Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	Cambridge Crystallographic Data Centre
Website:	https://www.ccdc.cam.ac.uk/
Certification Date:	31 January 2020
This repository is owned by:	Cambridge Crystallographic Data Centre

Cambridge Crystallographic Data Centre

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of Repository

The Cambridge Crystallographic Data Centre (CCDC) (https://www.ccdc.cam.ac.uk/) compiles and distributes the Cambridge Structural Database (CSD) (https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/), a repository for small molecule organic and metalorganic crystal structure data. The Centre also produces associated knowledge-based application software for the global community of structural chemists.

Originating in the Department of Chemistry at the University of Cambridge in 1965, the CCDC has been an independent institution since 1987, constituted as a non-profit company and a registered charity, with its operations overseen by an international board of trustees.

The CCDC supports structural chemistry worldwide by preserving crystal structure data and by curating and disseminating the CSD, which now contains over 1 million entries. The CCDC also retains close links with the University of Cambridge and is a University Partner Institution that is qualified to train postgraduate students for higher degrees (PhD, MPhil).

In July 2018, The CCDC began a collaboration with FIZ Karlsruhe – Leibniz Institute for Information Infrastructure (FIZ Karlsruhe) to allow the inorganic crystal structures stored by FIZ to be deposited and accessed through the CCDC data deposition and access services. This collaboration was developed to enable researchers to share data through a single deposition portal and provide worldwide access to all chemical structures for free.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository's Designated Community.

Members of the crystallographic and structural chemistry communities are the principal benefactors of CCDC's services. This includes researchers associated with academic institutions worldwide, as well as, various pharmaceutical and

chemical companies from the private sector who employ CCDC services.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

B. Basic curation – e.g. brief checking; addition of basic metadata or documentation, C. Enhanced curation – e.g. conversion to new formats; enhancement of documentation, D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

Data deposited at the CCDC receives various levels of curation.

By depositing data through the Deposit Structures service (https://www.ccdc.cam.ac.uk/deposit), a number of checks are run on the data, allowing the depositor to validate and enhance their deposited files (https://www.ccdc.cam.ac.uk/Community/depositastructure/correctingcifs/). Including checks for:
- Syntax errors
- Inclusion of processed data
- Scientific integrity
- Comparison with existing published data

Reports from the scientific integrity check are archived alongside the deposited data files and depositor responses to the errors which are reported by these checks are automatically embedded into the data file. Upon submission of data, all validated and enhanced data files can be downloaded by the depositor, along with an automatically generated chemical diagram file for their structure(s).

Following the deposition of data, further syntax checks are performed, and errors are fixed manually by CCDC staff. Each

individual structure which is deposited is assigned a persistent identifier, known as a CCDC Number.

Once data is published in a journal, the data will be updated with its associated metadata based on the publication literature and the dataset is made publicly available. An additional database identifier and a CCDC citation including a Digital Object Identifier (DOI) are also assigned to the data.

Datasets are curated into the CSD by the CCDC editorial team who review the scientific accuracy of the data and may add additional data items.

Each individual deposited dataset is made available to the public through the free Access Structures service (https://www.ccdc.cam.ac.uk/structures/) and through the CSD. Additional curated content is also made available for free alongside the underlying dataset to aide data discoverability and the interpretation of the data. The full curated content is available through the advanced CSD-System.

For inorganic structural data deposited at the CCDC as part of the CCDC/FIZ collaboration, additional metadata and data identifiers are added to the deposited data by the CCDC. Further curation is then performed by FIZ Karlsruhe who have the expertise required for curating the crystallographic data of inorganic compounds.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

FIZ Karlsruhe perform the scientific curation of the inorganic structural data deposited at the CCDC.

DataCite (https://datacite.org/) mints DOIs for CCDC datasets, therefore, supporting the curation and accessibility of datasets.

Microsoft Azure storage services are employed for backup storage of data.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

The entire database and associated software services are delivered to around 1,400 research sites worldwide, including academic institutions in 80 countries and all the world's top pharmaceutical and chemical companies.

In terms of usage of CCDC's free online data deposition and access services, in 2019 we received more than 50,000 data deposits from around 12,000 depositors and 20 million free searches of the database were performed.

