Assessment Information

CoreTrustSeal Requirements 2020–2025

Repository:	FaceBase
Website:	https://www.facebase.org/
Certification period:	March 26, 2025 - 25 March 2028
Requirements version:	CoreTrustSeal Requirements 2023-2025
This repository is owned by:	**FaceBase**

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Re3data Identifier

Please fill you Re3data identifier from the website: https://www.re3data.org/

Response:

https://www.re3data.org/repository/r3d100013263
http://doi.org/10.17616/R31NJMQ7

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Repository type

Please select your repository type.

Response:

- Specialist repository

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Overview

Provide a short overview of key characteristics of the repository, reflecting the repository type selected. This should include information about the scope and size of data collections, data types and formats. Further contextual information may also be added.

Response:

FaceBase was established in 2009 by the National Institute of Dental and Craniofacial Research (NIDCR) of the United States National Institutes of Health (NIH) with the goal of enabling the craniofacial research community to share and reuse research data. FaceBase is committed to FAIR data practices and provides access to various types of sequencing, array, 2- and 3-D imaging, and other data related to human, mouse, and zebrafish craniofacial development, as well as a web interface with detailed search and visualization capabilities and REST and Python interfaces.

FaceBase currently houses more than 1000 data sets, reflecting data describing more than 2800 experiments and 13,000 biosamples, with files containing more than 17 TB of data.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Designated Community

A clear definition of the Designated Community demonstrates that the applicant understands the scope, knowledge base, and methodologies—including preferred software/formats—of the group(s) of users at whom the curation and preservation measures are primarily targeted. The definition should be specific so that reviewers can assess whether that community is being served in the responses to other requirements.

Response:

FaceBase serves researchers and students involved in research across the full translational science spectrum of dental, oral, and craniofacial (DOC) research, as well as research into biologically and/or anatomically associated health and disease areas. FaceBase is funded and overseen by the National Institute of Dental and Craniofacial Research. Several members of FaceBase staff (including one of its two co-PIs) are members of the craniofacial research community. We work with a scientific advisory board of dental and craniofacial experts and hold annual community forums, boot camps, and monthly office hours open to anyone in the community to ensure that we stay abreast of the community's needs.

Materials from community forums and annual meetings from 2015 to the present can be found at https://www.facebase.org/community/annual-meeting/.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Levels of Curation

Please fill you level(s) of curation.

Response:

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

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Levels of Curation - explanation

Please add the description for your Level(s) of Curation.

Response:

Contributors initially ask for approval to contribute their data to FaceBase. Those requests are reviewed by a committee of FaceBase and NIDCR staff to determine whether the data is within scope of the repository, whether all required institutional approvals (e.g., IRBs) are met, etc.

After receiving approval, the contributor performs their initial data and metadata upload. We have separate approval processes for controlled-access data (protected human subjects data) and open-access data (everything else).

If their data is open-access, the contributor uploads all relevant data and metadata to the FaceBase public server. If the data is controlled-access, they upload the open-access portion of their metadata to the FaceBase public server for curation, and their data and any controlled-access metadata to a separate, secure server. All open-access data is curated to level D; because most curation occurs on the open-access server, controlled-access data may be curated to level B/C or D, depending on how much metadata the contributor is able to make public. As of this writing, we have 989 open-access

data sets and 61 controlled-access data sets.

The open-access curation process is as follows: As metadata is entered, foreign keys and required fields ensure that some of our standards (e.g., minimum metadata requirements, use of supported ontologies) are met. Separate automated processes perform more in-depth tests for incomplete (meta)data; once this is done, then either curation level C has been done or issues preventing curation to that level have been flagged. Human biocurators then review all submissions and work with the contributors to ensure that the data and metadata are complete and consistent with FaceBase standards. When that process is completed, curation level D has been achieved and the data is made available to the public (either immediately or after an embargo period related to a publication).

Because we don't run our complete software stack on our controlled-access server, less automated curation is performed. There are some basic automated checks to verify some amount of metadata completeness, that there's a metadata entry for each data file and vice versa, and there is some human examination of the metadata, but not to the same extent as on our open-access server. This process results in curation level B or C, depending on what metadata has been provided.

The underlying systems for the open-access data store and metadata catalog are versioned; the initial deposits of data and metadata, as well as any updates or edits, can always be retrieved. Data deposited to the controlled-access server are typically not modified; in the rare occasions where modification is necessary, the original data files are preserved.

Until data curation has been completed to the satisfaction of both FaceBase and the data contributor, data and metadata on are accessible only to the contributor (and possibly additional members of the contributor's team) and FaceBase staff.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Cooperation and outsourcing to third parties, partners and host organisations

Please describe any cooperation and outsourcing to third parties, partners and host organisations.

Response:

FaceBase is operated and managed by faculty and staff affiliated with the University of Southern California (USC) through a partnership between the USC Center for Craniofacial and Molecular Biology (CCMB) and the USC Information Sciences Institute (ISI).

Our open-access metadata server and data repository are hosted on Amazon Web Services (AWS). We are currently using Amazon Elastic Compute Cloud (EC2) and Amazon Elastic Block Store (EBS) for the computing and storage for the metadata server; Simple Storage Service (S3) for data repository storage (and backups of our EBS data). Our controlled-access server is also hosted on AWS, using EC2 and Amazon Elastic File System (EFS) for computing and storage.

Decisions regarding whether or not to include a potential data contribution in FaceBase are made in consultation with NIH (in general, regarding questions about whether the data falls within the scope of FaceBase and, for human data, whether consents gathered from subjects are compatible with sharing on FaceBase). Decisions regarding whether to approve access to controlled-access FaceBase data are made by the FaceBase Data Access Committee, which is comprised of representatives from NIH.

Our data and servers fall under the following service level agreements with Amazon:

EC2: https://aws.amazon.com/compute/sla/
EBS: https://aws.amazon.com/ebs/sla/
EFS: https://aws.amazon.com/efs/sla/
S3: https://aws.amazon.com/s3/sla/

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Applicants renewing their CoreTrustSeal certification: summary of significant changes since last application.

Please fill this field when you are renewing your CoreTrustSeal Certification.
This field can be marked with not applicable (N.A.) if you are acquiring a CoreTrustSeal certificate for the first time.

Response:

Updated "Community" to reflect changes related to our latest grant.
Updated "Levels of Curation - explanation" to show which levels are achieved at each step of our curation process.
Updated "Cooperation and outsourcing to third parties, partners, and host organizations" to include our relationship with NIH.
R03: Added information about disaster recovery procedures and about USC CCMB's commitment to house controlled-access data for at least five years.
R05: Added information about FaceBase's new 5-year renewal and 2024-2025 budget.
R06: Added detail about FaceBase communication with NIH and external advisors and about career development education for FaceBase staff. Also added information about a new co-investigator.
R07: Added detail about versioning.
R09: Added more detail about data/metadata format issues.
R10: Added link to data curation rules and more information about human biocuration.
R13: Added detail about community engagement.
R14: Added information about when data/metadata are deleted and how data are stored redundantly and integrity checked.
R16: Added clarification about what services are (and aren't) available on controlled-access servers; linked additional documentation.

Reviews

Reviewer 1:

Comments:

Reviewer 2:

Comments:

Organisational Infrastructure

R1 Mission & Scope (R01)

R01. The repository has an explicit mission to provide access to and preserve digital objects.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

One of the main goals of the FaceBase project is to provide a FAIR, comprehensive repository of craniofacial research data. FaceBase has been collecting, curating, preserving, and providing access to this data since 2009.

Links:

- FaceBase III grant
- FaceBase "about" page

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R2 Rights Management (R02)

R02. The repository maintains all applicable rights and monitors compliance.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

FaceBase data falls into two categories: controlled-access (containing identifiable human data) and open-access (all other data); this answer will address the two types of data separately.

Once released (after curation has completed and by the time it appears in a publication), open-access data is world-readable and covered by the FaceBase Terms of Use. Every metadata record in the FaceBase web portal contains a link to these terms of use. To summarize, the terms of use state that:

- Copyright to data and images on FaceBase are held by the individual contributors (each data record includes the identity of each contributor).
- Data users must acknowledge both the contributor and the FaceBase project (the web page for each data record includes a "share and cite" button, which includes versioned and unversioned permanent identifiers, as well as plain-text and BibTex citations).
- Users must not re-circulate original datasets to other users, but rather refer them to the original record on the FaceBase website.
- Users should inform FaceBase of any publications resulting from the use of FaceBase data (those publications will then be referenced on the website).

Controlled-access data is covered by a more restrictive access policy. In order to gain access to this data, users must submit an application that will be reviewed by our Data Access Committee, consisting of members from FaceBase and NIH. As part of the application process, the potential user must agree to the terms of our Data Use Certificate Agreement (DUC), submit a letter from the requestor's IRB approving this use and a copy of their IRB-approved protocol.

Under the terms of the DUC, access is granted for a period of one year, at the end of which the user must either reapply or submit a final report and attest to destruction of the data. The DUC also requires that the user notify FaceBase in the event of a data breach.

Anyone wishing to submit data to FaceBase must first fill out a form that includes an attestation that they have the right to share the data with FaceBase and that they agree to allow FaceBase to share data under the FaceBase terms of use.

