

ATLAS Search Knowledge Base

Jason Oliver

~

jaoliver@cern.ch



GMT+10½

What I'm Covering today

1. The ATLAS Collaboration as a case study
2. The variety of platforms they occupy
3. OpenSearch
4. The ATLAS Search Knowledge Base
    a. Search
    b. Chat
    c. MCP
5. Concluding Remarks

The ATLAS Collaboration

Unique, but shares many characteristics with large organisations

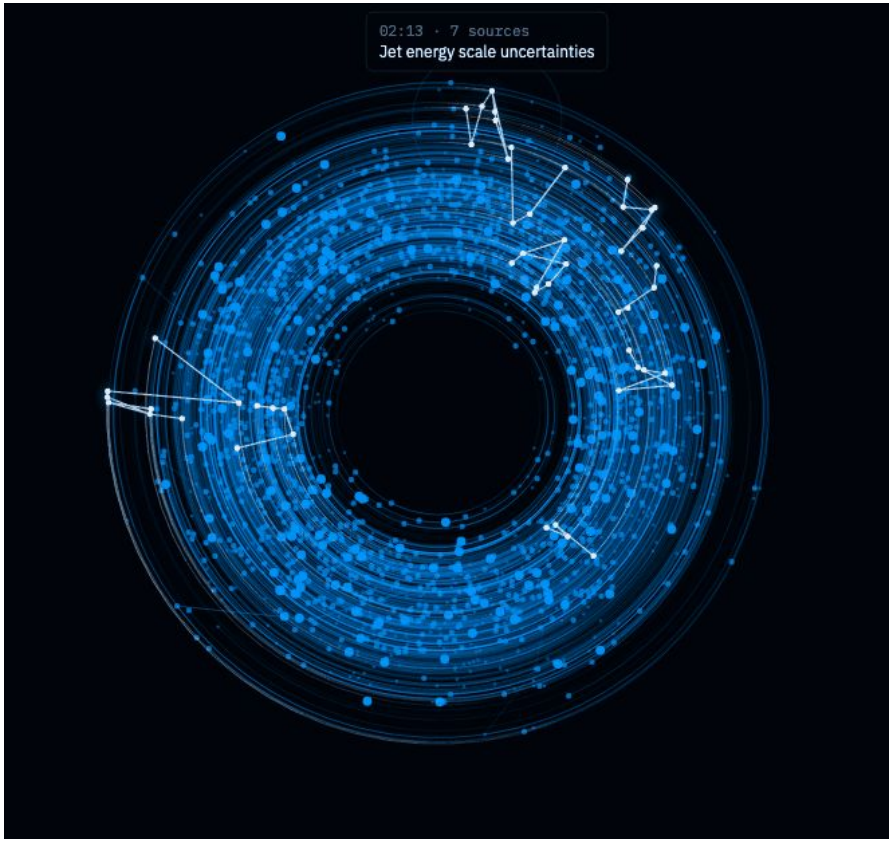
- Founded in 1992
- Over 6000 ATLAS members
- With 3000 physicists
- ATLAS has over 246 institutes
  - Universities
  - Technical Institutes
  - Clusters
  - Computing Sites etc
- Spread out over 40 different countries
- Information spreads via a number of platforms

ATLAS Institutes Map



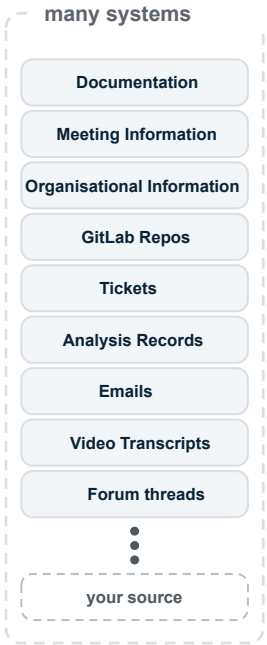
ATLAS Physics Office

Star map demo

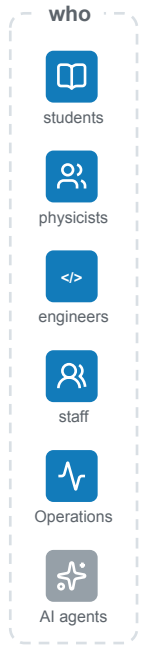


Some Considerations

The Key Question



How do we bridge this gap for users?