The CSD (Acta Cryst. (2016). B72, 171-179, https://doi.org/10.1107/S2052520616003954) has been cited over 7000 times (http://scripts.iucr.org/cgi-bin/citedin?bm5086).

As a research institute, the CCDC has produced more than 800 peer-reviewed publications (https://www.ccdc.cam.ac.uk/researchandconsultancy/ccdcresearch/ccdcpublications/).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

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.

To fulfil this mission, 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
- Publication of and access to data

These operations are overseen and signed off by a board of Trustees (https://www.ccdc.cam.ac.uk/theccdcprofile/trustees/) who are representative of the various stakeholder communities which the CCDC serves.

The CCDC's main mission of compiling and distributing data via the CSD is also outlined on the CCDC profile page (https://www.ccdc.cam.ac.uk/theccdcprofile/) and preservation policy (https://www.ccdc.cam.ac.uk/Community/depositastructure/scientific-data-preservation/ccdc-preservation-policy.pdf).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All deposited datasets are made available free of charge to anyone as soon as it is known that these can be made public. Access is via a structure summary page (e.g. http://dx.doi.org/10.5517/cc1jj92f) which allows researchers to view and interact with the data directly in a web browser. Datasets can be downloaded by the researcher and used for any purpose. Access and use of data is governed by the CCDC website Terms and Conditions (https://www.ccdc.cam.ac.uk/access-structures-terms/). To ensure users comply with the conditions of access and use, the CCDC has in place systems for monitoring data usage.

The CCDC also provides also software applications and services that enable systematic search and analysis across all data and application of derived knowledge to a range of scientific areas. We request a financial contribution for access to these products. The data sharing policy for these products is defined by a separate license agreement covering their use; a copy of this agreement can be provided on request. In summary, the license allows unrestricted use of products (including data) within an organisation for publishable or proprietary work, and in any collaboration that does not involve payment of fees. External distribution of systems derived from CCDC products is possible under a separate signed agreement.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC has been a self-sustaining not-for-profit organisation since at least 1987 when it was established as a UK registered charity. It has successfully demonstrated an ability to fund its activities through the provision of value-added software and services without reliance on direct grants from funding agencies.

To guarantee this continued sustainability, strategic reviews are undertaken every five years, in conjunction with the CCDC's Board of Trustees. This process includes extensive consultation with staff and user groups and analysis of industry trends, risks/threats and opportunities. From these reviews come validated or amended vision and purpose statements, core values, long term strategic goals and eventually, at a staff level for each year, organisational, team and individual objectives. Each person at the Centre therefore remains aligned with the long-term vision of the CCDC.

In the short to mid-term, this strategic plan is reviewed by the Board of Trustees each six months. Organisational requirements are reviewed every year in conjunction with setting an annual budget. Progress towards goals is reviewed throughout the year and priorities adjusted, if necessary, to satisfy the current needs of our user communities.

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.

These procedures for ensuring continuity of access to data are publicly available in section 8 of the CCDC preservation policy.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC takes its role as a data steward very seriously and takes every precaution to ensure that all data and personal information granted to the Centre by data producers and users is safely protected and managed.

On submitting data files, depositors are required to confirm CCDC's deposition terms and conditions. These terms and conditions 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 (https://www.ccdc.cam.ac.uk/Community/csd-communications/) if the data remains unpublished by the author. Although the CCDC is transferred the right to publish data after one year, we currently only publish data in this way after requesting additional consent from the data producer. Depositors can also extend the embargo date for their data beyond one year to further ensure that their data remains unpublished. This can be done on request to CCDC staff or via the My Structures (https://www.ccdc.cam.ac.uk/support-and-resources/support/case/?caseid=a567fad5-20b7-e611-837e-00505686f06e) service. Once datasets are deposited at the CCDC, these will therefore remain privately stored until we are made aware by depositors or the publication of an associated article that data can be made public.

