De-normalising data for archival preservation

Jan Rörden, University of Cologne

20th February 2015

University of Portsmouth

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO
ARHIV REPUBLIKE SLOVENIJE

DLM FORUM FOUNDATION

STATENS ARKIVER
THE DANISH NATIONAL ARCHIVES

DigitalPreservationCoalition

THE E-ARK PROJECT IS CO-FUNDED BY THE EUROPEAN COMMISSION UNDER THE ICT-PSP PROGRAMME

www.eark-project.eu

GOBIERNO DE ESPAÑA
MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS

ICTPSP

ama

KEEPSOLUTIONS
University of Minho SPIN-OFF

MAGENTA aps

ES Solutions

THE NATIONAL ARCHIVES OF NORWAY

RAHVUSARHIIV
THE NATIONAL ARCHIVES OF ESTONIA

archívum
Magyar Nemzeti Levéltár

TÉCNICO LISBOA

E-ARK

Content

1. Normalised databases
2. De-normalisation
3. Archival context – Problems
4. Archival context – Benefits

1. Normalised databases

- Optimised for use
  - Not optimal for reading/querying.
- Designed to avoid redundancies
  - Consistency – avoid anomalies
  - Optimize storage requirements
- Several levels of normalisation

First normal form (1NF)

CD_ID	Album Title	Artist	Published	Track Nr.	Song Title
001	Master of Puppets	Metallica	1986	1	Battery
002	Metallica	Metallica	1991	8	Nothing Else Matters
001	Master of Puppets	Metallica	1991	5	Disposable Heroes
003	Zeitgeist	Smashing Pumpkins	2007	1	Doomsday Clock

Second normal form (2NF)

CD_ID	Album Title	Artist	Published
001	Master of Puppets	Metallica	1986
002	Metallica	Metallica	1991
003	Zeitgeist	Smashing Pumpkins	2007

CD_ID	Track Nr.	Song Title
001	1	Battery
001	5	Disposable Heroes
002	8	Nothing Else Matters
003	1	Doomsday Clock

Third normal form (3NF)

CD_ID	Album Title	Published	Artist_ID
001	Master of Puppets	1986	11
002	Metallica	1991	11
003	Zeitgeist	2007	22

Artist_ID	Artist
11	Metallica
22	Smashing Pumpkins

CD_ID	Track Nr.	Song Title
001	1	Battery
001	5	Disposable Heroes
002	8	Nothing Else Matters
003	1	Doomsday Clock

- Normalised structure is best for usage, if usage means that data is added
  - Easy maintenance of data
- Unbiased regarding search pattern
- (rather) inefficient if data should be retrieved
  - Several/complex queries required to retrieve desired information

Preserve

- What should be preserved/what is important?
  - Only the content? Parts of it?
  - How the content was accessed/delivered?
  - Everything: content, transactions, behaviour?
  - (Maybe you will also need to preserve the application?)

Format normalisation

- Format normalisation != Database normalisation
- Store the database in an open format, suitable for preservation
  - Resemble original structure
  - Keep the content
- SIARD format

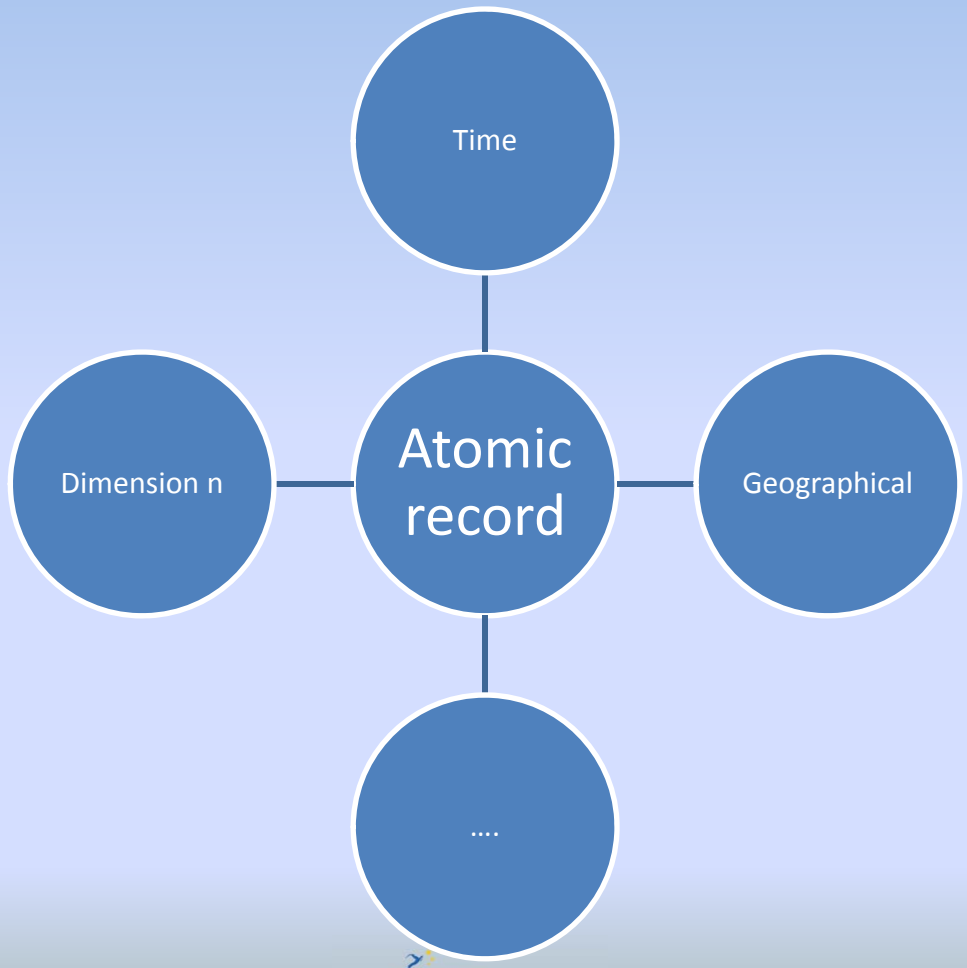
But:

- What happens if information is lost?
  - Tables that are no longer connected?
  - Context/documentation of the database is non-existant or lost?
    - Some Databases/Tables make no sense if you lack information.
- Why not go beyond „simply“ storing the database?

2. De-normalisation

- Time-space tradeoff: improve „read“ performance
  - (re-)introducing redundancy – more storage capacity is required
  - Materialised views: results of search queries stored in tables
  - Reorganize database

Example: Star schema



- Require a certain view of the database:
  - Which information is most important for queries?
  - Can be unflexible for varying analytics.

- Simpler queries + performance gains
- Fast aggregations
- Feed OLAP cubes

De-normalisation - questions

- How to do it?
  - Manually? Automatically? Which view?
- When to do it?
  - During ingest? When access is given? Never?
- How to handle BLOB/CLOB linked/inside the database?

3. Archival context - Problems

- Can be difficult to create
- De-normalisation comes with undesired effects:
  - Original context/structure is lost
  - Rendering authenticity decreases

- De-normalisation != fit for preservation

4. Archival context - Benefits

- Robust: complexity is reduced
  - Easier migration
  - Accessibility increased
- Archiving Service Oriented Architecture (SOA):
  - Database snapshots are useless, if they refer objects that are not going to be stored with it.

- Data mining as part of the lifecycle: dissemination as part of the archiving strategy
  - OLAP

OLAP

- Online Analytical Processing
- Multi-dimensional analytical queries
  - Analyze multidimensional data from multiple perspectives.
  - Get a lot of information very fast.

Thank you!

jan.roerden@uni-koeln.de