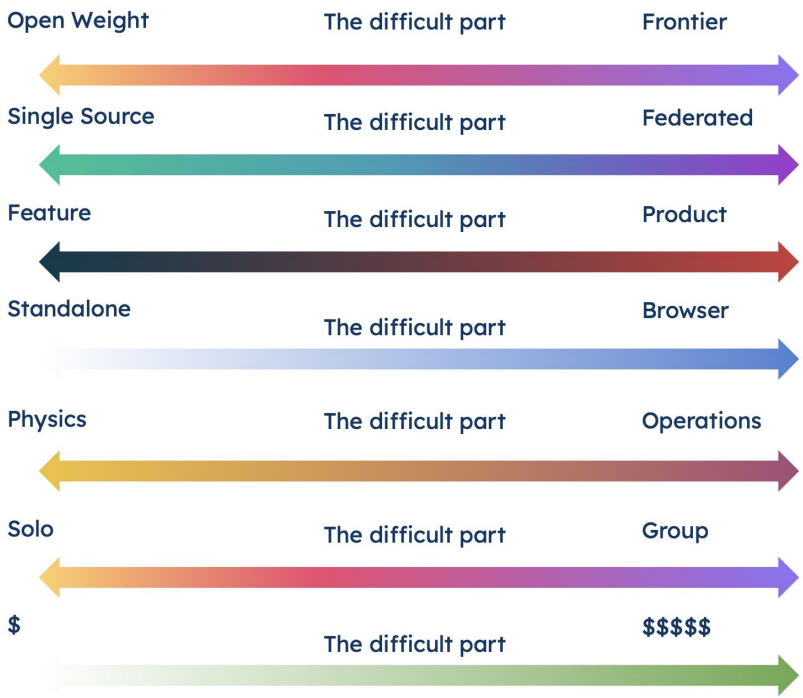


MCP via the AF: mcp-portal.af.uchicago.edu

Our Design Choices

Many considerations

- All designers have different constraints and desires
- In our case we need something federated, that can be compatible with openweight or frontier models
- That serves primarily groups
- That is ideally open source with incredible experts maintaining the core code
- It should also outlast one postdoc - turn over is a feature of HEP



Our Design Choices

Many considerations

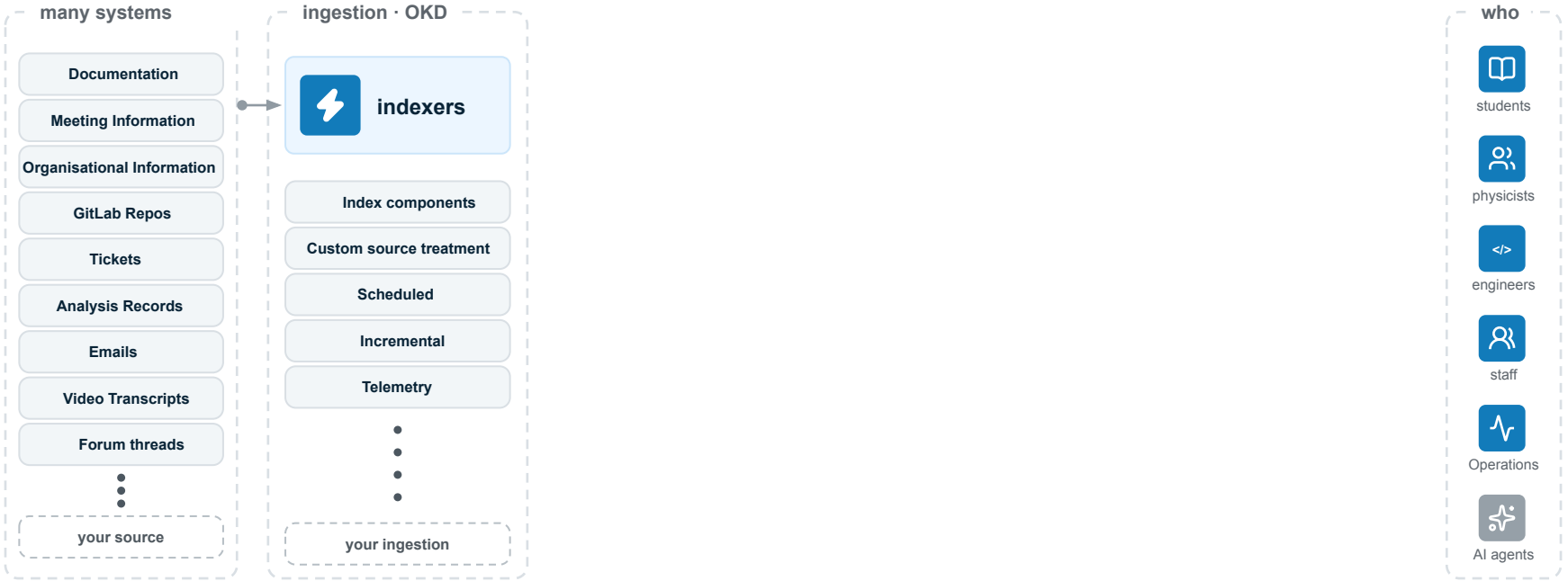
- All designers have different constraints and desires
- In our case we need something federated, that can be compatible with openweight or frontier models
- That serves primarily groups
- That is ideally open source with incredible experts maintaining the core code
- It should also outlast one postdoc - turn over is a feature of HEP



So how do we use it?