Links:

- FaceBase data submission form
- A longer explanation of the process for applying for access to restricted data
- FaceBase Data Use Certification Agreement
- FaceBase Terms of Use for Open Data

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R3 Continuity of Service (R03)

R03. The Repository has a plan to ensure ongoing access to and preservation of its data and metadata.

Compliance level:

In Progress: the repository is in the implementation phase - 0

Response:

FaceBase has been in operation since 2009 and is funded by the NIH, which is committed to FAIR and TRUST principals, and FaceBase is listed as an NIH-supported repository for craniofacial data. Although we have no formal succession agreement, the underlying structure (well-documented open-source software, cloud hosting, an open data schema) could ease the transition of our open-access data and metadata services. In the worst-case scenario where funding ended abruptly with no successor identified, our plan would be to bundle our existing open-access data and metadata, deposit them on a general-purpose repository, and update our permanent DataCite-generated DOIs with that information.

Disaster recovery: FaceBase is hosted on Amazon AWS EC2 instances, with data on EBS, S3, and EFS disks. We have well-documented procedures for bringing up a FaceBase deployment on a new EC2 instance, and we use standard mechanisms for backup and restore of our data.

If a successor for our controlled-access data is identified and all necessary agreements are in place, our cloud-hosting infrastructure and open data schema could facilitate that transition. Otherwise, we are committed to making a best effort to retain the data for at least five years. If no other successor can be found, we have a commitment from the University of Southern California's Center for Craniofacial and Molecular Biology to house the FaceBase controlled-access data in their secure locked facility.

Links:

- NIH statement on FAIR and TRUST principles

Reviews

Reviewer 1:

Compliance level:

In Progress: the repository is in the implementation phase - 0

Comments:

Reviewer 2:

Compliance level:

In Progress: the repository is in the implementation phase - 0

Comments:

R4 Legal & Ethical (R04)

R04. The repository ensures to the extent possible that data and metadata are created, curated, preserved, accessed and used in compliance with legal and ethical norms.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

Potential contributors of human subjects data are required to submit an institutional certification from an IRB recognized by the US Department of Health and Human Services. This and other relevant information are reviewed by a panel of data scientists, biocurators, and representatives from NIDCR before approval decisions are made. The panel, in consultation with the data contributor, will also determine any data use limitations for the data, which will then be recorded in our metadata catalogue and used when reviewing requests for access to this data.

All FaceBase staff with access to human subjects data have received human subjects training and are listed on the FaceBase IRB.

For non-human vertebrate animal data, we require contributors to attest that the data was collected with the approval of an institutional IACUC. FaceBase does not host any sensitive animal data (e.g., locations of endangered species).

Links:

- FaceBase data submission process

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R5 Governance & Resources (R05)

R05. The repository has adequate funding and sufficient numbers of staff managed through a clear system of governance to effectively carry out the mission.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

FaceBase has just begun its fourth five-year round of funding. It is funded by NIDCR and hosted at the University of Southern California; the current year's budget (2024-2025) is $2,500,000

FaceBase is organized into four teams, with some overlap of personnel between teams. These teams are Biological Experts, Data Integration and Management, Development and Operations, and Coordination and Collaboration. The Biological Experts take the lead on curation and provide input concerning quality control, data visualization, and data analysis. The Data Integration and Management team ensures proper storage of human subjects data and design improvements to the search/browsing experience and integration among different data types. The Development and Operations team takes primary responsibility for site maintenance and improvement. Finally, the Coordination and Collaboration team manages communications, outreach activities, and training opportunities. All of these efforts are guided by ongoing conversations with NIDCR program staff, a Scientific Advisory Group of domain experts in craniofacial biology, the craniofacial community and FaceBase user pool more broadly, and ad hoc working groups formed to advise the FaceBase team on how best to handle particular data types or other challenges that may arise.

FaceBase staff members and their roles are listed in the "About" page linked below.

Links:

- FaceBase award for 2024-2025
- FaceBase award for 2023-2024 fiscal year
- FaceBase "about" page

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R6 Expertise & Guidance (R06)

R06. The repository adopts mechanisms to secure ongoing expertise, guidance and feedback-either in-house, or external.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

FaceBase staff have extensive expertise both in information systems and in craniofacial biology.

Carl Kesselman (co-PI) is a William H. Keck Professor of Engineering in the USC Viterbi School of Engineering, has received numerous honors for his pioneering research including the Lovelace Medal from the British Computer Society and the Goode Memorial Award from the IEEE Computer Society, and is a Fellow of the British Computer Society and the Association for Computing Machinery. He leads the Information Sciences Institute's Informatics Systems Research division (ISRD), the work of which spans grid computing, information security, service-oriented architectures, and sociotechnical systems and reproducibility.

Yang Chai (co-PI) is the University Professor and the George and MaryLou Boone Chair in Craniofacial Biology at the University of Southern California. He serves as the Director of the Center for Craniofacial Molecular Biology (CCMB) and Associate Dean of Research at the Herman Ostrow School of Dentistry of USC. is a member of the National Academy of Medicine. He is an elected member of the American Academy of Arts and Sciences (AAAS). Dr. Chai has been continuously funded by the National Institutes of Health for more than 25 years. His work has earned him multiple awards including the 2011 IADR (International Association of Dental Research) Distinguished Scientist Award.

Dr. Parish P. Sedghizadeh (co-investigator) is a Professor of Clinical Dentistry at the University of Southern California, Herman Ostrow School of Dentistry, and Department Co-Chair of Diagnostic Sciences, Anesthesia & Emergency Medicine. As a clinician-scientist, Dr. Sedghizadeh conducts research, publishes, consults, and teaches oral and maxillofacial pathology, radiology, and medicine with an active intramural clinical practice. Dr. Sedghizadeh has over 85 peer-reviewed publications, and his research laboratory and clinical research projects at USC focus on the characterization and treatment of microbial biofilm infections, particularly osteomyelitis and osteonecrosis where he has developed novel bone-targeted antimicrobial therapeutics.

In-house staff provide additional expertise and experience in these two fields, and NIH staff and an external scientific advisory board are available for consultation. FaceBase and NIH staff meet biweekly; external advisors are consulted on an ad-hoc basis and more formally at the annual FaceBase Community Forum.

Career development education is primarily accomplished through a combination of self-study, lecture series, conference participation, and informal mentorships.

A complete list of FaceBase staff, advisory board members, and NIH personnel associated with the project are available on our "About" page.

Links:

- FaceBase "about" page

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Digital Object Management

R7 Provenance and authenticity (R07)

R07. The repository guarantees the authenticity of the digital objects and provides provenance information.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

Authentication for FaceBase open-access data (required for any operation that creates, modifies, or deletes data or metadata) is performed via OpenID Connect with the Globus Auth federated identity provider.

Each FaceBase metadata record includes system-generated fields that include the date/time that the record was created, the date/time that the record was last modified, and the identities of the users who created and last modified the record. In addition, the metadata records are versioned, so the complete history of changes to a record (including who made each change, and when) can be determined by examining all versions of that record. The complete history information is maintained in the metadata catalog (and is backed up nightly).

Each record can be reached through an unversioned REST URL (which returns the latest version); each version can be reached through a versioned REST URL. Similarly, the web user interface supports unversioned views of each record, as well as permalinks to both the unversioned view and the version being displayed.

There is currently no user interface or documentation for walking back through each version of a record, although FaceBase staff can help users who request it.

Open-access FaceBase data is stored in a versioned AWS S3 data store. Typically, a data update would be accompanied by a metadata update (to indicate the new version ID, size, and checksum of the updated data), and data changes would be reflected in metadata changes. In addition, data operations could be checked directly, as they are performed by REST requests, which are logged to AWS CloudTrail.

For controlled-access data, access is limited to a small number of approved FaceBase staff, and actions are tracked through github issues. System logs are preserved via AWS CloudTrail.

Links:

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R8 Deposit & Appraisal (R08)

R08. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for users.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

Prior to submitting data, potential contributors to FaceBase fill out an online form that asks for some basic information about the amount and type of data and the project for which it was collected. If this initial information indicates that the data is compatible with FaceBase's mission, a meeting is set up between the potential contributor and FaceBase staff. At that meeting, we make sure that the data is a format accepted by FaceBase, sufficient metadata is available, and the metadata can be expressed using ontologies accepted by FaceBase.

If these criteria are not met, then possible outcomes are that FaceBase will change its requirements (for example, some ontologies are species-specific; if the new data is from a new species, then we may add ontologies for that species) or the user will convert their data or metadata. If neither of those approaches work, however, the data will not be accepted.

After all details have been worked out, the contributor is granted access to upload their open-access data and metadata through the FaceBase web interface (for controlled-access human subjects data, the contributor uploads only the open-access portion of their metadata). During that process, foreign key constraints ensure that all metadata are expressed in FaceBase-accepted ontologies, and required fields contribute towards ensuring that all required metadata are entered for all uploaded data files.

Controlled-access data and the controlled-access portion of the associated metadata are uploaded through a separate process to our controlled-access server. This server does not run our full software stack and does not automatically enforce the same level of constraints; however, automated processes do check that each data file has a corresponding metadata record, and a biocurator performs some additional manual checks.

At the time of the initial upload, data and metadata are only accessible by the uploader (and possibly members of their team) and FaceBase staff. If, at the end of the curation process, the metadata is insufficient for long-term preservation, then the data will remain private -- however, in practice, those issues are ironed out before the upload process begins.

Supported data formats include:

Raw sequencing data: “raw” sequencing data (fastq files). These must be gzipped and use the ‘.fastq.gz’ file extension. If you use the common naming scheme to indicate the sequence read number, ‘example_1.fastq.gz’ or ‘example_R2.fastq.gz’, the system will automatically extract the read number from the file name.

Processed sequencing data: data that are derived from sequencing data through a particular pipeline. Usually fastqc reports (.fastqc.tgz or .fastqc.zip), count files (.count, .tpm, .fpkm), measures in tab-separated format (.tsv), and of course alignment mapping files (.bam) and indexes (.bam.bai).

Track data: data that are derived from sequencing or processed data and used in genome browsers, such as BED (.bed), bigBed (.bb), and bigWig (.bw) files.

Raw microarray data (CEL files).

Imaging Data: high-resolution 3D or 2D imaging data, such as micro-CT accepted in NIfTI format gzipped (.nii.gz), confocal or other microscopy sources in TIFF or OME-TIFF (.tiff or .ome.tiff), and other sources in JPEG (.jpg or .jpeg).

Surface Model / Mesh Data: 3D surface models (a.k.a., “polygon mesh” files) that are generally derived from hard tissue imaging data. Currently, we only accept Wavefront OBJ format and it must be gzipped (.obj.gz).

Supported ontologies include UBERON for anatomy and sex, HGNC, MGI, and ZFIN for chromatin modifiers, OBI for experiment types, NCBI (entrez) for genes, HPO and MP for phenotype, NCBI Taxon for species, MGI for strain, MONDO for syndrome, ZFIN, HGNC, and MGI for transcription factor. We accept several reference genomes for each species (e.g., hg19, hg38, mm10, mm39). A more complete list can be found in the "Key Concepts for Data Submitters" page linked below.

Links:

- Key Concepts for Data Submitters (includes supported data types)

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R9 Preservation plan (R09)

R09. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance level:

In Progress: the repository is in the implementation phase - 0

Response:

Our approach to long-term preservation (beyond the bit-level preservation strategies discussed in R14) is to, whenever possible, require that data be deposited in open, widely-used formats for which there is widely-used open-source software available. We use well-established and widely-used external metadata ontologies whenever possible. Our supported file formats and ontologies are listed in R08. Our metadata schema can be retrieved via a simple REST request. Our software stack is open-source and built on open-source frameworks.

We are constantly in contact with our user community regarding their requirements and practices for data formats and metadata ontologies. When community standards change or when a data format becomes obsolete, we will adopt a new data/metadata standard and do one of the following for existing data/metadata:
- Convert metadata to the new ontology, in consultation with experts (and possibly with the data contributor).
- Convert data to the new data standard, in consultation with experts (and possibly with the data contributor).
- Leave the existing data as-is, but extract as much as possible into a derived data object in a usable format.

Which of these options we choose is dependent on the nature of the change (i.e., how much effort is required for a conversion), how relevant the existing data is to present-day researchers, and how relevant it's expected to be in the future.

Permanent identifiers are never reused, even if the underlying data or metadata are removed. Data/metadata removal is very rare in FaceBase, but if an entry is removed, the versioned forms of its identifier continue to refer to the same versions. If a bare REST query is performed using the unversioned form of a permanent identifier, then an empty result set will be returned; however, additional queries can find previous versions (and, for example, our web interface will display the most recent version with a message explaining that the element has since been deleted).

We are currently working on posting a webpage with a more formal, detailed plan.

Links:

Reviews

Reviewer 1:

Compliance level:

In Progress: the repository is in the implementation phase - 0

Comments:

Reviewer 2:

Compliance level:

In Progress: the repository is in the implementation phase - 0

Comments:

R10 Quality Assurance (R10)

R10. The repository addresses technical quality and standards compliance, and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

FaceBase maintains standards for metadata elements required, ontologies, and data types. We update these periodically as needs arise, in consultation with subject-matter experts. In general, we try to ensure that there is enough metadata to effectively understand and use the data and that data is in a widely-used open format (or, in rare cases where no such format exists for a data type, a format that is widely-used and for which open-source tools exist). Whenever possible, we use widely-adopted standard ontologies, but where that's not possible, we generate our own ontologies rather than simply allowing free-text input.

Quality assurance is assured through a combination of automated features (constraints affecting data entry and automated nightly processes that flag possible issues) and human-based biocuration. When an issue is flagged by an automated process or found by a biocurator, we work with the data contributor to find a solution. Often that consists of adding additional metadata or adding detail to a project description. This is a collaborative process between FaceBase and the data contributor; no changes are made without the contributor's approval, and data are not made public until both the FaceBase team and the contributor agree.