To facilitate the publication and preparation of data stored in the repository, data may be made available pre-publication to publishers, referees and depositors. In these cases, checks are performed on requestors to confirm their role and identity. Subsequently, these stakeholders are invited to use the online Referee Service which allows data to be viewed and

downloaded before being made publicly available. Similarly, the My Structures service and its related functionalities serve to assist data publication workflows by allowing depositors to view and manage their own data pre-publication.

Regarding the personal information granted to the CCDC by data producers and users, the CCDC is committed to keeping this information private and confidential. This commitment is covered in the Centre's privacy policy (https://www.ccdc.cam.ac.uk/privacy/) and cookie policy (https://www.ccdc.cam.ac.uk/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.

The CCDC has developed systems and workflows which help to ensure that all data files stored in the repository comply with the organisation's ethical data values. For example, the CCDC has developed systems to help check for plagiarised data. This includes a duplicate check workflow and Unit Cell Check (https://www.ccdc.cam.ac.uk/News/List/post-12/) which compares deposited data against existing data. The CCDC is also currently conducting research into data fraud and plagiarism (https://www.ccdc.cam.ac.uk/Community/blog/in-crystallographic-data-we-trust/) and has become an associate corporate member of the Committee on Publication Ethics (COPE) (https://publicationethics.org/members/cambridge-crystallographic-data-centre), through which CCDC staff receive guidance and provide recommendations on publication ethics and data integrity issues. To ensure that depositor personal information is not inadvertently shared when data files are accessed by external users, the CCDC removes all non-scientific data/personal data from the information files when they are accessed or downloaded from the website. This guarantees that only the publication metadata and scientific data is made available to the public.

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 a Data Breach policy. All issues will be dealt with by the CCDC GDPR committee who meet regularly to review the organisation's data policies and actions.

Further information on CCDC's data preservation ethics and standards is publicly available in section 4 of the CCDC preservation policy.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively

carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC's sustainability model serves to guarantee the long-term functioning of the organisation. Overall, the revenue generated from financial contributions received for value-added products meets the costs of fulfilling the free data preservation and access activities. Furthermore, over the years the CCDC has built up financial reserves that provide flexibility to respond to changes in the wider environment. 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.

The CCDC employs around 70 people including scientific editors, software developers, applications scientists, technical support staff and finance and admin personnel. Many are based in Cambridge UK but we also have people located in the USA.

To guarantee that responsibilities are fulfilled and performed to the highest level, new and existing CCDC staff have opportunities year-round to receive internal and external training. To ensure this, part of the annual budget is allocated to staff training based on requirements identified through periodic personal development reviews.

To grow the expertise and knowledge within the organisation, CCDC staff also participate regularly in community and industry conferences. Attendance at these events is supported by annually budgeted funds.

Further information on CCDC's organisational infrastructure is publicly available in section 10 of the CCDC preservation policy.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The activities of the CCDC are overseen by a non-executive Board of Trustees (http://www.ccdc.cam.ac.uk/theccdcprofile/trustees/) who are selected so that they are representative of the CCDC's stakeholders and user communities. The Board of Trustees meet 4 times a year. Between meetings, they are provided with monthly updates and their advice is sought as need arises. Trustees serve a maximum term of 8 years. To gain further insights into community and industry needs, the CCDC has also assembled a scientific advisory board (https://www.ccdc.cam.ac.uk/News/List/ccdc-announces-formation-scientific-advisory-board/).

In order to stay abreast of developments within community, CCDC staff also attend and present at a broad range of scientific meetings and conferences across the globe (https://www.ccdc.cam.ac.uk/News/Events/). By attending and exhibiting at these events, the CCDC receives feedback on its tools and services from members of the community and learns about the current trends and preoccupations among its members. The CCDC website also has web-based forums (https://www.ccdc.cam.ac.uk/ideas/), through which members of the community can express which functionalities they wish to be added to CCDC tools.

Members of CCDC also acquire and share knowledge by participating in working groups for various global and community membership organisations (https://www.ccdc.cam.ac.uk/Community/collaborators/). These include crystallographic (IUCr,

ECA, BCA, BACG, ACA, USNCCr) and chemistry (IUPAC, ChIN, ACS, RSC) associations, and research data and scholarly communication groups (FORCE11, RDA, CODATA, WDS). A list of some of CCDC's collaborators and partners can be found on the website. By participating in these working groups and meetings, the CCDC gains insight into community needs, and technological and policy developments. The expertise and knowledge acquired from these engagement activities is then used to help develop the organisation's strategy.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Integrity:

Deposited datasets are identified by a unique accession ID, which is communicated to the depositor once data is processed. This ID is then used by CCDC staff and depositors for managing and tracking datasets.

To track changes to the stored data and metadata, the CCDC uses industry standard Microsoft technologies which log

automated and manual events. Automated checks are periodically run across the complete database to identify any data fields that may be in error. Should there be revisions to data files then different versions are separately stored.

The CCDC's data management system employs a duplicate check mechanism for comparing incoming deposited data with data already stored in its archive. When the mechanism identifies data as a duplicate, the CCDC has in place procedures for manually comparing data.

Detailed information on the processes for tracking changes to datasets and managing data is available to depositors in section 6 and 7 of the CCDC preservation policy.

Authenticity:

All primary deposited datasets are stored in Crystallographic Information Framework (CIF) files, a standard for information interchange in crystallography. This file type contains information fields for metadata, instrument details, software packages and parameters, and quality metrics, which are used by our user communities to check data integrity and authenticity.

To verify data depositor identity all depositors must provide a name and email address. Users are also prompted to register an account with CCDC when depositing data. Once data is ingested, it remains linked to the email address or registered account provided and only CCDC staff or the depositor (when signed in) can make changes to the data. As a requirement for depositing data, depositors must also confirm that they have provided correct metadata for their deposited files.

On publication, users accessing the data through Access Structures will be able to view the publication information under the "Associated publications" field (e.g https://doi.org/10.5517/ccdc.csd.cc1z1hgz). A data DOI is also assigned to the data using DataCite's DOI assignment service. The metadata for these DOIs contains the publication information so that a link is retained between the dataset and the article in which the data is published (e.g https://search.datacite.org/works/10.5517/ccdc.csd.cc1z1hgz).

Once data is published, the deposited data files are not changed except under unique circumstances. E.g A corrigendum to the article in which the data is published.

Links to repositories holding a dataset's raw data can also be added by the depositor at the point of deposition so that they become accessible to users. These links are shown under the "Raw Data DOI" heading in Access Structures (e.g https://www.ccdc.cam.ac.uk/structures/Search?Ccdcid=1586376&DatabaseToSearch=Published).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC requires that datasets are deposited in CIF format (https://www.iucr.org/resources/cif). CIF files are typically generated as a result of the experimental process, capturing the information needed to fulfil the chemistry and crystallographic community's reporting and reproducibility requirements. Additional file types for processed data and data integrity checks can be deposited alongside the CIF file as these are beneficial to our user communities.

The accepted file types and criteria for data can be found in information (https://www.ccdc.cam.ac.uk/Community/depositastructure/), support (https://www.ccdc.cam.ac.uk/support-and-resources/support/case/?caseid=45f14529-2002-4558-887b-a502fbb2f874) and guidelines (https://www.ccdc.cam.ac.uk/Community/depositastructure/cif-deposition-guidelines/) pages on the CCDC website. These criteria are updated in line with the changing community needs.

It is mandatory for depositors to provide alongside datasets a depositor name, email address and institution. This information is sufficient for verifying ownership of datasets and updating entries with the correct publication information. If clarification on data ownership is necessary, the depositor is contacted for additional information and the data may remain unpublished until data ownership is confirmed.

On submission of data, further automatic validation processes take place to check that the data is in the correct format

and to assign a reliability score based on curation status and the likelihood of complications in the automatic processing. Should the data fail the validation process, it will be manually checked by a member of the CCDC staff. If data is in the incorrect file format, the depositor will be contacted for further information or data.

The CCDC's data appraisal and validation processes are shown in the CCDC dataset workflow diagram (https://www.ccdc.cam.ac.uk/Community/depositastructure/scientific-data-preservation/ccdc_dataset_workflow.pdf).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC's data security and storage procedures are covered by an IT facilities policy which is reviewed annually in order to adapt to arising developments and risks. In conjunction with this, a disaster recovery plan exists that is reviewed annually.

