

# AusGeochem Preservation Policy

## Introduction/Purpose

The AuScope Geochemistry Network ([AGN](#)) concept was developed in response to open community workshops held across Australia in 2018-2019 involving academia, industry and geoscience surveys. One of the key issues identified was the need for improvement in the 'curation and preservation of sample and data collections of national significance'. Supported by not-for-profit company AuScope Ltd via funding from the Australian National Collaborative Infrastructure Research Infrastructure Strategy (NCRIS), the AGN has built the collaborative [AusGeochem platform](#) for collating, preserving, and disseminating geochronology and geochemistry (isotopic, major and trace) data. The AusGeochem platform aims to motivate a community change regarding geochemical data management by providing a research tool that preserves data following [FAIR](#) (Findable, Accessible, Interoperable, and Reusable) principles. By creating AusGeochem as a platform (an infrastructure or framework that offers a range of tools, libraries, and services going beyond just storage) rather than simply a data repository, the service provides a set of reusable components that enable the development of additional features or applications built on top of the existing infrastructure. This capability is critical in ensuring the long term and active preservation and re-utilisation of the invaluable sample materials and datasets within the platform.

With respect to analytical research laboratories, AusGeochem provides a centralised repository allowing laboratories to upload, archive, disseminate and publish their datasets. The intuitive user interface allows users to access national publicly funded data quickly through the ability to view an area of interest, synthesise a variety of geochemical data in real-time, and extract the required data, gaining novel scientific insights through multi-method data collation. Lithodat Pty Ltd has integrated built-in data synthesis functions into the platform, such as cumulative age histograms, age vs elevation plots, and step-heating diagrams, allowing for rapid inter-study comparisons. Data can be extracted in multiple formats for re-use in a variety of software systems, allowing for the integration of regional datasets into machine learning and AI systems.

## Preservation Strategy

The active preservation within AusGeochem is largely enacted through the development of and ongoing consultation with Expert Advisory Groups (EAGs) from a variety of institutions across Australia. The consultation with the EAGs consist of discussing and understanding the international data reporting best practices and data quality assessment for each type of data to be ingested and stored within AusGeochem. The resulting technique-specific data models within the AusGeochem platform are therefore fit for purpose and adhere to FAIR data principles. The structured and standardised way metadata is stored using these technique-specific data models in AusGeochem enables and allows for potential future developments and updates. Through continued consultation with EAGs, community feedback, outreach at national and international conferences, and involvement as an active board member in the OneGeochemistry initiative the AGN monitors the technical

environment to plan, then execute the required steps/platform developments to ensure that the needs of the community are met. The flexible architecture of the AusGeochem platform and continued dialogue with the EAGs allow for straightforward design and integration of bespoke data tables for differing analysis types, should they be desired/required by the community.

## Metadata Data Management

The (meta)data storage plan of AusGeochem data and the developed models is outlined in the contractual agreement between Curtin University and Lithodat Pty Ltd: “that at any one time a copy of all publicly made available data is shared, stored and kept up to date on a weekly/monthly basis”.

Preservation of the following components of the repository is outlined below:

1. Sample metadata, analytical data and associated analytical metadata
  - a. In flat tables (collated into a ZIP file of e.g., CSV files/excel spreadsheets with multiple sheets, shapefiles).
2. Data models developed by AGN
  - a. Core model
  - b. Sample
  - c. Person
  - d. Literature
  - e. Uranium-Lead (U-Pb)
    - i. instrument metadata (relevant to f-j)
    - ii. analytical metadata (relevant to f-j)
    - iii. reference material metadata (relevant to f-j)
  - f. Fission Track
  - g. Uranium-Thorium -Helium ((U-Th)/He)
  - h. Geochemistry (major and trace elements)
  - i. Thermal Histories Data
  - j. Argon-Argon (Ar-Ar)
  - k. Lutetium -Hafnium (Lu-Hf)
3. Lists and terms (published or referenced vocabularies)
  - a. Lists of terms for fields
    - i. e.g. uncertainty types
  - b. Mindat.org references (URL list)
  - c. Other attribute lists
4. Explanatory notes of how data, data models and lists interact in order to support users that only have access to the above.

Due to their intrinsic nature the components described above require individual and adapted preservation levels, which are presented below

1. Data and metadata - Needs to be regularly stored and updated,
2. Data models - Are not expected to change much once established only a few minor changes will be needed shortly after the production phase of the model, this should be recorded as changes occur and is of a more urgent matter than,

3. Lists and terms - Changes occur sporadically and at inconsistent intervals, archival of lists and terms should be done in a regular manner (half yearly).
4. Explanatory notes - Will only need intermittent revision as well as updating upon completion/addition of new data models or platform components regarding the interaction between data, data models and lists.

## Application

The preservation strategy applies to all data contributors who upload publicly available dataset collections (Data Packages) hosted within the AusGeochem platform. Data Packages that are designated private within the platform's data privacy controls, are not backed up for long term preservation (and thus fall outside of this preservation plan) due to the potential of the Data Package to change as the unpublished data collections are prepared for publication. Once the privacy control is set to public, the application of the preservation plan commences.

