

AusGeochem: An Open Platform for Geochemical Data Preservation, Dissemination and Synthesis

Samuel C. Boone (1)* , Hayden Dalton (1), Alexander Prent (2), Fabian Kohlmann (3), Moritz Theile (3), Yoann Gréau (4), Guillaume Florin (4), Wayne Noble (3), Sally-Ann Hodgekiss (4), Bryant Ware (2), David Phillips (1), Barry Kohn (1), Suzanne O'Reilly (4), Andrew Gleadow (1), Brent McInnes (2) and Tim Rawling (5)

(1) School of Geography, Earth and Atmospheric Sciences, The University of Melbourne, Melbourne, Victoria, 3010, Australia

(2) John de Laeter Centre, Curtin University, Bentley, Western Australia, 6102, Australia

(3) Lithodat Pty Ltd, Melbourne, Victoria, 3030, Australia

(4) Department of Earth and Environmental Sciences, Macquarie University, NSW, 2109, Australia

(5) AuScope, University of Melbourne, Melbourne, Victoria, 3010, Australia

* Corresponding author. e-mail: samuel.boone@unimelb.edu.au

To promote a more efficient and transparent geochemistry data ecosystem, a consortium of Australian university research laboratories called the AuScope Geochemistry Network assembled to build a collaborative platform for the express purpose of preserving, disseminating and collating geochronology and isotopic data. In partnership with geoscience-data-solutions company Lithodat Pty Ltd, the open, cloud-based AusGeochem platform (<https://ausgeochem.auscope.org.au>) was developed to simultaneously serve as a geosample registry, a geochemical data repository and a data analysis tool. Informed by method-specific groups of geochemistry experts and established international data reporting practices, community-agreed database schemas were developed for rock and mineral geosample metadata and secondary ion mass spectrometry U-Pb analysis, with additional models for laser ablation-inductively coupled-mass spectrometry U-Pb and Lu-Hf, Ar-Ar, fission-track and (U-Th-Sm)/He under development. Collectively, the AusGeochem platform provides the geochemistry community with a new, dynamic resource to help facilitate FAIR (Findable, Accessible, Interoperable, Reusable) data management, streamline data dissemination and advanced quantitative investigations of Earth system processes. By systematically archiving detailed geochemical (meta-)data in structured schemas, intractably large datasets comprising thousands of analyses produced by numerous laboratories can be readily interrogated in novel and powerful ways. These include rapid derivation of inter-data relationships, facilitating on-the-fly data compilation, analysis and visualisation.

Keywords: geochemical data platform, geosample registry, data repository, Big data, FAIR data.

Received 08 Sep 21 – Accepted 12 Jan 22

Geochemical analysis of rock and mineral specimens provides fundamental information about the composition, age and time–temperature–pressure evolution of our planet and solar system, yielding critical insights into how they evolved and operate. These data include major and trace element content, and stable and radiogenic isotopic concentrations acquired via a range of analytical techniques, such as electron probe microanalysis (EPMA), X-ray fluorescence spectrometry (XRF) and inductively coupled plasma-mass spectrometry (ICP-MS; Rollinson 2014). These geochemical, geochronological and thermochronological

information, often complemented by petrological and crystallographic observations acquired through a range of microscopy techniques, allow geochemists to address fundamental questions about the formation, composition, interaction and evolution of planetary bodies, namely Earth, as well as provide critical information needed for the discovery and utilisation of geological resources.

Yet, despite the significant financial resources and time invested into acquiring these high-value data by funding organisations, research institutions, geochemistry laboratories

doi: 10.1111/ggr.12419

© 2022 The Authors. *Geostandards and Geoanalytical Research* published by John Wiley & Sons Ltd on behalf of the International Association of Geoanalysts

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

and geoscientists across the globe, a large proportion of geochemical data remain inaccessible to the public (e.g., Federer *et al.* 2018, He *et al.* 2019). This lack of data accessibility and transparency impedes regional Big Data syntheses and hinders inter-laboratory analytical comparisons from being easily performed. In an effort to address the urgent need for changes in culture and infrastructure supporting the reuse of data, stakeholders across academia, industry, government, funding agencies and publishers around the globe have jointly established and endorsed the FAIR Data Principles (Wilkinson *et al.* 2016, Stall *et al.* 2017, 2018, 2019). These state that research data should be increasingly Findable, Accessible, Interoperable and Reusable by both humans and machines (Wilkinson *et al.* 2016). Since 2018, more than 100 research data infrastructure organisations, institutions, societies, publishers, repositories and individuals have committed to achieving the FAIR Data objectives by signing the Commitment Statement in the Earth, Space and Environmental Sciences (COPDESS 2014).

The wide variability in Earth Science data formats and models, their inherent multi-dimensionality (x , y , z , chemistry, petrology, time, temperature and pressure) and the complexity of big data processing necessitate the development of tailored analytical techniques and tools for FAIR geochemical data distribution (Baumann *et al.* 2016). To facilitate the accessibility, systematic collation, assessment and processing of geochemical and geochronological data acquired from geological specimens, numerous structured databases have been developed globally (He *et al.* 2019). They range from rock databases (e.g., NAVDAT, Walker *et al.* 2004, Walker *et al.* 2006), to geochemical surveys (e.g., China National Multi-Purpose Geochemical DB, Liu *et al.* 2012), archives of geochronology and isotope data (e.g., DataView, Eglington 2004) and reference materials (GeoReM, Jochum *et al.* 2005), varying in scope from national (e.g., Petlab, Strong *et al.* 2016) to global scales (e.g., EarthChem, Lehnert *et al.* 2004).

Within Australia, a number of geoanalytical databases have been built focussed around specific data types or regions, including OZCHEM (formerly ROCKCHEM, Hazell *et al.* 1995, Champion *et al.* 2007), the National Geochemical Survey of Australia (Caritat and Cooper 2016), a database of whole-rock chemical and geochronological data of igneous rocks in Queensland (Siegel *et al.* 2012) and the National Argon Map (Australian National University 2020). Additionally, a number of government-funded geological data portals have been developed to archive a range of geological data on a state to national scale, such as the Geoscience Australia Portal (Cropper and Sweeney 2021), AuScope Discovery Portal (AuScope 2019a),

GeoVIEW.WA (Government of Western Australia 2013) and South Australian Resources Information Gateway (SARIG 2021).

Despite these significant efforts, the current geochemical data ecosystem still falls well short of the FAIR Data goals, with the bulk of scientifically significant data never reaching the wider community. Instead, much of the high-value geochemistry data are lost along the typically protracted journey from the *Institutional*, to *Collaboration* and finally *Publication* domains (Figure 1). The transformation of analytical results along this Data Curation Continuum (Treloar *et al.* 2007, Treloar and Klump 2019) currently involves multiple stages of largely manual data curation and migration, requiring significant institutional investment in the form of time and resources. Results deemed unfit for publication, not necessarily due to poor quality but often due to circumstance, interpretation challenges or being regarded as 'unexciting', are then lost to the wider scientific community, and instead reside indefinitely on laboratory hard disk drives or other storage mechanisms. In a straw poll of leading geochemists across Australia at the 2018 TANG30 (Thermochronology and Noble Gas Geochronology and Geochemistry Organisation) Meeting, it was estimated that a mere ~5–10% of geochemistry data produced in Australian laboratories ever reaches the public domain. Even then, only approximately one-fifth of published papers make the supporting data publicly available via data repositories (Federer *et al.* 2018), and often without the detailed (meta-)data required for robust data integration and interpretation (He *et al.* 2019). This lack of data transparency stems, at least in part, from the fact that the scientific contribution of those that generate data, including funding bodies, sample collectors, lab technicians and dataset collators, is rarely recognised (Pierce *et al.* 2019). Instead, it remains common for data to be archived ad hoc on a per laboratory basis. While the aforementioned databases and portals have significantly increased the amount of publicly available geochemistry data, the fact remains that only a small proportion of the high-quality and valuable data produced by geochemistry laboratories ever become openly accessible.