ATLAS OpenSearch Outline

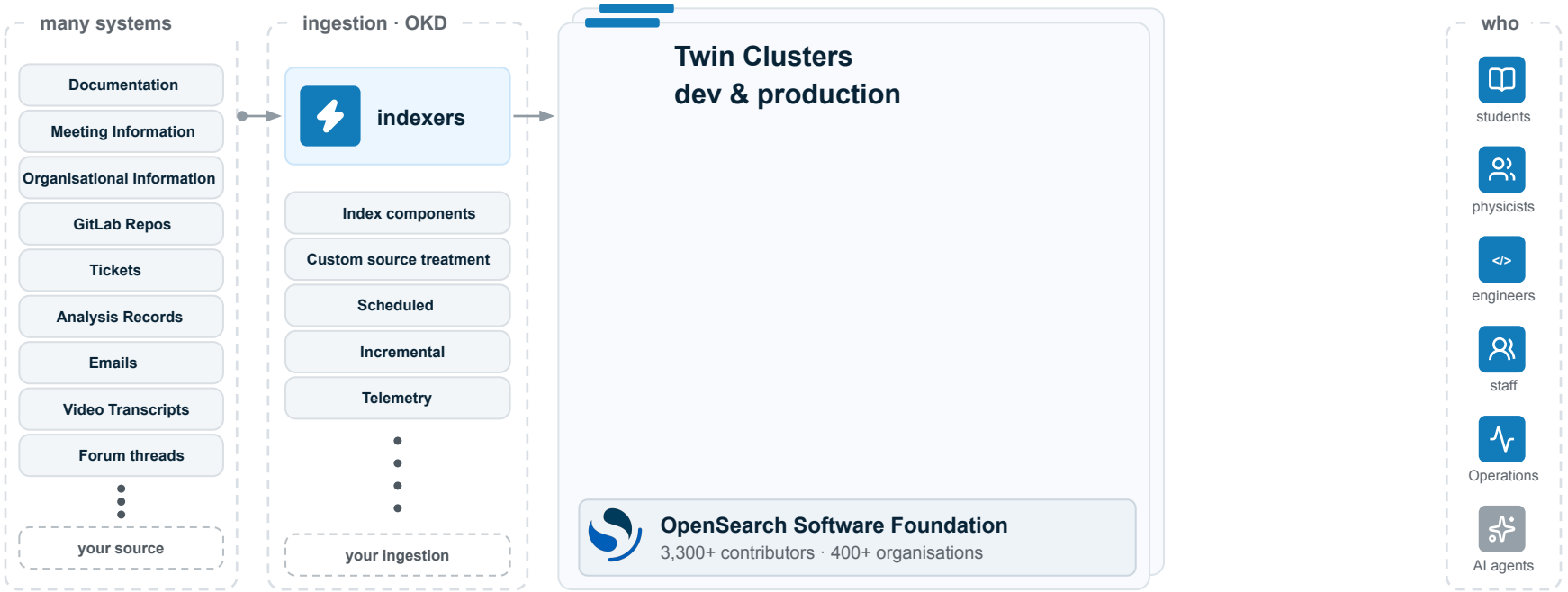
North Star: Information Accessibility — shared infrastructure for search, assistants, and agents



MCP via the AF: mcp-portal.af.uchicago.edu

ATLAS OpenSearch Outline

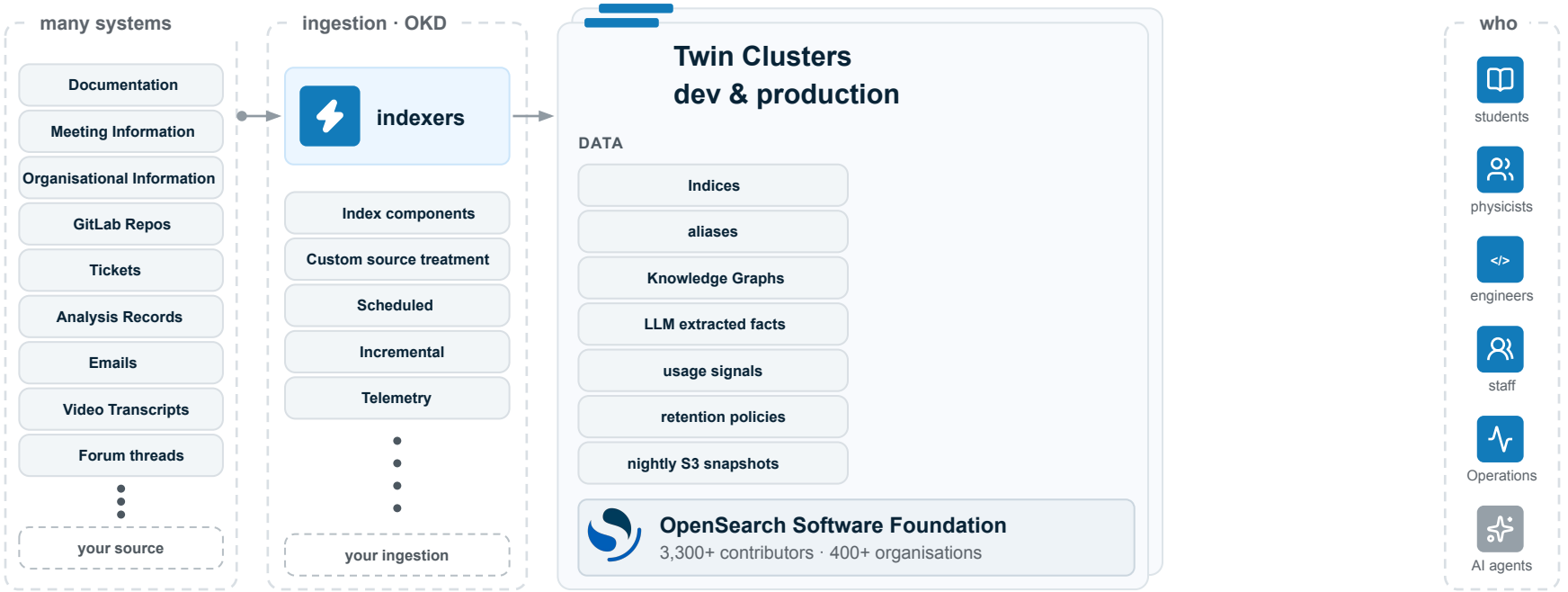
North Star: Information Accessibility — shared infrastructure for search, assistants, and agents



MCP via the AF: mcp-portal.af.uchicago.edu

ATLAS OpenSearch Outline

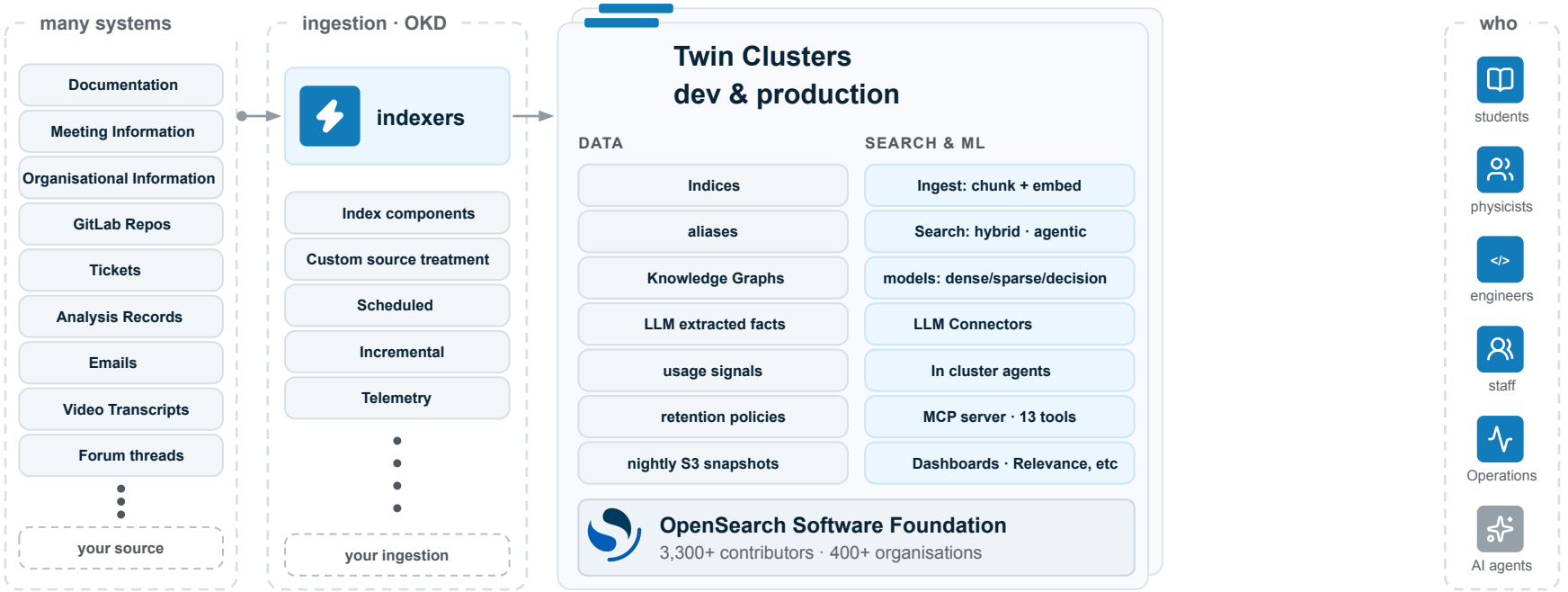
North Star: Information Accessibility — shared infrastructure for search, assistants, and agents



MCP via the AF: mcp-portal.af.uchicago.edu

ATLAS OpenSearch Outline

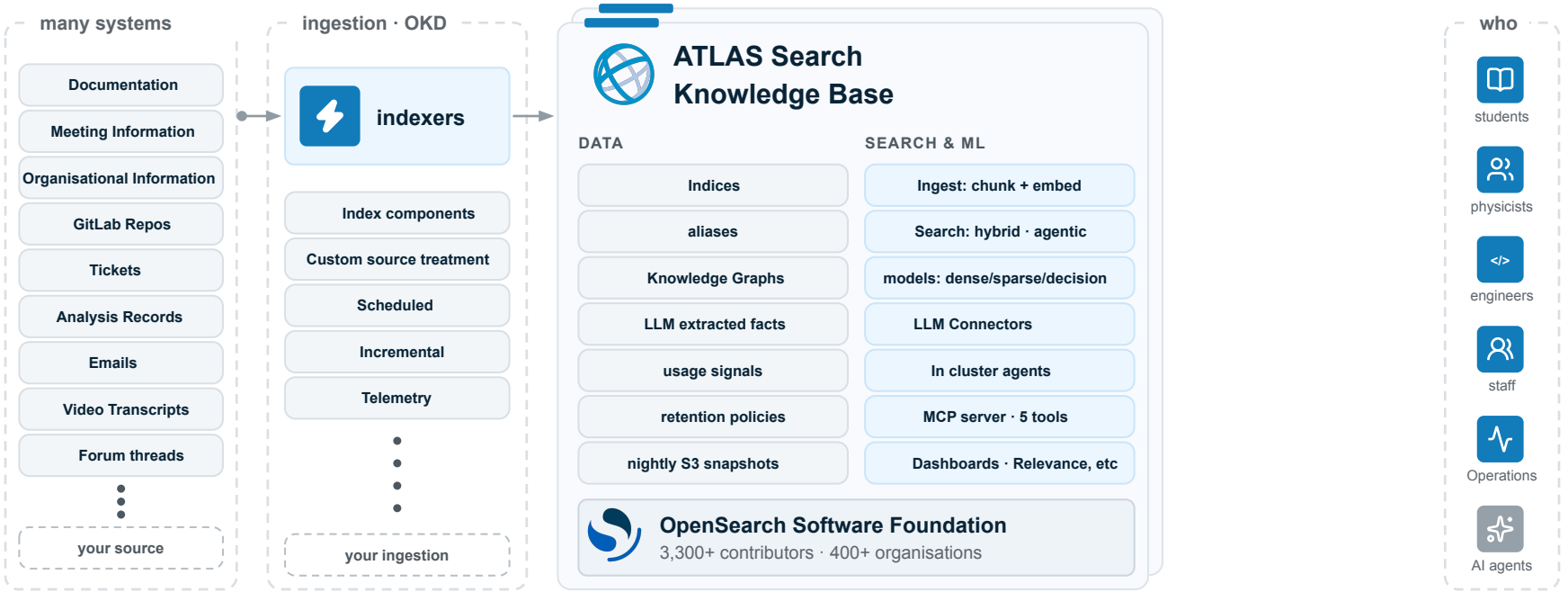
North Star: Information Accessibility — shared infrastructure for search, assistants, and agents