## Unique Persistent Identifier

A persistent identifier can be minted directly within the AusGeochem platform via two options: (i) a Digital Object Identifier (DOI) may be minted at the 'Data Package' level and (ii) an International Generic Sample Number (IGSN) may be minted at an individual sample level either individually or in bulk by selecting multiple samples (each obtaining their own IGSN identifier). The AusGeochem platform is able to offer this service due to Lithodat Pty Ltd's official membership in the IGSN organisation and DataCite.

Furthermore, if samples or data sets/collections already have associated IGSNs or DOIs the AusGeochem platform sample upload templates has a field available for inputting existing IGSNs and DOIs.

## Backup and Archive

All the data will be backed up to the AuScope Data Repository by an automated process of uploading data packages as ZIP files. Every month a process will run that checks if there are new or edited public data packages. If there is, the data is zipped and sent to the AuScope Amazon Web Services (AWS) S3 bucket for long term preservation. This process and long term preservation backup of the data is hosted by Commonwealth Scientific and Industrial Research Organisation (CSIRO) in collaboration with AuScope Ltd.

The AusGeochem platform makes use of AWS as a cloud computing platform and for scalable data storage in the cloud. Through the AWS, a robust back-up strategy has been implemented follow the below:

The technical disaster recovery plan is based on the backup process:

1. The complete state of the system is persisted in two places:
  - AWS RDS Postgres DB

- AWS S3 Bucket with Lithosurfer files
- 2. By having backups of these two storages the Lithosurfer system can be fully restored on most cloud infrastructures (e.g. AWS, Azure, Google, Heroku, etc.).
- 3. However: The environment variables need to be set.
- 4. The code doesn't need to be backed up due to it being managed by Git, therefore, there is a full copy, including all of the history, on each developer's laptop and also on Github.

#### Backup Process

1. Every day a snapshot is taken that is automatically kept for 35 days.
2. Every week a snapshot is copied and encrypted to S3.
3. Every month a routine creates a snapshot and copies it to local drives A and B.
4. Also every month a routine copies content of S3 Lithosurfer File Bucket to local drives A and B.
5. Drives A and B are never closer to each other than 1000km.
6. Drives A and B are replaced yearly or when full and taken into custody by a trusted party.
7. Keys for decryption are kept in a keepersecurity vault.
8. A restore of the complete system from the drives is exercised at least each year; from drive A in even years and drive B in odd years.

## Funding and resource planning

The AuScope Geochemistry Network (and by extension the AusGeochem platform) is dependent on funding from the not-for-profit company AuScope Ltd to carry out its activities. AuScope is "Australia's provider of infrastructure to the Earth and Geospatial Science community, aiming to foster open access to earth and geospatial science infrastructure to drive research across government, academia, and industry". The tools, data, services and analytics supported by AuScope Ltd enable scientists to understand Earth's evolution through time and explore how Earth resources may support growing demands. AuScope Ltd is an Australian Government [NCRIS program](#) platform dedicated to the Earth Sciences, and has received continuous funding support from the Commonwealth government since 2006. AuScope Ltd has produced a 10-year [strategic vision](#) (Strategy 2020 – 2030) and 5-Year Investment Plan following an extensive community engagement process between 2018 and 2020. AuScope Ltd operational funding is currently secured to the end of financial year 2028 through the NCRIS program. Within this latest round of the NCRIS Research Infrastructure Investment Plan (RIIP) process, the AGN has acquired additional NCRIS funding provided through AuScope through the end of the financial year 2028, as the EarthBank program. The 'EarthBank: An Australian Geochemistry Network activity enabling the analysis, preservation and discoverability of nationally significant Earth & environmental collections' program will support the AGN team to work alongside geoscience agencies, museums and CSIRO to develop national data reporting standards, preserve legacy sample collections and make associated data sets digitally discoverable using FAIR procedures through the AusGeochem platform.

## Roles and Responsibilities

Management and execution of the preservation plan rests on the coordinator of the AGN and is to be done in collaboration with the developers and service provider Lithodat Pty Ltd. The AusGeochem platform is a government funded initiative to gather, disseminate and preserve government/public funded data, and therefore, assume the right to undertake long-term preservation responsibility for the ingested geochemistry data.

## AusGeochem References/Useful Text:

Boone, S.C., Dalton, H., Prent, A., Kohlmann, F., Theile, M., Gréau, Y., Florin, G., Noble, W., Hodgekiss, S.A., Ware, B. and Phillips, D., 2022. AusGeochem: An Open Platform for Geochemical Data Preservation, Dissemination and Synthesis. *Geostandards and Geoanalytical Research*, 46(2), pp.245-259. doi: 10.1111/ggr.12419. This paper is open access and can be viewed [here](#).

Boone, S.C., Kohlmann, F., Noble, W., Theile, M., Beucher, R., Kohn, B., Glorie, S., Danišik, M., Zhou, R., McMillan, M. and Nixon, A., 2023. A geospatial platform for the tectonic interpretation of low-temperature thermochronology Big Data. *Scientific Reports*, 13(1), p.8581. <https://doi.org/10.1038/s41598-023-35776-3>. This paper is open access and can be viewed [here](#).

The McNaughton Collection GSWA record: The preservation of legacy collections project: a template for preserving high-value collections for future research. This paper is open access and can be viewed [here](#).