Data Management Policy

Committee on Information Technology (COIT)

PURPOSE AND SCOPE

Data is a key asset in meeting the demands of a modern government. Proper data management -- including management of systems that collect, store, transform, and share data -- is necessary for a government to operate efficiently and deliver services, especially when operations and services span multiple departments. The benefits of proper data management include:

- Consistent, discoverable, interoperable, secure, and trustworthy data
- Seamless data integration between departments
- Rapid deployment of new scalable services or enhancements to existing ones
- Shorter time-to-insight with access advanced data science tools
- Enhanced risk mitigation and compliance monitoring
- Reduction in cost

To deliver these outcomes, data must be proactively managed and maintained throughout its lifecycle. This policy applies to all information resources operated by the City and County of San Francisco as well as its departments and commissions, elected officials, employees, consultants, city-funded nonprofits, and vendors working on behalf of the City and County of San Francisco. The above listed entities are required to comply with this policy.

POLICY STATEMENT

This policy establishes a framework for the proper management of data as an asset across the City. Departments must adopt this framework, which includes:

- Ongoing modernization of data systems
- Inventory and classification of data systems, and the data within them
- Assessment of interdepartmental data uses
- Addressing security and privacy considerations
- Making data accessible for public use

POLICY REQUIREMENTS

MODERN DATA SYSTEMS

A data system, also referred to as an Enterprise System of Record as defined in California Government Code §7922.630, is the foundation of good data management. The City should prioritize modernizing its data systems whenever feasible. While this transformation won't happen overnight, departments should seek systems with the following capabilities when procuring new tools. These requirements should be added to contracts when possible:

- **Data Integration:** Systems should support secure read and write access to external sources through well-documented integrations (including but not limited to direct systems access or APIs).
- **Automation & Orchestration:** To reduce manual work performed by City staff, systems should make it easy to schedule and automate repeated tasks, such as sending notifications, updating records, or moving data between systems, or integrate with orchestration tools.

- **Repository & Version Control:** Systems should support integration with code repositories to allow for version control and collaborative development.
- **Collaboration:** Systems should promote team collaboration. Code, documentation, configurations, templates, scripts, and libraries should not be stored locally on individual machines but instead be maintained in shared, version-controlled environments.
- **Continuous Integration and Continuous Delivery (CI/CD):** Systems should support automated testing, validation, and deployment pipelines to enable rapid, reliable, and secure implementation of system enhancements, integrations, and updates.
- **Cloud-native Services:** Systems should utilize cloud-native and public cloud infrastructure to enhance scalability, interoperability, and efficiency, while reducing administrative complexity compared to legacy on-premises solutions.

Modern infrastructure lets cities choose the best tool for each part of the data lifecycle, whether for collection, storage, cleaning and transformation, orchestration, sharing, or visualization, without being locked into a single vendor or platform.

The City's standard contracts, such as the P-64X series, should be updated to more closely meet the standards listed above.

DATA SYSTEM AND DATASET INVENTORIES

Managing data as an asset starts with knowing the range and type of data under the control of the City. This process enables greater cyber- and information-security and enhances efficiency and transparency by identifying candidates for Interdepartmental sharing and Open Data. San Francisco Administrative Code Chapter 22D establishes the requirement to publish "a catalogue of the Department's data that can be made public". The inventory is not a requirement to publish data, only to catalog all existing systems and datasets owned by a department. The following requirements build on Chapter 22D and strengthen the foundation for good data management:

- **Data Systems and Dataset Inventory**. Departments must create and maintain a Data System Inventory and Dataset Inventory per procedures and templates set by the Chief Data Officer (CDO) and with guidance from the City Attorney's Office. This process should happen no less than once a year. See public Dataset and System Inventory.

- **Classification of Data Systems and Datasets**. As part of the inventory, Departments must classify data systems and datasets per procedures set by the CDO and City Chief Information Security Officer (CCISO) in the Data Classification Standard.

- **Publication of Data System and Dataset Inventory**. The CDO must publish the Data System and Dataset Inventories as datasets on the City's open data portal (or successor site) and update them whenever departments add, remove, or modify their inventory.

- **Data Retention**. Data retention will adhere to guidance of Chapter 8 of the Administrative Code.

The Data Inventory Explainer, incorporated for reference, provides more detailed direction on the data system and data inventory process.

INTERDEPARTMENTAL DATA

City services are deeply interconnected, but the data systems that support them remain siloed. To improve service delivery and operational efficiency, we must modernize how interdepartmental data is classified, stored, cleaned, discovered, integrated, analyzed, and shared.