MCP via the AF: mcp-portal.af.uchicago.edu

ATLAS OpenSearch Outline

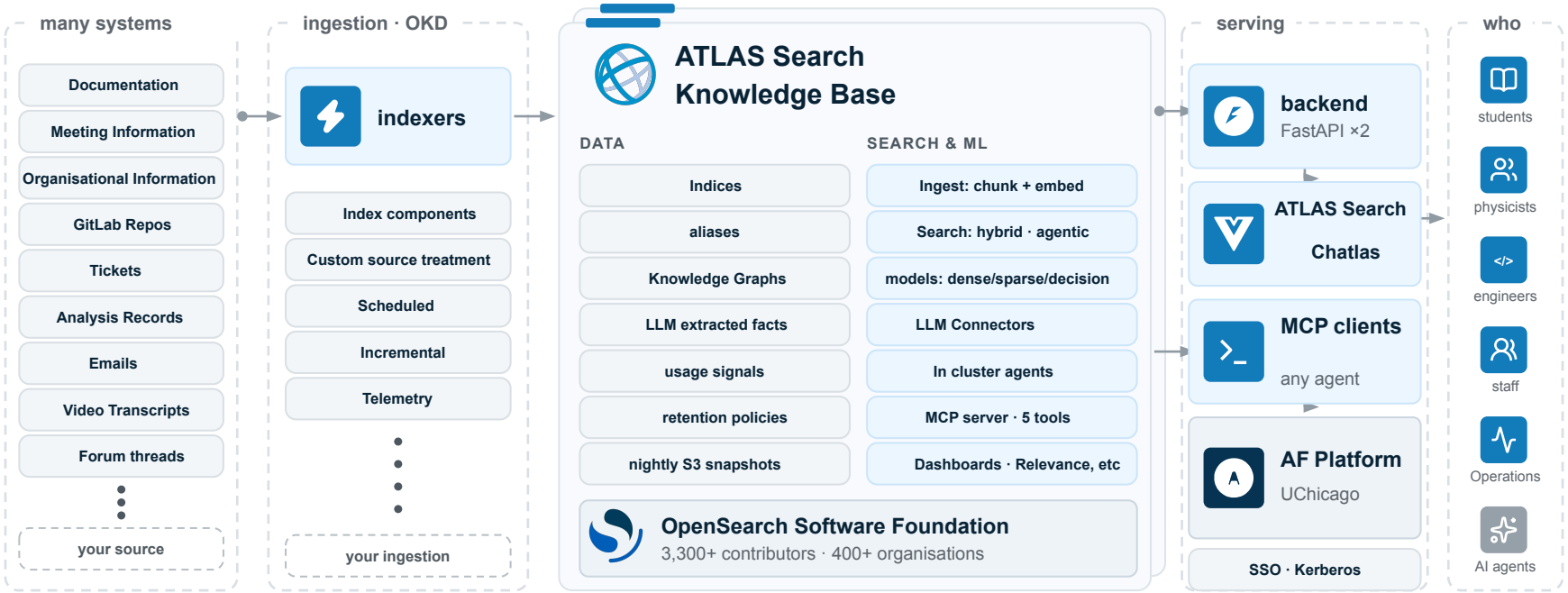
North Star: Information Accessibility — shared infrastructure for search, assistants, and agents



MCP via the AF: mcp-portal.af.uchicago.edu

ATLAS OpenSearch Outline

North Star: Information Accessibility — shared infrastructure for search, assistants, and agents



