SYSTEMS INTEGRATION

A community project at United Way

Systems Integration Project (SIP)

Data Architecture Strategy

Laura Reeves

6-25-2020

Contents

1 Executive Summary

The purpose of the Data Architecture Strategy is to provide a high-level understanding of what kind of data needs to be included in the Data Ecosystem. This does not define physical data structures or technology, which will be part of the solution selected through the RFP process.

1.1 Vision Statement

The work of this project is guided by the collective vision, which was established in 2018, by the founding members of the Systems Integration Team (SIT):

> "The greater Rochester community is working across a diverse network of committed providers to build an interconnected, person-centered system of health, human services, and education leveraged by a unified information platform (technology and consistent approach to service delivery processes), to improve the health and economic well-being of individuals and families, especially those who are vulnerable and/or impacted by poverty."

2 Critical Input

The data architecture reflects the work that has been done to date to outline the purpose and requirements for the data ecosystem. The following provide input into the design of the Data Architecture.

2.1 Business Requirements

The Business Requirements Document provides the overall context for what data will participate in the data ecosystem and the specific functionality that must be supported. Specific functional requirements are referenced as appropriate throughout this document.

2.1.1 Minimum Data Set Approach

The Systems Integration Team has identified several design goals for the proposed data ecosystem, including the use of an open-architecture framework that is interoperable, scalable and leverages existing digital infrastructure, and the identification and use of a minimum data set that aims to identify and share the minimum necessary information to coordinated service delivery and improve outcomes.

The minimum data set approach allows the SIP to leverage existing digital tools and information, without instituting a one-size fits all approach to screening, or an onerous self-reporting process, both of which have been rejected by a target population who feels they are over-screened, and providers who are overburdened by funder-driven screening protocols. This paradigm of data mapping and normalization also allows new data on service utilization, generated via the closed loop referral system, to adjust an individual's risk-score over time.

2.2 Data Governance Strategy

> ***Data Governance*** *is the specification of decision rights and an accountability framework to ensure the appropriate behavior in the valuation, creation, consumption and control of data and analytics.*

Data Governance is sometimes included in a Data Architecture Strategy, but is documented separately in the Integrated Data Governance Strategy document. Implementation of this Data Architecture must be done in conjunction and compliance with the Data Governance Strategy, including gaining informed consent and implementing appropriate levels of security.

3 FOUNDATIONAL CONCEPTS

The following are fundamental concepts that are important for the implementation and use of data, but are not explicit parts of the data architecture itself. These relate to how data should be handled and managed over time, but do not define the data itself. There are many ways to implement these concepts from simple to highly sophisticated systems and procedures. Each of these concepts should be included in the evaluation and selection of a solution.

3.1 DATA QUALITY

> ***Data Quality*** *is the assessment of the cleanliness, accuracy, and reliability of data that makes data appropriate for a specific use.*

There are several aspects of data quality that are critical to ensure data quality for SIP uses. Data Governance will drive the definition and implementation of data quality rules. There are three distinct areas of data quality that the project will address:

- **Selection of High-Quality Sources**
  The Systems Integration and Evaluation Workgroup has defined the following criteria to be considered when selecting data source to feed into the Data Ecosystem are:
  - Accessibility
  - Ability to Aggregate/Disaggregate (to a sub-group level of detail)
  - Frequency of data updates
  - Accuracy
  - Meaningful

  No fixed values have been set as a minimum required level of quality. There are wide variances between sectors and data sets. These criteria will be considered when comparing multiple candidate sources for each set or collection of data.

- **Maintain Quality Data within the Data Ecosystem**
  The Data Ecosystem is only responsible for data quality of data actually stored in the SIP environment. This means that the system must ensure accurate and appropriate handling and maintenance of data within the Data Ecosystem based upon business rules.

- **Measure and Monitor Data Quality**
  The system needs to include capabilities that will allow the quality of the data to be quantified and available. The following functional requirements clarify what level of data quality work is needed in the environment:

  Data Quality Rules (EBD-8)

a.) As a system, I want to control the rules that define data quality of the system entities and be able to audit the data quality against these rules on a periodic basis, so that users have a consistent way to express and validate quality data.
b.) As a provider, I want to provide missing or incorrect person information as determined by the data quality rules of a person, so that a complete and current view of a person is available to provide timely service.

