Moving from 3rd Normal Form to a web enabled world

Matthew West

Information Junction
Information Management Experts

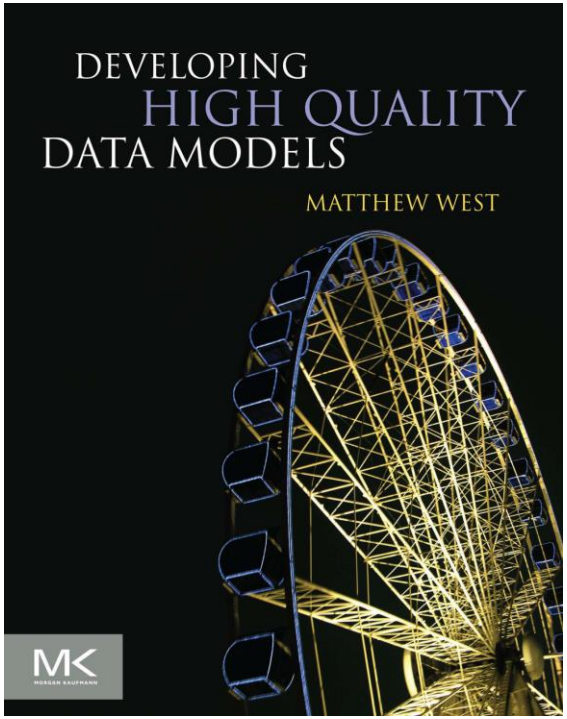
Matthew West

Shell (1978-2008)

- Initially Refinery Technologist
- Shell IT Planning for Manufacturing1989
  - Developed Refinery Data Models for Operations and Engineering
- Shell Group Data Management 1990
  - Developed Data Management Policy and Guides for Shell Group, especially on data modelling and data quality
- Founding Chair of EPISTLE 1993
  - EPISTLE – European Process Industries STEP Technical Liaison Executive
- Shell Global Asset Information Management 1995
  - Providing support for information management for major projects
- Technical lead in development of ISO 15926-2 Data Model
- Technical lead for Shearwater Project – first financially successful implementation of EPISTLE/ISO 15926
- Reference Data Architecture and Standards Manager for Shell Downstream 2003
  - Strategy and Policy for managing Master and Reference Data
  - Developed Downstream Data Model
- Contributor to ISO 8000 – Data Quality

Post Shell - Various projects undertaken including RSSB, UK MOD, and BP

Currently – Working for Platts with Datasmiths



Key Requirements for Information Systems

To manage information, you need to be able to meet the following requirements:

- know what information exists, and what it is about,
- extract portions of the information suitable for a particular purpose,
- exchange data between organizations and systems,
- integrate information from different sources, resolving what information is about things you already have information about, and what is about new things,
- share the same data between applications and users with different views, and
- manage the data, including history, for life.

It is not unusual for some or all of these requirements to be difficult and expensive to meet.

Desiderata

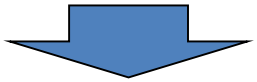
Integrating data models should:

- ✓ meet the data requirement,
- ✓ be clear and unambiguous to all (not just the authors),
- ✓ be stable in the face of changing data requirements,
- ✓ be flexible in the face of changing business practices,
- ✓ be reusable by others,
- ✓ be consistent with other models covering the same scope,
- ✓ be able to reconcile conflicts between other data models, and
- ✓ It should be possible to develop data models quickly.

3NF (and 4&5NF) Current State Model

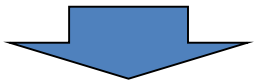
Record for Jack at 15/3/1985

Name	Date of Birth	Marital Status	Weight (kg)	Height (m)
Jack	15/03/1985	Single	4	0.45



Record for Jack at 15/3/1995

Name	Date of Birth	Marital Status	Weight (kg)	Height (m)
Jack	15/03/1985	Single	30	1.3



Record for Jack at 15/3/2015

Name	Date of Birth	Marital Status	Weight (kg)	Height (m)
Jack	15/03/1985	Married	65	1.85