MCP via the AF: mcp-portal.af.uchicago.edu

The Embedding space



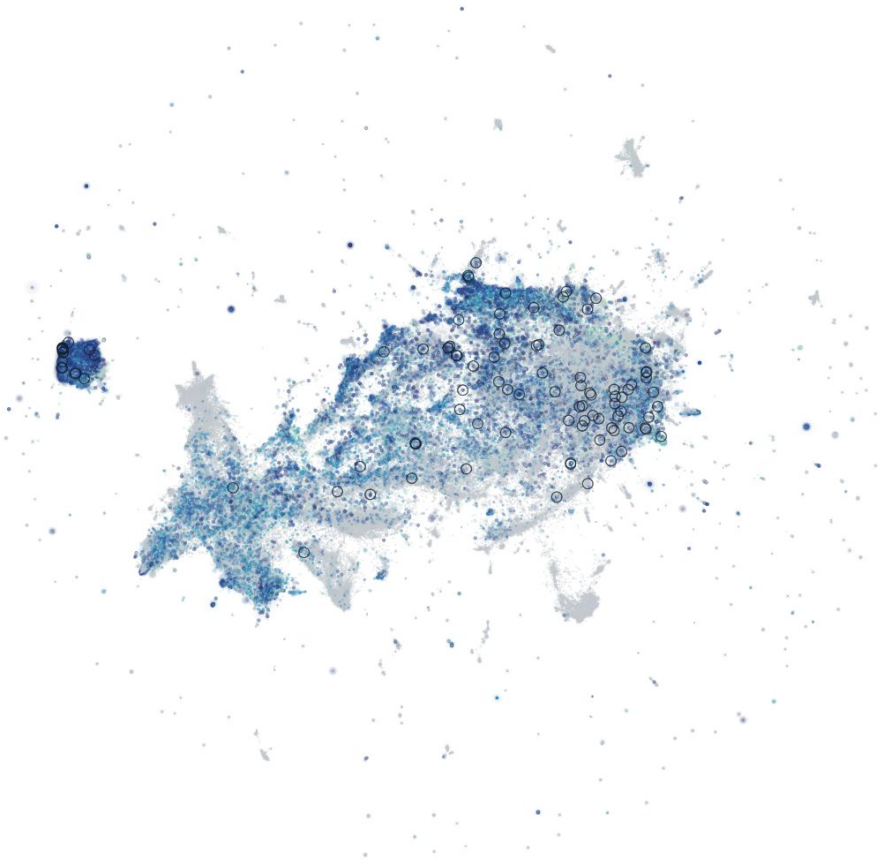
213,597 pages from **68 indices** (20,443,726 documents in all), one mark each, placed by meaning: nearby marks are pages about nearby things. 39,634 of them are from derived indices (generated from the others).

UMAP keeps each page's neighbours and pulls clusters apart; how far apart two distant clusters sit means little.

INDEX FAMILIES

- **TWiki & Confluence** 339,236
  - wiki · 339,236
- **Documentation sites** 1,565,194
  - derived · web-items · 831,596
  - derived · web · 405,354
  - web · 327,756
  - rnd-athena-prose · 488
- **Publications** 1,404,373
  - inspire-jobs · 388
  - proceedings · 168
  - pubs · 1,403,817
- **Analysis notes & Glance** 3,526,068
  - analysis-docs · 740,306
  - derived · po-notes-items · 2,192,558
  - derived · po-notes-knowledge · 16,171
  - derived · po-notes · 565,070
  - glance · 11,743
  - derived · kb-variable-cards · 220
- **Meetings** 1,144,933
  - codimd · 108
  - conference-talks · 123
  - conferences · 39
  - derived · meetings-knowledge · 1,922
  - derived · meetings · 14,378
  - indico-attachments · 28,594
  - indico-contributions · 103,271
  - indico-global · 16,968
  - indico · 979,530
- **JIRA** 2,707,667
  - derived · jira · 1,636
  - jira · 2,706,031
- **Code & GitLab** 6,286,082
  - code-src · 4,493
  - code · 5,437,626
  - gitlab-commits · 23,394
  - gitlab-docs · 660,980
  - gitlab-mrs · 12,460
  - mcp-code · 144,889
  - rnd-athena-code · 2,240
- **Talk, e-groups & news** 1,012,783
  - atlastalk-authors · 103
  - atlastalk-categories · 18
  - centralpage · 205
  - communication · 138,238
  - egroups · 874,219
- **HEPData, PDG & registries** 461,778
  - derived · hepdata · 44,118
  - dsid · 278,313

The Embedding space



PAGERANK (PERCENTILE)

PageRank over the whole links registry (authored links), as a percentile among linked pages.

highest ninth	5,322
	5,480
	5,193
	6,609
middle	4,120
	5,267
	79
	15,912
lowest ninth	0
not linked	165,615

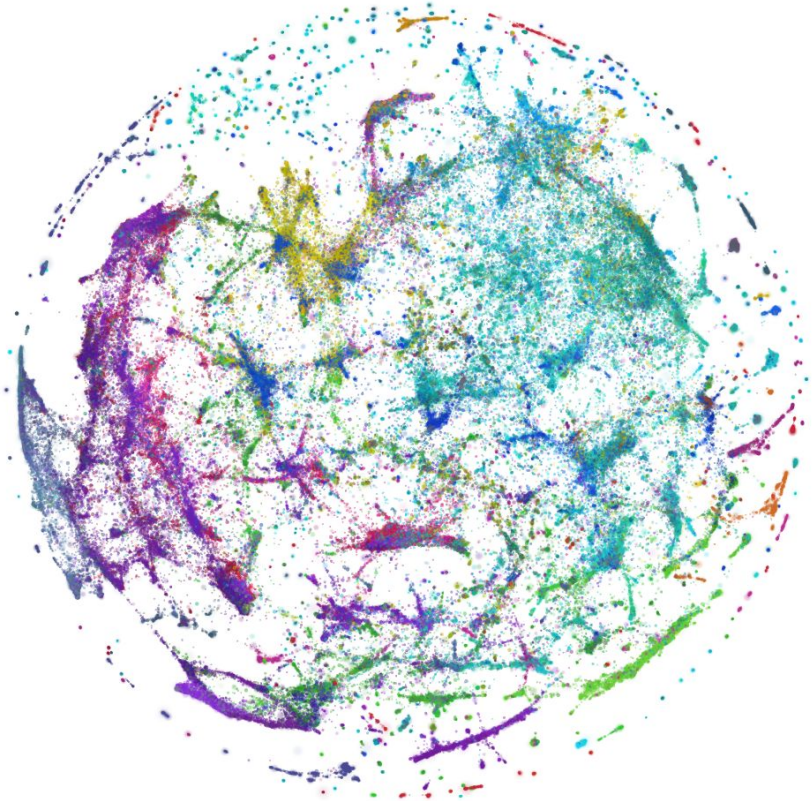
The 100 highest are ringed.

47,982 of 213,597 pages have a value.

213,597 pages from **68 indices** (20,443,726 documents in all), one mark each, placed by meaning: nearby marks are pages about nearby things. 39,634 of them are from derived indices (generated from the others).

UMAP keeps each page's neighbours and pulls clusters apart; how far apart two distant clusters sit means little.

The Embedding space



213,597 pages from **68 indices** (20,443,726 documents in all), one mark each, placed by meaning: nearby marks are pages about nearby things. 39,634 of them are from derived indices (generated from the others).

Force-directed, links + meaning: links as strong springs, meaning-neighbours as weak ones, so an unlinked page still settles beside what it is about.

INDEX FAMILIES

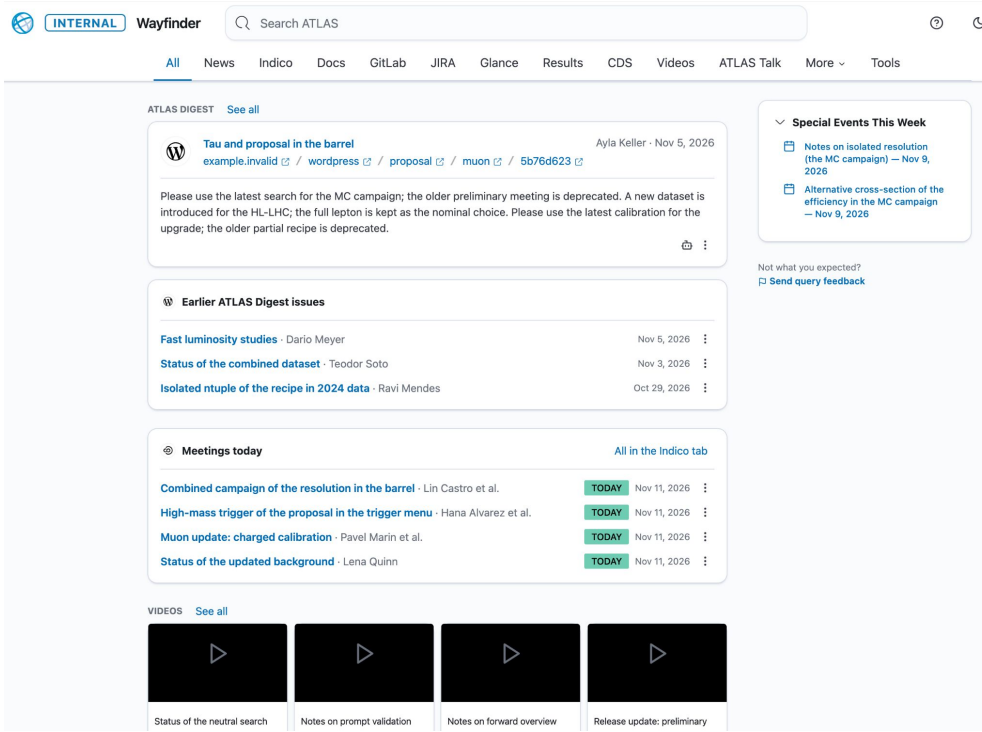
- TWiki & Confluence 339,236
  - wiki · 339,236
- Documentation sites 1,565,194
  - derived · web-items · 831,596
  - derived · web · 405,354
  - web · 327,756
  - rnd-athena-prose · 488
- Publications 1,404,373
  - inspire-jobs · 388
  - proceedings · 168
  - pubs · 1,403,817
- Analysis notes & Glance 3,526,068
  - analysis-docs · 740,306
  - derived · po-notes-items · 2,192,558
  - derived · po-notes-knowledge · 16,171
  - derived · po-notes · 565,070
  - glance · 11,743
  - derived · kb-variable-cards · 220
- Meetings 1,144,933
  - codimd · 108
  - conference-talks · 123
  - conferences · 39
  - derived · meetings-knowledge · 1,922
  - derived · meetings · 14,378
  - indico-attachments · 28,594
  - indico-contributions · 103,271
  - indico-global · 16,968
  - indico · 979,530
- JIRA 2,707,667
  - derived · jira · 1,636
  - jira · 2,706,031
- Code & GitLab 6,286,082
  - code-src · 4,493
  - code · 5,437,626
  - gitlab-commits · 23,394
  - gitlab-docs · 660,980
  - gitlab-mrs · 12,460
  - mcp-code · 144,889
  - rnd-athena-code · 2,240
- Talk, e-groups & news 1,012,783
  - atlastalk-authors · 103
  - atlastalk-categories · 18
  - centralpage · 205
  - communication · 138,238
  - egroups · 874,219
- HEPData, PDG & registries 461,778
  - derived · hepdata · 44,118
  - dsid · 278,313

