1) Policy purpose

Given its role conserving, curating and disseminating data since 1965, the Cambridge Crystallographic Data Centre (CCDC) endeavours to maintain its high standards of data preservation and management. In this regard, the objective of this preservation policy is to outline the standards and procedures which the CCDC aims to uphold in order to guarantee the long-term preservation of data deposited and stored at the Centre.

2) Repository purpose

The Cambridge Crystallographic Data Centre exists to support the advancement of structural chemistry worldwide through the development of the Cambridge Structural Database (CSD), and related software. This objective is underpinned by CCDC's dedication to the promotion of chemistry and crystallography for public benefit by providing high quality information services and resources to be used for research, teaching and learning.

The CCDC considers the chemistry and crystallography research community as the principal benefactors of its services. This includes researchers associated with academic institutions worldwide, as well as, pharmaceutical and chemical companies from the commercial sector. For the research activities of these communities, the organisation represents an important resource as one of the main repositories for structural data, where researchers can deposit and access data sets. The CCDC's core activities include the preservation, curation and dissemination of crystallographic data with the aim of guaranteeing that all data entrusted to it by depositors remain suitable for the needs of its primary users now and in the future. In this capacity, the Centre actively works to promote and facilitate:

- Secure storage of data
- Reliability and usability of data
- Quality and integrity of data
- Publication of and access to data

3) Policy scope

This document describes the preservation processes for all X-ray crystallographic data deposited and stored at the CCDC which falls within the Centre's criteria for accepted data types and formats. These include experimental crystal structure determinations and supporting data in the following file types: CIF, FCF, HKL and checkCIF reports. This document also covers the preservation of any accompanying metadata associated with deposited data, which the CCDC aims to preserve, keep up to date and enrich where necessary.

4) Preservation standards and ethics

The CCDC actively seeks compliance with community driven initiatives for best practice in data preservation. In this regard, The CCDC data preservation policy is broadly guided by the Open Archival Information Systems (OAIS) reference model (DPC and Brian Lavoie, 2014), which provides standards and frameworks of concepts for digital repositories and archives. The Cambridge Crystallographic Data Centre is also a member of the World Data System (WDS) and as such has been certified as being compliant with international standards for digital repositories.

As a proponent of FAIR data policies and initiatives, the CCDC aims to function in line with European Commission recommendations on FAIR data by making data as open a possible within the sustainability model of the Centre (EU Commission, 2018: 21). In this respect, the CCDC endeavours to adhere as closely as possible to FAIR guiding principles (Wilkinson, M. D.et al., 2016) and provide findable, accessible, interoperable and reusable datasets to its user community.

The CCDC participates in global data initiatives such as the Research Data Alliance (RDA) and the ICS World Data System, as well as, domain specific groups, such as the International Union of Pure and Applied Chemistry (IUPAC), InChI Trust and various Crystallography Unions and Associations. By adopting outputs from these initiatives, the CCDC seeks to conform to, as well, help promote community best practices and standards for research data. Likewise, as an associate corporate member of the Committee on Publication Ethics (COPE), CCDC staff receive guidance and provide recommendations on publication ethics and data integrity issues.

The CCDC is unreservedly committed to keeping the personal information granted to the CCDC by data producers and users private and confidential. This commitment is covered in the Centre's privacy policy and cookie policy, which have been developed to comply with the General Data Protection Regulation (GDPR). To ensure compliance with these policies and regulations, all existing and new CCDC staff receive training in the GDPR policy and practices. Any issues relating to data or confidentiality breaches can be reported to the CCDC by users at gdpr@ccdc.cam.ac.uk. The CCDC's processes for dealing with data breaches are governed by its Data Breach policy. All issues raised will be dealt with by the CCDC GDPR committee who meet regularly to review the organisation's data policies and actions.