Data Quality (EBD-10)

a.) As a user, I want the system to support data quality workflows (e.g., Inspection, Cleaning, Verifying, Reporting); so that I can ensure consistency of the data stored in accordance with the implemented rules.
b.) As a system, I want to configure and employ business rules that systematically enforce data quality; including, but not limited to, rules around data validity, accuracy, completeness, consistency, and uniformity.

Note: Data validity may be defined through a combination of data-type constraints, range constraints, set-membership constraints, regular expression constraints, and cross-field validation.

3.2 Change Data Management

Change Data Management *is the set of practices and technology to determine and track changes to data so that other related systems can make appropriate changes and necessary actions can be taken. Change Data Management is critical to ensure compliance with government and industry regulations.*

The key system capabilities for change data management to support the SIP are:

Interoperability: End-state solution easily communicates between systems (e.g., uses application programming interfaces to receive and provide data) thus reducing the need for organizations to enter and view the same data in multiple solutions

Notifications (EPS-2): As a user, I want to configure my notifications, so if I choose, I can be informed when my information in the system changes.

These indicate that individuals and/or data participants be made aware when data changes. This allows appropriate notifications/messages to be generated. These can be made available to those who "subscribe" to these notifications/messages. The overall solution architecture will dictate the how this is to implemented.

3.3 Data Dictionary

Data dictionary *is the place where information about data that exists in the organization is stored. This is important to capture and communicate business knowledge of the data.*

Data Definitions *describe what the data element represents and may include sample values and/or a list of all possible values (domain values).*

Clear communication of what data is available in the data ecosystem and what it means is critical to the overall success. The project team is carefully defining vocabulary and shared language. Data element definitions are included in the data requirements documentation. This valuable content needs to be made available as the system is implemented. A common method to make this type of information available is to have a data dictionary, which may be a sophisticated searchable database with an interface or simply publishing a PDF that can be searched. A data dictionary is often found on a web site that provides documentation and support for any system or environment.

3.4 Master Data Management

> ***Master Data Management (MDM)*** *is a technology-enabled discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, semantic consistency and accountability of the enterprise's official shared master data assets. Master data are the consistent and uniform set of identifiers and extended attributes that describes the core entities of the enterprise including customers, prospects, citizens, suppliers, sites, hierarchies and chart of accounts.*

The principles of master data management are essential to the main goals of the SIP. These practices, people and technology, will be applied to the development and maintenance of:

- Unique Person
- Directory of Organizations and the Services they provide

The data architecture notes the need for the data but the specifics of how MDM will be implemented will be a key part of the chosen solution.

3.5 Data Lineage

> ***Data Lineage*** *is a description of the path along which data flows from the point of its origin to the point of its use. - Irina Steenbeek, Ph. D*

There are many interpretations of data lineage and with many different vendor offerings available in the marketplace. There are two common ways that data lineage technology capabilities are used:

a. Managing Data Across Enterprise Systems: Organizations that have a large number of application systems utilize data lineage capabilities to crawl though the web of enterprise data systems to survey and trace data. This helps to identify and catalog data, and may also build data flow diagrams.

b. Support for Data Lakes: Organizations that have loaded detailed data from many different places into a data lake often find themselves with a need to understand the context of that data. Data lineage capabilities support the need to trace the source of individual data elements.

SIP is not facing either of the challenges listed above, and does not have a requirement for full data lineage capabilities. Since SIP is a new environment, the primary challenge is to ensure that the environment is built to provide transparency as data flows in and is stored. There is a need for SIP to:

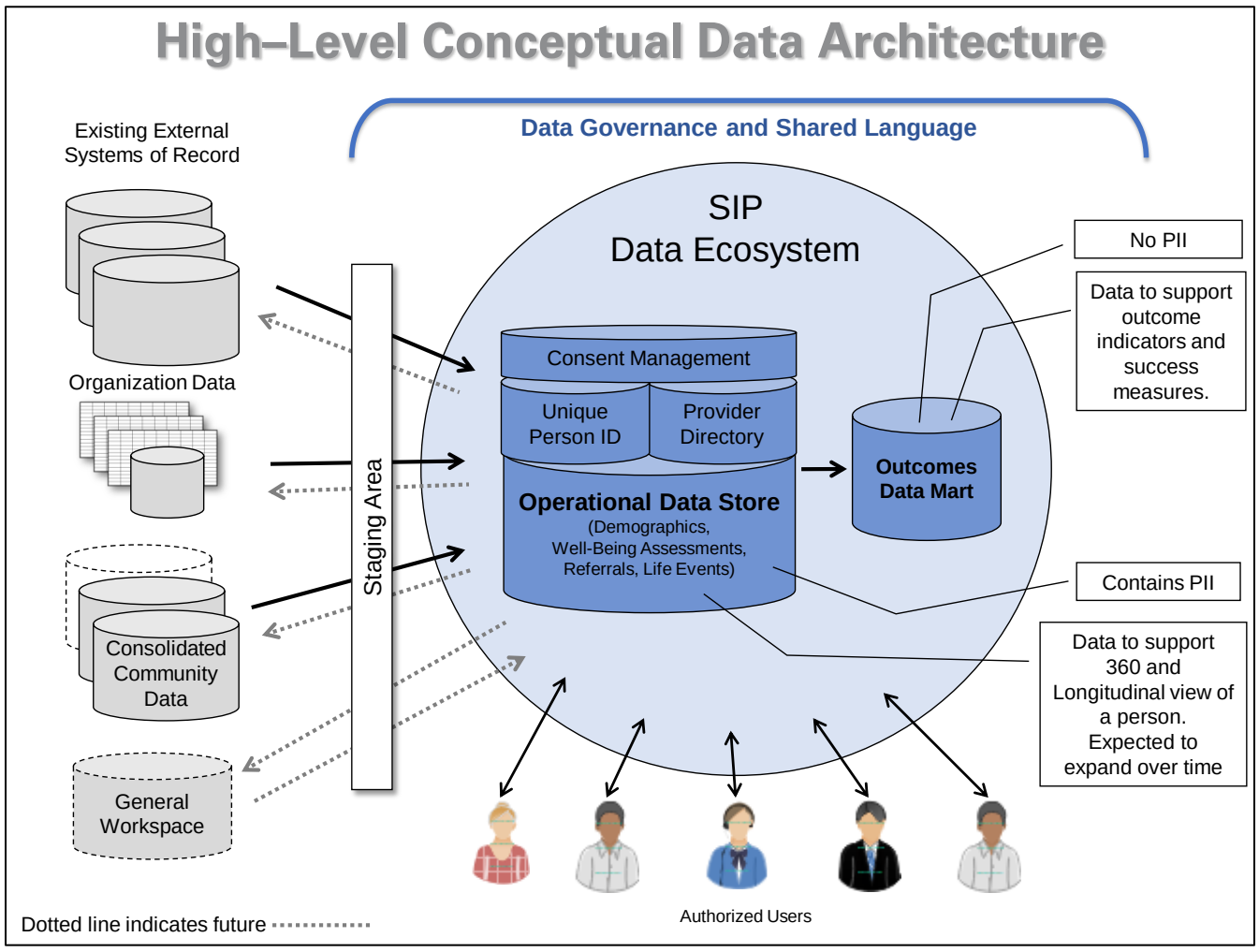
- Provide historical data for the longitudinal record of a person (see section 4.4.4)

- Track changes to master data, which is expected to be part of standard Master Data Management capabilities.

4 The Data Ecosystem Data Architecture Vision

This section describes the high-level conceptual Data Architecture vision for the SIP. This outlines the different kinds of data that are involved in the data ecosystem and describes the purpose of that data. This ensures that the right data is available, in the right place, at the right time to support the required functionality.

The diagram below shows the high-level conceptual data architecture that is being recommended for the SIP Data Ecosystem. Each of the components is described in the following sections.



4.1 Existing External Systems of Record

Data is captured or created by community organizations during their regular operations in some type of system. There are many different kinds of systems in use, ranging from large, sophisticated applications to basic spread sheets. The major types of sources available across the community are described here.

4.1.1 Organization Data

These are the fundamental systems used to run an organization. These are also called business applications or operational systems. These may be third party applications or built in-house. Some of these are hosted by third party vendors, while others are run and managed within the organization.

Some data may not be part of a full-blown business operational system, but is still valuable to the data ecosystem. A mechanism must be available to allow that data to be uploaded to a staging area and then processed to load into the Operational Data Store.

4.1.2 Consolidated Community Data

There may be some instances where data from multiple organizations is already being collected, integrated and is available for the data ecosystem to tap into. There may be cases where a consolidated source of data would be valuable to the community but does not exist at this time. It is not in the scope of this initiative to create such a data source.

4.1.3 General Work Space

Some of the organizations that are participating in the SIP do not have full IT resources to build and maintain their own applications. It may be valuable for the SIP to offer a base platform for data storage and minimal application support. This also provides a shared place to store data in support of other collaborative initiatives in the community.

4.2 Data Flow

Data can be exchanged between participating organizations and the data ecosystem in a variety of ways. The technical architecture of the selected vendor/solution will drive the interface specifications for this data exchange. The data architecture is simply identifying that data from these underlying sources is needed to support operational functions and/or outcomes. In the future, the vision is for data that has changed (modified or added) to be made available to the external systems.

4.3 Staging Area

> *The **Staging Area** is a place where data is stored while it is being prepared for use. This is where raw data is uploaded, then processed and loaded into the data ecosystem.*

The Data Architecture Vision includes a staging area to support data quality checks, audit tracking and backout/recovery of data in case there are issues. This allows the data ecosystem to better managing incoming data.

4.4 Operational Data Store

> ***Operational data store (ODS)*** *This is a collection of data from operational systems, most often integrated together, that is used for some operational purpose. The most critical characteristic here is that this is used for some operational function. This operational dependency takes precedence and the ODS should not be considered as a central component of the data warehousing environment. An ODS can be a clean, integrated source of data to be pulled into the data warehousing environment.*

The data to be included within the ODS includes:

4.4.1 Consent Management

Data needed to properly handle the collection, renewal and management of a person's consent to store and share their data across constituents who participate in the data ecosystem.

This supports the following functional requirement:

Management (EPS-4):
a.) As a system, I want to ensure that only persons that consent for services are served, so that privacy and trust is maintained.
b.) As a system, I want to ensure the person has a simple way to change their consent to share information so that I can maintain current preferences and uphold system rules.

4.4.2 Unique Person Identifier

A *Unique Person Identifier* that provides the ability to uniquely identify an individual across the data ecosystem. This includes a cross reference for different ways that a person can be known so that these alternate identifiers map to the unique person identifier.

This supports the following functional requirement:

Single instance of a person (EBO-1): As a system, I want to provide a unique system identifier to each person and a cross reference to other provider systems so that I can match and deduplicate person data.

4.4.3 Provider Directory

This is an integrated Directory of Organizations and the Services that they provide. This supports the following functional requirement:

Resource directory (EBO-5): As a provider, I want to be able to identify the available and situationally appropriate resource and understand the eligibility and capacity of the appropriate resource so that the person can engage the resource for help.

4.4.4 360 View / Longitudinal Record of a Person

*The **360 Degree View of a Person** makes available summary and detailed information about an individual that is known to the data ecosystem.*

*The **Longitudinal Record** is a record that presents a picture of a person's care history. This record includes multiple data points over time, including assessments, referrals and referral outcomes that enables providers throughout the care continuum to assess the person accurately to provide faster and more specific service plans.*

The data ecosystem needs to include the following information about a person:

- Demographics
- Well Being Assessments
- Referral Tracking
    - Support interconnection between referral systems by offering sharing of data.

- A shared digital closed-loop referral system that offers multi-sector providers a method for referring individuals to services offered outside of their organization
- Ability to have complete view of all referrals

- Life Events
  - Ability to gather a significant Life Events for a person over time

This supports the following functional requirements:

> *360-degree view of a person (EBO-1): As a provider, I want a 360-degree view of a person so that I can understand "current state" and identify future needs of and individual and/or family.*
>
> *Longitudinal record (EBD-2): As a system, I want to log events and changes in state of well-being of a person so that users can analyze and derive business intelligence from this data.*
>
> *Shared data hub (EBD-3): As a system, I want to collect state of well-being of a client over time and collect the right information that supports cross sector well-being assessment and growth so that users can extract their information and analyze it.*

4.4.5 Internal System Data

The ODS will also include other data needed to support the functions and capabilities of the system itself such as user ids and passwords, internal auditing and logging.

4.5 Outcomes Data Mart

> *A **Data Mart** is a storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The data mart then combines that data in an aggregate, summary form suitable for data analysis and reporting for predefined business needs.*

The Outcomes Data Mart will include:

4.5.1 Outcome Indicators

Data needed to support outcome indicators at an aggregate and sub-group level of detail.

4.5.2 Unidentifiable Person Data

Data about individuals to support reporting and analysis. This includes data collected over time. Initially, this will be limited to data from the ODS, within the consent framework. Other person level data could be added to the data mart from external sources in the future as business requirements dictate.

This supports the following functional requirement:

> *De-identification (EPS-5): As a system, I want to de-identify data used in analysis and reporting activities, so that the identity of a person is shielded and not exposed via the reporting and analysis activities.*

4.6 Data Access and Use

Data will be available for a variety of uses, both operational and analytical. All access will follow Data Governance policies and practices, which are documented separately. The types of access and use of data include, but are not limited to:

- Supporting integrated workflows
- 360-Degree View Shared Dashboard
- Reporting and Analytics
- Exporting

The data mart with data access capabilities (query, reports, dashboards, business intelligence and analytics) will support the following functional requirements:

> *Reporting Capabilities (EBD-5) As a system, I want the ability to provide cross sector quality reports, so that users can reduce duplicated provider agency reporting efforts.*
>
> *Exports (EBD-6): As a system, I want to be able to extract information by provider, so that each provider can easily get out the information they contribute to the system.*
>
> *Analytics (EBD-4): As a system, I want to provide cross sector analytics capabilities so that users can manipulate data to evaluate performance, make informed decisions, etc.*

4.6.1 CAUTION: Statistically Significant Data Set

The initial project direction is to take a grass roots approach to building up data on individuals. As individuals give consent to be included in the data ecosystem, their data will begin to accumulate. It will take time for a statistically significant percentage of the population being studied to agree and have their data be collected. Basic reporting and presentation of the data will be possible, but users must understand the limitations of the data and be cautious performing analytics until there is sufficient data available.

5 DATA REQUIREMENTS AND MODELING

The Business Requirement Document articulates what the system needs to do, and does not explicitly define the data required to support those functions. Detailed data requirements have been compiled from existing project documentation, discussion with the SIP project team members, pilots, project work groups and data focused sub-groups. Where possible, any future data needs will be considered and potentially included in the data requirements. The intent is to capture the creative ideas now for reference in the future.

> *NOTE: The data modeling effort is not tied to specific technology or data modeling approach. The selected solution provider will either:*
> *A. Include an existing model that is part of their application/solution*
> *- OR -*
> *B. Build their own data model appropriately aligned with the chosen data storage technology*

Specific data requirements can be found in:

- **Data Requirements Spread Sheet**: List of specific data elements needed in the data ecosystem to support required functionality. The minimum required set of data elements are identified.
- **Conceptual Data Model**: This provides a high-level visual depiction of major groupings of data and relationships needed to support operational functions.
- **Business Dimensional Model (BDM):** A data model that provides a business abstraction of data needed in the Outcomes Data Mart to support reporting and analysis. The model shows diagrams of dimensions and

facts so that the details can be reviewed and discussed in business terms. This also separates the business perspective from technical implementation details.

6 GLOSSARY OF TERMS

7 CHANGE LOG

Version	Date	Who	What
1.0	2020-05-28	L. Reeves	Initial Draft of Data Architecture Strategy.
2.0	2020-06-15	L. Reeves	Updated with edits from team.
2.1	2020-06-25	L. Reeves	Updated with feedback from data work sessions.