The ATLAS Search Frontend

Demo

https://atlas-search.docs.cern.ch *[ATLAS Internal Only]*

- ATLAS Search is one of a few front ends that is built on the Knowledge base
- It priorities rapid hybrid based search
- For each source we have a custom
  - result card / source styling
  - structure/relationships
- On Staging we have custom agents for each card type, as well as for combinations

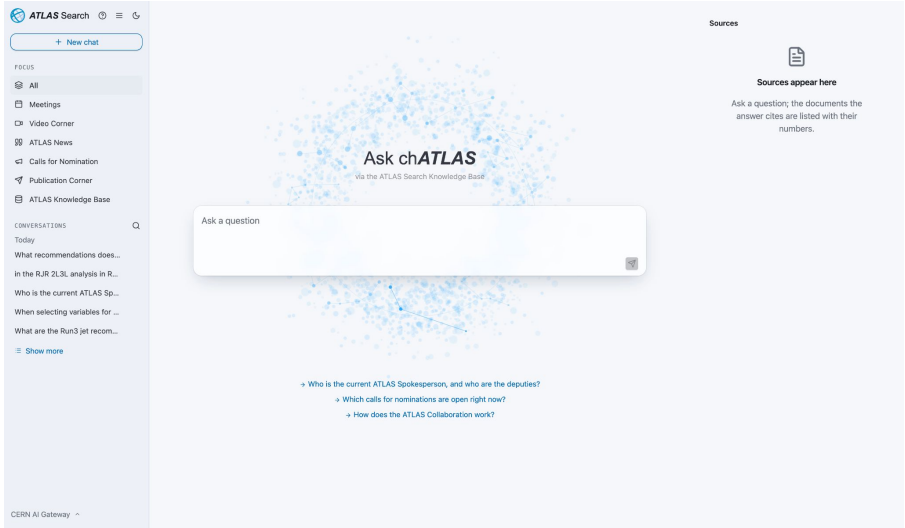


Censored Demo

The Chatlas + ATLAS Search Project

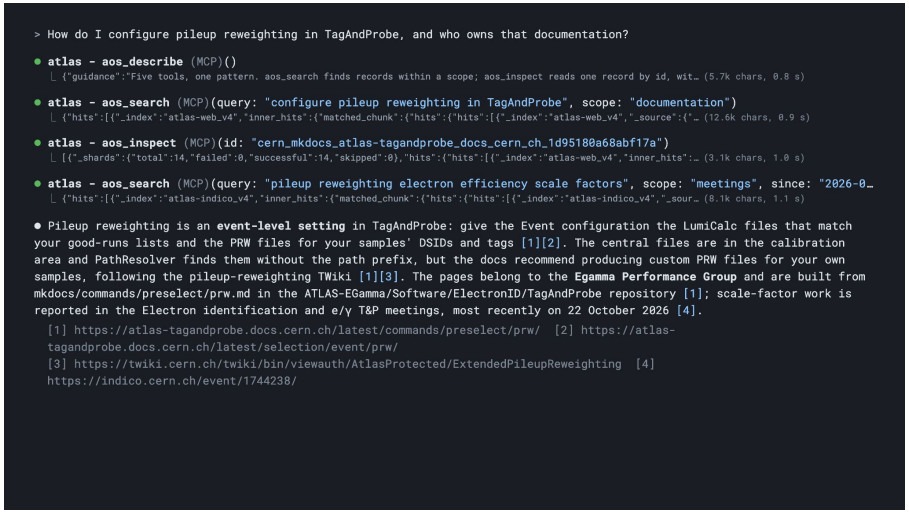
- Chatlas was the first agentic platform in ATLAS - beginning approximately in 2023
- Chatlas focused on a fine tuned RAG model - while focusing on wiki sources
- Recently, Chatlas shifted to more agentic workflows and we've decided to integrate it in the OpenSearch Knowledge Base - combining efforts

Staging area RAG / LLM integration



The MCP Server

- OpenSearch has a rich support of agents and MCP features
- We expose search, semantic, graph search, via the UChicago Analysis Facility and via a bridge w/ kinit
- Graph-aware retrieval - relationship lookups (2.3M curated edges linking analyses to papers, meetings, repos, people) alongside text search, answering questions keyword search can't.
- Agentic query planning - an LLM translates a natural-language request into structured query DSL;



Concluding Remarks

- The LLM era opens up the opportunity to vastly improve the information accessibility across all of our communities
- As we explore agentic flows we should remember that standard search and agentic search are both worthwhile services
- OpenSearch provides an incredible breadth of capability
- Organisations should consider spooling up their own OpenSearch clusters while providing broad options.

>750
unique visitors

2,300
sessions

7.5 k
distinct searches

~2,700
captured ‘url’ clicks
* unreliable metric

Concluding Remarks

- The LLM era opens up the opportunity to vastly improve the information accessibility across all of our communities
- As we explore agentic flows we should remember that standard search and agentic search are both worthwhile services
- OpenSearch provides an incredible breadth of capability
- Organisations should consider spooling up their own OpenSearch clusters while providing broad options.

Thats all for today