6 Things You Need to Know About Data Governance

John Klein

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213
1. A data set produces benefits only when it is used to make decisions.
2. \( \text{value} \equiv \sum \text{benefits} - \sum \text{costs} \)
3. Moving Parts

- Sensor
- Open source feed (e.g., Twitter)
- 3rd Party Source (e.g., geo map)
- External system

Producer → Publisher → Consumer → Decision-Maker

- Data
- Acquires
- Stores
- Makes available
- Develops decision support application
- Uses application to make decisions
Our Focus

Producer → Publisher → Consumer → Decision-Maker
Easy!

Publisher → Consumer

Challenging!

Publisher → Consumer

Publisher → Consumer

Publisher → Consumer
4. Governance constrains the data publisher, in order to help the data consumers
Constrains the publisher...

May *reduce* the costs to the publisher

- Reduce types and numbers of interfaces
- Restrict backward compatibility on interfaces
- Restrict technology options

Usually *increases* the costs to the publisher

- Different schema
- Data quality
- Quality of Service (e.g., availability)
...to help the consumers

Reduce local and global costs of consuming the data
• Standard interfaces
• High quality data
• Service level agreements

Enable consumers to deliver more benefits
• New decision support applications
• Higher quality decisions
5. Apply governance only when it increases value (benefits > costs)
6. Focus your governance on the things that data consumers want
Data Consumer Concerns

Category

1. Existence and Appropriateness
   - What data sets are available?
   - Privacy? Proprietary Data? Data use restrictions?
   - Is the data set managed?
   - Will it be available for as long as I need it?

2. Data Set Semantics
   - What information does the data set represent?
   - Raw?
   - Cleansed/processed?
   - Derived?
   - What type of data does it contain?
   - Where does it come from?
   - Provenance? Trustworthiness?
   - Timeliness? When is the data set updated?
   - How does this data set relate to other data sets?

3. Data Record Semantics
   - Schema/Vocabulary
     - Indexes/Views?
     - What queries are possible?

4. Data Access and Processing
   - Location
     - Where is it?
     - Can I reach it?
   - Technology constraints - special client library, etc.?
   - APIs/Protocols?
   - Can I process it in-place or do I need to copy?

5. QoS of access
   - Batch/Interactive?
   - Availability?
   - Request latency?
1 Existence and Appropriateness

- What data sets are available?
- Privacy? Proprietary Data? Data use restrictions?
- Is the data set managed?
- Will it be available for as long as I need it?
Data Set Semantics

- What information does the data set represent?
- What type of data does it contain?
  - Raw?
  - Cleansed/processed?
  - Derived?
- Where does it come from?
  - Provenance? Trustworthiness?
- Timeliness? When is the data set updated?
- How does this data set relate to other data sets?
3 Data Record Semantics

Schema/Vocabulary
Indexes/Views?
What queries are possible?
Data Access and Processing

- Location
  - Where is it?
  - Can I reach it?

- Technology constraints - special client library, etc.?

- APIs/Protocols?

- Can I process it in-place or do I need to copy?

<table>
<thead>
<tr>
<th>QoS of access</th>
<th>Batch/Interactive?</th>
<th>Availability?</th>
<th>Request latency?</th>
</tr>
</thead>
</table>
Data Consumer Concerns

1. Existence and Appropriateness
   - What data sets are available?
     - Privacy? Proprietary Data? Data use restrictions?
     - Is the data set managed?
     - Will it be available for as long as I need it?

2. Data Set Semantics
   - What information does the data set represent?
     - Raw?
   - What type of data does it contain?
     - Cleansed/processed?
     - Derived?
   - Where does it come from?
     - Provenance? Trustworthiness?
   - Data Quality
     - Timeliness? When is the data set updated?
   - How does this data set relate to other data sets?

3. Data Record Semantics
   - Schema/Vocabulary
     - Indexes/Views?
     - What queries are possible?

4. Data Access and Processing
   - Location
     - Where is it?
     - Can I reach it?
   - Technology constraints - special client library, etc.?
   - APIs/Protocols?
   - Can I process it in-place or do I need to copy?

5. QoS of access
   - Batch/Interactive?
   - Availability?
   - Request latency?
Playbook for Data Governance

Identify your high benefit decisions
  • Infrequent but high impact, high frequency but low impact, …
Identify the data sets that support your highest benefit decisions
What is the producer/consumer relationship for each high benefit data set?
  • 1-1 → Probably no governance needed
  • 1-Many, Many-Many → Consider applying governance
What to govern = what constraints to impose on the producer?
  • Use Data Consumer Concerns checklist – where are your gaps?
  • Balance cost and benefit to keep value positive
Work down the list of high benefit decisions
Review periodically – have your high benefit decisions changed?
Contact Information

John Klein
Senior Member of the Technical Staff
Architecture Practices Initiative
+1 412-268-7378
jklein@sei.cmu.edu

U.S. mail:
Software Engineering Institute
Customer Relations
4500 Fifth Avenue
Pittsburgh, PA 15213-2612
USA

World Wide Web:
www.sei.cmu.edu
www.sei.cmu.edu/contact.html

Customer Relations
customer-relations@sei.cmu.edu
+1 412-268-5800