Proposed changes to schemas, required metadata elements, accepted ontologies, or data formats are discussed during weekly FaceBase hub meetings and biweekly steering committee meetings.

A more detailed description of our quality assurance process and lists of accepted data types and ontologies can be found in R08 (adherence to our ontology requirements is enforced at the time metadata is entered, through foreign key constraints). The linked "Key Concepts" document includes detailed instructions for submitters, including required metadata elements and quality control rules enforced by our automated processes (specifically, the seven sections from "Key Concepts for Data Contributors" through "Releasing Your Dataset").

While there is no documentation for the QC done by the human biocurators, the process includes close reading of free-text project descriptions and working with the data contributor to clear up any ambiguity or inconsistencies, comparing the structured metadata to free-text descriptions and (again, in consultation with the data contributor) resolving any apparent inconsistencies, and so on.

Links:

- FaceBase Quality Control Rules
- Key Concepts for FaceBase Data Submitters

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R11 Workflows (R11)

R11. Digital object management takes place according to defined workflows from deposit to access.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

As discussed in R8, the data submission process starts when the potential contributor fills out a form describing their project and data. The approval process for open-access data (data that is not protected human subjects data) is discussed in R8; the approval process for controlled-access (protected human subjects) data is discussed in R8 and R4.

The progress of each data submission request is tracked internally via github issues.

Once a data submission is approved and FaceBase staff have met with the potential contributor to iron out any data format or metadata issues, the contributor creates a FaceBase account (if they don't have one already) and the FaceBase team sets up a group for the contributor's project and authorizes that group to create a dataset. The user logs in, creates a dataset record, and starts adding metadata (this is true regardless of whether the data consists of protected human data or not, although no metadata that would itself reveal protected human data is uploaded).

For open-access data, the contributor enters all relevant metadata on our public server and uploads their data using our web interface or other tools. At the time the metadata is entered, a relational database enforces basic structural quality using foreign keys, uniqueness constraints, and not-null constraints (a high-level diagram of the data model is linked below - this diagram does not show ancillary tables, such as vocabularies, that are used to constrain metadata values).

Additional requirements are checked by an automated process that runs nightly (see "Quality control rules" linked below).

Human biocurators and data scientists review metadata and work with contributors to clarify language, resolve any discrepancies between metadata and textual descriptions, add additional links when appropriate, etc. No changes are made without the contributor's approval.

Open-access data and metadata, as well as the open-access portion of metadata associated with controlled-access data, are uploaded directly by the contributor, using the FaceBase web interface or other tools. Controlled-access data and metadata are encrypted by the client and uploaded to a separate, protected server (see R16 for details).

OME-TIFF versions of imaging files are generated nightly to accommodate our image viewer application (the original files remain in the repository, available for download).

For studies involving controlled-access data, the contributor enters as much relevant metadata as possible on our open-access server (at a minimum, a description of the study involved, the principal investigator, assay types used, any syndromes being studied, etc.) and encrypts and uploads additional metadata and data to our controlled-access server (see R16 for details). For the open-access portions, automated and human biocuration processes proceed as described above. For the controlled-access portions, automated processes perform some consistency checking (but not as extensive as for unrestricted data), and a human biocurator performs some additional checks. Because we run a minimal set of software on our restricted server, we perform less automated checking there and don't run restricted data through our imaging and bioinformatics pipelines.

Change management is tracked through github.

Finally, when the contributor and FaceBase both agree that the data and metadata are sufficient, the unrestricted metadata and data are made public. For open-access data, this means that anyone can browse the metadata and download the data. For controlled-access data, this means that anyone can browse the public portion of the metadata and request the data; this request then follows the approval process described in R2.

Links:

- Quality control rules enforced by our automated nightly process
- Data submission process

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R12 Discovery and Identification (R12)

R12. The repository enables users to discover the digital objects and refer to them in a persistent way through proper citation.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

FaceBase uses DERIVA as its underlying service framework. Most users search the FaceBase catalog using the web interface, which provides faceted searches based on metadata elements (for example, users can search for datasets based on experiment type, species, gene, developmental stage, anatomy, phenotype, syndrome, investigator, etc.) or combinations of those elements. Users can also search the catalog directly using a REST or Python interface or use those interfaces to retrieve all or part of the catalog schema.

Each data record (of any data type) is assigned a persistent unique ID by the metadata catalog; these IDs are versioned; a bare id will always refer to the latest version of the record, while an id with a version suffix will refer to a specific version. In addition, datasets are assigned DataCite DOIs. We recommend that users use DOIs when citing datasets.

The web page for each data element includes a "share and cite" button that displays versioned and unversioned identifiers, as well as plain-text and BibTex citation information.

FaceBase is included in NIH's list of supported repositories, re3data, FAIRsharing, and Google datasets.

Links:

- DERIVA documentation
- Google dataset entry for FaceBase
- FAIRsharing entry for FaceBase

- re3data entry for FaceBase
- NIH list of supported repositories

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R13 Reuse (R13)

R13. The repository enables reuse of the digital objects over time, ensuring that appropriate information is available to support understanding and use.

Compliance level:

In Progress: the repository is in the implementation phase - 0

Response:

FaceBase maintains an extensive set of metadata and data format standards, developed and updated through consultation with our scientific advisors and community members. Whenever possible, we require data formats that are open standards and for which free and widely used tools are available. Similarly, we have adopted external ontologies for most metadata elements. When there is no suitable external ontology, we will create managed term lists rather than allowing widespread free-text metadata entry. We will sometimes replace one of our internally-managed term lists. In that case, our curators will update any affected existing metadata. Creating term lists, adopting new ontologies, and adjusting existing metadata are always done in close consultation with subject-matter experts in the project or in the larger community.

Supported data formats are listed in the linked "Key Concepts" document. Supported ontologies include UBERON for anatomy and sex, HGNC, MGI, and ZFIN for chromatin modifiers, OBI for experiment types, NCBI (entrez) for genes, HPO and MP for phenotype, NCBI Taxon for species, MGI for strain, MONDO for syndrome, ZFIN, HGNC, and MGI for transcription factor. We accept several different reference genomes for each species (e.g., hg19, hg38, mm10, mm39).

We engage with the community through several mechanisms:

- An annual FaceBase Community Forum, featuring presentations from FaceBase staff and community members who have contributed data to or used data from FaceBase in their research, tutorials, and discussions between FaceBase staff and community members.

- Monthly online "office hours", open to the general public, in which FaceBase staff answer questions from and have discussions with current or potential FaceBase users.

- Attendance, booths, presentations, and symposia at scientific conferences, such as the American Association for Dental and CranioFacial Research (AADOCR) and the Society for Craniofacial Genetics and Developmental Biology (SCGDB).

- Tutorials and "bootcamps" throughout the year.

- One-on-one discussions with each new data contributor.

- Documentation in our website, including a description of our current data priorities with request for feedback, a "key concepts" guide describing how FaceBase data is structured and accessed, and a guide to how (and why) to contribute data to FaceBase.

In addition, we have undergone a usability review from a third-party consultant, involving user testing and interviews, and have made improvements to our user interface as a result. We are currently working with the NSF SGX3 project to undergo a second round of user interface review and user testing, this time aimed at data contributors. Note: each of these has been focused on our web-based user interface, not on the data itself.

Links:

- FaceBase office hours
- FaceBase training events

- FaceBase annual Community Forums
- FaceBase data priorities
- Contributing data to FaceBase
- Key Concepts for FaceBase Data Submitters

Reviews

Reviewer 1:

Compliance level:

In Progress: the repository is in the implementation phase - 0

Comments:

Reviewer 2:

Compliance level:

In Progress: the repository is in the implementation phase - 0

Comments:

Information Technology & Security

R14 Storage & Integrity (R14)

R14. The repository applies documented processes to ensure data and metadata storage and integrity.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

To preserve bit-level integrity on our public server, our web server, command-line interfaces, and APIs generate, check and store cryptographic checksums whenever data is uploaded and verify those checksums when data is downloaded. Open-access data is stored on, and open-access metadata is backed up nightly to Amazon S3 buckets; each bucket is versioned and stored redundantly across multiple availability zones. All open-access data and metadata are versioned, and previous versions can be retrieved using our service interfaces.

Data and metadata are deleted only if the contributor of that data requests it.

The Amazon S3 FAQ details how S3 data are replicated and periodically checked for integrity: "S3 has end-to-end integrity checking on every object upload and verifies that all data is correctly and redundantly stored across multiple storage devices before it considers your upload to be successful. Once your data is stored in S3, S3 continuously monitors data durability over time with periodic integrity checks of all data at rest. S3 also actively monitors the redundancy of your data to help verify that your objects are able to tolerate the concurrent failure of multiple storage devices."

Controlled-access data is stored on AWS EFS file systems and backed up using the AWS Backup service.

Links:

- Amazon S3 FAQ with data redundancy information

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R15 Technical Infrastructure (R15)

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

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

Our open-access data is stored in AWS S3 buckets; our controlled-access data is stored in AWS EFS filesystems. We run linux-based operating systems (currently RHEL 9.2 for open-access data and Amazon Linux 2023 for controlled-access data).

Our open-access server runs our software stack (DERIVA), which is open-source and built on Apache, Flask, and Postgres (currently Apache 2.4, Flask 3.0, and Postgres 14). Most of our services and pipelines are written in Python. Our frontend web interfaces are written in JavaScript/REACT. Change management and issue tracking is handled through github. Python client libraries and command-line interfaces are distributed through PyPI. We do continuous integration testing on a development server and final testing on a staging server before updating our production server.

We monitor usage with AWS CloudWatch. We handle capacity planning through consultation with potential contributors; the new NIH Data Management and Sharing Plan requirements will ensure that we're notified of potential data additions even earlier than we are currently.

There is significant overlap between the DERIVA development group and FaceBase staff; however, DERIVA is also used (and funded) by a number of other projects, including the NIH Common Fund Data Ecosystem, ATLAS-D2K, and Synapse projects.

Links:

- DERIVA documentation site
- DERIVA git repository
- NIH Data Management and Sharing policy
- Synapse
- ATLAS-D2K
- Common Fund Data Ecosystem

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

R16 Security (R16)

R16. The repository protects the facility and its data, metadata, products, services, and users.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

FaceBase

Data can be considered to fall into two categories, controlled-access and open-access. Controlled-access data is human-subjects data that can't be widely shared; open-access data is everything else.

1. Controlled-access data is intended to be seen only by researchers who have a legitimate need for it and have received all necessary approvals (as described in R02). This data resides on an encrypted Amazon Web Services EFS filesystem, and Internet access is be protected behind a firewall. Data transfer to and from the storage infrastructure is encrypted and transferred using ssh, and access to the server is limited to FaceBase staff who are included in the FaceBase IRB. Relevant roles are FaceBase IRB-approved curators, end-user contributors (for uploads) and consumers (for downloads), and FaceBase IRB-approved systems staff (whose only role is to maintain the system).

Currently, controlled-access data is accessible only to:
- A small number of FaceBase staff members, who log in to the controlled-access servers via ssh, and
- Approved data contributors and consumers, whose access is limited to uploading or downloading data.
No external services other than ssh (limited to FaceBase staff) and data transfer run on those servers.

Data transfer for controlled-access data is handled through a second "transit" server that is set up similarly to the permanent controlled-access data server. For uploads, a FaceBase curator grants the contributor temporary access to the transit server and walks the contributor through the process of encrypting and uploading their data; after the data is uploaded, the FaceBase curator moves the data onto the permanent server. For downloads, the same process is followed in reverse.

2. Open-access data is world-readable once the curation process is complete and any embargo period has expired. Relevant roles for unrestricted data are FaceBase system staff, FaceBase curators, all approved contributors (people who have been approved to create datasets), project-specific contributors (people who have been approved to operate on data within a specific project), and the public (anyone, including people who have not authenticated to the system). The levels of metadata access depend on the lifecycle stage of the dataset:

a) Upon a project's approval, FaceBase systems staff create authorization groups and metadata entries for the project, and add the contributors to the all-approved-contributors group and the project-specific contributors group.

b) A contributor then begins creating metadata for a dataset belonging to the project. At this stage, metadata can be read, created, updated, and deleted only by members of the project-specific contributors group and by FaceBase systems staff and curators.

c) Project contributors continue working on metadata; curators may also make some changes. At the same time, they upload data files. Once a data file has been created, only the person who created it can update it (technically, create a new version of it).

d) When curation is complete and agreed upon, a curator sets the "released" flag on the dataset, which changes the policy: the dataset is now publicly readable, and only FaceBase curators can make additions or changes. (Note: the contributor may set an embargo period for the data to coincide with publication of a paper; in this case, the "released" flag is set after curation is complete and the embargo has expired).

Technically, FaceBase systems staff could override the permissions restrictions in c) or d), but our practice is not to do so.

Authentication to the open-access system is performed via OpenID Connect using the Globus federated identity provider; group membership is maintained through Globus groups. Authentication to the controlled-access server and controlled-access transit server is performed via ssh. The ERMRest (metadata server) and Hatrac (object store) documentation describes how access control rules are enforced by those servers.

A new working group is being formed to evaluate and update security policies and procedures for controlled-access data on an ongoing basis.

Links:

- ERMRest (metadata catalog) documentation
- Controlled Access to Human Data
- Hatrac (object store) documentation
- Globus

Reviews

Reviewer 1:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Reviewer 2:

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Comments:

Applicant feedback

R17 Applicant Feedback

We welcome feedback on the CoreTrustSeal Requirements and the Certification procedure.

Compliance level:

In Progress: the repository is in the implementation phase - 0

Response:

N/A

Links:

Reviews

Reviewer 1:

Compliance level:

In Progress: the repository is in the implementation phase - 0

Comments:

Reviewer 2:

Compliance level:

In Progress: the repository is in the implementation phase - 0

Comments: