Evolving the Data Central Archive
Scaling and Optimising Archival and Automated Data Reduction Systems

1. About the Data Central Archive

The Data Central Archive is an ongoing archival service that provides online access to data observed with the Anglo Australian Telescope (AAT) and the ANU 2.3m Telescope, with AAT data dating back to 1974 including with standard instrumentation, photographic plates and most IPCS detector observations available. The 2.3m archive hosts data taken since automated operations began in 2022. Data Central also hosts an automated data reduction system for the WiFeS instrument on the 2.3m, reducing data shortly after observations have taken place.

2. Data Volumes and Querying Methods

As more observations are added, the querying performance of the archive decreases. Through initial testing, queries of 4 degrees of angular radius were averaging around 150ms. The current archive interface takes the RA, dec and radius arguments, constructs a square around the target and returns any objects within that square. MongoDB uses a B-tree structure sorting objects into range buckets before finding objects within those buckets that fit within the query values. This works for one dimensional queries (time complexity of O(log N + k) with k being the number of objects returned) but for a two dimensional query used by the cone search method this grows to a time complexity of essentially O(N) scaling with the number of objects in the database not the number of results as the region can only be narrowed by a single dimension at first, before having to check if each object found lies within the constraints of the other dimension. The box approach also would include objects in the corners that would not be match in an equivalent cone search and ignores how RA converges at each pole, leading to missing results at high declinations. The targets were also stored in array, compounding the problem as MongoDB can't combine upper and lower bounds on array fields, meaning the query scans a larger range than requested. After the box search, further filtering with SkyCoord separation is performed over the targets which adds performance overheads and did not recover objects near the poles or at high declinations.

Original Cone Search Method	GeoJson + 2dsphere
2 dimensional lookup: O(N)	Cone covered by a few cell ranges: O(log N + k)
1st step box search misses objects at high Dec, near poles	Search performed on the sphere, whole cone is covered
Target Array: 4 limits are checked individually	One condition per target, no partial matches returned
SkyCoord match: extra performance overhead	Angular-distance checked inside database, no 2nd pass

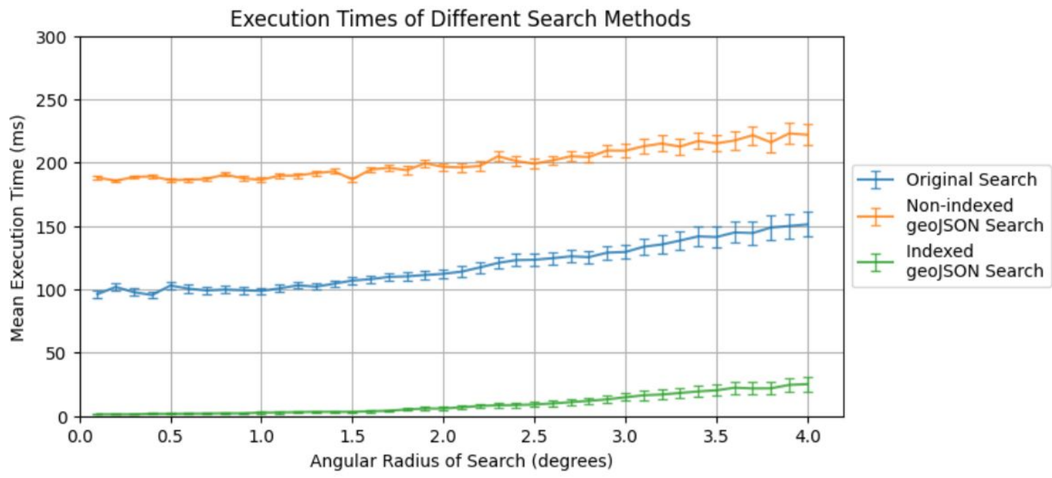
3. GeoJSON + 2dsphere Improvements

To improve these performance bottlenecks, each target's position is stored as a GeoJson point, a standardised spherical coordinate system, and then indexed using MongoDB's 2dsphere index. 2dsphere divides the sphere into hierarchical cells and gives them an ID so that objects close to each other on sky will have ID's close to each other. A cone search can now be covered by a small number of cells, meaning a two dimensional lookup is covered by a few one dimensional range lookups. This is the strength of the B-tree lookup that MongoDB uses reducing the time complexity back down to O(log N + k).

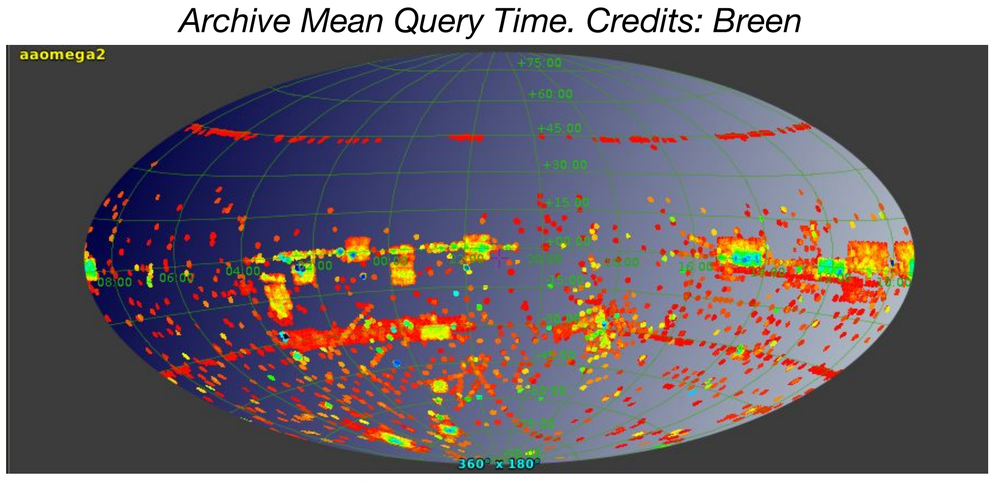
4. Future Directions

Other ongoing work includes fibre table matching using MongoDB's aggregation pipelines. An RA and Dec position query finds observations and extracts exact coordinate matches for individual fibres. The first step of the pipeline is a spatial search, benefiting greatly from the improvements describe above. A set of HiPS visualisations of 2dF observations have been developed alongside this work allowing exploration of the sky coverage of the instrument from individual fields up to full sky providing a useful visual interface for exploring the archive's data.

Also in progress is work on modernising the archive's backend to support direct API access and improve download performance of requested files. A smaller archival collection of a monitoring programme of Rubin Deep Drilling Fields using the Australian SKA Pathfinder (https://arxiv.org/abs/2610.13011) will be deployed using this new backend before the AAT and ANU 2.3m archives are migrated to this backend.



Archive Mean Query Time. Credits: Breen



AAOMega HiPS Spatial Overview. Credits: Kuhnke-Barakauskas

James Coulson james.coulson@mq.edu.au

James Coulson[1], Brent Miszalski[1], Marcus Breen[2], Nicholas Kelly[2], Jonah Kuhnke-Barakauskas[2]

[1]Australian Astronomical Optics, Faculty of Science and Engineering, Macquarie University, NSW 2109, Australia
[2]Faculty of Science and Engineering, Macquarie University, NSW 2109, Australia

Research Data and Software
rds.org.au

Data Central
datacentral.org.au