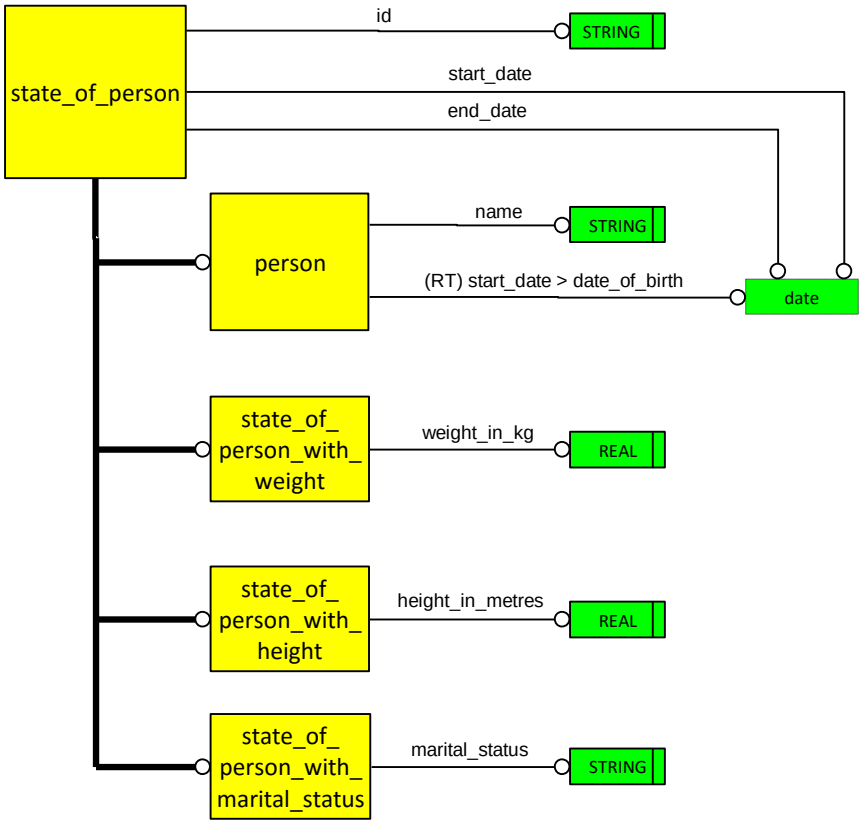
Managing change over time

Just add a date?

Name	Date of Birth	Marital Status	Weight (kg)	Height (m)	Date
Jack	15/03/1985	Single	4	0.45	15/3/1985
Jack	15/03/1985	Single	30	1.3	15/3/1995
Jack	15/03/1985	Married	65	1.85	15/3/2015

Unfortunately, this now has repeating groups, so it is not in 3NF.

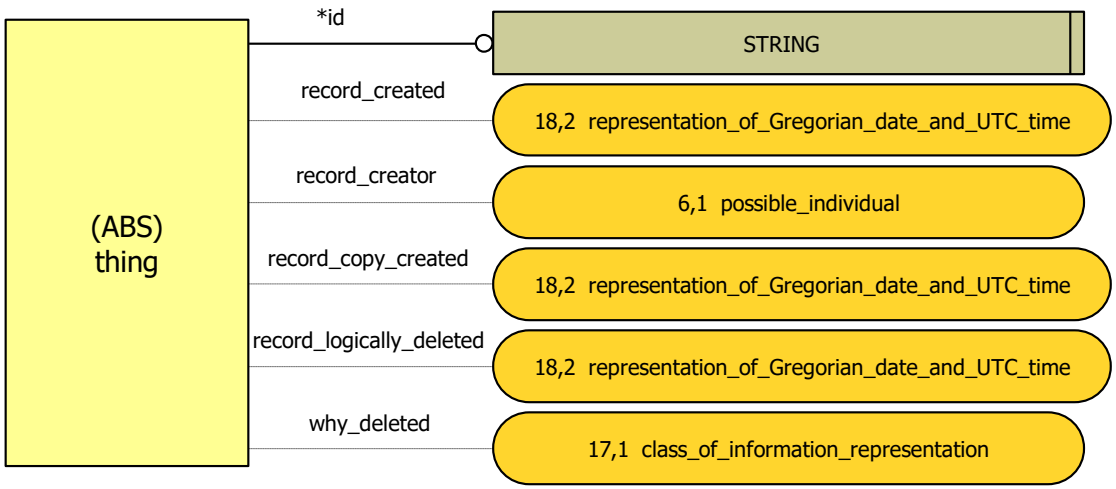
We need to renormalize



Here is one way you can do it.
We now have 6NF (change over time and 5NF)

We need to add metadata for e.g. provenance

Here is an example from ISO 15926

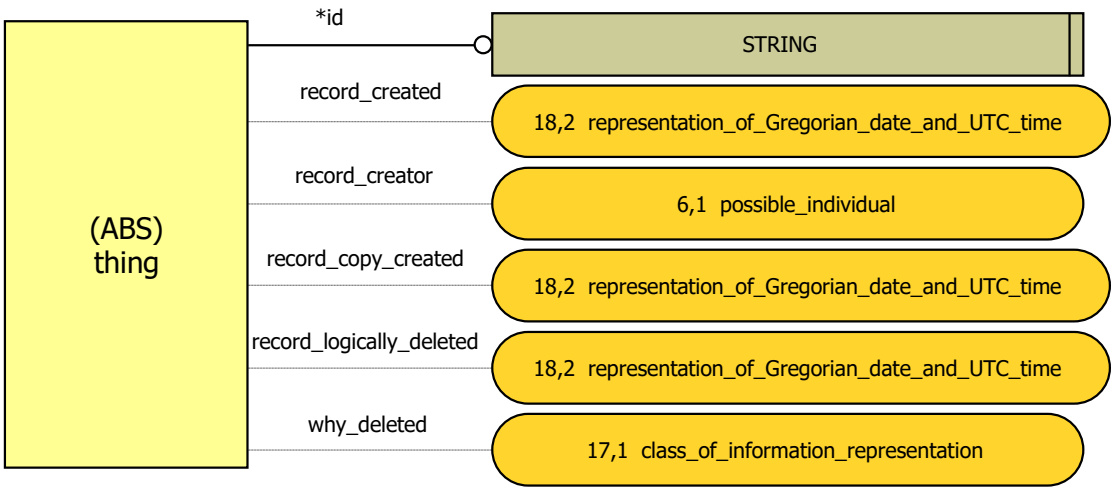


What more can we do?