Data files are stored on Microsoft Azure Datacentres which have many layers of local redundancy, providing at least 99.9% durability of objects over a given year. All data is encrypted both in transit and at rest and a continuous local

backup copy of data files are taken.

The remainder of the internal infrastructure is built around NetApp storage and VMware which has multiple levels of built in redundancy. Onsite backups of important data are taken daily and full offsite backups of the data infrastructure, including all virtual machines to a secure server, are taken 3 times per week.

All backup data are archived monthly to tape and stored securely offsite in a temperature-controlled environment. Source code is archived on a weekly basis to a cloud-based provider. All offsite files and backups are encrypted.

Information on CCDC's archival storage procedures are covered in section 6 of the CCDC preservation policy.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC's plan for the long-term preservation of datasets is outlined in section 8 of the CCDC preservation policy. This describes how the CCDC's sustainability model and regularly reviews of its strategy ensure that the data stored at the

Centre remains preserved and accessible to its user communities.

As a steward of data, the CCDC does not acquire ownership of datasets upon submission. Ownership of data deposited at the CCDC remains with the data producer/depositor but held in trust by the CCDC for long-term preservation. When data is ingested, depositors are made aware that once published their data will be curated into the CSD if it meets the criteria for inclusion. Users are also required to accept a publication embargo period of one year for their data, after which the CCDC are granted the right to publish the deposited data as a CSD Communication if the data remains unpublished.

Sections 5-7 of the preservation policy communicate to the user community the actions which the CCDC undertake for the long-term management and stewardship of the data stored at the repository. Section 4 outlines the archival information standards which the CCDC adheres to.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC data deposition service encourages depositors to fix errors and add scientific information and metadata which is relevant to the community when depositing their data. Firstly, checks on the syntax of the deposited data files take place and errors are reported to the depositor to fix. Any files that are deposited with broken syntax are then fixed by CCDC staff. The CCDC also strongly encourages researchers to deposit processed data in addition to derived data to aid with validation and reproducibility. When processed data is detected as missing among the deposited files, the depositor is therefore made aware of this so that they can add this data or give a reason for its omission. Through Deposit Structures, depositors are also able to submit their data to the IUCr CheckCIF service that checks the completeness and crystallographic integrity of the data file. Reports from these checks are archived alongside the data and made available to end users (see for example http://dx.doi.org/10.5517/cc1k1ywf).

Data quality is further checked by the CCDC during the data curation process. All entries curated into the CSD are reviewed by a CCDC scientific editor to ensure that a correct chemical representation of the substance studied is associated with the crystallographic data. Should any inconsistencies or errors be identified as a result of this, they are reported to the depositor/publishers. Additional data items and chemical compound names may also be added by editors if evident from the published literature.

When an article reporting the data is published, the CCDC updates the metadata provided by the depositor with the full article citation. To aid this process, agreements are in place with publishers who communicate publication information for datasets to the CCDC. Once the citation is added, the Crossref REST API is automatically run and performs further checks on the citation details. Depositors can also update the citation information for their structures themselves using the My Structures service.

The user community can provide feedback on the quality of data entries by contacting the CCDC through the general enquiries page (https://www.ccdc.cam.ac.uk/theccdcprofile/contactus/Enquiry/generic), with members of the scientific editorial team in place to respond to these queries.

Metadata for datasets is openly accessible through DataCite infrastructure in various formats. E.g https://search.datacite.org/works?query=*Cambridge+Structural+Database&data-center-id=ccdc.csd. Through DataCite’s services, CCDC metadata records are incorporated into the Clarivate Data Citation Index (http://wokinfo.com/products_tools/multidisciplinary/dci/).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC Database workflow diagram outlines the archival and storage activities which occur starting from pre-ingest of data and terminating with user access to data. This document along with the preservation policy is publicly available on the CCDC website (https://www.ccdc.cam.ac.uk/Community/depositastructure/scientific-data-preservation/) to inform depositors and users about CCDC's storage and archival procedures.

When data is deposited it is appraised throughout the deposition workflow through a number of automated and manual validation steps. The criteria for appraising and selecting data for preservation is communicated to users in the deposition guidelines and information pages. If deposited data does not meet the criteria for preservation, CCDC staff will contact the depositor to request further information.

