

Software Solutions Symposium 2017

March 20–23, 2017

6 Things You Need to Know About Data Governance

John Klein

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213



Copyright 2017 Carnegie Mellon University

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[Distribution Statement A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

DM-0004340

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

Producer



Publisher



Consumer



Decision-Maker

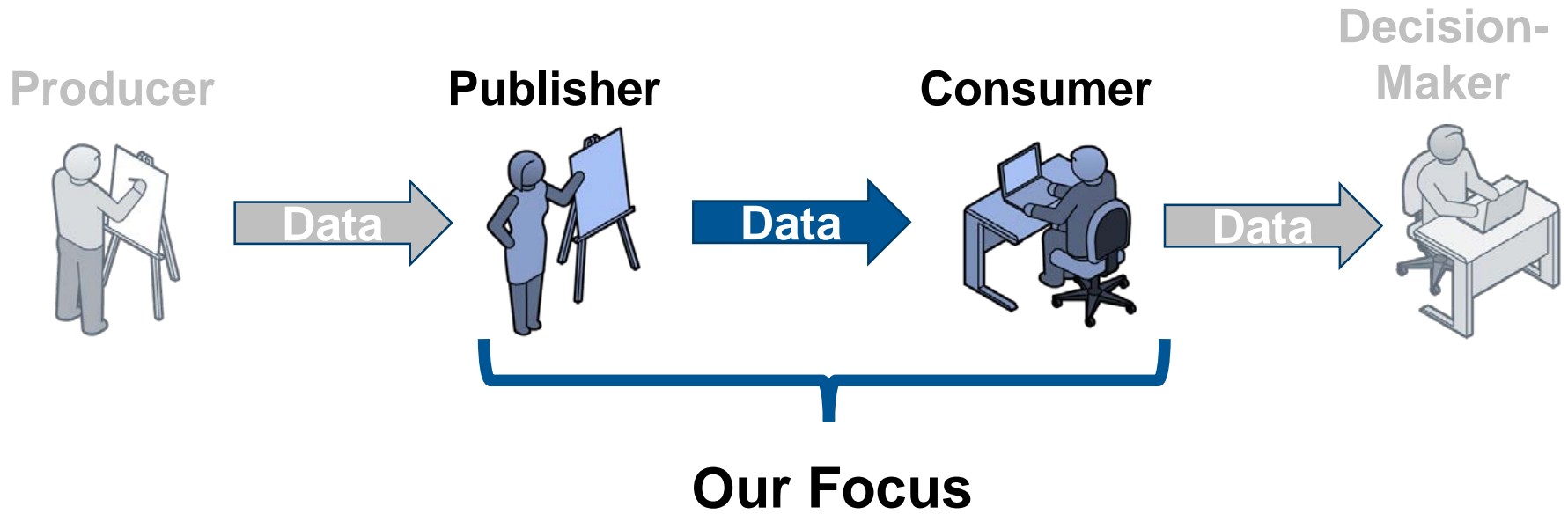


- Sensor
- Open source feed (e.g., Twitter)
- 3rd Party Source (e.g., geo map)
- External system

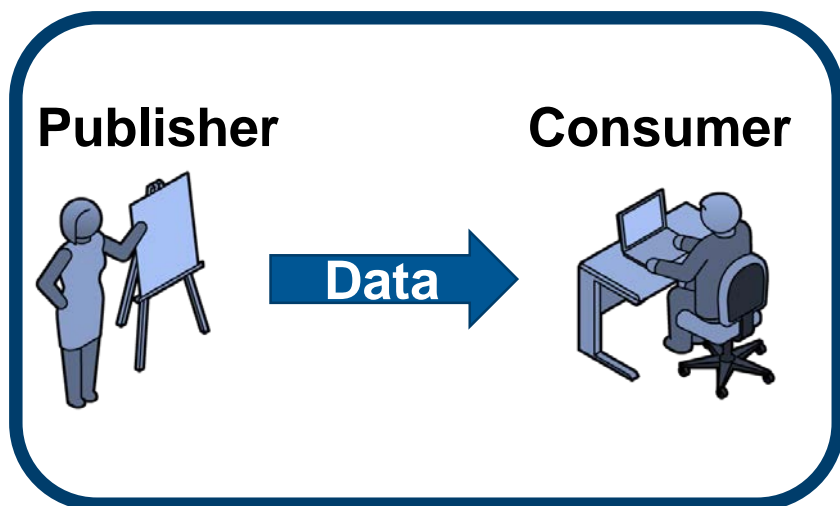
- Acquires
- Stores
- Makes available

- Develops decision support application

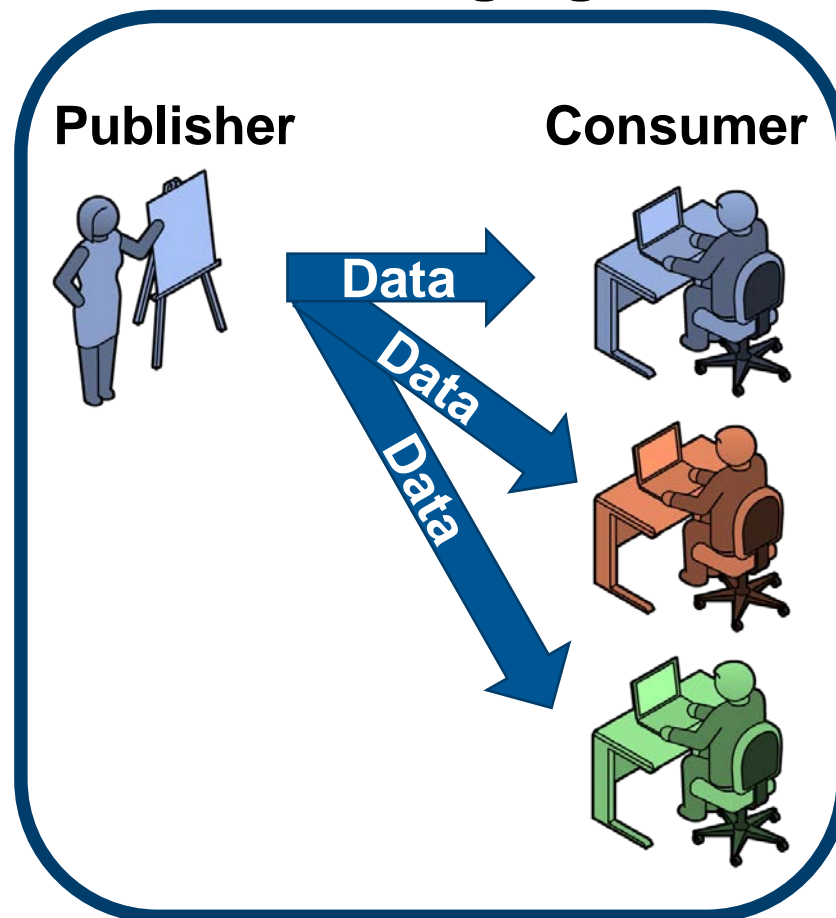
- Uses application to make decisions



Easy!



Challenging!



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

Category

Typical Questions



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?

Data Quality

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?**
- Authorization? Credentials? Access Controls?**

5 QoS of access

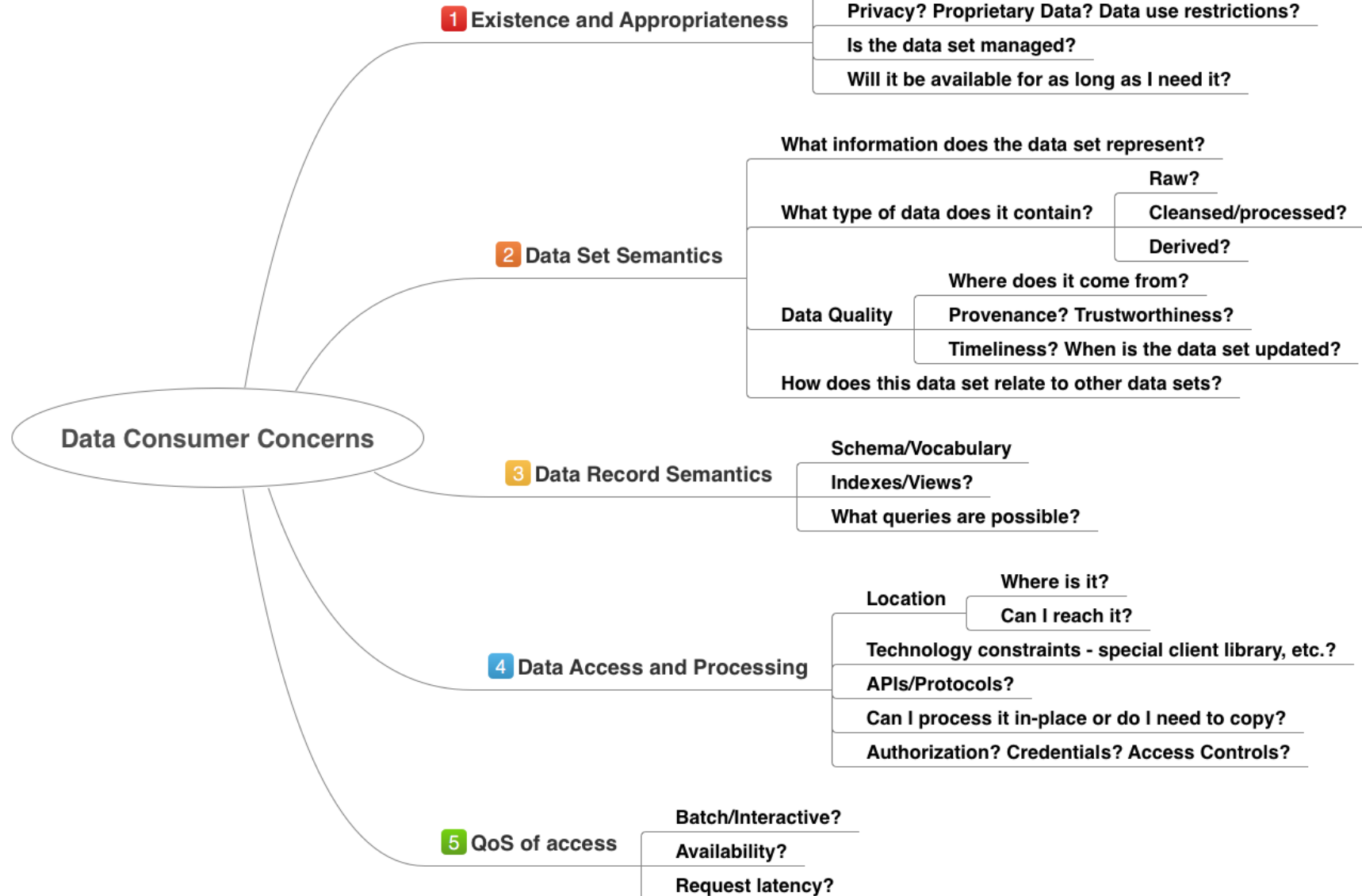
Batch/Interactive?

Availability?

Request latency?

Category

Typical Questions



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