Speaker: Anastasia Laity

- Short overview of what IRSA's been up to this last year, what's going on, what's giving us a little bit of trouble and what is going to be up to next year.
- Focused on a couple of large missions, they have added support for:
  - SPHEREx
    - Which they started supporting weekly data releases for over the summer.
    - Custom SPHEREx data link service that returns the images and cutouts.
    - Spectrophotometry tool runs as a UWS service (avoids latency, can run 2+ hours).
    - They saved a lot of effort developing a separate bulk cutout tool for SPHEREx.
    - There are ways programmatically and in the GUI to utilize those DataLink table service descriptors to package up requested smaller cutouts.
  - Euclid
    - At their site, using fully VO models and protocols.
    - Custom Euclid DataLink service can return everything associated with a Euclid Tile ID or an object ID, including spectra available in spectrum DM format.
    - Everything available in SSA and Q1 catalog is available in cloud HATS parquet format.
- A challenge they're facing: addressing cloud copies of data. They've added a cloud access column in DataLink/SSA results to help clients build cloud addresses. For cloud-only datasets: access_url = AWS virtual-hosted style URL. This may seem redundant with cloud access column, but allows future metadata
- This year they quietly released an obscore table and ObsTAP service (not in registry yet).
- It includes many datasets and uses new DataLink service for cutouts; planning to register and add more data next year.
- Plan for 2026:
  - More data in obscore table
  - Transition SSA/SIA2 to DataLink URLs instead of direct file links; minimizing client/user impact.
  - Additional UWS services for Euclid/SPHEREx; more cloud-only data.
  - Registry improvements and cleanup.

QUESTIONS:

Q1 (room): How is inspector served through SSA?
A1: Via IRSA SSA calling a web API; returns combined spectra FITS files for requested object; outputs spectrumDM VOtable format.

Q2 (room): Only one record in SPLAT for IRSA SSA; likely registry situation; SSA registered by dataset. Needs verification if UQIT sector added.
A1: Good catch

ESDC Statistics and LLM Crawlers

Speaker: Rach Bhatawdekar / Deborah Baines

- ESDC (Science Data Center in Madrid) supports 30+ missions across astronomy, planetary science, heliophysics, and human/robotic exploration.
- Statistics tracked for years: shared with missions, funding bodies, and ESA Science Program Committee.
- Defining usage metrics is challenging due to varying community sizes, data volumes, and user interactions.
- They track three metrics: total archive size, number of active users (unique IPs), and data download volume.
- Over the past 3 years, they see unexpected increases in usage metrics from automated large-scale access, likely LLM crawlers. These distort web metrics and inflate usage statistics.
- Other centers, including planetary and heliophysics, report similar issues and tried mitigations like logins, or nightly reboots.
- Historically, archive usage showed credible trends, with peaks around data releases. Starting summer 2024, monthly unique IP addresses showed unexplained peaks exceeding the number of astronomers worldwide.
- Amazon Science Cloud users create thousands of IP addresses per user. Large-scale AWS users generate massive numbers of queries across many IPs.
- Thousands of apparently normal users, particularly from Brazil, complicate statistics.
- By 2024, spikes were seen in August and September, daily TAP queries from AWS rising sharply.
- Pre-2023 usage trends reasonable and predictable; post-2023, cloud and web calls alone no longer explain peaks.
- Planetary Science Archive and TAP services largely normal, but web directory easily scraped, millions of unique IPs.
- Potential solutions:
  - Rate limiting (challenging due to IP overlaps)
  - External service to blacklist the IPs (effective but expensive)
  - Machine learning to detect/filter users
  - They're looking at restricting access to just registered users. When you want to download the data, you need to log in across all the interfaces (easiest midterm solution; conflicts with ethos of data access and space science archives and some communities might find this easier than others others)
- Interested in hearing about other experiences. What solutions they've tried and how they'll handle this going forward.

QUESTIONS:

Q1 (room): I know that the Google bots do tend to respect robots that text at the root level, and these AI bots don't basically they just crawl everything.
A1: Google bots obey robots.txt; AI bots generally do not and behave like users.
A2: They don't have to use your agent that we can't identify them.

Q2 (room): What possible use could these companies have for this data to train an LLM?
A1: Lots of different LLMs. Purpose of LLM crawlers unclear; likely training from publicly available data. Crawlers access data, sometimes briefly, may not truly "learn" from it.
A2: It looks like they need all the identity files, and we are scanning how they're sending information to you.

Q3 (room): What I would like to understand also is, are you affected only in querying to the databases or also in data access?
A1: We are not affected by data access issues, but we are for databases and that could cause problems to our performances.

Q4 (room): So should we encourage AI access or should we try?
A1: Good question. I know this is the problem there are genuine users that are training. Crawlers access data, sometimes briefly, may not truly "learn" from it.
A2: So basically, what I'm seeing is that when you hook up an AI model to an MCP server—whether it's hosted locally or through GPT—it tries to grab the resources you point it to. Most of the time, simpler models don't know how to handle the schema properly, so they hit an error and just give up. That means only a few requests actually succeed, but over time, the number of successful accesses is going to grow.
A3: What we probably need is a way to make legitimate access easier: every server should have MCP-aligned protocols and clear resource descriptions so models can interact correctly from the start. That works fine for short bursts of queries, but when someone starts crawling tons of data sequentially across all the databases, that's when system performance could really take a hit. So short bursts = manageable with MCP; heavy crawling = might need extra protections.