Increasing Interdepartmental Data Availability

All data within the City and County of San Francisco should be assessed for Interdepartmental use cases. Data often has a network effect – the more who use it, the more valuable it becomes. It is often easy to evaluate the risks of sharing data, but the risk of sharing data should not be assessed in a vacuum. It must be balanced against the value of increased access, usage and standardization.

This policy establishes a three-part check to verify that data can be shared with another City entity. Specifically, a department should not withhold data from another city department as long as three requirements are met:

1. There is a valid business reason to request data, such as supporting public services, internal operations, regulatory compliance, reporting, analysis, or decision-making
2. Regulations allow the requesting department to access the data
3. The requesting department has the systems and processes in place to comply with data regulations

As outlined in "Modern Data Systems", to facilitate the sharing of Interdepartmental data, all departmental and vendor data systems should support data sharing through direct system access, APIs or other methods that enable easier access and automation.

Interdepartmental Data Standards

To realize the full value of Interdepartmental data, a shared data standard as outlined in this section below must be applied to ensure proper usage, efficiency, and interoperability. Without these standards, the risk of incorrect data usage, inefficiency when collaborating across departments, and inconsistent reporting grows. Specifically:

- Metadata Standards: Applying COIT Metadata Standards to all Interdepartmental data, ensuring all datasets are properly documented.
- Field & Column-Specific Standards: Applying best practices outlined in DataSF's Data Standards Guide, including formatting for addresses, street segments, parcels, business ID, and other common identifiers. This list is meant to grow as more standards are established.

Custodianship and Ownership of Interdepartmental Data

Interdepartmental data stored on Department of Technology (DT) infrastructure is covered by the Data Custodian and Stewardship Policy. When data is replicated from a system of record onto DT systems, ownership remains with the originating department. All Sunshine requests must be directed to the department that owns the enterprise system of record where the data originated. In all other cases, ownership and custodianship should be defined in the data-sharing MOU (see "sharing non-public data" below).

SHARING NON-PUBLIC DATA

When managing and sharing non-public data, departments must follow best practices, listed below, to protect privacy and ensure the security of City data, datasets, replica datasets, derived datasets, and systems. Maintaining public trust requires strong safeguards and full compliance with applicable privacy laws. Secure and lawful data sharing not only reduces the risk of breaches but also improves efficiency by limiting duplicate data collection. To minimize risk:

- **Follow Cybersecurity Best Practices.** Abide by the guidance in the City and County of San Francisco Cybersecurity Requirements, and the guidance of the Office of the Chief Information Security Officer (CCISO), and departmental CISOs.

- **Assess Privacy and Security Risks.** When sharing data, departments can utilize the Privacy Toolkit and Security Toolkit established by the CISO and CDO. These resources should be applied alongside any department- or domain-specific regulations and laws, with additional guidance from the City Attorney's Office as needed.
- **Establish Data Sharing Agreements.** Standardized Memoranda of Understanding (MOUs) for data sharing are available from the City Attorney's Office and should be used when appropriate.
- **Utilize Secure Systems.** Sensitive data should only be shared using secure, approved systems that meet the City's privacy and security requirements. Departments should avoid using email or unencrypted file transfers for sensitive information and instead rely on tools designed for secure data exchange, such as encrypted portals or file-sharing platforms vetted by the Department of Technology or the CISO.
- **Avoid Dataset Replicas.** Whenever possible, departments should provide secure access to data in their systems or a central system rather than generating static copies or replications for dispersal.

OPEN DATA

Data yields the most value when it is free to access, use and share – consistent with safety, privacy, and security considerations. Making open data available means it is published on the City's Open Data Portal (https://data.sfgov.org/ or successor site).

Publishing Process and Prioritization

Once departments have identified potential Open Data candidates through the annual inventory process, they must:

- **Create or update their publishing plan**. Departments must prioritize datasets for publication using the Prioritization Guidance created by the CDO
- **Follow the publishing process**. Departments must follow the Publishing Process established by the CDO, when sharing data on the Open Data Portal

Other Publishing Considerations

- **Non-Public Data.** Departments must use the Open Data Release Toolkit and work with the CDO to transform non-public data so that a relevant and appropriate view of that data (e.g. limited fields, aggregate data) can be published as Open Data. Non-public data includes anything properly classified as Level 2 or higher per the Data Classification Standard.
- **Licensing Standard.** The default license for all Open Data is the Public Domain Dedication License (PDDL). While this license is the default, other licenses may be used for specific datasets and as approved by the CDO.
- **Recording.** The CDO must update the publishing status of datasets in the Dataset Inventory when a dataset becomes public.
- **Language.** Data sets, including metadata, are not required to be published in additional languages beyond those used at the source. The City may opt to implement automated language translation of data sets in the future.

EVOLUTION OF DATA MANAGEMENT

As technology evolves, so too must the data management policy. This section outlines specific use cases for data management, and will expand as new technologies require:

Artificial Intelligence

The use of Artificial Intelligence (AI) within the City is governed by Chapter 22J of the Administrative Code. In the context of data management, several key principles warrant particular attention:

- **Data Provenance.** The origin, sources, and any transformations of data used in the training and operation of AI and machine learning (ML) models should be thoroughly documented to ensure traceability and accountability.
- **Data Quality.** All data utilized in AI/ML systems should be evaluated for accuracy, completeness, and potential bias.
- **Algorithmic Transparency.** To the extent practicable, the City shall promote transparency in the design, inputs, and functioning of AI/ML algorithms, including the data they rely on and the decisions they generate.
- **Ethical Considerations.** AI/ML applications should be reviewed for potential ethical risks. This includes proactively addressing concerns related to bias, fairness, individual rights, and privacy.

The AI Advisory Board—convened by the Office of Emerging Technologies, in partnership with the Chief Data Officer and the City Attorney's Office—shall be responsible for developing and maintaining guidelines that promote ethical AI/ML data practices. AI Advisory Board shall also establish a framework for assessing and mitigating risks associated with the use of AI/ML technologies.

ROLES AND RESPONSIBILITIES

Chief Data Officer

The CDO is responsible for the following:

- Develop and oversee the process for creating, maintaining, and publishing the annual Data System and Dataset Inventory, including data classification;
- Develop processes and resources and support the publishing of open and internal data consistent with Citywide policies and standards
- Develop, with the City Attorney and input from departments, template agreements, supporting guidance, and a standard and timely review process for confidential data and information-sharing agreements across departments or with external partners;
- Develop, with the CISO minimum data quality and protection standards;
- Define data architecture modeling standards and tools that departments will use;
- Engage with Data Coordinators, Stewards, and Custodians to strategically plan, and provide support and training for the publishing of data; and
- Lead & implement citywide data modernization via consolidation and integration into a unified data platform to achieve the objectives of the data management policy and the city's technology goals (done in partnership with the CIO and the Department of Technology).

City Chief Information Security Officer

The CISO is responsible for the following:

- Develop minimum data protection standards;
- Complete and publish the annual Open Data Program Risk Assessment; and
- Advise departments on data management.

Data Coordinators

Department Data Coordinators are responsible for the following:

- Coordinate the annual Data system and Dataset Inventory process, including classification, in their department and per CDO procedures;
- Coordinate the department prioritization of data for publication;
- Coordinate publication of data on the open data portal per procedures and standards set by the CDO
- Identify and record Interdepartmental data originating from their department
- Alert CDO to any changes of the coordinator or stewards within their department

Information Security Officer

Department Information Security Officers are responsible for the following:

- Apply the 3-part check to assess Interdepartmental data requests when received
- Make final decision on data access for Interdepartmental data requests

Data Stewards

Department Data Stewards are responsible for the following:

- Work with the Department Data Coordinator to properly document data for which they are the steward according to the inventory requirements;
- Follow the requirements set out in the Data Classification standard and properly document the classification of data in the inventories; and
- Provide proper documentation of shared datasets to support the responsible use of data and metadata for all data, including unstructured documents, is accurately recorded and maintained.

Data Custodians

Department Data Custodians are responsible for the following:

- Work with the Data Coordinator and Data Stewards to provide appropriate documentation to support the data system inventory;
- Work with Data Stewards and Data Coordinator to support efforts as needed to make data available through the publishing process referenced in this Policy; and
- Adequately support their department's Data Stewards, Data Coordinators, and Information Security Officer in conducting their responsibilities in this Policy and in the Data Classification Standard.

Data Users

Data Users are responsible for the following:

- Responsibly use open data or data obtained from departments by reading or requesting documentation on data and applying analysis understanding any constraints on the data;
- Follow any constraints on use as specified in MOUs or other agreements where applicable;
- Be familiar with federal, state and local confidentiality or privacy laws pertaining to the data they collect, access, use, or maintain in conducting their work; and
- Adhere to data retention schedules and disposition procedures when handling data, and report any potential issues or discrepancies to the Data Steward.