The CCDC holds all unpublished data in trust for depositors. However, as part of publisher workflows, datasets may need to be made available to reviewers, depositors and publishers before publication. For this release of data pre-publication, the CCDC therefore has in place strict measures which oblige users seeking to access data pre-publication to provide sufficient information about deposited data. The CCDC also tracks and monitors any external access to unpublished data.

An overall review of deposition, publication and access processes occurs on a monthly basis to check that workflow processes are functioning as expected. If integral changes are made to the workflow based on this monitoring, the workflow documentation is then updated to reflect this.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All individual datasets are free to search and access via Access Structures. Datasets are retrievable using accession IDs, dataset and article metadata and DOIs, and chemical compound names. Users can choose to search uniquely the CSD or the ICSD, or both databases together.

A free utility (https://www.ccdc.cam.ac.uk/Community/csd-community/cellcheckcsd/) is available for integration with diffractometer software so that crystallographers can check for prior determinations of a crystal structure before proceeding with an experiment.

The metadata for datasets with DOIs is openly accessible in various formats, including OAI-PMH, via DataCite and is searchable using the DataCite Rest API.

Recommendations on how to cite data are available on information (https://www.ccdc.cam.ac.uk/support-and-resources/support/case/?caseid=b709ea14-188b-44b2-9232-2b9d18771c1d) and policy pages (https://www.ccdc.cam.ac.uk/access-structures-terms/#citation) across the CCDC website.

Dataset landing pages uses schema.org for dataset metadata markup, aiding the discoverability of datasets.

The CSD is registered as a trusted resource for data preservation by re3data (https://www.re3data.org/repository/r3d100010197) and FAIRsharing (https://fairsharing.org/FAIRsharing.vs7865). The CCDC is also recommended by a number of journal publishers as a trusted repository for depositing supplementary data for articles. Examples being the Royal Society of Chemistry (RSC) (https://www.rsc.org/journals-books-databases/journal-authors-reviewers/prepare-your-article/experimental-data/) and the American Chemical Society (ACS) (http://pubsapp.acs.org/paragonplus/submission/acs_cif_authguide.pdf?).

The CCDC has partnerships with the main publishers of crystallographic data to develop services which enable article-data links. For example, links to CCDC data are available in Elsevier articles using the Scholix mechanism for data link exchange. E.g https://www.sciencedirect.com/science/article/pii/S0022286019302625?via%3Dihub#ec-research-data. This system was developed through the RDA/WDS Publishing Data Services Working Group of which the CCDC is an active participant.

The CCDC also has links with other repositories, such as the Pesticide Properties DataBase (PPDB), Protein Data Bank (PDB), ChemSpider, DrugBank and PubChem, to create cross-references between related datasets and expose crystal structure data in other resources. E.g https://www.ccdc.cam.ac.uk/structures/Search?Ccdcid=204267

For more systematic search and analysis based on a wider range of data items and features, CCDC value-added products and services are available (https://www.ccdc.cam.ac.uk/solutions/csd-system/components/). The CSD software applications provide easy access to knowledge derived from the database and for this to be applied in areas such as drug discovery and materials design.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All deposited data is made accessible in CIF file format, a standard for the crystallography community which contains information fields pertaining to creation and provenance of the data. The CCDC's deposition guidelines also encourage depositors to enhance the reusability of data by suggesting additional information to be added.

The CCDC provides a free visualiser programme (https://www.ccdc.cam.ac.uk/solutions/csd-system/components/Mercury/) that enables anyone to view structure data files and maintain a subset of structures selected to support teaching activities in chemistry and crystallography.

The data curation undertaken by the CCDC editorial team aims to ensure the comprehensibility and applicability of data by researchers in chemistry, biology and other domains. This is done by making data consistent with community standards for representing and interpreting the scientific information and by improving the scientific accuracy of the data.