Managing automated access: Experience from Observatorio Astrofísico de Javalambre (OAJ)

Speaker: Tamara Civera

- Presentation covers the OAJ (Javalambre Astrophysical Observatory) experience with managing automated access.
- OAJ = Spanish astronomical ICTS (Unique Science & Technology Infrastructure). Located in the Javalambre mountain range (Teruel, Spain), managed by CEFCA.
    - Designed for large-scale photometric sky surveys. Current surveys: J-PLUS (12 filters) and J-PAS (57 filters) (> 8000 square degrees).
    - As an ICTS, >20% of observing time is offered to the community (legacy surveys + open time).
    - Surveys generate large data volumes.
    - Data is served through the CEFCA Catalogues Portal, hosted at CEFCA/OAJ.

It provides advanced tools for data search, visualization, and download.
    - Two main service categories:
        - Web interface tools: sky navigator, image/object search, visualization, downloads.
        - VO services: TAP, SIAP, SCS, HiPS.
    - Data access includes:
        - Public data releases (open to all)
        - Private data releases (restricted to project/survey members)
- Large-scale surveys → massive datasets → astronomers increasingly use automated tools.
- However, open access means non-astronomical automated clients also reach the services: bots, IA crwalers.
- Difficult to distinguish scientific automated access from AI training or non-scientific usage.
- Automated clients could cause: Unexpected load and performance degradation, even if not malicious.
- Protective strategies are needed to maintain balance between usage and service reliability.
- OAJ benefits from security services provided by RedIRIS, the Spanish academic/ research network.
    - RedIRIS supplies: Security services and coordinated incident response.
    - Example service: SinMalos ("Nonmalicious")
        - Helps reduce malicious traffic collaboratively.
        - Two types of institutions:
            - One type analyze their own traffic, detect malicious IPs and share reputation data through this service.
        - Other type import these IP reputation lists and integrate them into local security systems for blocking/mitigation.
- Even with protections, unexpected situations can still occur. Example Incident (September 2024):
    - Observed service slowness + some 403 Forbidden errors.
    - Log analysis showed: A single IP making extremely high numbers of requests to a resource-intensive service.
    - IP belonged to a university → likely a legitimate user, not a bot.
    - To handle it, OAJ updated security policies. Added per-IP rate limits, different for each service. If limit exceeded → user receives a friendly: “Max retries exceeded” message.
    - Result: User contacted OAJ after a few days. Allowed OAJ to explain how to use the service more efficiently / less aggressively.
- Conclusions:
    - Astronomical services are often accessed by automated clients; this requires thoughtful management.
    - Experience showed that automation can affect performance even for legitimate users.
    - Collaboration between institutions is essential for proactive protection.
    - Key tools: Proactive monitoring, service-specific rate limiting, clear communication with users. Policies are not trivial to define and must be tailored per service.
    - Challenges to consider:
        - Distinguishing attacks, intensive AI usage, and legitimate users.
        - Deciding whom to block without generating false positives.
    - Finding the right balance between protecting services and supporting legitimate users is key, although sometimes it is not easy to do.

Managing intense web activities

Speaker: Gilles Landais

- Simbad:
    - They note an important and dramatic increase of IP beginning in 2024.
    - The number of IP has been multiplied by 10 or more (> 2 million IPs per month)
    - Difficult to identify query origin: Queries may come from AI systems, bots or manual code. Some are legitimate queries, others not.
    - Difficult to identify true queries.
    - Possibility of use user-agent: sometimes it is good, sometimes it is not enough.
    - For the moment, Simbab manage well this increasing number of queries.
    - Many queries from Alibaba were managed by updating robots.txt. And now works better.

- Vizier:
  - Different access To access to that part.
  - Some critical activities can possibly compromise the service.
  - High queries cadence usually it concerns a small list of abusive users. Users can generate hundreds of thousands of queries/day.
  - Heavy queries: includes big data operations, conversational queries.
  - This year there are more bots, there are more AI, but there is also this kind of users or queries.
  - The number of IPs has increased a lot, multiplied by 15. They come from cloud, and it is sometimes difficult to manage.

- Mitigation Strategies

  - Indexing & filtering: Optimizing data indexing to handle heavy/bot traffic.
  - Blacklist is another thing they did with a new interface code, sometimes, for IP, but also for a range of IP. Limitation that cloud IPs are dynamic.
  - Apache filtering they also did. You can extract some information from the header.
    - They also use a mode evasive based on the core cadences. It was not well for they. They have a lot of queries coming from different IPs, and this module is clearly not useful today.
    - Multi-processing module is another module they I discovered this year. It allows to manage better the resources to increase the number of fork of thread. Improved resource usage and reduced unavailability caused by socket connections waiting.
  - They enforced all those architecture with more CPU and balancing using mirrors. It can be very useful, but it is interesting if it is transparent for the user. Redirecting queries transparently to mirrors improves server availability.
  - They have a pretty huge black list.
- How to filter AI ?
  - Modify the robot.txt with different user agents

- User-agent differentiation: user agent for cloud to collect the data to train the model, a user agent which is used by the user request, and another one of the point.
- AI are important but they need to be controlled. So it requires to adapt regularly the regulations.
- They begin to think about authentication (maybe too early). So historically, it is not in the CDS spirit to add the authentication, but it seems that it will be maybe a solution. If you log, if you authenticate, you are responsible for queries.

QUESTIONS

Q1 (room): Are other LLM operators, besides Cloud, implementing measures to block certain crawlers, and is there a way to do this in a unified way rather than per LLM?
A1: For the moment, I don't know. because the origin is not given always.

Q2 (online): Do you know the financial cost of the bandwidth used by the bots? Could we start to charge for access beyond a free tier?
A1: No answer because not listened.

Open discussion

- The first approach to identify automated clients is using the User-Agent. Legitimate users usually have a recognizable User-Agent, which allows identification.
    - But, many bots either do not include a User-Agent or use one that cannot be identified.
    - Request all users and applications to register a standard and identifiable User-Agent.
    - This can apply to scripts, applications, or custom code.
    - Consider creating a recognizable User-Agent to distinguish robots from legitimate users.
    - A note on operational identification of software components was published by Mark and Markus a few years ago.
        - It indicates if you are harvesting or if you are validating, put this or that string into your user agent string. This can help people can do statistics more easily.
        - Original recommendation: Do not include identification strings in user agents if performing scientific queries.
            - Markus, propose change it. Suggestion to include “IVOA-” in user agent strings. That could protect the APIs by only letting in request who have that "IVOA-" in their user agent. And that could be a better solution than authorization.
    - IAs can generate and execute code (e.g., AstroPy), making the user agent appear as AstroPy.
    - While sophisticated crawlers might bypass this, basic measures can filter out less careful or “idiotic” crawlers, making it easier to manage more

advanced ones.

- About using authentication as solution:
  - All users accessing data will be required to authenticate. Because it's not only data, but also put all other resources.
  - Authentication can help to manage load and protect services.
  - Should authentication use a global SSO (single login for all services) or per-service logins? Multiple service logins (e.g., four separate logins) would make the system impractical. Proposal: A global login would be acceptable if technically and socially feasible.
  - Implementing a global login has technical challenges.
  - Sociological aspects: Acceptance of a general login depends on the broader community.
  - Social arguments against authentication are less compelling if service viability is at risk.
  - Authentication may become mandatory to ensure fair use and prevent service disruption.
  - Suggestion for using ORCID or some other mechanisms to access public data and own mechanisms for private data.
  - Requiring registration can act as an initial barrier to prevent uncontrolled access. Without any control, bots can overwhelm services, similar to a DDoS attack.
  - Authenticated users are known, enabling service operators to monitor usage and coordinate if excessive load occurs.
  - (chat discussion) JHWhat's ultimately preventing LLM crawlers to aout-create a million users?
    - C, JH Maybe could be solved with using phone number in registration.
    - TD At least you can automatically block one without impacting other users.
    - TD You can email them in case they are legitimate users who need guidance.
    - JF Buy should I have to fork over my email to access astronomical data?
    - C 'Yes', if that's viable solution to maintain the service from going down.

- Has anybody interact with major AI companies?
  - Verify the identity of the companies. Can they help us?
  - It was noted that large AI models themselves are not necessarily the problem. It's most startups and people that really have no idea what they're doing.

- Example issue: Open-source Python packages can create expensive queries in tight loops, potentially overloading servers, whether triggered by AI or humans.
- Data will increasingly be hosted in the cloud, making cost-driven management more critical.

- Does anyone know if sticking a basic or in front of all the services with a very dumb password, like 12345, how much that would cut down the robots?
    - It would reduce some bots but only modestly, as more sophisticated bots could bypass it.
- Use filters to temporarily restrict heavy or suspicious usage, while still allowing genuine users to continue. Prompt users who trigger filters to register for continued access.
- Propose a tiered access policy: open access for general users, registration for frequent users, and elevated permissions for trusted institutions.
- Astronomers, not AI bots, are the main contributors to heavy data access in terms of volume, frequency, and number of service hits. Certain users, when working on specific problems, can generate extremely high loads, potentially overloading systems.
- In CANFAR: Over the past two years, the rate of data being accessed from outside the CANFAR data center has decreased by about 50%. Providing compute resources directly on the platform reduces the need to download files externally, keeping internal hits very high. The decrease in external hits contrasts with other observed trends of increasing access elsewhere.
- In ESA: Anonymous users: Limited resources; can perform queries and downloads but restricted in capabilities. Authenticated users: Gain access to advanced services, including disk space, database spaces, and computing infrastructure.