

C3 AI has established the C3 AI COVID-19 Data Lake, a federated data lake containing a multitude of different datasets all related in some way to the COVID-19 pandemic. Public access to the Data Lake is granted through a RESTful API interface (see the official documentation [here](#) to learn more).

While for the public this provides a rich interface to a large conglomeration of data without the need to integrate multiple databases, members of the C3.ai DTI can have more direct access. Behind the C3 AI Data Lake is a package containing the definitions of every Type, as well as instructions for how to fetch that data and integrate it into the Data Lake's data model.

Below, for the benefit of the C3DTI researchers, we share an up-to-date diagram of the entire C3 AI COVID-19 Data Lake data model.

The diagram illustrates a comprehensive data model for COVID-19 research and public health monitoring. The entities and their attributes are as follows:

- HospitalPrediction (TimeValue)**: parent, value, timestamp.
- Hospital**: icuBeds, icuBedsAdult, icuBedsPed, bedUtilization, icuBedsPotential, ventilatorUsage, origin.
- HospitalPredictionSet (TimeValueHistory)**: parent, hospitalPrediction, predictionType, timestamp.
- OutbreakLocation (PhysicalAsset, MetricEvaluatable)**: pointMeasurements, lineList, biologicalAssets, location, denomAncestors, denomDescendants, hospitals, hospitalPredictionSet, policy.
- Diagnosis**: id, diagnostics, image, url, source, asImageFile, location, OutbreakLocation, idPatient, age, temperature, testResults, clinicalNotes.
- DiagnosisMetadata**: parent, key, value, source.
- LaborDetail**: parent, timestamp, year, month, laborForce, employedPopulation, unemployedPopulation, unemploymentRate, origin.
- LocationExposure**: asset, locationVisited, OutbreakLocation.
- PopulationData**: location, value, gender, race, ethnicity, ageRange, range, year, estimate.
- PointPhysicalMeasurementSeries**: asset, measurementType, treatment, data, earliest, latest.
- PointMeasurement**: parent, quantity, start.
- PatientRoute**: location, OutbreakLocation, coordinates, Lat, Long.
- PolicyDetail**: location, OutbreakLocation, startDate, endDate, entryDate, policyType, policySubType, name, value, flag, details, origin.
- LocationPolicySummary**: location, policyName, details, startDate, endDate, easingOrder, stayAtHome, mandatoryQuarantine, nonEssentialBusiness, largeGatherings, restaurantLimit, schoolClosure, primaryElectionPostponement, emergencyDeclaration, waiverTreatmentCost, freeVaccine, waiverOfOtherAuthorizationRequirements, prescriptionRefill, premiumPaymentGracePeriod, marketplaceSpecialEnrollmentPeriod, section1135Waiver, paidSickLeaves.
- LineListRecord (MetricEvaluatable)**: location, locationType, age, ageRange, gender, symptoms, symptomStartDate, exposureStartDate, exposureEndDate, hospitalAdmissionDate, caseConfirmationDate, caseCountry, isGroundZero, livesInGroundZero, traveledToGroundZero, relevantTravelHistoryLocation, relevantTravelHistoryDates, groundZeroExposure, notes, chronicDisease, sequencesAvailable, outcome, didDie, didNotCover, outcomeDate, recordSource, lineListSource.
- ClinicalTrial**: id, url, pdf, location, startDate, endDate, trialStatus, design, blinding, arms, covidStatus, severity, outcome, treatmentType, size.
- TherapeuticAsset**: id, productType, description, clinicalTrialsOtherDiseases, developer, stageOfDevelopment, fundingSources, nextSteps, therapyType, links.
- ExternalLink**: id, urlType, url, anchor, therapeuticAsset, origin.
- VaccineCoverage**: location, vaxView, year, timestamp, value, lowerLimit, upperLimit, difference, sampleSize, confidenceInterval, vaccineDetails, demographicClass, demographicClassDetails, totalPopulation, kindergartenPopulation, target, percentSurveyed, surveyType, other.
- BiologicalAsset**: id, location, sequence, assetType, status, host, species, genus, family, bioSample, isolationSource, club, publications, authors, genBankTitle, collectionDate, releaseDate.
- Sequence**: id, parent, subsequence, sequence, length, find, count.
- Subsequence**: parent, sequenceType, startindex, endindex, secondStartindex, secondEndindex, value.
- AminoAcidLookup**: id, name, abbreviation.
- NucleotideLookup**: id, name.
- SurveyData**: location, id, annualHouseholdIncome, birthYear2020, coronavirusConcern, coronavirusEmployment, coronavirusIntent_Mask, coronavirusIntent_SixFeet, coronavirusIntent_StayHome, coronavirusIntent_WashHands, coronavirusLocalCommunity, coronavirusSupportSystem, coronavirusSymptoms, ratioAdultHospitalization, coronavirusWhenShouldReopen, education, ethnicity, gender, hasCoronavirusBelief, politicalParty, politicalParty, religion, religiosity, trumpApproval, zipCodePrefix, start time, coronaSimilarFlu, coronaOnlyElderly, youngInvulnerable, elderlyMoreRisk, coronaAllHospitalize, coronaAllBakos, ethnicitySpreadsCovid, allSpreadsCovid, nonNativesSpreadsCovid, asymptomaticSpread, onlySickSpread, infectFromAnimal.