Therefore, in order to transform the geochemistry data ecosystem into one that is more aligned with FAIR data objectives, a transformation in the way in which geochemistry data is curated, migrated and stored is needed (Chamberlain *et al.* 2021). This includes adopting or, where needed, establishing standardised community agreed upon (meta-) data reporting templates and the routine reporting of reference material analyses, without which data users are unable to independently assess data quality. Such a

Data Curation Continuum

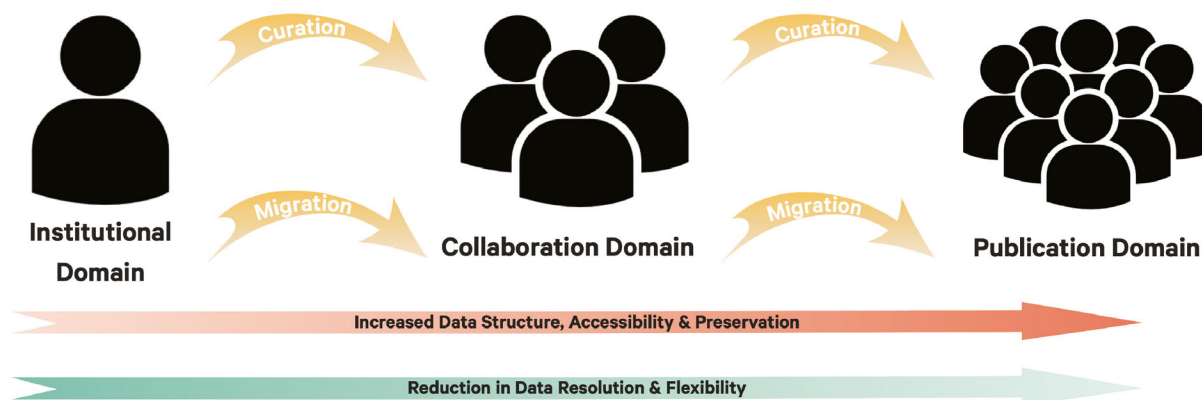


Figure 1. The Data Curation Continuum (after Treloar and Klump 2019). Under this model, analytical data are produced in the Institutional Domain. Researchers may then choose to share these data with collaborators within or outside of their institution to produce innovative and novel science. Only once the data are in a state deemed ready for public consumption are the analyses migrated into the Publication Domain through publication in scientific journals or data repositories. Data typically undergo significant transformation along this continuum, becoming more structured, more accessible and better preserved while simultaneously losing resolution and flexibility. Data migration between domains generally requires significant manual curation (e.g., reformatting, computation, synthesis, formatting), making the current Data Curation Continuum inefficient and laborious. Consequently, only a small portion of geochemistry data makes it into the Publication Domain.

transformation will also require data-producing laboratories to be incentivised to make significantly more of their data publicly available. This urgent change in data-sharing culture will, thus, require a novel data migration system that (a) enables easy and efficient data upload, (b) facilitates powerful new ways of synthesising and interpreting data and (c) rewards data sharing by increasing scientific recognition of data producers – from sample collectors and analysts to data interpreters and dataset collators.

Towards these goals and with the support of AuScope (2019b), the AuScope Geochemistry Network (AGN) and collaborators Lithodat Pty Ltd have developed AusGeochem (Figure 2), a cloud-hosted open geochemistry data platform that simultaneously serves as a geosample registry, a geochemistry data repository and an active research tool. Equipped with an intuitive user interface and bulk data uploader, AusGeochem allows laboratories and individuals to readily *upload*, *archive*, *disseminate* and *publicise* their geosample (meta-)data and associated geochemistry data while maintaining *privacy control*. This will ensure the preservation and re-utilisation of these invaluable sample materials and datasets, saving funding agencies, institutions and laboratories significant resources by preventing

unnecessary duplication of efforts. Once uploaded into AusGeochem, its relational database and range of on-the-fly data visualisation functions enable users to perform data synthesis and analysis across data types within the context of large volumes of publicly funded geochemical data. Users can then expand on the existing capabilities of the platform through the use of its open application programming interface (API), enabling them to take full advantage of a global eResearch infrastructure that is rapidly strengthening. Through the API, external software components, operating systems and applications can thus access and interact with AusGeochem-hosted data to facilitate a variety of tasks that include automatic machine-to-database (meta-)data upload, automated data retrieval and the incorporation of additional data synthesis functions and machine learning algorithms.

In the following sections, we describe the collaborative development of AusGeochem, the platform architecture, its various components (Geosample Registry, Geochemical Data Repository and Geochemistry Research Tool) and functionalities. We then discuss how AusGeochem can help reshape the Geochemistry Data Ecosystem into one that is more interconnected and FAIR. Readers are encouraged to

AusGeochem

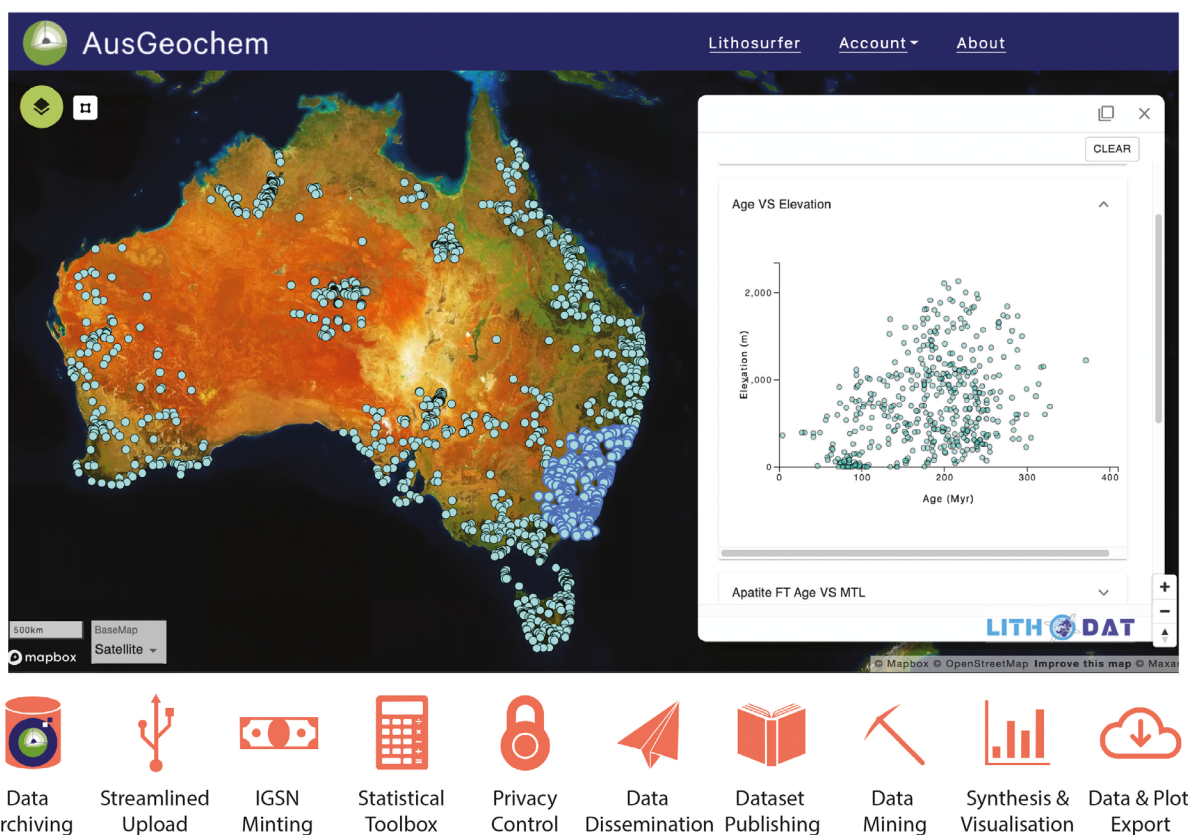


Figure 2. The multi-purpose AusGeochem platform allows laboratories to upload, archive, disseminate and publish their datasets, as well as perform data synthesis and analysis within the context of large volumes of publicly funded geochemical data aggregated by the AGN.

register to use AusGeochem at <https://ausgeochem.auscope.org.au>, with which they can freely explore, disseminate and interrogate geosample data and associated geochemical analyses in the context of thousands of data from across the globe.

Collaborative development of the AusGeochem platform

Lithodat partnership

To develop AusGeochem, the AGN has entered into a collaborative partnership with Lithodat Pty Ltd (Lithodat 2021a), a cloud-hosted geoscience data provider comprising a group of international senior geologists, data scientists and software engineers. This relationship leverages the unique expertise offered by this hybrid group of geo- and data scientists with a firm understanding of use-cases relevant to the exploration industry and wider geoscience

community. Lithodat has already developed and maintained LithoSurfer (Lithodat 2021b), a cloud-hosted platform for geospatial data. The AGN and Lithodat utilised LithoSurfer, an established geospatial data platform previously built by Lithodat, as the foundation upon which AusGeochem was built, implementing new data types and requirements.

