PSDI
PHYSICAL SCIENCES DATA INFRASTRUCTURE

Case Study 8
Summary Report

The Role of Structure in Physical Sciences Data Management

Ian Bruno, Suzanna Ward

Cambridge Crystallographic Data Centre, Cambridge, CB2 1EZ

Area of Physical Sciences covered

At the core of this case study are the Chemical Sciences where the representation of molecular structure is critical to the design, execution and communication of research. It extends to the Materials Sciences where representation of atomic structure is key to the understanding of properties of materials. It applies to any experimental or computational study where a chemical substance or material is being studied or modelled and as such is relevant to a vast majority of research areas in the Physical Sciences.

Related research areas

Aspects of this case study will be relevant to all other case studies, either because of the need to represent molecular and atomic structure, or because of the more general data management considerations that the study highlights.

This study further links to domains outside of the Physical Sciences including Biological Sciences and Earth Sciences where an understanding of molecular structure, chemical composition and material properties is essential.

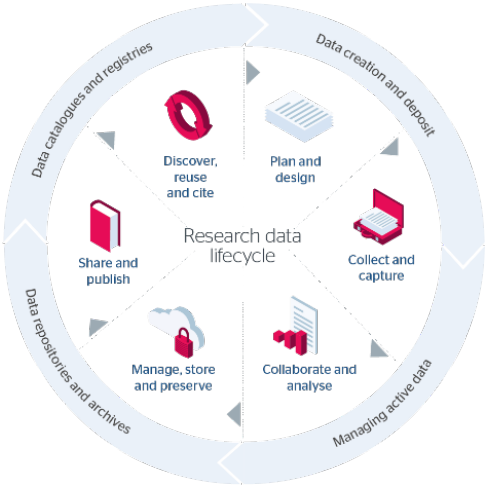
Applicability to the Research Data Lifecycle

Plan and design: Retrieving information to inform experimental design. Designing inputs into experimental and computational studies.

Collect and capture: Capturing reaction schemes and outcomes. Modelling 3D structure. Mining data resources and building predictive models.

Collaborate and analyse: Sharing results with collaborators. Refinement and validation of models.

Manage, store and preserve: Associating chemical identity, molecular structure and atomic composition with data, samples, and models.



Share and publish: Reliably communicating structures and related data in forms understandable by researchers and machines. Linking data across domains and resources.

Discover, reuse, and cite: Reliably retrieving data and models associated with specific substances to provide a foundation for future research.

Key focus and activity

A review of the importance and challenge of reliable structure representation, the need for enabling workflows and infrastructure, the importance of having curated structure-based data resources, and the opportunities for global collaboration with industry and academia.

Main outputs

The main output of this case study is a report that covers the following areas:

- Structure in Context – importance of structure to the Research Data Lifecycle
- Structure Representation – state-of-the-art and current challenges
- Workflows and Infrastructure – considerations for reliably capturing structure
- Publication and Curation – need for and role of data repositories, importance of trust
- Related Initiatives – opportunities for global engagement

Throughout, the report highlights considerations that should be factored into the design and implementation of a UK Physical Sciences Data Infrastructure at both a technical and a social level.

Outcomes and Recommendations

Embrace standards for structure representation

Outcome

Consistent and reliable representation of structure is a critical enabler of data reuse and discovery throughout the research data lifecycle, across the physical sciences and beyond. Structure representation should thus be a core consideration in the design of a Physical Sciences Data Infrastructure.

Recommendations

Current research practices tend towards structure representations that are not conducive to interpretation by machines and use in data-driven science. A Physical Sciences Data Infrastructure should enable and encourage researchers to provide representations of structure that can be readily interpreted by machines.

A plurality of structure representations should be supported and encouraged. Traditional representations of structure (diagrams, names) are helpful but should ideally be accompanied by more machine-readable representations.

Existing machine-readable representations work well for organic structures but even then, may be prone to ambiguity of interpretation. Consideration should be given to the technical and scientific limitations of current structure representation methodologies.

Structure-based classifications that enable semantic interoperability and faceted search should be supported. Identifiers such as registry and database accession IDs are also important for enabling links between structural data resources.

Wherever possible, a standard International Chemical Identifier (InChI) should be generated and stored for all structures as a key enabler of findability and interoperability. This does not necessarily mean requiring researchers to provide InChIs but does require a representation from which an InChI can be reliably generated.

IUPAC guidelines for the depiction of 2D chemical diagrams should be followed where appropriate.

Consideration should be given to the representation and storage of multi-component systems including reactions and mixtures and not just individual molecular structures.

Accommodation must be made for the handling of 3D structures that have been determined computationally or experimentally. The provenance of a 3D structure – whether experimental or computational – and the methods used to generate it should be clearly indicated. For aspects of the infrastructure that involve human interaction, the ability to easily visualise key features of 3D structure should be provided.

The software packages involved in the generation of structure representations and models should be captured in line with community recommendations for software citation.

There is significant intersection between the physical sciences and biological sciences, particularly when it comes to understanding biological structure and mechanisms. A physical sciences data infrastructure should consider storage of biological macromolecular structures and/or links to biological resources based on structure. Partnership with relevant bioinformatics organisations is advised.

Recognise that change takes time and that a future infrastructure may have to support legacy representation formats for longer than might be desirable. Be prepared to invest in tools and education that will enable communities to embrace change without fear of disruption.

Insofar as possible, separate out services, tools and workflows from underlying formats to enable these to evolve independently. Provide opportunity for ongoing investment and experimentation with new paradigms and approaches for structure representation.

Support the discovery and interoperability of structures and associated data

Outcome

A Physical Sciences Data Infrastructure should facilitate the future reuse of structural data by providing technical enablers that support the discovery and interoperability of relevant data and metadata across resources.

Recommendations

Recognise the digital representation of the structure of a substance studied or a material modelled as an essential piece of metadata that should be stored and tracked alongside physical sciences datasets.

Contribute to the development of a community chemical structure validation service to support and encourage best practice and enable assessment of the reliability of a digital structure representation. Help to develop benchmarking reference sets that can be used to judge compliance of tools with structure representation standards.

If sophisticated storage, search and analysis and management of structural data is required as part of the infrastructure, partner with organisation(s) who have developed solutions to satisfy these needs.

Encourage and enable the registration of metadata in open registries, including appropriate structure identifiers and links to related objects in order to contribute to wider networks of open science knowledge graphs.

Connect structure representation to sample identification, taking advantage of standard identifiers for samples and structures.

Provide services that enable the retrieval of data from across resources based on structure identity, composition, and connectivity, partnering with existing solution-providers where possible. In particular expose interfaces that enable the lookup and linking of data based on standard identifiers such as InChI.

Support the ability to faithfully exchange individual datasets and aggregated subsets of structure-based data and metadata between systems within and external to a future infrastructure. Enable citation of datasets and aggregated subsets to support reproducibility, provenance and credit.

Advocate for and enable access to curated and trusted structural data

Outcome

A Physical Sciences Data Infrastructure should promote the importance of high-quality curated data, enable access to existing sources of curated structural data in support of UK academic research, and cultivate expertise and criteria that can encourage the increased availability of FAIR and trusted structural data.

Recommendations

Consideration should be given to how a future infrastructure can provide access to high quality curated structural data that is available in third party resources as well as within the infrastructure itself.

Building on the tradition of EPSRC support for access to highly curated data resources through the PSDS and its predecessors, explore how this can be extended to incorporate increased access to richly curated sources of structural data and properties, crucially ensuring access via machine APIs.

Identify opportunities to invest in making data in public resources more valuable and available in more structured forms through richer APIs. Also consider APIs that enable federated access to data across public and proprietary resources based on common languages of structure representation.

Champion the importance of and requirements for high quality data publishing and curation in the physical sciences. Adopt existing curation guidelines where available – work with wider communities to develop new ones where these are not available.

Promote the importance of machine interpretable data formats to support efficient publication workflows and enable data reuse – advocate against practices that result in non-semantic publication of data that cannot be reliably interpreted by machines.

Consider how a future infrastructure can provide the motivation and means for publication of structural data associated with doctoral theses and dissertations in machine-accessible and reusable forms – not just to ensure data are available for future research, but also to train the next generation of researchers in best practices for data management and publication.

Identify and highlight gaps in domain-specific data storage and curation and cultivate communities to address these. Consider providing a publication platform for physical sciences data that does not have a natural domain-specific home.

Become a centre of expertise and best practice in research data management and curation in the physical sciences alongside provision of any technical infrastructure.

Recognise that investment in data infrastructure requires investment in expertise as well as technology. Invest for the long term, not the short term and ideally to enable all data and services to be openly available.

Learn from the experiences of existing data repositories when considering benefits of investment in data storage and curation. Consider commissioning case studies that demonstrate the return on investment of established physical sciences data resources for the wider economy. Consider where there may be opportunity for national and international cooperation to pool resources and minimise costs.

Recognise that if data storage and curation activities have to be self-sustaining some restrictions or barriers are likely. Adopt the Principles of Open Science Infrastructure to guide the governance and sustainability of data infrastructure.

Identify criteria for characterising a physical sciences data repository as trusted and a dataset stored within that as FAIR from a domain perspective. Base this on existing frameworks for establishing trustworthiness and assessing FAIR maturity.

Invest in and support change in partnership with global communities

Outcome

A Physical Sciences Data Infrastructure should partner with and invest in international initiatives that aim to develop the standards, infrastructure and guidelines needed to advance the management of structural data specifically and research data generally.

Recommendations

Align with other organisations and initiatives looking to address data representation, publication and management challenges relevant to the physical sciences. Be willing to contribute time to efforts aimed at the development of shared infrastructure and standards.

Cultivate partnerships with industry to inform priorities and identify funding opportunities for development of a physical sciences data infrastructure that will deliver value for industrial and research sectors in the UK.