The CCDC's software and applications have been developed for providing targeted scientific solutions that enable data and derived knowledge to be easily applied in industry as well as academia. The Python and REST Application Programming Interfaces (APIs) (https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd-python-api/) offer a way for users to embed access to data and knowledge into their own workflows and access this from within third party systems. These products and services also allow export of data and metadata in a range of formats widely used by the user communities.

As a requirement for data to be made public, it must have complete metadata. This includes a full author list, year of publication and article publication information (e.g DOI or page/volume). This metadata is either communicated to the CCDC by publishers/depositors, or updated by CCDC staff from the published literature.

The reuse policy for datasets is covered by the website Terms and Conditions (https://www.ccdc.cam.ac.uk/access-structures-terms/), which allows researchers to freely access and use deposited datasets.

The CCDC actively engages with its user communities and has partnerships with the key stakeholders to ensure that the

Centre stays abreast of the requirements of the chemistry community.

Work is currently in progress to develop the CCDC’s underlying database formats. This would give the CCDC the flexibility to adapt to new file types were the file formats used by the community ever to change.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC's IT infrastructure is developed and maintained in-house by the CCDC Systems Team. The CCDC’s procedures for developing and maintaining the IT infrastructure relate to the Archival Storage element of the OAIS reference model and are described in section 6 and 10 of the CCDC preservation policy.

Developments made to the IT infrastructure are guided by the CCDC’s strategic objectives discussed in R3. These objectives are then reflected in departmental objectives to be fulfilled by the relevant teams.

The IT infrastructure is based on Microsoft, Linux and MacOS X operating systems. Core informatics systems are built on Microsoft Dynamics CRM and SQL. External web services are hosted on Microsoft Internet Information Services and Apache. Software development is primarily undertaken in Microsoft Visual Studio and takes advantage of third-party cross-platform libraries such as Qt. Source code is maintained in the Mercurial version control system and documented using Doxygen.

For testing key production services, the CCDC uses an extensive range of unit and integration tests, a continuous integration build environment and staging and integration servers.

Support contracts are in place for critical hardware systems.

Internet connectivity is provided by the UK academic network, Janet, which has historically been very reliable. Many of our value-added services are packaged such that data and software can be installed locally on a researcher’s own machines thus avoiding dependence on the CCDC's own infrastructure.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The CCDC has various measures in place to protect data and services against downtime and potential disaster scenarios, which are covered in the disaster recovery plan. The disaster recovery plan outlines the IT services which are vital to the CCDC and would form the principle focus of recovery in the event of a disaster to ensure a quick recovery. As part of the disaster recovery procedure, the information and resources necessary to recreate services in the event of a major disaster are stored securely off-site so that these are accessible even if CCDC servers are not.

Deposited data is stored in Microsoft Azure Datacentres, across three data centres in a secure and highly available environment. The remainder of CCDC's infrastructure is stored in a server room equipped with a number of uninterrupted power supply systems that will keep systems running in the event of a brief power outage and also ensure that systems can be shut down cleanly and thus efficiently restored when power returns.

Microsoft Azure and the main servers can only be accessed by selected members of staff. Internal servers are securely housed in an air-conditioned server room protected by a fire suppressant system and high temperature cut off safeguards. Entry to the building in general is controlled by an entry card system and the building is monitored overnight by the University of Cambridge security team.

Backups are stored internally in a fireproof safe and offsite with a recognised third-party provider in a climate-controlled environment. Full backups of key machines are taken nightly, and full offsite backups taken three times per week to be stored remotely on a server hosted by the University of Cambridge. Nobody other than selected CCDC employees can access this remote machine.

All CCDC staff can access the main internal network but only those whose job function requires it can modify our data holdings. Any guest access to our network is strictly limited and when staff members leave the organisation, their accounts are deactivated immediately.

A firewall controls what services can be accessed from within and outside of the CCDC, with firewall rules being reviewed regularly. The CCDC engages security specialists to routinely probe for vulnerabilities in our network. Various layers of spam and virus protection are in place to limit the impact of malicious files. Machines are patched on a regular basis and we have change control procedures that ensure external facing services are not compromised by updates and patching.

The CCDC has in place a risk register which is reviewed annually as part of the organisation's strategy reviews by senior management and trustees. The threats identified by these reviews are then used to inform IT security policies.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: