Kentucky Health Data Trust Initiative

Key Considerations for Planning a Data Integration Approach

Deliverable 4.3.3 and 4.4.2

Prepared for the Kentucky Health Data Trust Interagency Governance Workgroup

Freedman HealthCare
July 30, 2015

Introduction

Data integration within a Data Trust requires careful planning and resource allocation from the beginning of the Data Trust development process. The purpose of this document is to highlight the necessary decision points and critical requirements for data integration, by presenting a set of best practices from other Data Trust states. This document also offers a checklist for CHFS and its partners as they develop a data integration approach for the Kentucky Health Data Trust (KyHDT).

Best Practices for Planning the Data Integration Approach

1. Determine the essential data elements needed to assign the Master Patient Identifier.

The Data Trust must be able to reliably identify individuals across all files residing in the database. A reliable, consistent Master Patient Identifier is essential for analysis, research and potentially care management. The KyHDT must build upon the foundation that CHFS has already established with Release 5. All incoming data, including information from external sources, should include as much information as needed to either directly match individuals or meet probabilistic thresholds in the MPI methodology.

2. For data from external sources, clearly define and document data submission requirements. A key challenge among existing APCDs has been developing a standardized format for data submission (e.g. ensuring that all tabs and data elements align across all files and all data submitters). It is difficult and time consuming to integrate data from multiple submitters when the data elements do not match. Taking the time to clearly define and communicate the submission format during the Data Trust development/planning stage will help the Data Trust team streamline the data integration process later on.

In the context of the KyHDT, CHFS should base the data submission specifications on the project's identified use cases. Understanding how data from the Data Trust will ultimately be used is critical for developing data specifications that capture the necessary data elements in the correct format.

Once the Data Trust team develops appropriate data specifications, it should clearly document them in a Data Submission Guide and share the guide with all data submitters. However, disseminating written guidelines does not suffice: the team should also engage data submitters in an ongoing dialogue to confirm that they understand the specifications and are able to meet them. A key lesson learned from existing APCD states is that there is no "shortcut" for this step. Data Trust teams should create open communication back and forth with data submitters before data intake begins, and then maintain that communication on an ongoing basis.

Clearly documenting, and then communicating, expectations for data submission at the outset is often the biggest hurdle for data integration in the Data Trust. If all data submissions are able to meet the required specifications, then integrating the data within the data warehouse becomes a much easier endeavor.

Note that even the most successful voluntary multi-payer claims data initiatives report that data submitters are less likely to provide timely data sets that conform to requested specifications. Data Trust leadership must effectively communicate the mutual benefits of providing high quality data files and broker a commitment to correcting errors in a timely manner.

3. For data from internal (state agency) sources, clearly establish timelines, formats and problem resolution protocols. State agencies have multiple reporting and data supply commitments. The Data Trust should minimize additional reporting burden as much as possible through collaboration with state agency submitters. Data Trust data developers will create links and crosswalks to ensure that cross-agency queries are accurately and appropriately extracted.

In the long run, the Data Trust could relieve state agencies of some public reporting demands and routine information requests through standardized reports and self-service access to aggregated data.

4. Let the business needs drive the data integration approach. For systems that pull in disparate data sets, each containing different data fields, the Data Trust team must first decide which type of data integration approach will best suit the project's business needs. There are two basic approaches for this: project-specific data integration, and automated data integration.

With a project-specific approach, disparate data sets live in a data warehouse but are not automatically integrated or linked upon intake. As researchers identify a specific use case (e.g. a research question they want to answer), the Data Trust team grants them access to the data warehouse to pull only those data sets that they need for their specific project. The Data Trust team may provide the researchers with a basic algorithm for linking the data sets, but the researchers are responsible for the majority of the data integration and analyses related to their project.

In the second approach, the data warehouse architecture is built to integrate the disparate data sets automatically as they enter the warehouse. Consequently, all data stored within the warehouse is already linked through a Master Patient Identifier (MPI). This is a much more robust approach for data integration, but requires an upfront technical build and ongoing technical support to load and validate the data and maintain ongoing communication with data owners.

Therefore, a critical step in planning the data integration approach is determining which option will best fit the Data Trust's business needs. This decision will influence the necessary resource allocation, timeline, and other considerations for planning the data integration approach.

5. Identify dedicated resources for building the technical infrastructure. States building a Data Trust should allocate the appropriate staff resources to building and implementing the Data Trust. At minimum, States require the following staff resources for data integration:

- ETL developers
- Architects
- Project Manager

The number of resources will depend on a variety of factors, including the number of commercial carriers submitting data; the level of usage for the Data Trust; and the resources that are currently available within the agency. Experience from other states suggests that it is better to create a dedicated team for Data Trust data integration – including data architecture, intake/loading, and communication with submitters – rather than add these functions on to an existing development team within the state agency. Allowing a dedicated team to focus fully on the Data Trust from the beginning allows the project to gain momentum.

To determine the necessary resource allocation, states should inventory the staff resources that they currently have available and identify the remaining resource gaps. Then, states should determine whether they can fill these gaps with additional staff, or if they would benefit from a vendor.

6. Identify a feasible timeline for data integration. The timeline for Data Trust data integration is typically a minimum of six months, assuming that the state has already disseminated a data submission guide. Payers typically require three-to-six months to prepare a data submission, followed by additional time for the Data Trust data integration team to check the data and troubleshoot issues with the submitters.

The data integration timeline varies greatly for large-scale data warehouses that bring in multiple disparate datasets, as each system is unique. States building such an integrated warehouse should examine the similarities and differences across all the incoming datasets to determine a feasible timeline for data integration.

7. Determine the most appropriate analytics environment that will meet business requirements. Just as states must decide the data integration approach (i.e., project-specific or fully automated data integration), states must also decide what analytics environment best meets the business requirements of its state agency users.

Using APCDs as an example, some states choose to create a research and reporting team to run all analyses and produce custom reports, extracts, and dashboards for agency users. With this approach, the agency business users identify research questions for the team to answer. Other states create a "self-serve" analytics environment in which trained business users within the state agency have access to the APCD's business intelligence (BI) tool to interact with the data themselves. Typically, these agency users are non-technical staff who understand their agency's business requirements and are interested in and comfortable with data. Role-based permissions within the BI tool allow them access to the data while still maintaining data security. While the self-serve analytics approach does not require a research team to produce many reports, it does require a strong customer service structure with qualified, accessible analysts who can train state agency users and provide them with ongoing technical assistance and support.

Historically, self-serve environments provide one of the biggest value-adds for a Data Trust, provided that the state has sufficient users who would benefit from this level of access. A self-service format provides agency users with more frequent and real-time access to Data Trust data; in addition, business users can provide valuable feedback and insights into the BI tool and allow the Data Trust team to improve and enhance the system.

While the self-serve environment provides multiple benefits, it may not be the right approach for some states. Initially, the Data Trust should have a dedicated reporting resource to assist with report development and delivery. As the Data Trust matures and reporting needs are clarified, the Data Trust can identify and train "power users" who can spearhead agency-specific reporting projects.

Action Items for Data Integration within the Kentucky Health Data Trust (KyHDT)

CHFS and its partners should complete the following activities as they develop their approach for integrating data within the KyHDT. The actions listed below are in no specific order, and can take place concurrently.

- Define ongoing data use cases for the KyHDT.
- Inventory the analytics and dashboard requirements of agency users (e.g. summary reports, custom extracts, interactive dashboards, and/or decision-support tools).

- Document data submission requirements (data elements and format) in a Data Submission Guide.
- Communicate submission requirements to all data submitters and maintain ongoing dialogue.
- Determine the data integration approach (project-specific data integration and analysis, or fully automated integrated data environment), based on use cases. Build flexibility for later development.
- Determine existing staff resources (including architects, analysis, ETL developers, and project managers) and identify gaps.
- Determine a feasible timeline for data integration based on the project scope.
- Identify the most appropriate environment for analytics and dashboards, based on business requirements.