RCSB Protein Data Bank: Architectural Advances Towards Integrated Searching and Efficient Access to Macromolecular Structure Data from the PDB Archive

Yana Rose [4], Jose M. Duarte [4], Robert Lowe [1,2], Joan Segura [4], Chunxiao Bi [4], Charmi Bhikadiya [1,2], Li Chen [1,2], Alexander S. Rose [4], Sebastian Bittrich [4], Stephen K. Burley [1,2,3,4,5] and John D. Westbrook [1,2,3]*

*1 - **Research Collaboratory for Structural Bioinformatics Protein Data Bank,** Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*

*2 - **Institute for Quantitative Biomedicine,** Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*

*3 - **Cancer Institute of New Jersey,** Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA*

*4 - **Research Collaboratory for Structural Bioinformatics Protein Data Bank,** San Diego Supercomputer Center, University of California, La Jolla, CA 92093, USA*

*5 - **Department of Chemistry and Chemical Biology,** Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA*

***Correspondence to John D. Westbrook:** Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. john.westbrook@rcsb.org (J.D. Westbrook)*

https://doi.org/10.1016/j.jmb.2020.11.003

Edited by Michael Sternberg

Abstract

The US Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) serves many millions of unique users worldwide by delivering experimentally-determined 3D structures of biomolecules integrated with >40 external data resources via RCSB.org, application programming interfaces (APIs), and FTP downloads. Herein, we present the architectural redesign of RCSB PDB data delivery services that build on existing PDBx/mmCIF data schemas. New data access APIs (data.rcsb.org) enable efficient delivery of all PDB archive data. A novel GraphQL-based API provides flexible, declarative data retrieval along with a simple-to-use REST API. A powerful new search system (search.rcsb.org) seamlessly integrates heterogeneous types of searches across the PDB archive. Searches may combine text attributes, protein or nucleic acid sequences, small-molecule chemical descriptors, 3D macromolecular shapes, and sequence motifs. The new RCSB.org architecture adheres to the FAIR Principles, empowering users to address a wide array of research problems in fundamental biology, biomedicine, biotechnology, bioengineering, and bioenergy.

Introduction

Established in 1971 as the first open-access, digital data resource in biology with just seven protein structures, the Protein Data Bank (PDB)[1] is universally regarded as a fundamental and core data resource essential to the basic and applied research in the life-sciences, fundamental biology, biomedicine, biotechnology, bioengineering, and energy communities. Now in its 50th year of continuous operation, the PDB serves as the singular global repository for 3D structural information, making >170,000 experimentally determined structures of proteins, DNA, RNA, and their complexes with drugs and/or other small molecules freely available without usage limitations.

Since 2003, the PDB has been managed jointly by the Worldwide Protein Data Bank (wwPDB)

partnership,[2,3] (including US-funded Research Collaboratory for Structural Bioinformatics Protein Data Bank or RCSB PDB,[4,5] Protein Data Bank in Europe,[6] Protein Data Bank Japan,[7] and the Biological Magnetic Resonance Data Bank.[8] wwPDB partners provide global deposition-validation-biocuration services to guarantee that archived structures are as complete as possible and receive consistent validation and expert biocuration.[9–11] In addition to wwPDB data acquisition and archiving, the RCSB PDB provides a variety of data delivery services packaging the primary archival data integrated with additional content from >40 leading biological and life science resources. These services include tools enabling data search, browsing, custom report generation, visualization, and analyses tailored for both programmatic and web interactive users. Collectively these activities and services strengthen our enduring commitment to the **FAIR** Principles of **F**indability-**A**ccessibility-**I**nteroperability-**R**eusability.[12]

The RCSB PDB (hereafter RCSB) assumed responsibility for the PDB within the US in 1999. Since then, RCSB data delivery services have undergone substantive design changes. Redesigns have been motivated by proactive efforts to address challenges arising from new primary data entering the PDB archive,[3] growth in related community domain data resources, and dramatic growth and broadening of the diverse array of PDB data.[13–15] Among the challenges coming from new primary data sources are growing numbers of depositions, increasing size and molecular complexity of deposited structures, and the rapidly evolving technology landscape in structural and computational biology. External resources targeted for data integration in the biological and life sciences have grown significantly in both number and scale, making it ever more demanding to maintain data correspondence and mapping information for interoperability across and beyond this bioscience data ecosystem. In addition to challenges arising from primary and integrated data, the growing capacity, performance, and feature requirements coming from both web-based interactive and programmatic RCSB users represent ongoing drivers for data delivery service redesign. Concurrently, regular improvements have been necessary to maintain IT infrastructure implemented using contemporary tools and modern software engineering best-practices. The April 2020 release of new RCSB data services represents the most significant upgrade in overall design to date, implementing sweeping changes in both data and software architecture. The following sections present features and capabilities of this new RCSB service architecture.

Methods

Architecture overview

The new RCSB data and search service architecture is illustrated in Figure 1. With this implementation, we transitioned from a large feature-rich multi-purpose web application to a modern architecture within which services are delivered by a collection of loosely coupled collaborating applications each with narrow, well-defined responsibilities. A major software overhaul decomposed the previous implementation into small single-purpose services, with logical service boundaries, and well-defined connecting APIs.

Decomposition yielded services supporting both search and data access functions. New back-end services are individually responsible for searches of text and structured attributes, sequence similarity, sequence motifs, structure similarity, and chemical similarity. These search services are orchestrated by an aggregation application responsible for dispatching search tasks to appropriate back-end search services and combining results therefrom.

Deposited primary data integrated with external data supporting search and data access services are managed in a data warehouse, the document store populated with structured data that acts as the authoritative source of content in the architecture. Structured data are indexed for text and attribute searches. Raw data artifacts are delivered by a content delivery network (CDN) service. Data access services (data.rcsb.org) are provided through both REpresentational State Transfer (REST)[16] and GraphQL[17] APIs.

Data flow along the pipeline is illustrated in Figure 1. Data from the regional wwPDB deposition sites are added to the PDB archive and shared with the wwPDB partners on a fixed weekly schedule. The jointly managed archive provides the source of primary PDB data for the RCSB data delivery pipeline. The ETL (*i.e.*, Extract, Transform, Load) and service deployment operations supporting this update workflow are orchestrated by a Luigi workflow management agent (github.com/spotify/luigi). The search aggregator service (search.rcsb.org) and data access service (data.rcsb.org) provide entry points for search and access operations shared by both the RCSB web front-end application (rcsb.org) and public-facing programmatic web services. In this new architecture, the RCSB.org website front-end has adopted a modern and extensible front-end web framework, while retaining the familiar look and feel of the resource.

Data and schema

An important feature of overall data management in the new RCSB architecture (Figure 1) is continuity in metadata management spanning the full data lifecycle. Use of metadata starts at the beginning of the pipeline with acquisition of depositor-provided information, followed by validation and biocuration of these data by the wwPDB OneDep system.[9] To ensure consistency and extensibility in all of its data process operations,

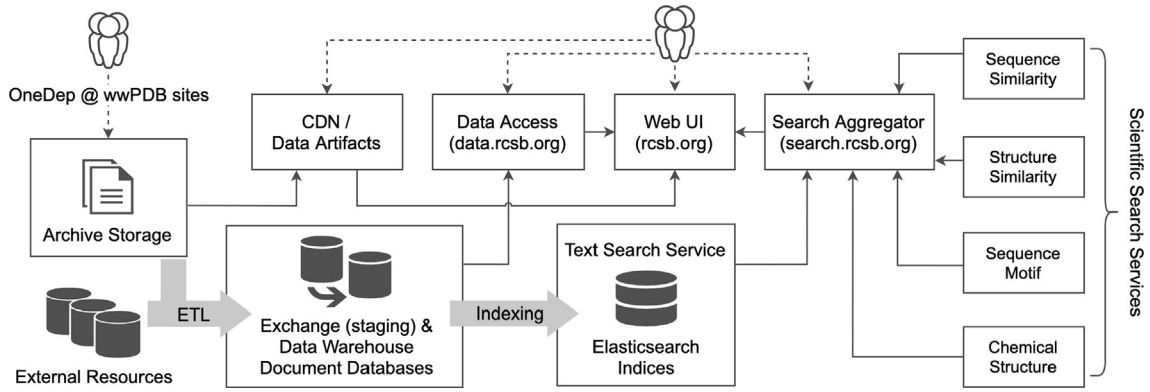

Figure 1. Data management and delivery system underpinning the new RCSB architecture.

OneDep and its supporting tools rely on a digital metadata dictionary as the authoritative reference source for information about PDB archive data. This metadata reference dictionary is a product of the PDBx macromolecular Crystallographic Information Framework (mmCIF),[18–23] which evolved from standardization efforts of the International Union of Crystallography that began in the 1990s.[24] In 2014, PDBx/mmCIF became the internationally-recognized metadata standard for the PDB archive[25] and in 2019 PDBx/mmCIF was required for deposition of atomic coordinates generated using macromolecular crystallography (MX).[26]

The PDBx/mmCIF data model is highly structured, defining a rich collection of biological, molecular, chemical, structural data, and data quality features. PDBx/mmCIF is a dynamic data model that grows continuously with technology evolution and methodological advances in structural biology. PDB chemical and molecular reference data[27,28] are also managed within the PDBx/mmCIF schema. This data model further provides metadata required to perform precise semantic alignment of data integrated from external data resources.