AGN expert advisory groups

To ensure that technique-specific data models are fit for purpose, foster scientific collaboration and adhere to FAIR data principles, the AGN has set up a number of Expert Advisory Groups (EAGs) to assist in the development of data models for the secondary ion mass spectrometry (SIMS) U-Pb, laser ablation-inductively coupled plasma-mass spectrometry (LA-ICP-MS) U-Pb and Lu-Hf, fission track, (U-Th-Sm)/He and Ar-Ar techniques. Comprising internationally recognised geochemical specialists from across Australia and adopting established international reporting guidelines where

possible (e.g., Walker *et al.* 2008, Horstwood *et al.* 2016, Schaen *et al.* 2021), the EAGs provide invaluable advice regarding data reporting best practices, data quality assessment and visualisation tools incorporated into AusGeochem. The 28 members of the EAGs come from a variety of institutions across Australia (Boone and Manifold 2021), with this number likely to increase as additional method-specific data models are developed for AusGeochem in the future.

Legacy datasets as test cases

To test the robustness of method-specific data models prior to release, large legacy datasets of national (and international) scientific importance are uploaded into AusGeochem in collaboration with EAGs and other AGN contributors (e.g., Geological Survey of Western Australia, Museums Victoria). These test cases both ensure that each model is fit for purpose and meets technique-specific international data reporting requirements, with the added benefit of preserving these unique data resources in a FAIR and open manner. Legacy datasets to date include the McNaughton sample and SIMS U-Pb data compilation (2765 samples, 701 with corresponding U-Pb results), the Geological Survey of Western Australia SIMS U-Pb data compilation (1404 analyses) and the Australian-wide apatite fission-track survey (2594 samples) of Kohn *et al.* (2002) and Gleadow *et al.* (2002), the analytical results of which will soon follow when the corresponding fission track model is released in early 2022. Technique-specific data models currently under development for the fission track, (U-Th-Sm)/He, Ar-Ar and LA-ICP-MS U-Pb and Lu-Hf techniques will undergo similar testing using large legacy datasets, to be made publicly available upon release of the corresponding data models.

AusGeochem platform architecture

Platform v portal architecture

AusGeochem is differentiated from existing geoscience data repositories and portals by its relational *platform* architecture, which equips it with a number of key benefits and added functionalities (Figure 3). While these existing databases are able to house a range of extractable data, which in the case of portals are geospatially displayed, there are often disparities in terms of data field naming, data field reporting and general data structures. This portal architecture inhibits the possibility of performing analytics within the portal or repository, and constrains the user's interaction with the data to simply viewing and extracting (Sherratt 2013). In contrast, the relational database architecture of the

AusGeochem platform ensures that the various data inputs are 'cleaned' and standardised upon upload and then archived in a structured and consistent way. This persistent data structure allows for searching across disparate data types and performing live cross-data statistical analysis and will enable potential future developments for performing on-the-fly computing, such as re-calculating isotopic ages based on updated decay constants. Once data analysis is complete, the user can then export related data and figures for later use. A further key difference between the AusGeochem platform and typical portals is the ability to allow users to maintain privacy control, choosing from several options regarding how widely their uploaded data is shared.

AusGeochem platform architecture

There are four integral components of the AusGeochem platform architecture (Figure 4). The first foundational component is the *Ownership and Access Control Model*, which enables AusGeochem's user-defined privacy control system, allowing users to enter their own data in the global database whilst reserving permission options. Users have the option to keep their unpublished data private, disseminate their data to select collaborators, or publicise their data to the wider community. This permission system gives users full control over their data whilst still getting to take advantage of the AusGeochem data synthesis and analysis functions. To facilitate collaboration, users can also give other AusGeochem users access to Data Packages of their choosing, assigning individuals roles, such as administrator, data curator or user, allowing them to work with their data within AusGeochem or extract for use in external software.

Privacy settings are controlled on a per dataset, or 'Data Package', basis. By default, data uploaded to AusGeochem will be set as *publicly available*, unless changed by users. Datasets that are attached to a publication must be uploaded as *publicly available* datasets and/or will be treated as *publicly available* data. The *private* option is intended to keep data out of the public view for a limited embargo period. However, because AusGeochem is designed as an open data platform, there is a limit of 2 years from the date of data upload during which a user may nominate to keep their data private, unless an embargo extension is requested.

The *Core Model* builds the heart of the platform architecture, as it contains all entities required for frequent use of geoscientific datasets. Registered samples are stored here as 'Data Points' along with their associated information.

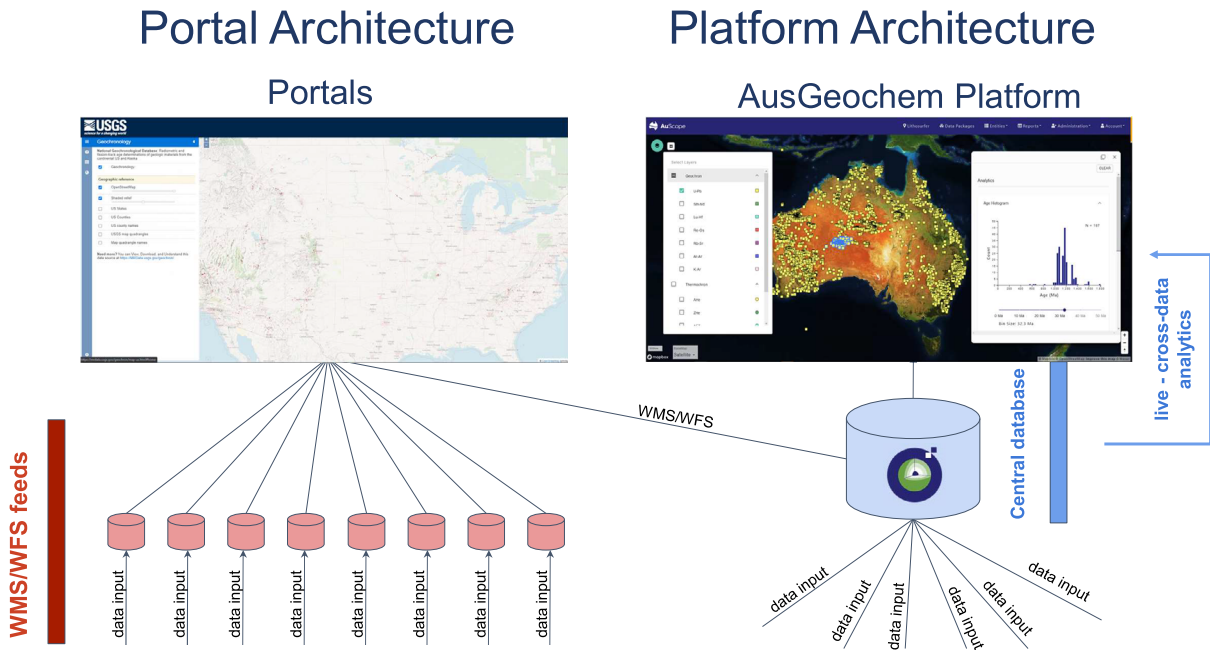


Figure 3. In contrast to conventional data portals, the AusGeochem platform relational database architecture allows for simultaneous on-the-fly synthesis of a variety of data of different types, while maintaining full privacy control. WMS, Web Map Service; WFS, Web Feature Service.

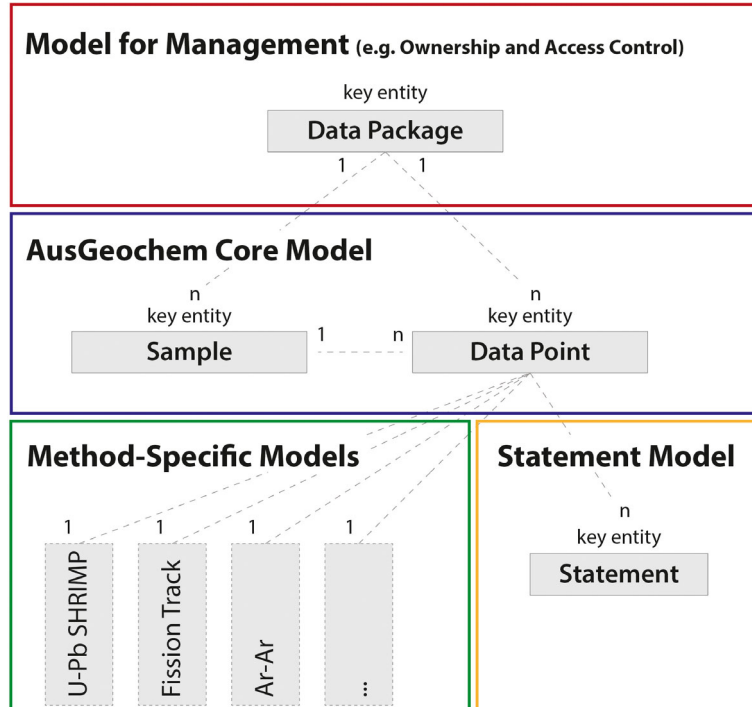


Figure 4. AusGeochem model overview.

The 'Data Points' can then be linked to *Method-Specific Models*, where each model contains data fields specified by the associated EAG for that geochemistry technique. Along

with recording the parameters required for each method-specific data model, the schemas accommodate secondary reference material results and error reporting, allowing

AusGeochem users to independently perform quality assessment of data archived in the database. The persistent archiving of reference material analyses will also enable inter- and intra-laboratory precision to be assessed over time in novel and powerful ways.

The fourth and final portion of the AusGeochem portal structure is the *Statement Model*, which facilitates the advanced 'on-the-fly' analytics across all datasets and data types. The *Statement Model* holds the 'statement(s)' derived from a 'Data Point', for example, age, chemical composition, isotopic ratio or, in the case of thermochronology data, time-temperature history, allowing data from different geochemical methods to be compared directly.

The AusGeochem platform utilises the Lithodat technology stack. The Frontend, or client-facing side of AusGeochem giving users access to a browser-based user interface, is coded using React, a JavaScript library for building modern user interfaces. The Backend, or server-side of AusGeochem, is coded using Java Spring Boot with all data housed in a PostgreSQL architecture, an open-source object-relational database system. The platform is built via Git to allow versioning control and a 2-stage Amazon Web Services (AWS) pipeline for maintainability. The platform runs via an AWS Elastic Beanstalk cloud service, which supports easy deployment, scalability and reliability.

Components of the AusGeochem platform

Geosample registry

The need for archiving geosample information arises immediately upon specimen acquisition. The AusGeochem platform is designed, in the first instance, to act as a geosample registry, where users can archive and publicise information about mineral and rock samples, maximising potential collaboration opportunities and preventing redundant sampling efforts.

The sample metadata scheme (Table 1) is consistent with those required for IGSN minting and those utilised by other geoscience repositories, such as EarthChem (Lehnert *et al.* 2004) or the Geoscience Australia Portal (Cropper and Sweeney 2021). Of the sample metadata attributes, five are required for sample registration: Sample ID, Sample Kind, Lithology or Mineral Type and Latitude and Longitude in the WGS84 global reference system. For its lithology and mineral lists, the platform utilises the comprehensive Mindat.org database (Mindat 2021) and associated hierarchical data structures, enabling more efficient and intelligent

querying of the AusGeochem database. In addition to the five required sample metadata, users should also upload the IGSN (International Geo Sample Number, Lehnert *et al.* 2011, 2019, Klump *et al.* 2021) of a sample, if previously minted. The use of globally unique and persistent IGSN identifiers for physical samples prevents unintentional sample ID duplication and enables samples to be discoverable on the internet (Klump *et al.* 2021).

International Geo Sample Number (IGSN) minting function: AusGeochem uses the unique IGSN identifier to ensure that geochemical data are related to the appropriate rock or mineral geosamples from which they were derived and prevent database redundancies. Therefore, if no IGSN exists for a given sample upon upload, AusGeochem strongly encourages users to utilise its automatic IGSN minting function.

Leveraging Lithodat's status as an official IGSN allocating agent, the AusGeochem platform provides the option for users to automatically assign IGSNs to their samples during the final steps of sample metadata upload, or at any subsequent time via their account's My Data tab. In addition to generating a unique IGSN alphanumeric identifier for a geosample, the minting function uses the related sample metadata uploaded into AusGeochem to generate an IGSN sample metadata profile webpage and corresponding QR code (Figure 5).

Geochemical data repository

In addition to registering geosample information, the AusGeochem database is designed to archive geochemical data generated by a variety of analytical methodologies of certain degrees of data processing (Table 2). In general, the database is designed to archive processed or 'hard' geochemical data, such as digital image analyses, reduced mass spectrometry data and electron probe microanalyses (Level 1), as well as calculated parameters and ages (Level 2). However, the database has the flexibility to accommodate some select 'raw', unprocessed instrumental data (Level 0) or digital imagery, required for certain method-specific data models. For instance, the U-Pb SIMS data model allows users to upload raw data output files generated with the SQUID data processing software (Bodorkos *et al.* 2020). In addition, AusGeochem will store some types of modelling data (Level 3 data), such as thermal history modelling parameters and results.

AusGeochem is currently equipped with a functioning data model for the SIMS U-Pb technique. Additional data models for the fission-track, (U-Th-Sm)/He, Ar-Ar and laser

Table 1.
AusGeochem geological sample metadata

Data	Datatype	Description
Sample ID	string	Sample ID assigned by collector
IGSN	string	International Geo Sample Number; to be assigned by the collector, analyst or, in cases where no IGSN exists at the time of data upload, automatically by AusGeochem. In order to assign an IGSN to a sample using AusGeochem, the user should be the 'owner' or have the permission from the sample owner
Sample Kind	list	The type of sampled material
Sample Method	list	The sample collection method
Lithology	list	Lithology type of rock sample
Mineral Type	list	Mineral type of mineral sample
Sample Comment	string	Additional information about sample lithology or mineralogy
Latitude (WGS84)	numeric	Latitudinal sample coordinate
Longitude (WGS84)	numeric	Longitudinal sample coordinate
Lat/Long Precision [m]	numeric	The precision of the reported latitude and longitude, determined from GPS or estimated on the method and vintage of a lat/long determination
Elevation	numeric	Sample elevation (auto-determined for surface samples from lat/long, if not defined)
Vertical Datum	list	Vertical datum (e.g., mean sea level, measured depth, etc.)
Location Kind	list	Type of sampling location (e.g., outcrop, well location, etc.)
Location Name	string	Name of sampling location
Location Comment	string	Additional information about the sample location, such as outcrop description or physiographic feature
Unit Name	list*	Geological name of the unit from which sample was collected. AusGeochem uses the Stratigraphic Units Database (Geoscience Australia and Australian Stratigraphy Commission 2017) for localities within Australia. In cases where samples are from outside Australia, users can manually enter the unit name
Chronostratigraphic Unit Age	list	Chronostratigraphic age of the unit sample was collected from
Unit Comment	string	Additional information about the sampled unit
Age (minimum)	numeric	Minimum age of rock sample
Age (maximum)	numeric	Maximum age of rock sample
Depth (minimum)	numeric	Minimum depth of rock sample; important for core, well or dredge samples
Depth (maximum)	numeric	Maximum depth of rock sample; important for core, well or dredge samples
Date Collected (minimum)	date	The minimum date sample was collected. The range of dates is only required if the exact sample collection date is unknown
Date Collected (maximum)	date	The maximum date sample was collected. The range of dates is only required if the exact sample collection date is unknown
Person	list*	The person who performed a particular task associated with the sample
ORCID	list*	ORCID (2012) of person
Person Role	list*	Role of person (e.g., sample collection, chief investigator, investigator, analyst, etc.)
Last Known Sample Archive	list*	Last known location where sample material is archived
Archive contact email	list*	Email associated with sample archive
Archive Comment	string	Additional information about the sample archive
Grant ID	string	ID of the funding grant that supported the collection, processing, analysis or collation of the associated data
Funding Body	list	Funding body that supported sample collection/preparation
Funding Comment	string	Comment about funding
Reference DOI	string	DOI of associated reference(s), when available, such as a publication, thesis, book, report, etc.
Reference Type	list	Type of reference (e.g. article, book, thesis, etc)
Reference Authors	string	Authors of publication or dataset. Names of each author must be input separated by commas or the word 'and'
Reference Title	string	Title of publication or dataset (auto-populated from DOI)
Reference Journal	string	Journal of publication or dataset (auto-populated from DOI)
Reference Year	numeric	Year of publication or dataset (auto-populated from DOI)
Reference Volume	numeric	Publication volume
Reference Issue	numeric	Publication issue
Reference Pages	string	Publication pages
Reference Publisher	string	Name of the publisher (e.g., Elsevier, Springer, etc)
Reference Publication Month	string	The month of publication, if applicable
Reference School	string	University of publication for MSc and PhD theses
Reference Book Title	string	Title of book, if applicable
Reference Editor	string	Editor(s) of book
Reference Keywords	string	Keywords relevant for publication. Can be converted to tags at a later stage for easier searching of literature
Reference Chapter	string	Name or number of the book chapter in which publication sits
Reference Series	string	Number of series, if applicable

Table 1 (continued).
AusGeochem geological sample metadata

Data	Datatype	Description
Reference Organisation	string	Name of organisation or institute publishing the reference, in the case of technical reports or special publications
Reference ISSN	string	International Standard Serial Number (ISSN), if relevant
Reference URL	string	URL, if known. Not necessary if DOI is provided
Reference Abstract	string	Associated abstract, if applicable
Reference Comment	string	Comment about reference

Required data fields are highlighted in red. Users must provide either a Lithology or Mineral Type (highlighted in purple) depending on the type of geosample being uploaded. Highly recommended fields are in blue, while the upload of all other metadata is strongly encouraged. Note, multiple Persons, Funding Grants and References can be added to any one sample, as these have an *n*-to-1 relationship with Sample. list* indicates that new entries can be manually added to lists by the user.

ablation-inductively coupled-mass spectrometry (LA-ICP-MS) U-Pb and Lu-Hf techniques are currently in testing and will be subsequently released. However, the AusGeochem vision is not limited to just these analytical methods and the AGN is actively seeking collaboration with additional geochemistry laboratories to facilitate the development of other method-specific schemas, such as for major and trace element geochemistry, thermal ionisation mass spectrometry and LA-ICP-MS Sm-Nd and Rb-Sr.

Data export: All data within AusGeochem can be filtered and extracted in multiple formats to be used in a

variety of software systems. This ensures that state-of-the-art techniques such as artificial neural networks can take full advantage of data stored in AusGeochem. Publication-quality plots, maps and images produced within AusGeochem can also be exported.

Geochemistry research tool

Due to its unique relational database structure, AusGeochem acts as an active research tool, providing geoscientists with new and powerful ways of synthesising and interrogating large, multi-system datasets. Using the

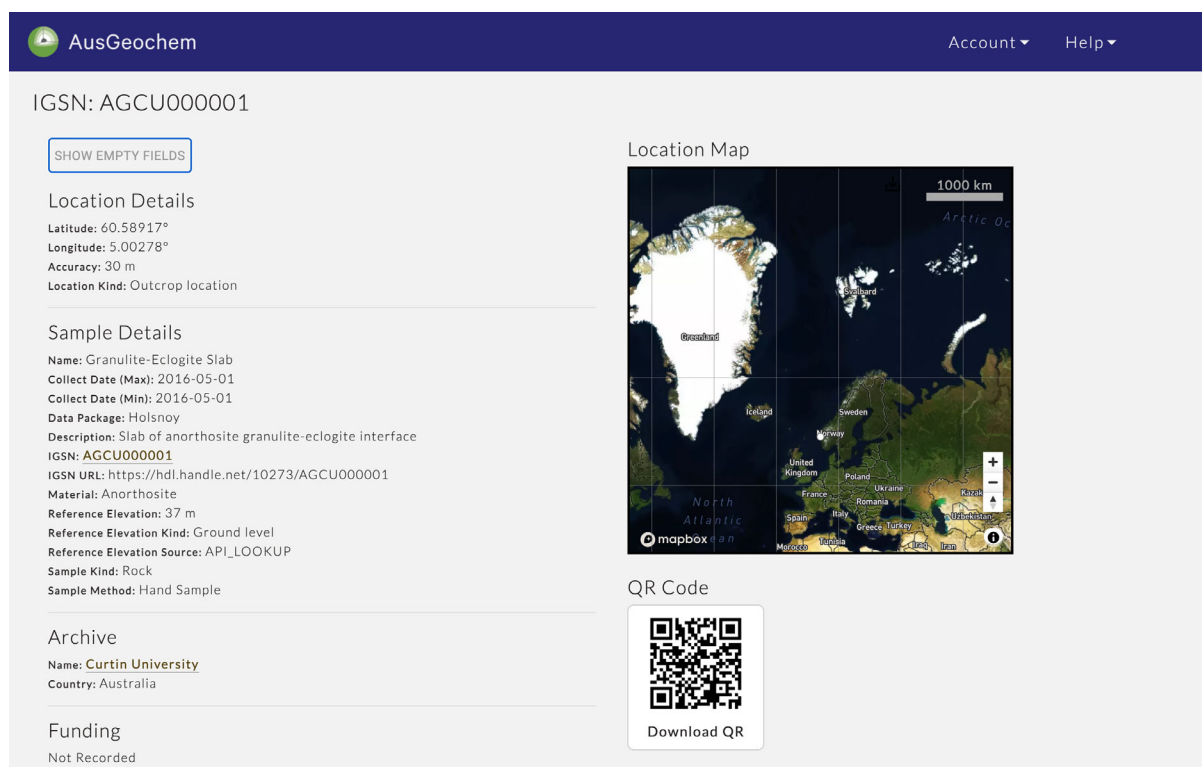


Figure 5. Example of an IGSN sample metadata profile webpage for a sample minted within AusGeochem.

Table 2.
The AusGeochem data hierarchy [derived and modified from NASA’s Data Processing Level hierarchy (NASA 2021)]

Data type	Data level	Description
Sample	Sample ^a	Sample information (e.g., IGSN, Sample ID, location info, lithology, etc.)
Raw	Level 0 ^b	Unprocessed instrumental data and metadata (digital imagery, raw mass spectrometry counts, etc.)
Hard	Level 1 ^a	Reduced or analysed raw data (e.g., calculated isotopic ratios, chemical concentrations, digital image analyses, etc.)
	Level 2 ^{a,c}	Calculated parameters and ages using Level 1 data
Soft	Level 3 ^{a,c}	Derivative models determined using Level 1 and/or 2 data (e.g., thermal history models)
	Level 4 ^c	Derivative models combining/interpolating Level 1, 2 and/or 3 data over an area (e.g., spatial interpolations, regional cooling/heating maps)

^a Data archived in AusGeochem.

^b Data archived in AusGeochem in some instances.

^c Data derived within AusGeochem.

sample, level 1, 2 and, in some cases, level 3 data stored within the relational database, AusGeochem is able to generate certain types of sample, hard and soft data. At the most basic level, this includes operations that streamline data upload, such as automatically determining the elevation of outcrop samples based on their coordinates using a Digital Elevation Model API (Google Elevation API). However, the real strength of the structured platform architecture is its ability to perform analytics across disparate data types. By defining the relationship between

different parameters and data tables, method-specific data models can be developed with the ability to determine elemental ratios, calculate radiometric ages and perform other types of on-the-fly statistical analyses using the detailed analytical (meta-)data archived in the AusGeochem database. This could include the ability to recalculate and remodel data using the most up-to-date constants and kinetic algorithms, enabling analyses determined using different parameters to be equated and compared across regional to global scales.

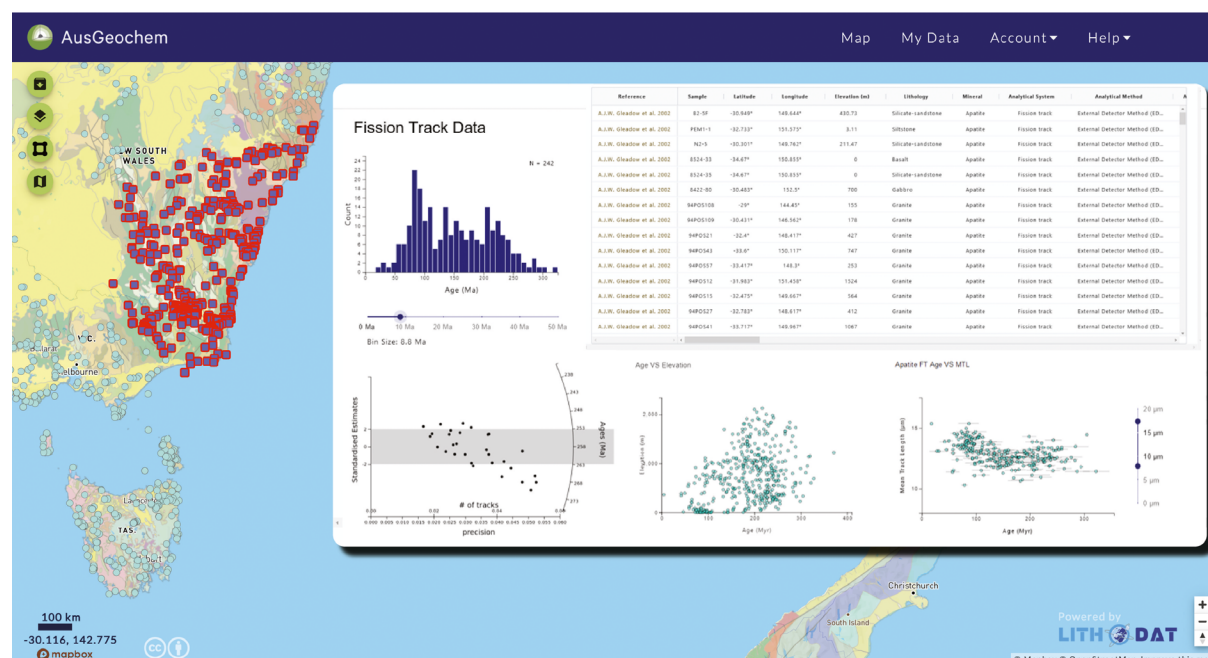


Figure 6. Apatite fission-track data dashboard displaying a data table, age histogram, age versus elevation plot and age versus mean track length plot generated on the fly for a subset of results from Kohn *et al.* (2002) and Gleadow *et al.* (2002). Users can choose from a selection of graphs relevant to each geochemistry technique. Most graphs in AusGeochem are interactive, allowing users to readily interrogate their data in the context of thousands of archived analyses. All data and figures are downloadable.

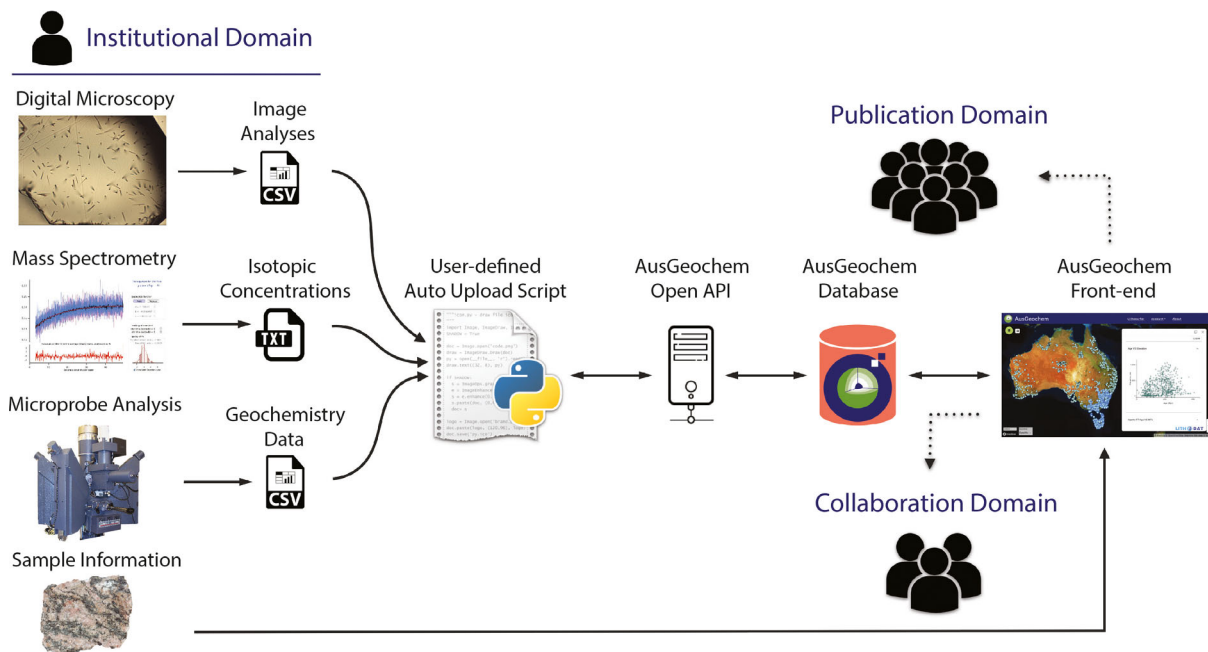


Figure 7. Schematic illustration of the AusGeochem data continuum. Laboratories can upload data using AusGeochem’s built-in user interface. However, to fully automate the data upload process, laboratories can develop programming scripts to populate the database directly from their scientific instruments, data reduction packages and image processing software utilising AusGeochem’s open-source API. The analytical data are then related to their corresponding sample information within the AusGeochem database using their unique IGSN. Once in AusGeochem, users are then able to disseminate their data to collaborators of their choosing or make them publicly available.

One of the most powerful features of AusGeochem is its ability to rapidly derive inter-data relationships, facilitating on-the-fly data synthesis and visualisation across multiple datasets and data types (Figure 6). By dragging a polygon over an area of interest, users can produce a range of on-the-fly comparative plots and derivative maps synthesising their results with the wealth of data stored in AusGeochem. These soft, Level 4 data derived within AusGeochem include interactive age histograms, radial plots, concordia diagrams and a range of scatter plots.

AusGeochem Open Application Programming Interface (API): The AusGeochem platform utilises an Open API, allowing any developer to build clients which interact with the platform to, for example, automatically retrieve data from its database, add in enhanced data visualisation tools and create direct links to analytical equipment. Importantly, the Open API can also be used to automatically upload data into AusGeochem from analytical software and in-lab hard drives, enabling a flexible and automated data migration process that accommodates diverse data types while minimising manual data handling for platform users (Figure 7). This function is critical for the geosciences where a large array of in-house or off-the-shelf data reduction

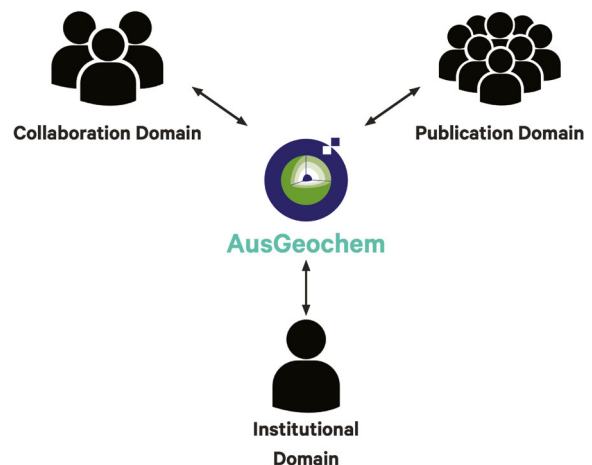


Figure 8. AusGeochem as a mechanism for linking Institutional, Collaboration and Publication Domains.

software are routinely used to produce final data tables for any given technique.

AusGeochem is also able to utilise the APIs of external clients to auto-populate certain fields (prior to a manual

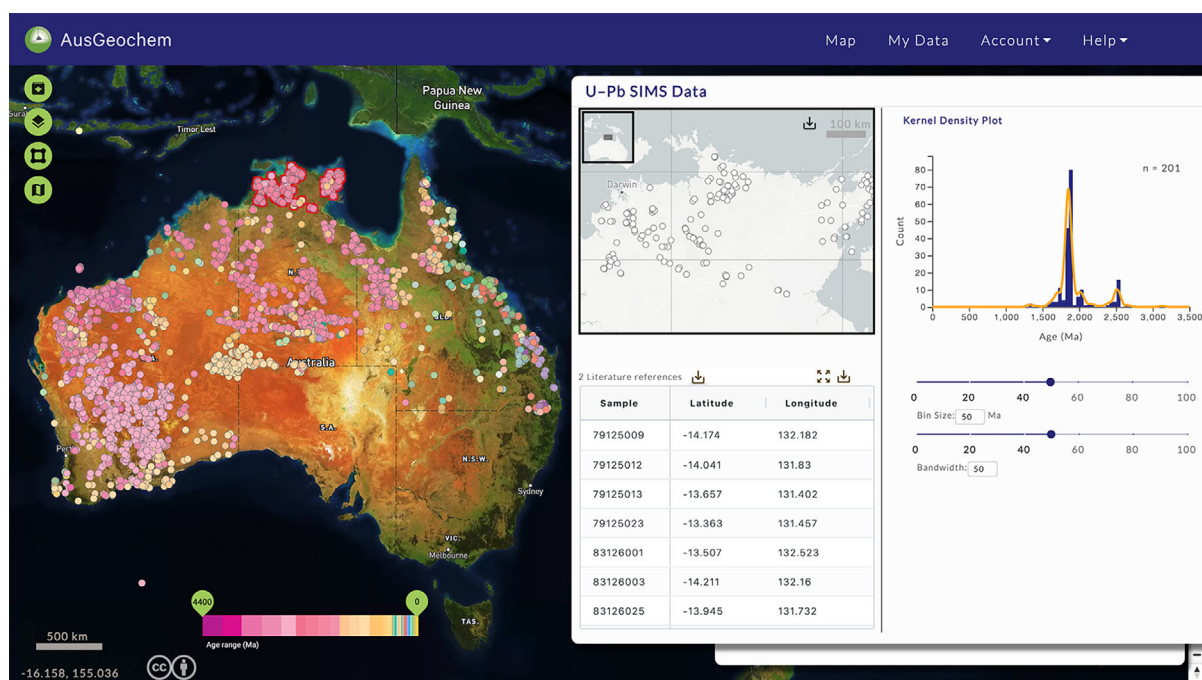


Figure 9. AusGeochem map view geospatially displaying SIMS U-Pb data from the Northern Territory of Australia.

check). Examples include deriving geopolitical information (e.g., region, country) and elevation (via digital elevation model API).

Discussion

Towards reshaping the geochemistry data ecosystem

AusGeochem provides a mechanism with which to transform the broadly sequential geochemistry data continuum to a more interconnected data nexum, centred around the open data platform (Figure 8). From there, geochemical data will be simultaneously accessible to the Institutional, Collaboration and Publication domains alike. With AusGeochem at the fulcrum of this network, detailed analytical data can be interrogated, curated, synthesised and disseminated all from within AusGeochem, removing the need for manual data migration.

Once archived in AusGeochem, huge compilations of structured scientific data from a growing variety of geochemical techniques are readily comparable from *within* the platform in a way never before possible. This wealth of open and relational geochemistry data could spur a pivotal transformation for Big Data in geoscience, enabling more efficient and robust investigations into Earth processes. Through on-the-fly data synthesis, visualisation and analysis

within AusGeochem, users can rapidly interrogate intractably large datasets comprising thousands of analyses in novel and powerful ways (Figure 9). The structured archival of reference material results acquired during analyses of unknowns will facilitate greater data quality transparency and could provide previously unavailable insights into inter-laboratory analytical variability through time.

With the development of its core architecture and the public release of AusGeochem in October 2021, future plans for the platform will focus on expanding its data capabilities while striving for greater compliance with FAIR data principles (Wilkinson *et al.* 2016). These efforts will include greater collaboration with the international geochemistry and data science communities, the development and release of more method-specific data models, implementing a function for minting DOIs for datasets and obtaining CoreTrustSeal certification to become an accredited data repository, recognised by global funding bodies and the world's leading scientific journals.

Conclusions

AusGeochem is an open relational geochemistry data platform, designed to simultaneously serve as a geosample registry, a geochemical data repository and a data analysis tool. To incentivise a FAIR data reporting culture in the geochemistry community, the platform is designed to service

individuals and groups spanning the Data Curation Continuum, from data producer to data user. AusGeochem provides geochemistry laboratories with a unique tool for improving their data management, facilitate efficient data dissemination, assist with research and big data analytics, increase the visibility of their work and promote cross-community collaboration. Structured around community-agreed (meta-)data reporting standards, and geared towards archiving data of defined processing levels, the development of AusGeochem marks an important step towards a more FAIR geochemistry data ecosystem. Users can register to explore, disseminate and extract geosample and geochemistry (meta-)data from around the globe at <https://ausgeochem.auscope.org.au>

Acknowledgements

AusGeochem development is funded by the AuScope program of the Australian National Collaborative Research Infrastructure Strategy (NCRIS). The AuScope Geochemistry Network member research groups and laboratories would also like to acknowledge the significant support received over the last many decades from AuScope, the Australian Research Council and Australian Geodynamics Cooperative Research Centre for the procurement of analytical instrumentation and production of geochemical data. We would like to thank the Australian Research Data Commons, the International Geo Sample Number e.V. Executive Board, Jolyon Ralph of Mindat.org and the Expert Advisory Group members for their on-going support and assistance during the formation of the AGN and development of AusGeochem. We kindly thank Kathryn Linge, Lesley Wyborn and one anonymous reviewer for their thoughtful and constructive feedback. Open access publishing facilitated by The University of Melbourne, as part of the Wiley – The University of Melbourne agreement via the Council of Australian University Librarians. [Correction added on 18 May 2022, after first online publication: CAUL funding statement has been added.]

Data availability statement

Data sharing is not applicable as no new data were generated.

References

AuScope (2019a)

AuScope discovery portal. Accessed 1 December 2021. <http://portal.auscope.org.au>

AuScope (2019b)

AuScope website. Accessed 1 December 2021. <https://www.auscope.org.au>

Australian National University (2020)

AuScope initiative: National argon map. Accessed 1 December 2021. <https://earthsciences.anu.edu.au/research/facilities/auscope-initiative-national-argon-map>

Baumann P., Mazzetti P., Ungar J., Barbera R., Barboni D., Beccati A., Bigagli L., Boldrini E., Bruno R., Calanducci A., Campalani P., Clements O., Dumitru A., Grant M., Herzig P., Kakaletris G., Laxton J., Koltsida P., Lipskoch K., Mahdiraji A.R., Mantovani S., Mericariu V., Messina A., Misev D., Natali S., Nativi S., Oosthoek J., Pappalardo M., Passmore J., Rossi A.P., Rundo F., Sen M., Sorbera V., Sullivan D., Torrisi M., Trovato L., Veratelli M.G. and Wagner S. (2016)

Big data analytics for earth sciences: The EarthServer approach. *International Journal of Digital Earth*, 9, 3–29.

Bodorkos S., Bowring J.F. and Rayner N.M. (2020)

Squid3: Next-generation data processing software for sensitive high resolution ion micro probe (SHRIMP). Geoscience Australia (Canberra, Australia).

Boone S.C. and Manifold P. (2021)

Meet the large collaboration network behind AusGeochem, AuScope, 16 November 2021. <https://www.auscope.org.au/posts/2021/03/29/blank-story-lsdmr-s47h8>

Chamberlain K.J., Lehnert K.A., McIntosh I.M., Morgan D.J. and Wömer G. (2021)

Time to change the data culture in geochemistry. *Nature Reviews Earth and Environment*, 2, 737–739.

Champion D.C., Budd A.R., Hazell M.S. and Sedgmen A. (2007)

OZCHEM national whole rock geochemistry dataset. Geoscience Australia.

COPDESS (2014)

Commitment statement in the Earth, space, and environmental sciences. Coalition for Publishing Data in the Earth and Space Sciences. <https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and-environmental-sciences/>

Cropper J. and Sweeney M. (2021)

Community and education data portal user guide. Geoscience Australia (Canberra).

De Caritat P. and Cooper M. (2016)

A continental-scale geochemical atlas for resource exploration and environmental management: The National Geochemical Survey of Australia. *Geochemistry: Exploration Environment Analysis*, 16, 3–13.

Eglington B.M. (2004)

DateView: A windows geochronology database. *Computers and Geosciences*, 30, 847–858.



references

- Federer L.M., Belter C.W., Joubert D.J., Livinski A., Lu Y.L., Snyders L.N. and Thompson H. (2018)**
Data sharing in PLoS ONE: An analysis of data availability statements. *PLoS One*, 13, e0194768.
- Geoscience Australia and Australian Stratigraphy Commission (2017)**
Australian stratigraphic units database.
- Gleadow A.J., Kohn B.P., Brown R.W., O'Sullivan P.B. and Raza A. (2002)**
Fission track thermotectonic imaging of the Australian continent. *Tectonophysics*, 349, 5–21.
- Government of South Australia (2021)**
South Australian resources information gateway (SARIG). Accessed 1 December 2021. <https://map.sarig.sa.gov.au/>
- Government of Western Australia (2013)**
GeoVIEW.WA – Interactive geological map. Department of Mines, Industry Regulation and Safety. Accessed 1 December 2021. <https://www.dmp.wa.gov.au/GeoView-WA-Interactive-1467.aspx>
- Hazell M., Kilgour B., Wyborn L.A.I., Sheraton J.W. and Ryburn R. (1995)**
ROCKCHEM dataset version 2 documentation. Australian Geological Survey Organisation Record, 26.
- He Y., Bai Y., Tian D., Yao L., Fan R. and Chen P. (2019)**
A review of geoanalytical databases. *Acta Geochimica*, 38, 718–733.
- Horstwood M.S.A., Kosler J., Gehrels G., Jackson S.E., McLean N.M., Paton C., Pearson N.J., Sircombe K., Sylvester P., Vermeesch P. and Bowring J.F. (2016)**
Community-derived standards for LA-ICP-MS U-(Th-) Pb geochronology – Uncertainty propagation, age interpretation and data reporting. *Geostandards and Geoanalytical Research*, 40, 311–332.
- Jochum K.P., Nohl U., Herwig K., Lammel E., Stoll B. and Hofmann A.W. (2005)**
GeoReM: A new geochemical database for reference materials and isotopic standards. *Geostandards and Geoanalytical Research*, 29, 333–338.
- Klump J., Lehnert K., Ulbricht D., Devaraju A., Elger K., Fleischer D., Ramdeen S. and Wyborn L. (2021)**
Towards globally unique identification of physical samples: Governance and technical implementation of the IGSN global sample number. *Data Science Journal*, 20, 33.
- Kohn B.P., Gleadow A.J.W., Brown R.W., Gallagher K., O'Sullivan P.B. and Foster D.A. (2002)**
Shaping the Australian crust over the last 300 million years: Insights from fission track thermotectonic imaging and denudation studies of key terranes. *Australian Journal of Earth Sciences*, 49, 697–717.
- Lehnert K.A., Klump J., Arko R.A., Bristol S., Buczkowski B., Chan C., Chan S., Conze R., Cox S.J., Habermann T. and Hangsterfer A. (2011)**
IGSN eV: Registration and identification services for physical samples in the digital universe. AGUFM 2011, IN13B-1324.
- Lehnert K., Klump J., Wyborn L. and Ramdeen S. (2019)**
IGSN: Trustworthy and sustainable services for FAIR samples. *Geophysical Research Abstracts*, 21, 1.
- Lehnert K., Walker J.D., Carlson R.W., Hofmann A.W. and Sarbas B. (2004)**
Building the EarthChem system for advanced data management in igneous geochemistry. AGUFM 2004, SF41A-0767.
- Lithodat (2021a)**
Lithodat website. Accessed 1 December 2021. <https://lithodat.com>
- Lithodat (2021b)**
Lithosurfer data platform. Accessed 1 December 2021. <https://app.lithodat.com>
- Liu R.M., Xuan W.U., Xiang Y.C. and Geng Y.T. (2012)**
China national multi-purpose geochemical database development and application prospect. *Geoscience*, 26, 989–995.
- Mindat (2021)**
Mindat.org. Accessed 1 December 2021. <https://www.Mindat.org>
- NASA (2021)**
Data processing levels. Accessed 1 December 2021. <https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy/data-levels>
- ORCID (2012)**
ORCID – Connecting research and researchers. Accessed 1 December 2021. <https://orcid.org>
- Pierce H.H., Dev A., Statham E. and Bierer B.E. (2019)**
Credit data generators for data reuse. *Nature*, 570, 30–32.
- Rollinson H.R. (2014)**
Using geochemical data: Evaluation, presentation, interpretation. Routledge (London), 384pp.
- Schaen A.J., Jicha B.R., Hodges K.V., Vermeesch P., Stelten M.E., Mercer C.M., Phillips D., Rivera T.A., Jourdan F., Matchan E.L., Hemming S.R., Morgan L.E., Kelley S.P., Cassata W.S., Heizler M.T., Vasconcelos P.M., Benowitz J.A., Koppers A.A.P., Mark D.F., Niespolo E.M., Sprain C.J., Hames W.E., Kuiper K.F., Turrin B.D., Renne P.R., Ross J., Nomade S., Guillou H., Webb L.E., Cohen B.A., Calvert A.T., Joyce N., Ganerød M., Wijbrans J., Ishizuka O., He H., Ramirez A., Pfänder J.A., Lopez-Martínez M., Qiu H. and Singer B.S. (2021)**
Interpreting and reporting $^{40}\text{Ar}/^{39}\text{Ar}$ geochronologic data. *Geological Society of America Bulletin*, 133, 461–487.
- Sherratt T. (2013)**
From portals to platforms – Building new frameworks for user engagement. IANZA 2013.
- Siegel C., Bryan S.E., Purdy D., Gust D., Allen C., Uysal T. and Champion D. (2012)**
A new database compilation of whole-rock chemical and geochronological data of igneous rocks in Queensland: A new resource for HDR geothermal resource exploration. *Proceedings of the 2011 Australian Geothermal Energy Conference (Geoscience Australia)*, 239–244.

references

Stall S., Robinson E., Wyborn L., Yarmey L.R., Parsons M.A., Lehnert K., Cutcher-Gershenfeld J., Nosek B. and Hanson B. (2017)

Enabling FAIR data across the Earth and space sciences. *Eos*, 98.

Stall S., Yarmey L., Boehm R., Cousijn H., Cruse P., Cutcher-Gershenfeld J., Dasler R., de Waard A., Duerr R., Elger K., Fenner M., Glaves H., Hanson B., Hausman J., Heber J., Hills D., Hoebelheinrich N., Hou S., Kinkade D., Koskela R., Martin R., Lehnert K., Murphy F., Nosek B., Parsons M., Petters J., Plante R., Robinson E., Samors R., Servilla M., Ulrich R., Witt M. and Wyborn L. (2018)

Advancing FAIR data in Earth, space, and environmental science. *Eos*, 99.

Stall S., Yarmey L., Cutcher-Gershenfeld J., Hanson B., Lehnert K., Nosek B., Parsons M., Robinson E. and Wyborn L. (2019)

Make scientific data FAIR. *Nature*, 570, 27–29.

Strong D.T., Turnbull R.E., Haubrock S. and Mortimer N. (2016)

Petlab: New Zealand's national rock catalogue and geoanalytical database. *New Zealand Journal of Geology and Geophysics*, 59, 475–481.

Treloar A., Groenewegen D., and Harboe-Ree C. (2007)

The data curation continuum. *D-Lib Magazine*, 13, 1082–9873.

Treloar A. and Klump J. (2019)

Updating the data curation continuum. *International Journal of Digital Curation*, 14, 87–101.

Walker D., Renne P., Deino A., Koppers A. and Hodges K. (2008)

EarthChem workshop on geochronology for Ar-Ar. Geochron Workshop reports sponsored by EarthChem and EARTHTIME. <https://doi.org/10.5281/zenodo.4313859>

Walker J.D., Bowers T.D., Black R.A., Glazner A.F., Farmer G.L., Carlson R.W. and Sinha A.K. (2006)

A geochemical database for western North American volcanic and intrusive rocks (NAVDAT). *Geological Society of America Special Paper*, 397, 61.

Walker J.D., Bowers T.D., Glazner A.F., Farmer A.L. and Carlson R.W. (2004)

Creation of a North American volcanic and plutonic rock database (NAVDAT). *Geological Society of America Abstracts with Programs*, 36, 9.

Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L.B., Bourne P.E., Bouwman J., Brookes A.J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C.T., Finkers R., Gonzalez-Beltran A., Gray A.J.G., Groth P., Goble C., Grethe J.S., Heringa J., † Hoen P.A.C., Hooff R., Kuhn T., Kok R., Kok J., Lusher S.J., Martone M.E., Mons A., Packer A.L., Persson B., Rocca-Serra P., Roos M., van Schaik R., Sansone S.-A., Schultes E., Sengstag T., Slater T., Strawn G., Swertz M.A., Thompson M., van der Lei J., van Mulligen E., Velterop J., Waagmeester A., Wittenburg P., Wolstencroft K., Zhao J. and Mons B. (2016)

The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 1–9.