The CCDC assumes the indefinite retention of all information resources deposited at the Centre, and with 5 decades of experience preserving data has proven its ability to do so. This meets the criteria of the UK Research and Innovation principals for Open Research Data which advocates the retention of a research data resource for up to 10 years after its last used. Usage in the case of CCDC data resources relates to all direct or indirect searches and downloads of data through CCDC online and desktop services and software, as well as, references and citations of CCDC data and publications containing underlying CCDC data. Given the sustained usage of CCDC data by its designated community via these mediums, the CCDC therefore has in place systems to ensure that data granted to it by data depositors can be preserved indefinitely.

5) Data ingest

When depositing data via the CCDC data deposition service, depositors are made aware that once published their data will be curated into the CCDC's Cambridge Structural Database if it meets the criteria for inclusion. On submitting data, users are required to confirm CCDC's deposition terms and conditions which request that depositors provide correct metadata for their deposited files and accept a publication embargo period of one year for their data, after which the CCDC have the right to publish the deposited data as a CSD Communication if the data remains unpublished.

Around 90% of data is deposited at the CCDC through the web-based deposition Service, CCDC Deposit. This page provides information on the file types and data formats which the Centre is able to process and preserve and links to areas of the CCDC website with more detailed support and guidelines. These pages detail the expected scientific content of deposited data and are presented in language tailored to CCDC's designated crystallographer and chemist user community. Since the CCDC receives up to a third of its deposits from China, Chinese translations of key contents are also available on the website to further aid comprehensibility. Prior to deposition, depositors would be expected to read the available information to confirm that their data meets the Centre's remit for preservation.

The CCDC web-based deposition service has been designed to enhance data processing efficiency from deposition through to publication by giving depositors the chance to fix errors and add additional scientific or publication metadata at the point of deposition. This service contains eight steps for the depositor to complete before the final submission of files: Login, Upload, Check Syntax, Validation, Add Publication, Enhance Data, Review, Submit.

Using this deposition service, various checks are run on the data allowing the depositor to validate and enhance their deposited files:

- **Processed data (or structure factors) check** - If these data are not included in the upload, the user is alerted and can provide an explanation for their omission. Depositors can also provide a link to raw data if this is archived at another repository.

- **IUCr checkCIF** - Depositors can run the International Union of Crystallography (IUCr) checkCIF service on their data to produce a report on the completeness and crystallographic integrity of the scientific data. These reports are archived alongside the data files and depositor responses to the errors which are reported by these checks are automatically embedded into the data file. The reports and responses are then made available for download to the public and journal reviewers/editors, as part of the wider peer review process.

- **Unit Cell Check** - Reduced cell checking software is run on the deposited structures and compared with existing data in the CSD. The function of this mechanism is to help depositors identify whether their new crystal sample has been published before and if they have accidentally crystallised unexpected materials. Thanks to this check, depositors may avoid wasting valuable research time and resources on unintentional collection of a full dataset for a known crystal structure.

In cases where depositors are unable to use the online deposition service, alternative methods for depositing data are available. Primarily, this can be done by sending data by email to deposit@ccdc.cam.ac.uk, or for data which exceed the limits of our systems, via third party file sharing systems. These alternative methods of deposition have been put in place to ensure that the Centre is flexible in the file types it ingests and processes. Furthermore, this system is adapted to users depositing from regions where they may have difficulties accessing and using web-based services.

Following the submission of data, further automatic validation processes take place to allow for the conversion of the data into a format which can be processed by CCDC systems. A duplicate check is run on the deposited data to establish whether this has already been added to the CCDC archive. Deposits which fail the validation process or are found to be duplicates and have not been identified as revised datasets by the depositor are then put in queue to be processed manually by CCDC staff. Datasets which pass all checks are assigned a unique Deposition Number, which is communicated by email to the depositor immediately after processing.

6) Archival storage

As a repository for the global scientific community, the CCDC seeks to ensure the indefinite retention of all information resources stored at the Centre. As such, the CCDC's archive system aims to achieve secure long-term storage of data by conducting a number of data preservation and security activities. These procedures are covered by an IT facilities policy which is reviewed annually.