Departments

Department leadership and program management are responsible for the following:

- Assure that staff handling confidential data are sufficiently trained and aware of their duties with respect to securing and protecting private information including Data Stewards, Data Custodians and Data Users
- Make good faith effort to follow Interdepartmental data standards; and
- For those having Interdepartmental data or data standard candidates, work to meet expectations for management of Interdepartmental data and data standards as resources allow and by priority of value to the enterprise.

City Chief Information Officer

CIO and the Department of Technology are responsible for the following:

- Advise department leadership and program staff in technologies to support and build Interdepartmental data and data standards;
- Publish and promote standards and patterns on how to best implement the requirements for managing Interdepartmental data and data standards;
- Drive central procurement strategies to increase adoption of modern technology systems as described in Section 1.0; and

- Partner and support citywide data modernization, in partnership with the CDO, via consolidation and integrations into a unified data platform to achieve the objectives of the data management policy and the City's technology goals.

Controller City Services Auditor

CSA Audits is responsible for the following:
- Audit this policy for compliance as needed

Office of Contract Administration

OCA is responsible for the following:
- Explore updating the language in the Software as a Service (SAAS) contract to include stronger language related to Modern Data Systems

METRICS FOR POLICY EFFECTIVENESS

To ensure the ongoing effectiveness of this Data Management Policy and to drive continuous improvement, the Office of the Chief Data Officer will publish the following metrics:

Inventory Metrics

- **Completion Rate:** Percentage of departments that have submitted a complete and up-to-date Data System and Dataset Inventory by the annual deadline.
  - Target: 100% compliance.

Interdepartmental Metrics

- **Participation Rate:** Percent of departments with MOU or other data sharing agreement in place with other department(s).
  - Target: 70%

Open Data Metrics

- **Publishing Rate:** Percentage of departments that are publishing data on the Open Data Portal.
  - Target: 100% compliance.

EXCEPTIONS

This policy does not override any State or federal regulations such as the Health Insurance Portability and Accountability Act ("HIPAA"), Criminal Justice Information Services Policy ("CJIS"), Sunshine Ordinance, Data Retention, etc. Further, this policy does not apply to data provided to the City by the State or Federal government, or by other municipal or regional jurisdictions.

DEFINITIONS

Data System

Also referred to as an Enterprise System of Record, as defined in California Government Code §7922.630

Dataset

A dataset is defined as a structured or unstructured collection of digital information that is used in the delivery of public services, enables internal operations, regulatory compliance, reporting, analysis, or decision-making. For the purposes of this inventory, a dataset must be:

- **Official and authoritative.** Recognized as the source of truth for its purpose.
- **Production-level**. Used in finalized, downstream applications—not temporary, intermediary, or raw data tables used only during processing or development.

This includes structured data (e.g., databases, spreadsheets, digital maps, and CAD files) and select unstructured materials (e.g., internal knowledge bases, SOPs, photos, audio, videos, PDFs, or scanned files).

Dataset Replica

A dataset replica or replications is a copy of a dataset that is functionally the same as a dataset, but does not serve as the official and authoritative source of truth. It is commonly referred to as a: Snapshot, clone, mirror, or backup.

Derived Dataset

A derived dataset is a dataset that is created by combining one or more source datasets or replica datasets to create a new asset. If a derived dataset is Official and Authoritative and Production-level, it can be treated as a new dataset with shared stewardship between the source dataset owners.

Interdepartmental Data

Interdepartmental Data is a dataset that meets a two-part definition:

- **Process dependencies**. At least one department is dependent on data from another department to conduct its work. This includes (but is not limited to) scenarios where it would be costly, redundant, or inefficient for the dependent to collect the data itself.
- **Authoritative Source**. The data is the authoritative source. Authoritative sources can be codified in federal, state or local law or through administrative policy or practice.

Nonpublic Data

Any dataset not classified as Level 1 (Public) using the COIT data classification standards.

Open Data

Open Data means that the dataset or replica dataset is:

- Free to download, use, and share
- Discoverable and accessible to all over the internet
- Published to minimize time between the creation and dissemination of the data
- Documented
- Provided under terms that permit re-use, redistribution, and the mixing with other datasets
- Provided in an open format that is machine-readable and accessible via API