The organization of the PDBx/mmCIF data model[20–23,26,29–30] is tabular, allowing facile management of data represented in this framework using relational database tools. In addition to providing data content specification, the PDBx/mmCIF metadata framework contains data typing, data provenance, validation, and organizational details required for automated checking of data consistency. The wwPDB partners in collaboration with community domain experts (wwPDB PDBx/mmCIF Working Group) coordinate development and maintenance of the PDBx/mmCIF dictionary. Working Group deliberations and data dictionary content are published on the GitHub platform (github.com/pdbxmmcifwg) and a data portal site (mmcif.wwpdb.org), respectively.

Metadata inform the next step of the new data delivery workflow (Figure 1), in which the largely tabular archival primary data is projected onto a document hierarchy that reflects the underlying macromolecular structural hierarchy. Biological macromolecules have a natural hierarchy, building from units of different granularity extending from atoms to polymer components (*e.g.*, amino acids) to polymer chains to assemblies of interacting polymer chain macromolecules and ligands. The document data model is well-suited to handling the hierarchical data representation for macromolecular structures. Data organization is formalized in a document schema within which features describing a particular level in the macromolecular hierarchy are grouped into document collections. The current document schema includes collections describing the following: deposited data (or entry); macromolecular assemblies generated from the entry; decomposition of the entry in terms of distinct polymer, branched (*e.g.*, oligosaccharide), and non-polymer molecular entities (*e.g.*, small-molecule ligands such as co-factos and enzyme inhibitors); and the observed atomic structure instances of these molecular entities in the deposited data set. The schema does not include raw atomic-level coordinates. Information about atomic-level coordinate data including counting statistics, completeness, and a range of experimental and geometrical data quality metrics are summarized within the molecular hierarchy. The document schema includes a subset of available primary (PDBx/mmCIF) data content that is well-populated across the entire archive.

Operationally, transformation to the document organization takes place as the archival data products of the OneDep system are loaded into an intermediate staging document store (Figure 1: Exchange DB). This step requires both data and schema transformations with the latter being encoded in a standard JSON (JavaScript Object Notation) schema representation (json-schema.org) with some local RCSB extensions. Some essential data integration tasks closely tied to the primary data and chemical reference data are also performed at this stage. These ETL operations

include updates to reference sequence database correspondences plus reference database correspondences for small molecules in PDB chemical and molecular reference data.

The Data Warehouse (DW, Figure 1) provides the authoritative source of data for all RCSB data access and search services. Primary archival data loaded into the staging database (Exchange DB) with data from external resources processed by local ETL operations are federated into the DW document store. The DW document store is in turn responsible for supporting the data access needs of the RCSB.org website and the public-facing RCSB programmatic data access API services.

A document organization was chosen for the new DW because this approach most closely resembles the predominant data access patterns employed in assembling content for the RCSB.org website and in delivering programmatic web services (data.rcsb.org). Having the data organized in this readily accessible document format improves retrieval performance and avoids computationally expensive JOIN operations on normalized tabular data that were required with our previous relational database architecture.

The DW data schema extends the primary data schema used in the staging database (Exchange DB) with additional annotations coming from external resources. Addition of these annotations to the data schema allows the origin of each annotation to be clearly defined. These schema revisions and extensions within the new architecture are managed and automatically deployed through a GitHub (github.org) version control workflow.

In the new architecture, the main communication medium for data exchange for the different service APIs is JSON, making JSON Schema a convenient format choice to store and exchange schema information. The DW data schema, encoded using JSON Schema language, includes attribute descriptions, examples, and validation keywords such as data type, controlled vocabularies, and boundary values. The DW is hosted in a MongoDB (mongodb.com) document-oriented database, which supports document validation using a flavor of JSON schema.

Data schema represent essential vehicles for sharing data specifications among different services. Knowledge of the data schema is essential in a modular architecture composed of many collaborating services wherein ensuring data integrity and consistency between the services is critically important. Central consolidation of data schema in a technology-agnostic format allows this information to be shared among components of the data management and delivery system.

Figure 2 illustrates how the JSON schema is used to establish contracts between services in our data management and delivery system: (i) JSON schema validation constraints are used by ETL processes to check data before loading the DW; (ii) search indexing processes use JSON schema data types to automatically create an indexing configuration; (iii) the Search Aggregator service uses JSON schema metadata describing searchable attributes and possible search operations to validate search requests; (iv) the front-end RCSB.org module uses the same metadata to dynamically construct the RCSB.org Advanced Search query builder, and supporting pages of documentation describing searchable attributes; and (v) the data access module uses metadata type details to enable type-safe data parsing and automatic documentation generation for the data access services.

Data access services

The data access service (data.rcsb.org) provides the gateway to the DW datastore (Figure 1). This service is implemented as a lightweight application delivering both GraphQL and traditional REST-based APIs over an HTTP/S protocol.

