

EBI Search: Engineering and Sustaining Metadata Infrastructure for Life Sciences



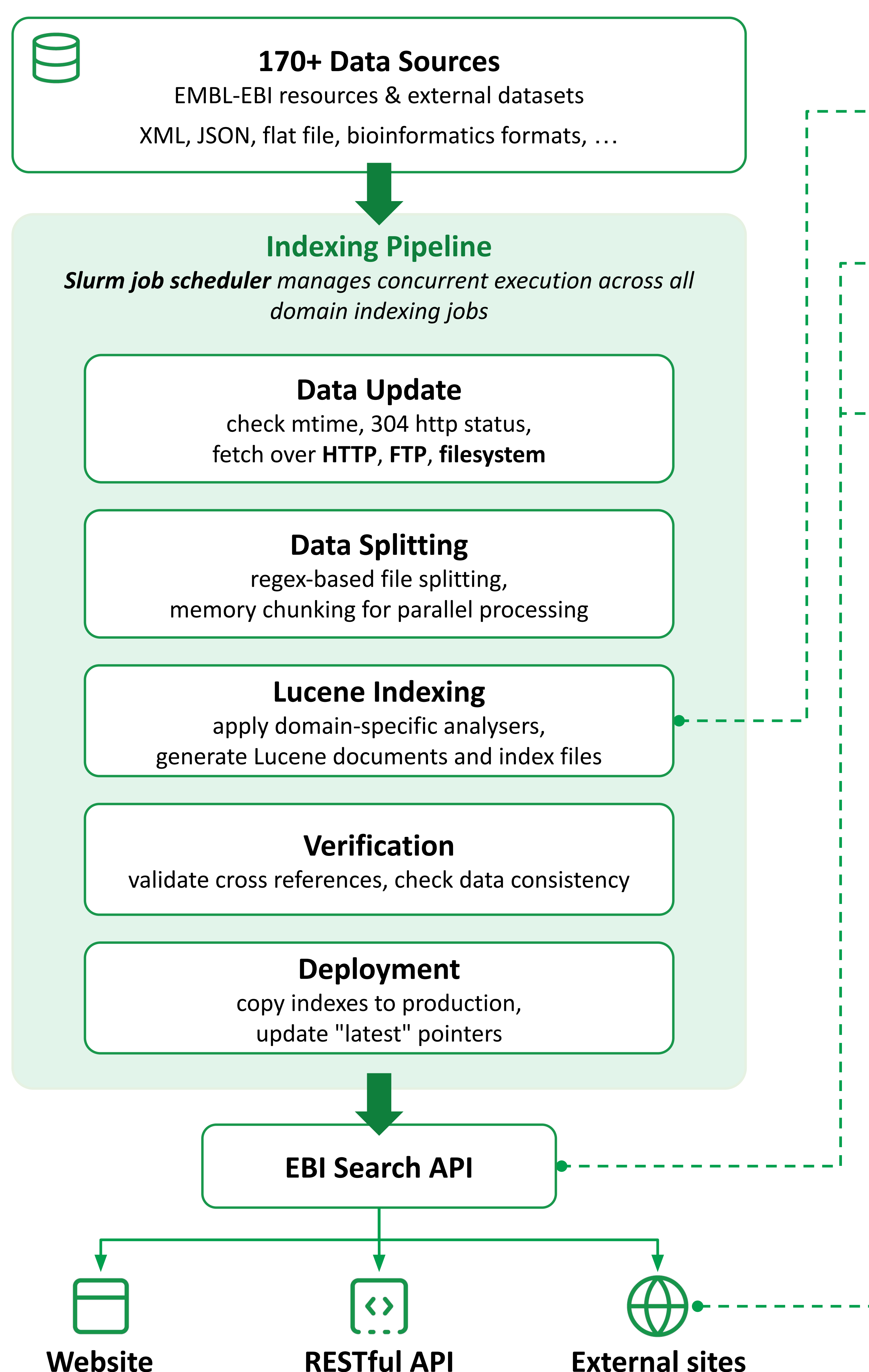
Dalya Al-Shahrabi, Prasad Basutkar, Renato Caminha Juaçaba Neto, Vijay Subramoniam, Rose Neis, Iva Tutis, Henning Hermjakob, Matt Pearce

EBI Search (www.ebi.ac.uk/ebisearch/) is a unified metadata search engine that indexes biological data from 170+ internal and external sources, providing users with data discovery tools such as **free-text search**, **multi-level faceting**, **cross-referencing capabilities**, and **bulk querying** across the indexed metadata. Built on **Lucene Core** [1], it offers both **web interface** and **RESTful APIs** for programmatic access [2].



Nightly Indexing Pipeline

The nightly indexing pipeline processes updated data from internal and external sources using parallelized execution managed by Slurm.



Engineering Decisions

Increasing data volumes and user demands create competing technology requirements. The engineering decisions we make in response shape the EBI Search infrastructure.

Index Partitioning

Work around Lucene's 2.1 billion document limit through hierarchical domain splitting with transparent query distribution.

Query rewriting

Avoid storage duplication by rewriting text searches to target multiple fields at query time rather than creating combined fields.

API Optimisations

Separate filter queries from relevance scoring for better search performance, and enable bulk streaming for results exceeding 1 million entries.

Reusable Portal Components

Reusable templates and components streamline EBI Search portal creation, avoiding duplicated effort while preserving performance and feel.

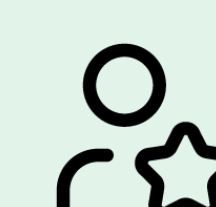
Beyond the Code

EBI Search is run and maintained by a small team of software engineers. Daily decisions, provider relationships, and systematic troubleshooting are critical for long-term sustainability.



Daily monitoring

Email alerts, failure investigation, immediate issues



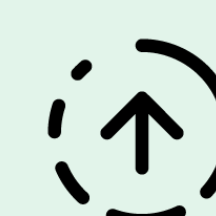
Provider relationships

Negotiating formats, ongoing support, proactive communication when APIs fail



Domain management

Manual domain retirement and rollback decisions



Technology evolution

Weighing infrastructure improvements against risk factors for continued operation

Future Directions

Scaling infrastructure

To better support the indexing of metadata for tens of billions of records, we are migrating to **Apache Solr**. This will enhance both scalability and efficiency, and enable more seamless search experiences across multiple datasets.

DocBot

Developing **DocBot**, a chatbot that provides user-friendly access to EBI's extensive documentation resources. This pilot project develops internal expertise in LLM deployment, laying the foundation for a comprehensive SearchBot integrated with EBI Search and spanning most EBI content.

Acknowledgements

EMBL-EBI is indebted to its funders, including the EMBL member states and the European Commission through the H2020 Programme under EOSC-Life [824087], BY-COVID [101046203], and EarlyCause [84815].

References

1. Apache Lucene Core [Internet]. Lucene.apache.org. 2022 [cited 30 August 2022]. Available from: <https://lucene.apache.org/core/>
2. Matthew Pearce, Prasad Basutkar, Renato Caminha Juaçaba Neto, Vijay Venkatesh Subramoniam, Kelsey Neis, Iva Tutis, Henning Hermjakob, *EBI Search: providing discovery tools for biological metadata in 2025*, Nucleic Acids Research, 2025;, gkaf359, <https://doi.org/10.1093/nar/gkaf359>