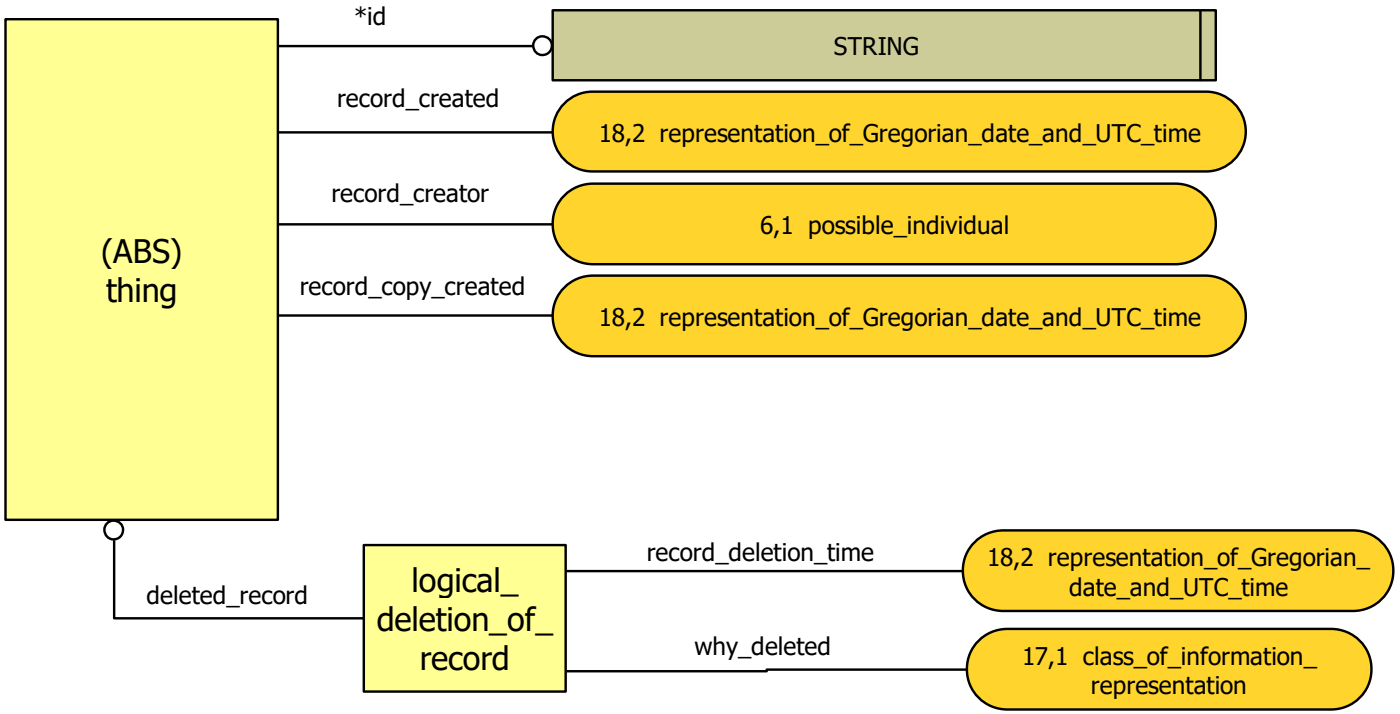
- In data bus based systems it adds complexity if you have to make changes to records.
- On the web it is complex to manage changes to records when the users maybe unknown.

- The next step is records that never change. Data is only ever added.

Web Normal Form (or 7NF)



Modified model requiring no updates



A Web Normal or 7NF?

Triple stores and beyond

- **The problem with triple stores**
  - They break your data into tiny bits where what you wanted was something like the record we started with which had all the information about Jack in one record
- **Named Graphs**
  - A named graph is a way of collecting together a number of triples. Nominally it makes a record one of the related objects in a triple
- **Quad stores**
  - Quad stores are acknowledged by W3C, but there is no definition of what the fourth element is for.
  - Some implementers of quad stores are likely to have used the 4th element to support named graphs
  - In ISO15926 we found named graphs useful for what we called Object Information Models, which effectively gives you a view on the database, so you can have a named graph that consists of all the data about Jack.
- **Quint Stores?**
  - We are still left with what to do with data about a record, record creation date etc.

Questions?