Storage and data backup

To ensure safe archival and prevent loss of data, the CCDC stores data files on storage platforms with many layers of redundancy, where all data is encrypted both in transit and at rest. In parallel to this, continuous local backup copies of the data files are taken. Onsite backups of important data are taken daily and full offsite backups of the CCDC infrastructure, including all virtual machines, are taken multiple times per week. All backup data, including copies of data files representing published data files in CIF format and database entries in XML, are archived monthly to tape and stored securely offsite in a temperature-controlled environment. Additionally, source code is archived on a weekly basis to a cloud-based provider. All offsite files and backups are encrypted.

Data Monitoring

The CCDC data management systems are built on industry standard Microsoft technologies and track automated and manual events that may result in changes to the data stored. Automated checks are also periodically run across the complete database to identify any data fields that may contain errors.

Disaster recovery policy

The information and resources necessary to recreate services in the event of a major disaster are stored securely off-site so that these remain accessible.

7) Data management

To maintain the integrity of the information resources deposited at the CCDC, data management activities are performed on a daily basis. Descriptive data, such as unique persistent identifiers (Accession Numbers, CCDC DOIs) assigned to data, are preserved and checked. Similarly, bibliographic details deposited with datasets are preserved, but may be modified by CCDC staff to reflect CCDC formatting rules and published citations, or when requested by the data producer.

The CCDC promotes the involvement of its designated user community in the management of their data through its web-based My Structures service. This service's functionalities include viewing, editing and publishing data, as well as, the ability to share data amongst colleagues prior to publication. Through this service, depositors can also extend the embargo date for their data beyond one year to ensure that their data remains unpublished.

In line with the OAIS archive model, the CCDC has systems in place for producing and monitoring "performance data" and "access statistics" (DPC and Lavoie, 2014:12). These systems produce documentation regarding actions such as:

- Modifications made to data or metadata
- Changes to data status
- External access to data

These logs and documentation allow CCDC staff to run reports and queries on the performance and status of all deposited information resources on a daily basis.

The CCDC's deposit mechanism also plays an important role in the Centre's data management function by performing the duplicate check on all information files it receives, as described in Section 5 of this policy. When a depositor sends a revised version of data which they or a collaborator have previously deposited at the CCDC, the duplicate check system analyses the scientific contents of the file and ensures that the data is not assigned a new persistent identifier. Once the system identifies the duplicate data, the mechanism interrupts processing before a Deposition Number is assigned. A member of CCDC staff then manually compares the files to confirm whether the file is a revision of existing data or should be assigned a new number. When a revised version of dataset is identified, the Deposition Number assigned to the data already in the archive will be assigned to the new dataset. The old dataset will be archived and deactivated. In this way, any modifications or revisions to datasets are logged and tracked within the system and different versions of data remain associated.

Within this system, all copies of data which have been deposited are preserved, along with their descriptive data. Likewise, in most circumstances the CCDC will not delete data from the data archive but deactivate the documentation. Data stored in this deactivated state cannot be further processed or accessed externally.

8) Preservation planning

To safeguard against external impacts which could affect the security of data stored at the CCDC, the Centre has in place various initiatives to ensure the organisation's sustainability and ability to preserve deposited information resources now and in the future.

The CCDC's sustainability model serves to guarantee the long-term functioning of the organization. Overall, the revenue generated from financial contributions received for value-added products meets the costs of fulfilling the free data preservation and access activities. As a UK registered charity, the CCDC is not permitted to make any profit and any surplus is reinvested into the activities of the organisation. Furthermore, over the years the CCDC has built up financial reserves that provide flexibility to respond to changes in the wider environment.

In conjunction with the Board of Trustees, the CCDC undertakes strategic reviews every five years. This process includes extensive consultation with staff and user groups, as well as, analysis of industry trends, risks/threats and opportunities. Frequent strategic objective updates are also provided to the Board of Trustees giving the organisation considerable agility in the face of changing technological and scientific trends. In addition to strategic reviews, the CCDC has assembled a scientific advisory board which will provide up-to-date insight into scientific trends, as well as, the developing requirements of our designated user community. Finally, should the CCDC be unable to continue its operations, a Safeguarding and Continuity Fund has been established to ensure that there will be financial resources remaining for the transfer of data and other assets to the stewardship of an appropriate organisation.