The GraphQL query language provides composable access to the full range of content within the DW, allowing users to craft custom requests for subsets of data that match particular needs. The new GraphQL API was implemented using an open-source SPQR (Schema Publisher & Query Resolver) Java library (github.com/leangen/graphql-spqr). The scope of content available to GraphQL is defined in a GraphQL schema, which is automatically generated from

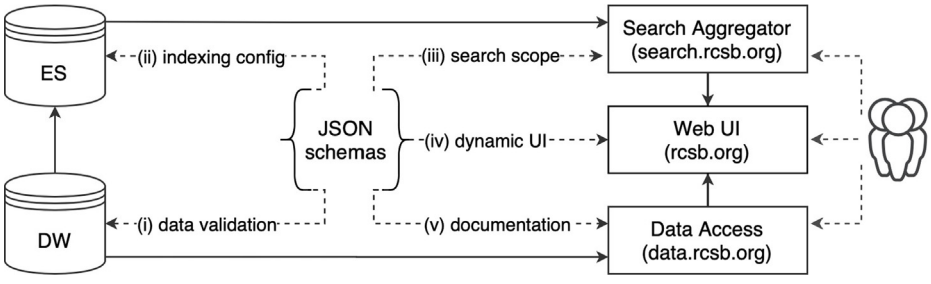

Figure 2. Schema usage by different components of the data management and delivery system.

our JSON schemas (Figure 2). Programmatic GraphQL API requests are served through a single URL/endpoint (data.rcsb.org/graphql). In addition to the API, the GraphQL service provides an interactive browser-based user interface (UI), *GraphiQL* (data.rcsb.org/graphql) that facilitates composing, validating, and testing GraphQL API queries. This UI also exposes the rich documentation defined in the data schema.

The new REST API is implemented using the open-source JAX-RS support in the Jersey Java framework (eclipse-ee4j.github.io/jersey). The defined collection of URLs/endpoints are organized to reflect the underlying data hierarchy of the new data architecture. For example, the REST API entity group provides access to features of distinct polymer, branched entity (*e.g.*, oligosaccharides), and non-polymer entity data. REST API service responses are delivered as JSON payloads. REST endpoints documentation (data.rcsb.org/redoc) including input parameters and output data schema is generated as the OpenAPI specification (www.openapis.org) and rendered using ReDocUI tool (redoc.ly).

CDN services

Static data assets are delivered through content delivery network services (Figure 1: CDN). The file service (files.rcsb.org) provides access to data hosted in the PDB FTP repository. This body of information includes primary archival data files containing atomic coordinates and files containing supporting experimental data. Macromolecular structure images (*e.g.*, assemblies, structures, sub-structures), small-molecule chemical diagrams, carbohydrate SNFG[31] diagrams, and other static content used by the RCSB.org website are served by the CDN service (cdn.rcsb.org).

Specialized services have been developed to efficiently deliver large data artifacts to web-based molecular graphics and analysis tools. These services, built on the open-source Mol* library,[32] enable interactive manipulation and analysis of the largest macromolecular assemblies, 3D Electron Microscopy (3DEM) map volumes, and electron density maps from MX in the PDB repository. The model service (models.rcsb.org) delivers atomic coordinates together with the annotations in the primary data files in a compressed binary CIF encoding (BCIF).[33] Structure data can be served at different levels of granularity (*e.g.*, assembly, polymer chain, ligand), and ligand data may also be delivered in popular chemical informatics formats (*e.g.*, SDF, MOL, MOL2). The volume service (maps.rcsb.org) provides access to volumetric data from MX electron density and 3DEM volume maps. This service can also deliver volume subsets and optionally downsample the volumetric data to reduce data transfer bandwidth requirements.

Search aggregator service

The responsibilities of the search aggregator service (Figure 1: Search Aggregator) include: (i) providing the single-entry point for all search operations, (ii) routing requests to the appropriate underlying search services for processing, and (iii) applying basic Boolean logic operations to combine the multiple search results.

This service enables searches across elements of macromolecular structure data at different levels of granularity. A common example involves searching for macromolecular assemblies containing a protein similar to a target sequence bound to a ligand similar to a drug target. Here, the results from the two search modes, for small molecules and protein sequences, are first merged at the granularity of the assembly result set type and then combined with a Boolean AND operator.

The aggregator service provides a uniform API layer that abstracts details required to merge and combine the results of the underlying search services. To support this API, a custom domain-specific language (DSL) has been developed to describe search queries. Queries in this custom DSL are represented as a graph. Nodes in this graph can represent either individual or group searches. A simple terminal node describes a single search operation (*e.g.*, an attribute-based search or sequence search). A group node combines multiple nodes with a Boolean combination operation. Nodes in the query graph may be arbitrarily nested allowing for the construction of complex search patterns. While the introduction of our custom DSL added some additional software development effort required to build the search aggregator service, it has been more than offset by the greater flexibility and extensibility that the DSL has afforded. In particular, the custom DSL provides a simple abstraction for query construction that encapsulates the implementation differences in the underlying search services.

The search aggregator application is implemented as a lightweight stateless REST-style web service. The service endpoint (search.rcsb.org/rcsbsearch/v1/query) provides a JSON-based API implemented using the Java Jersey framework (eclipse-ee4j.github.io/jersey). The API accepts HTTP/S GET and POST and returns a JSON response.

Search services

The new architecture (Figure 1) implements searches for text and structured attributes, sequence similarity, sequence motifs, structure similarity, and chemical similarity as independent services all orchestrated by the search aggregator.

The Attribute and Text Search service enable composable queries on the content of the DW

document store. This service is built on the open-source Elasticsearch search engine (V7: https://www.elastic.co). Elasticsearch transforms JSON documents from the DW into *Inverted Indices* optimized for type-specific structured attribute searches or unstructured text searches. Structured attribute queries enable matching numbers, Boolean values, dates, and exact text values. Keywords and phrases can be matched as unstructured text. Search results are returned as identifiers for DW documents at the desired granularity in the DW data model. Matching documents are returned with an internally calculated relevance score allowing rank ordering of results by significance.

The Sequence Similarity Search service enables performant queries for protein, DNA, and RNA one-letter-code sequences in the PDB archive by employing the sequence comparison tool MMseqs2.[34,35] For each matching sequence, the service returns a unique identifier for the PDB polymer sequence, matching scores (*e.g.*, sequence identity, E-value, and bit-score), and the residue boundary positions of the matching sequence.

The Sequence Motif Search service enables queries on protein or nucleic acid polymer sequences, using three different types of input format simple one-letter-code sequence patterns, regular expressions of one-letter-code patterns, and PROSITE[36] patterns. The service returns PDB polymer entity identifiers and residue boundaries for matches in the sequences.

The Structure Similarity Search service enables queries for global similarity in spatial arrangements of macromolecules and macromolecular assemblies. The service employs a computationally-efficient BioZernike method developed by RCSB.[37] To perform fast shape comparisons the method uses pre-calculated rotationally invariant descriptors of the volumetric and geometric representations of each 3D shape. The service outputs identifiers of the matching structural elements along with similarity scores calculated for volumetric and geometric descriptors.

The Chemical Search service enables queries of small-molecule constituents of PDB IDs or structures, based on chemical formula and chemical structure. Both molecular formula and formula range searches are supported. Queries for matching and similar chemical structures can be performed using SMILES[38] and InChI[39] descriptors as search targets. Graph and chemical fingerprint searches are implemented using tools from the OpenEye Chemical Toolkit (www.eyesopen.com/oechem-tk).

Infrastructure support

The new architecture (Figure 1) is deployed in a private cloud environment built on the open-source OpenStack cloud platform. Local tools have been developed using the OpenStack API for creating and deploying custom cloud virtual machine instances. Customizable instance configurations enable efficient management of overall physical resources which can flexibly adapt to changes in user demands. The RCSB manages cloud resources in geo-redundant data centers located on the campuses of the University of California, San Diego and Rutgers, The State University of New Jersey.

Results and discussion

To illustrate the power of the new search and data access capabilities, we present the example of finding 3D structures for a particular subset of ligand-protein complexes related to SARS-CoV-2 and then accessing a wide range of metadata associated with the matching structures. Analogous examples are also available[40] describing how these capabilities have been used to develop new search and data access tools for the RCSB PDB web resource (RCSB.org).

The search service API request for this example is depicted in Figure 3(a). It combines full-text (a1), sequence (a2), structure (a3), and chemical similarity (a4) search modes. The text search targets structures of the *Coronaviridae* family taxonomy. The sequence similarity search (a2) targets the sequence of the SARS-CoV Nsp5 domain (PDB ID 1Q2W)[41] with a comparison threshold of 50% sequence identity. The structure similarity search targets the shape of the first biological assembly similar in PDB ID 6LU7 (SARS-CoV-2 Nsp5).[42] The chemical similarity search targets the 3C-like protease inhibitor 7J,[43] represented as a SMILES chemical descriptor. At the time of writing, performing each of these searches and applying a Boolean AND operation to select the only common results yields 3 structures satisfying all of the search criteria (PDB IDs: 6W2A, 6XMK, 6VH3).[43] This example showcases how accessing four different search services is achieved through a single API request to the Search Aggregator service (Figure 1).

The companion data access API request is shown in Figure 3(b) illustrating the GraphQL API service request to fetch details spanning the DW schema hierarch. This query includes entry-level attributes such as the entry title, experimental methods, depositors, deposition and release dates (b1); primary citation data (b2); taxonomy details for polymeric entities (b3), descriptive information for branched entities (*e.g.*, oligosaccharides) (b4); and names, formulae, and formula weight for small-molecule ligands (b5). The attributes in this example can also be accessed using the REST API; however, multiple different requests are required to retrieve the same data. The advantage of the GraphQL API is the ability to craft a single API request for all of the desired data content.

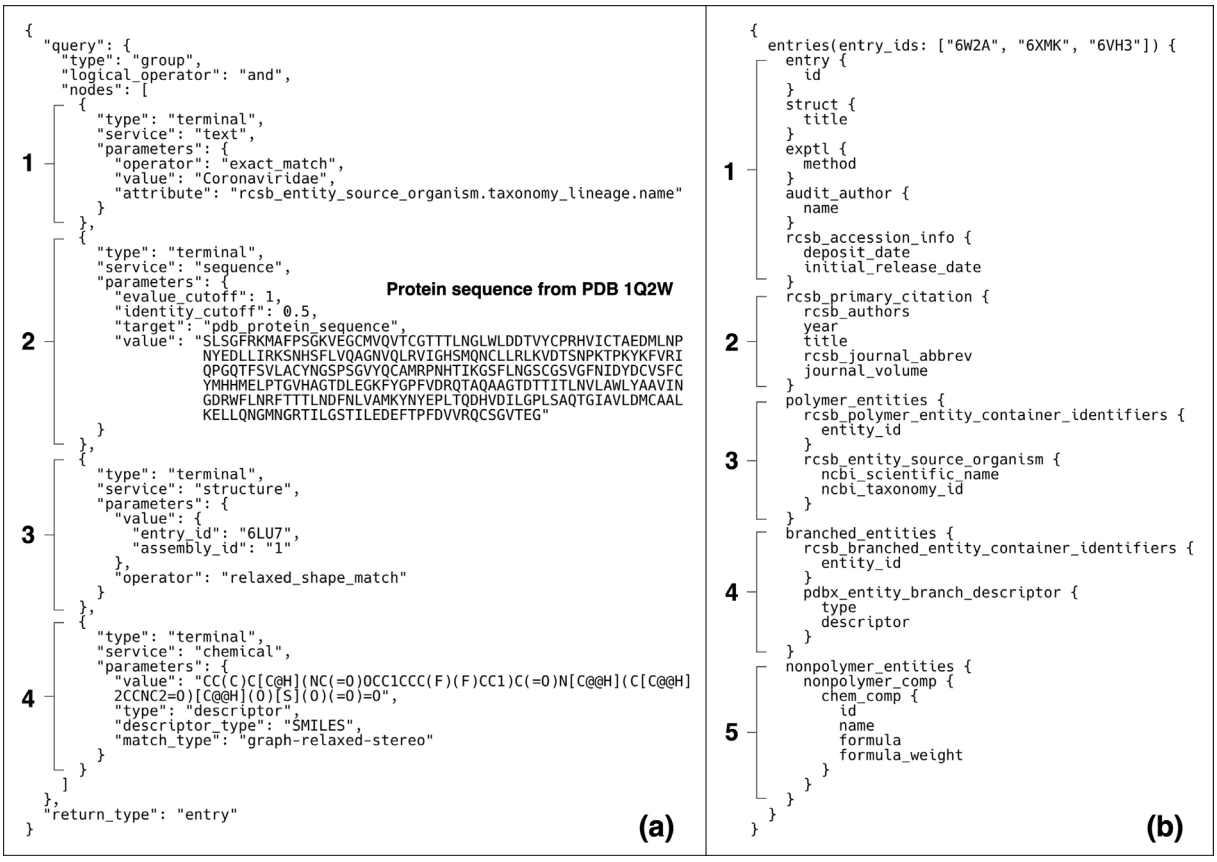

Figure 3. Example search and data access queries: (a) query that combines text (1), sequence (2), structure shape (3), and chemical similarity (4) searches; (b) GraphQL API query including essential entry details (1–2), information details of the macromolecular entity data hierarchy (2–4) and small-molecules (5).

A key objective of our architectural redesign has been improving the overall FAIR-ness of the RCSB data delivery services. As the preceding example demonstrates, improvements in search and data access services significantly advance the Findability and Accessibility of the PDB structure data. Consolidation of data schema and adoption of community standards such as OpenAPI, JSON Schema, and JSON for API data exchange collectively improve both data Interoperability and Reusability.

We hope the outcome of the architectural redesign will enhance operational efficiencies, improve deployment scalability, reduce the time for the rollout of new features and bug fixes, and enable more proactive targeted monitoring of service health. The architectural redesign is also expected to pave the way for more economical future deployments in publicly hosted cloud resources.

CRediT authorship contribution statement

Yana Rose: Writing - original draft, Conceptualization, Methodology, Software. **Jose M. Duarte:** Conceptualization, Methodology, Software, Supervision. **Robert Lowe:** Conceptualization, Methodology, Software. **Joan Segura:** Conceptualization, Methodology, Software. **Chunxiao Bi:** Conceptualization, Methodology, Software. **Charmi Bhikadiya:** Methodology, Software. **Li Chen:** Conceptualization, Methodology, Software. **Alexander S. Rose:** Conceptualization, Methodology, Software. **Sebastian Bittrich:** Conceptualization, Methodology, Software. **Stephen K. Burley:** Conceptualization, Supervision, Funding acquisition, Writing - review & editing. **John D. Westbrook:** Writing - original draft, Writing - review & editing, Conceptualization, Methodology, Software, Supervision.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

RCSB PDB is funded by the National Science Foundation [DBI-1832184], the US Department of Energy [DE-SC0019749], and the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health [R01GM133198].

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2020.11.003.

Received 30 September 2020;
Accepted 5 November 2020;
Available online xxxx

Keywords:
structural biology;
databases;
computer architecture;
FAIR principles

Abbreviations used:
3DEM, Three-dimensional Electron Microscopy; API, Application Programming interface; FAIR, Findability, Accessibility, Interoperability, and Reusability; MX, Macromolecular Crystallography; REST, REpresentational State Transfer

References

1. Protein Data Bank, (1971). Crystallography: Protein Data Bank. *Nature (London), New Biol.*, **233**, 223.
2. Berman, H.M., Henrick, K., Nakamura, H., (2003). Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.*, **10**, 980.
3. wwPDB consortium, (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, **47**, D520–D528.
4. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., et al., (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
5. Burley, S.K., Berman, H.M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., et al., (2019). RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464–D474.
6. Velankar, S., Best, C., Beuth, B., Boutselakis, C.H., Cobley, N., Sousa Da Silva, A.W., et al., (2010). PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **38**, D308–D317.
7. Kinjo, A.R., Bekker, G.J., Suzuki, H., Tsuchiya, Y., Kawabata, T., Ikegawa, Y., et al., (2017). Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids Res.*, **45**, D282–D288.
8. Ulrich, E.L., Akutsu, H., Doreleijers, J.F., Harano, Y., Ioannidis, Y.E., Lin, J., et al., (2008). BioMagResBank. *Nucleic Acids Res.*, **36**, D402–D408.
9. Young, J.Y., Westbrook, J.D., Feng, Z., Sala, R., Peisach, E., Oldfield, T.J., et al., (2017). OneDep: Unified wwPDB system for deposition, biocuration, and validation of macromolecular structures in the PDB archive. *Structure*, **25**, 536–545.
10. Gore, S., Sanz Garcia, E., Hendrickx, P.M.S., Gutmanas, A., Westbrook, J.D., Yang, H., et al., (2017). Validation of structures in the Protein Data Bank. *Structure*, **25**, 1916–1927.
11. Young, J.Y., Westbrook, J.D., Feng, Z., Peisach, E., Persikova, I., Sala, R., et al., (2018). Worldwide Protein Data Bank biocuration supporting open access to high-quality 3D structural biology data. *Database.*, **2018** bay002 1–17.
12. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al., (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 1–9.
13. Goodsell, D.S., Zardecki, C., Di Costanzo, L., Duarte, J.M., Hudson, B.P., Persikova, I., et al., (2020). RCSB Protein Data Bank: Enabling biomedical research and drug discovery. *Protein Sci.*, **29**, 52–65.
14. Feng, Z., Verdiguel, N., Di Costanzo, L., Goodsell, D.S., Westbrook, J.D., Burley, S.K., et al., (2020). Impact of the Protein Data Bank across scientific disciplines. *Data Sci. J.*, **19**, 1–14.
15. Burley, S.K., Berman, H.M., Christie, C., Duarte, J.M., Feng, Z., Westbrook, J., et al., (2018). RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Sci.*, **27**, 316–330.
16. Fielding, R.T., (2000). Architectural Styles and the Design of Network-based Software Architectures. University of California, Irvine.
17. GraphQL Foundation/Linux Foundation. GraphQL API Oriented Query Language, 2019.
18. Fitzgerald, P.M.D., Westbrook, J.D., Bourne, P.E., McMahon, B., Watenpaugh, K.D., Berman, H.M., (2005). 3.6 Classification and use of macromolecular data. In: Hall, S.R., McMahon, B. (Eds.), *International Tables for Crystallography*,. Springer, Dordrecht, The Netherlands:, pp. 144–198.
19. Westbrook, J., Berman, H.M., (2005). Ontologies for three-dimensional molecular structure. In: Jorde, L.B., Little, P.F. R., Dunn, M.J., Subramaniam, S. (Eds.), *Encyclopedia of Genomics, Proteomics, and Bioinformatics*,. John Wiley & Sons Ltd, Chichester, pp. 3474–3480.
20. Westbrook, J.D., Fitzgerald, P.M.D., (2009). Chapter 10 The PDB format, mmCIF formats, and other data formats. In: Bourne, P.E., Gu, J. (Eds.), *Structural Bioinformatics*,. Second Edition. John Wiley & Sons, Inc., Hoboken, NJ, pp. 271–291.
21. Fitzgerald, P.M.D., Westbrook, J.D., Bourne, P.E., McMahon, B., Watenpaugh, K.D., Berman, H.M., (2005). 4.5 Macromolecular dictionary (mmCIF). In: Hall, S.R., McMahon, B. (Eds.), *International Tables for Crystallography G Definition and exchange of crystallographic data*,. Springer, Dordrecht, The Netherlands, pp. 295–443.

22. Westbrook, J., Yang, H., Feng, Z., Berman, H.M., (2005). 5.5 The use of mmCIF architecture for PDB data management. In: Hall, S.R., McMahon, B. (Eds.), *International Tables for Crystallography*,. Springer, Dordrecht, The Netherlands, pp. 539–543.
23. wwPDB, PDBx/mmCIF Resource Site, 2017.
24. Bourne, P.E., Berman, H.M., McMahon, B., Watenpaugh, K.D., Westbrook, J.D., Fitzgerald, P.M., (1997). Macromolecular crystallographic information file. *Methods Enzymol.*, **277**, 571–590.
25. Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J. L., (2013). The future of the protein data bank. *Biopolymers*, **99**, 218–222.
26. Adams, P.D., Afonine, P.V., Baskaran, K., Berman, H.M., Berrisford, J., Bricogne, G., et al., (2019). Announcing mandatory submission of PDBx/mmCIF format files for crystallographic depositions to the Protein Data Bank (PDB). *Acta Crystallogr. Sect. D, Struct. Biol.*, **75**, 451–454.
27. Westbrook, J.D., Shao, C., Feng, Z., Zhuravleva, M., Velankar, S., Young, J., (2015). The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics*, **31**, 1274–1278.
28. Dutta, S., Dimitropoulos, D., Feng, Z., Persikova, I., Sen, S., Shao, C., et al., (2014). Improving the representation of peptide-like inhibitor and antibiotic molecules in the Protein Data Bank. *Biopolymers*, **101**, 659–668.
29. Westbrook, J., Bourne, P.E., (2000). STAR/mmCIF: An extensive ontology for macromolecular structure and beyond. *Bioinformatics*, **16**, 159–168.
30. Westbrook, J., Fitzgerald, P.M., (2003). The PDB format, mmCIF formats and other data formats. In: Bourne, P.E., Weissig, H. (Eds.), *Structural Bioinformatics*,. John Wiley & Sons, Inc., Hoboken, NJ, pp. 161–179.
31. Neelamegham, S., Aoki-Kinoshita, K., Bolton, E., Frank, M., Lisacek, F., Lutteke, T., et al., (2019). Updates to the symbol nomenclature for glycans guidelines. *Glycobiology*, **29**, 620–624.
32. Sehnal, D., Rose, A., Koca, J., Burley, S., Velankar, S., (2018). Mol*: Towards a common library and tools for web molecular graphics. *MolVa: Workshop on Molecular Graphics and Visual Analysis of Molecular Data 2018*,. https://doi.org/10.2312/molva.20181103.
33. Sehnal, D., Bittrich, S., Velankar, S., Koča, J., Svobodová, R., Burley, S.K., et al., (2020). BinaryCIF and CIFTools – Lightweight efficient and extensible macromolecular data management. *PLOS Comput. Biol.*,. https://doi.org/10.1371/journal.pcbi.1008247 (in press).
34. Steinegger, M., Soding, J., (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
35. Steinegger, M., Soding, J., (2018). Clustering huge protein sequence sets in linear time. *Nat. Commun.*, **9**, 2542.
36. Sigrist, C.J., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., et al., (2013). New and continuing developments at PROSITE. *Nucleic Acids Res.*, **41**, D344–D347.
37. Guzenko, D., Burley, S.K., Duarte, J.M., (2020). Real time structural search of the Protein Data Bank. *PLoS Comput. Biol.*, **16**, e1007970.
38. Weininger, D., (1988). SMILES 1. Introduction and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
39. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., Pletnev, I., (2013). InChI – The worldwide chemical structure identifier standard. *J Cheminform.*, **5**, 7.
40. Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G., et al., (2020). RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering, and energy sciences. *Nucleic Acid Res*,. https://doi.org/10.1093/nar/gkaa1038 (in press).
41. Pollack, A., (2003). Company Says It Mapped Part of SARS Virus. The N.Y. Times, p. C2.
42. Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., et al., (2020). Structure of M(pro) from SARS-CoV-2 and discovery of its inhibitors. *Nature*, **582**, 289–293.
43. Rathnayake, A.D., Zheng, J., Kim, Y., Perera, K.D., Mackin, S., Meyerholz, D.K., et al., (2020). 3C-like protease inhibitors block coronavirus replication in vitro and improve survival in MERS-CoV-infected mice. *Sci. Transl. Med.*, **12**, eabc5332.