DataSF Research Brief:

Data Governance, Quality and Integration

Author: Joy Bonaguro

1. Overall Summary

For this project, I investigated data governance, quality and integration. In the document, we review for each of these:

- What it is and why we care
- Summary of the research, including tools
- Analysis and recommendations

Below are some high level findings across all three areas.

The three concepts are connected. In practice, data governance, quality and integration meaningfully interact. A data governance program can be used to support a data quality program or an integration effort. If data governance doesn't exist, it emerges in the course of a data quality or integration project. An integration project probably needs to incorporate a consistent approach to data quality, and cleaning data is essential for integrating data from disparate systems.

We are already practicing all three. The very existence of the open data program and the practices, roles and processes we have introduced have created the basis of a data governance program. ShareSF is another example of data governance work. We have several data quality projects in the works. And SF OpenData is a form of data integration, albeit a basic one.

We should formalize and accelerate this work. Based on this research, there are several next steps that we can take. This may actually inform several of our year 3 efforts and can feed our strategic planning. Some of them, we have already identified via our other projects (e.g. creating a data standards gitbook), and we should systematically list, define and scope a set of projects based on this research. The recommendations in each section below are a good starting point to develop portions of our year 3 plan.

2. Data Governance

2.1 What and Why

Data governance has a variety of definitions. These are my favorites:

- System that decides:
    - Who is responsible for what data
    - What is expected of those who are responsible for managing data
    - The systems and roles that support the technical management and dissemination of data
- data governance is about the exercise of authority, control and shared decision-making over the management of data assets.
- data governance is the process by which a company manages the quality, consistency, usability, security and availability of its data.
- Set of approaches for how we obtain/produce, manage, share, and use data

Not having any explicit approaches to support data governance is in fact a form of governance. We are interested in data governance:

- To support the open data program
- To facilitate cross department data sharing
- To ensure that cross department needs are met in the course of generating data
- To increase the quality, usability and consistency of the City's data

2.2 Research Summary

Literature and Practices

While I read a variety of papers and articles, I primarily settled on the book "Non-Invasive Data Governance." Unlike some of the formal and heavy approaches in other readings, this book seemed to be a good fit for our City culture. Most of the practices below come from this book. Similar to data quality, I think it will be useful to have a go-to book that we supplement with other resources. Below are some high level takeaways from the book.

Frame data governance as simply formalizing work we are already doing. One of the reasons data governance projects can fail is that they come across as big and scary. Framing it this way is not threatening.

Data governance does not need to be a huge cost. A lot of the traditional vendors sell this as a huge endeavor, but the book emphasizes how it can be quite cheap, esp if you recognize it is simply a part of people's existing job. For example, if we were building software, we wouldn't see creating requirements as a different job - similarly defining data requirements (who does this and how) is a form of data governance that just needs to be formalized.

Most successful data governance programs are implemented incrementally. An incremental approach can be achieved via both the data domains and the level and detail of governance.

Governance is about behavior. For example, governance should formalize work that people are

already doing versus adding new workloads.

Apply governance to existing processes, don't create new ones. This is another approach to keep it light and cheap.

Technology is not necessary. Some governance programs have a technical component but some do not.

Frame data governance in business terms. No one cares about data governance in the abstract, just what it does for them. Same thing applies to data quality. How you should frame the business value:

- What we can't do because of the data doesn't support it
- The business value to be expected if we address the data issue

References

- https://www.amazon.com/Non-Invasive-Data-Governance-Robert-Seiner/dp/1935504851
- http://www.datagovernance.com/wp-content/uploads/2014/11/dgi_framework.pdf
- The Forrester Wave™: Data Governance Tools, Q2 2014
- http://www.oracle.com/technetwork/articles/entarch/oea-best-practices-data-gov-400760.pdf
- http://www.isaca.org/chapters3/Atlanta/AboutOurChapter/Documents/GW2014/Implementing%20a%20Data%20Governance%20Program%20-%20Chalker%202014.pdf
- https://www-01.ibm.com/software/es/events/doc/pdfs/Big_Data_Needs_Agile_Information_And_Integration.pdf
- http://www.ciosummits.com/CEO-WP-Series-Information-Governance.pdf
- https://studentaid.ed.gov/sa/sites/default/files/fsawg/static/gw/docs/ciolibrary/ECONOPS_Docs/DataGovernancePlan.pdf
- http://web.stanford.edu/dept/pres-provost/cgi-bin/dg/wordpress/wp-content/uploads/2011/11/StanfordDataGovernanceMaturityModel.pdf

2.3 Analysis and Recommendations

In practice, data governance will be an evolving, low investment activity that should not be a separate initiative but something we feed with projects and as needed. One of the things I am not recommending is a new data governance council but instead continue using working groups or committees, such as COIT, tied to specific goals or initiatives. More of a "just in time" approach to data governance.

1. Develop a governance "workplan" based on the implications in Chapter 4 of the book. Chapter 4 sets out principles and implications of those. For example, create data standards and record them in a common place. This chapter also includes a data governance test that can essentially be turned into a list of deliverables. At least some of the workplan elements should include evolving the definition of roles and responsibilities, which later chapters can help inform. This includes the notion of data domains and r/r related to those domains. This should also include the notion of a formal governance escalation (e.g. council) and how to use COIT and how it may need to be altered to reflect this extension.

2. Tie data governance workplan to clear data projects in high priority areas. We are already

doing this in practice (housing data pipeline, facilities dataset, ShareSF, etc). We should also focus on cross-department areas as that has the greatest demand for data governance. This is in contrast to announcing a large data governance initiative. Some projects may not be tied to other data projects, e.g. a data standards guidebook.

3. Incorporate clear governance into the geo strategy. In addition to high priority data projects in the prior recommendation, data governance in the geo strategy can be put to immediate work. We should consider using this as the testing bed for standard governance tools.

4. Continue to gently introduce governance principles via the open data program. By instituting data coordinators, an inventory and basic roles and responsibilities, we have already introduced data governance. We've even introduced a lightweight data classification scheme. We should continue to build this out over time. Some ways to do this may be through offering training or guidebooks on governance to data coordinators and firming up responsibilities around data requests and data concierge.

5. Create one or more data domain working groups. Another way to gently introduce data governance is to establish one or more domain working groups, ideally centered on a project. This is where the data governance work can flex into the cross-department data groups called for under our other strategy. We should investigate how to leverage the F$P as one of these key domains.

6. Monitor but do not invest in governance tools at this time. While some of the templates and tools we are developing may not scale well, we should first assess the need, demand and potential uptake of tools. While word documents are not elegant, they will do the job until we better understand how to manage this and the level of commitment and buyin.

7. Create a set of guiding policies via the COIT policy working group and ShareSF. Data policies are missing entirely from the city. We should shepherd the creation of data policies and partner both with DT, esp the CISO, as well as records managers in departments. A good set of principles to guide this is in chapter 4 of the book.

8. Research and strongly consider buying/developing a metadata management tool. Given the weakness of our metadata tools and the time consuming activities with generating and managing disparate metadata repositories, we should investigate alternatives. Our need for better metadata management will only grow and is needed to improve governance, quality and integration.

3. Data Quality

3.1 What and Why

Data quality is the degree to which information can be a trusted source for any and/or required uses. It can be further defined a number of dimensions. McGilvray offers 12 dimensions:

#	Dimension	Definition
1	Data specifications	The data models, documentation, business rules, etc
2	Data integrity fundamentals	Completeness/fill rate, validity, list of values and frequency distributions, patterns, ranges, min/max, referential integrity. This is often where most data quality stops.
3	Duplication	Measure of unwanted duplication existing within or across systems for a particular field, record or dataset.
4	Accuracy	Measure of the correctness of the data (requires an authoritative source to be available)
5	Consistency and synchronization	Measure of the equivalence of data stored across systems
6	Timeliness and availability	Measure of the degree to which data are current and available for use in the timeframe expected
7	Ease of use and maintainability	Measure of the degree to which data can be accessed and used and degree to which maintained.
8	Coverage	Measure of the completeness of the data
9	Presentation quality	Measure of how information is presented to and collected from those who use it; should support appropriate use
10	Perception, relevance and trust	Measure of the perception of and confidence in the data
11	Decay	Measure of the rate of negative change to the data
12	Transactability	Measure of the degree to which data will produce the desired business transaction or outcome

We care about data quality because bad quality can:

- Contribute to bad decisions
- Increase costs (staff time, mailing, duplication etc)
- Create compliance/legal risk
- Lead to sub-optimal use of resources

Loshin defines poor quality as falling into the following groups:

1. Hard impact - things that can be measured (e.g. customer attrition, cost from rework, cost fixing problems, delays in processing etc)
2. Soft impact - evident but hard to measure (e.g. difficult decision-making, employee morale)

3. Decreased revenue
4. Increased costs
5. Increased risk
6. Lowered confidence

3.2 Research Summary

Data quality is relative. There is no such thing as perfect data quality. Data is collected for a purpose and the quality needs for that purpose may be different than for other purposes. This is is particularly relevant to open data or even internal sharing of data as we are often repurposing or broadening the use of data from its original intent.

Data quality is a relatively young discipline and there is some disagreement on what it consists of. Most organizations don't have designated staff working on data quality, treating it has a distinct discipline. And there weren't many "official" approaches to model.

Data quality issues are associated with large costs. While the numbers are hard to verify, numerous reports cite data quality as costing billions to trillions to the economy.

Start with specifications and data integrity. It is very hard to address other dimensions of data quality if the data is poorly specified and lacks integrity. A data quality project probably needs to start with and include this if not in place.

Much like design and accessibility, data quality should be integrated into the beginning and many issues are preventable. If data is validated against business rules at the point of collection and data collectors are consistently trained, data quality will improve as many issues are preventable.

Data quality metrics can be established and tracked. Ongoing tracking and monitoring is probably more cost-effective than episodic data purges.

Modern architectures should offer data quality rules as a service. This could include validation as a service etc. It could also be simply having a shared set of rules to adopt (e.g. in the form of a guidebook).

3.3 Analysis and Recommendations

1. Continue working on targeted data quality improvements and use the 10 steps process from McGilvray as our playbook; document ROI when possible. This book had the highest reviews and really is a clear playbook for doing data quality projects. During the course of projects, we should clearly document and communicate the ROI of our projects. Note that this recommendation is about diagnosing existing problems. We need more approaches to move to a preventative stance.

2. Investigate and invest in a data profiling tool. A data profiling tool should be part of our repertoire and will allow us to deepen and expand our data quality work.

3. Develop and conduct data quality and data profiling trainings. Once we acquire a tool, we should offer classes on that or an open source tool that also includes an overview of the basics of

data quality.

4. Partner with the DT Applications group and the new digital services team to:

1. ***Define a professional set of roles/responsibilities for data quality.*** Deliberately cultivating data quality expertise as part of our core development approach should be another piece of product quality.
2. ***Develop a data specifications and integrity guidebook.*** This should be a consistent methodology and approach for developing standards, data models, business rules, metadata, reference data, assessment and monitoring, and training. Use the dataset management framework and requirements worksheet as a starting point.
3. ***Develop a data quality guidebook.*** This can be the basis for us to consistent data standards (or starting defaults) over time.This guidebook should include the contents of a future training offered through Data Academy.
4. ***Investigate developing data validation and quality services to be consumed by new digital services team.*** This team can initially develop and then offer these as a service to other departments and digital teams. The guidebook can serve as the bridge strategy and for when things cannot be serverized.
5. ***Pilot data quality scoreboards/cards to shift to ongoing versus periodic purges.*** Many data quality projects are episodic. Having live dashboards of data quality versus targets can reduce the overall maintenance effort.

5. Incorporate data profiling as a standard and initial service for publishing. Data profiling can quickly identify problems in the dataset that may emerge during the publishing process. Catching this at the beginning of the process can reduce back and forth and provides a structural way to improve data quality in partnership with departments.

6. Leverage the open data portal as a repository for reference data. As a transitional strategy, the open data portal can be used to store common reference data (e.g. department names and codes, various geo data etc). We can consider adding an indicator or flag to reference data sets, e.g. (Reference data). In addition to storing reference data, during the publishing process we should also use it to clean up data during the ETL process.

3.4 Appendix

Tools and Vendors

Expected functionality. Tools related to quality and governance have a high overlap. Data quality tools can include:

- Profiling
- Cleansing, deduplication
- Stewardship
- Metadata management
- Automation
- Modeling
- Matching
- Many tools also include ETL tools

Vendors. The list of vendors is cross listed with the data integration tools plus additional tools

from Gartner's magic quadrant.

- Informatica
- Talend
- Information Builders
- Experian
- SAS
- SAP
- IBM
- Trillium
- Ataccama
- Tamr
- DataMartist

References

- http://data-informed.com/five-steps-to-ensure-data-quality-success/
- https://www.edq.com/uk/blog/how-to-improve-data-collection-accross-multiple-departments/
- https://www.edq.com/uk/blog/3-2-1-start-measuring-data-quality/
- http://www.damauk.org/RWFilePub.php?&cat=403&dx=1&ob=3&rpn=catviewleafpublic403&id=106193&sid=136eb0aa1540985a4cc04a6c75c45ffc
- https://www.edq.com/uk/blog/from-headless-chickens-to-lean-and-agile-data-quality-practitioners/
- https://www.edq.com/uk/blog/data-quality-whats-your-role/
- https://www.edq.com/uk/blog/data-quality-maturity-a-customer-perspective/
- Executing Data Quality Projects This book should be our bible.

4. Data Integration

4.1 What and Why

Data integration is a mix of tools, processes, and architecture for extracting, matching and loading data into systems and applications to:

- Acquire data for data warehousing
- Control data lists/sets (reference and master)
- Move/synchronize data across applications
- Share data
- Match and reconcile data

A key flavor of integration is that the data is unified in some fashion, often to be consistent with a shared data model. So in the process of integrating data it is also transformed to be standardized and combined and then available via a data warehouse. In contrast, our work, i.e. SF OpenData, is closer to a data hub where disparate data sources at different levels of granularity are made available, but not intentionally linked. SF OpenData could be considered a basic level of data integration.

The common functions are data extraction, cleaning, standardizing and joining and reconciling then providing the data either via data stores or to applications etc.

We are primarily interested in data integration to support:

- Technical solutions for confidential data sharing, which has more specific requirements around matching and protecting confidential data.
- Options to develop shared data systems for internal use. This is more generic and opportunistic seeking - what is state of the art or opportunities to explore in greater detail.
- Generally, as part of data quality and open data, including enforcing data standards via data as a service.

4.2 Research Summary

Definitions

Integration. Infrastructure for enabling efficient data sharing across incompatible applications that evolve independently in a coordinated manner to serve the needs of the enterprise and its stakeholders.

Data Categories. Organizations can divide their data into 4 types.

1. Master data. People, places and things that are part of organization's business processes.
2. Transactional. Describes business events such as buying products from suppliers.
3. Reference data. Sets of values or classification schemes used by business applications, e.g. lists of valid values.
4. Metadata. Data about data.

I think of it as master data = nouns, transactions = verbs and reference, & metadata = dictionary.

Master Data Management (MDM). Set of disciplines and methods to ensure the currency, meaning, and quality of master (and reference) data.

Customer Data Integration. A customer domain specific version of MDM.

Literature and Practices

Below are summary points from the literature and suggested practices. Some of the references recommended different approaches and the largest area of disagreement was on data and reference models.

Most organizations don't have a disciplined approach to integration, leading to high costs and complexity. The costs and complexity are driven by many point solutions, cumulating in a collection of works of arts. This is because many platforms/apps maintained and evolving distinctly based on local business needs. When integrations happen, they are point to point and a hodgepodge of middleware may be deployed over time.

Integration should be a discipline with a dedicated team as part of an enterprise strategy - not a project. Most references encouraged approaching integration as an expertise to be staffed, ideally centrally as a shared service. This allows for:

- Build up in expertise
- Repeatable and teachable processes, and
- Treatment of integration as an ongoing activity requiring maintenance and continuous improvement.

The literature also suggests that it's straightforward to develop a business case for a data integration team.

That team should have an integration lifecycle framework, distinct but related to SDLC. The literature notes that this is simpler than software and should include:

1. Identify data requirements
2. Data analysis
3. Integration design
4. Develop integration
5. Verify and validate
6. Deploy validation

The integration team should have a core set of tools/processes. Those tools will likely include:

1. Customer engagement request system
2. Requirements capturing tool
3. Integration feed design tool
4. Build/dev tool
5. QA
6. Production configuration
7. Scheduler
8. Knowledge base repository

MDM integration projects are more involved than other integrations. Their focus is about

providing reconciled and authoritative data to other systems. In business speak, these types of projects are often used to develop a "360" degree view of the customer. In this case a central hub is normally used that:

- Becomes the single point where other apps retrieve data
- Ensures consistent value representation
- Merges data across sources
- Cleans and reconciles data

These types of projects can be deployed in 3 ways:

- Registry - data stays in source system and pointers or linkages in integration hub
- Persistent - copies data as a physical record
- Hybrid

Integrations are an excellent point to enforce data quality standards/metrics/monitoring. Data quality, gov and integration end up being highly related/dependent.

Avoid common sources of "waste" in integration programs. Lean integration identifies 5 key contributors to waste in integration programs.

1. Gold Plating. Building features before they are needed. Instead, build on a project basis and refactor as needed. Integration can be costly, so instead of overbuilding, refactor.
2. Using too much middleware. Middleware costs money and sometimes P2P integrations are the best answer.
3. Reinventing the wheel. Integrations fall into a handful of patterns over time and should not be treated as "works of art".
4. Unnecessary complexity. This can come from a mix of tools, platforms, protocols, formats etc.
5. Not planning for retirement. Not all orgs get rid of old, unused integrations.

Common Architectures. Data integration architectures commonly consist of the following functions:

- Data stores (the source production systems)
- ETL engine
- Data staging areas
- Data warehouses (which keep the historical data)
- Data marts (smaller version of warehouses)
- Data cube
- Operational data stores
- Reporting tools etc

The particular architecture depends on the needs of the business users coupled with data latency and performance issues.

Expected Tool Functionality

Gartner does a decent job of summarizing what functionality you should expect from at least the tool side of data integration. This makes a nice "starter checklist" for requirements. Below are

some edited examples from their document:

- Connectivity/adapter capabilities (data source and target support). The ability to interact with a range of different types of data structure, e.g. RDMS, xml, Saas, etc
- Support different ways to interact, e.g. bulk/batch, trickle, change data capture, event-based
- Data delivery. Ability to provide data to apps, processes and databases in multiple ways (batch, federated, message, replication etc)
- Transformation tools (standard and custom
- Metadata and modeling
- Design tools for creating the data integration processes
- Data governance support

Tools

The vendor space feels split between major investments or very lightweight tools. The list of vendors below is mostly from Gartner's magic quadrant but also tools captured over time. Very expensive, heavy solutions are likely not appropriate at this time. I cross referenced these tools with data quality/governance tools.

- Informatica
- Talend
- Information Builders
- SAP
- SAS
- IBM
- Tamr
- Mulesoft
- Starfish ETL
- Zapier
- SnapLogic
- Boomi from Dell

References

Below is a list of readings used.

- Numerous online articles
- http://www.slideshare.net/roddickerson/dickersondataarchgovernancex-1223017
- http://www.oracle.com/us/products/middleware/data-integration/odi-ingredients-for-success-1380930.pdf
- Getting Data Right via O'Reilly
- http://www.amazon.com/Lean-Integration-Factory-Approach-Business/dp/0321712315/
- http://www.amazon.com/Customer-Data-Integration-Reaching-Institute/dp/0471916978

4.3 Analysis and Recommendations

As a city, we do not have a strategy or approach for data integration and this project was consciously exploratory - should we be doing something more deliberate and if so, what? Creating a business case for a complicated data integration project without a very strong business driver

feels premature. At the same time, our general lack of significant investment also means we can learn from others mistakes and potentially leapfrog older, more costly approaches. We recommend:

1. **Establish a data services and infrastructure learning group with other key departments and COIT.** A cross department learning group provides an opportunity for us to learn together and when ready, make the appropriate investments. This learning group should leverage expertise in other organizations and sectors. In addition, we should reach out to IT departments with existing data integration and infrastructure including PUC, MTA, DPH, in particular the CCMS deployment, and DPW.
2. ***Work with the new digital services team to:***
    a. ***Develop a professional set of roles/responsibilities and tools for data integration.*** We should deliberately create and designate data integration expertise as part of our digital services strategy to increase product quality and contribute to scalability.
    b. ***Develop a set of data services and architecture.*** As part of streamlining development and contributing to data consistency and quality, we should develop a suite of data services, including repositories, reusable components, APIs and strategy, and supporting data infrastructure and architecture.
3. **Conduct a more in depth research/analysis on data matching tools for confidential data sharing and per initial requirements.** Ideally, we should develop a business case and pilot and test one or more tools.