9) Access

The CCDC takes its role as a data steward very seriously and takes every precaution to ensure that data remains private until we are made aware by depositors or published literature that data can be made public.

Upon publication of data, organic and metal-organic experimental structures will be curated into the Cambridge Structural Database and inorganic experimental structures will be curated into the Inorganic Crystal Structure Database.

Once made public, the deposited data can be searched for using the persistent identifiers or publication details associated with the data via CCDC's Access Structures service. Data Consumers can then view, download and interact with the data directly in the web browser free of charge. The access and use of data available through this service is governed by our Access Structures Terms and Conditions legal and regulatory framework.

To facilitate the publication and preparation of data stored in the repository, data can also be made available pre-publication to publishers, referees and depositors. When data is requested pre-publication, checks are performed on requestors to confirm their role and identity. Once their identity is confirmed, they are invited to use the CCDC Referee Service which allows data to be viewed and downloaded before being made public. Similarly, the My Structures service and its related functionalities serve to assist data publication workflows by allowing depositors to view and share their own data pre-publication.

10) Administration and staff responsibilities

The administration of the database is undertaken by various members of CCDC staff who are responsible for maintaining different elements of the database and its functioning. These administrative roles for data preservation and management are summarized in the table below and are classified based on the administrative responsibilities for archival storage as outlined in the OASIS model (DPC and Lavoie, 2014: 13).

Role	Responsibility
- Deposition Coordinators - Scientific Editors	Interacting with data producers and consumers Providing customer service support
- Head of Database - Editorial Team Leader - Head of Strategic Partnerships	Monitoring compliance with data policies and standards Reviewing data preservation policies
- Systems Team Leader - Systems Administrators - Systems Development Team Members	Monitoring system performance, and coordinating updates to the system

To guarantee that preservation responsibilities are fulfilled and performed to the highest level, new and existing CCDC staff in the roles outlined above receive internal and external training in data protection knowledge and CCDC preservation activities.

11) Preservation policy review

This preservation policy is to be reviewed annually to allow for modifications when necessary in light of technological or external developments. The decision to update the policy is the responsibility of the CCDC managerial staff.

This document will be shared with all CCDC staff and published on the CCDC website, available for all data producers and consumers to view.

Questions relating to the preservation policy can be directed to the Cambridge Crystallographic Data Centre's Database Group at: deposit@ccdc.cam.ac.uk

12) Definition of terms

Data - Within the OAIS framework, data represents all information resources deposited and archived at the Centre. This policy covers primarily the information resources deposited in CIF, FCF, HKL and checkCIF report file formats, for which CCDC has the operating systems and expertise for preserving.

Data Consumers – Data consumers are considered as users of the data stored at the CCDC. This includes all individuals who interact, view or download data via CCDC online services.

Data Depositor - Using OAIS terminology, depositors can be considered as data “Producers” or “individuals, organizations, or systems that transfer information [..] for long-term preservation” to the Centre. These are not always the creators of the data (DPC and Brian Lavoie, 2014: 6).

Designated user community – This is the main user group of data stored at the CCDC, towards whom the primary CCDC data depositing, preservation and access services are targeted and aim to benefit.

Metadata – This is understood as the information which describes the information resources deposited at the CCDC.

Bibliography:

Digital Preservation Coalition and Brian Lavoie (2014) *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)*, Digital Preservation Coalition**.** DOI: 10.7207/twr14-02

EU Commission Expert Group on FAIR Data (2018) *Turning FAIR into reality*
Final Report and Action Plan from the European Commission Expert Group on FAIR Data, EU Commission. DOI: 10.2777/1524

Wilkinson, M. D.et al. (2016) *The FAIR Guiding Principles for scientific data management and stewardship*, Sci. Data 3:160018, DOI: 10.1038/sdata.2016.18