In praise of inconsistency
the long tail of small data

Alan Dix
Talis and Lancaster University

www.hcibook.com/alan/
alandix.com/blog
today I am not talking about ...

- intelligent internet interfaces
- visualisation and sampling
- situated displays, eCampus, small device – large display interactions
- fun and games, virtual crackers, artistic performance, slow time
- physicality and product design
- creativity and Bad Ideas
- modelling dreams and regret

... or even lots of lights

http://www.hcibook.com/alan/projects/firefly/
back in the 1980s ... Codd and all that

• in theory:
  – normalisation, atomicity
  – illusion of single use & strong internal consistency

• in practice
  – de-normalise for efficiency
  – maintain consistency through controlled transactions
  – business logic, APIs

the IS ideal
the more things change ...

... the more they stay the same
limits of consistency

consistency not always possible

- distribution and caching
- multi-user update (Alison and Brian)
- view-based updates

ordering problems
(race conditions)

Alison: It's a beautiful day. Let's go out after work.
Alison: perhaps not, I look awful after the late party
Brian: I agree totally

Alison: It's a beautiful day. Let's go out after work.
Brian: I agree totally
Alison: perhaps not, I look awful after the late party
limits of consistency

consistency not always possible

- distribution and caching
- multi-user update (Alison and Brian)
- view-based updates

--

view based update

complimentary functions

view / display

central state / data base
view based update
complimentary functions

view / display

central state / data base

... always

goal is eventual consistency
sometime not possible

distributed garbage collection
– various algorithms ...
  aim to make sure referenced items not lost
– but always storagee node can die

• options:
  – prevent loss of referenced item
  – accept loss of referenced item
    • leases or “ref not found” exceptions

what is consistent?

• conflicting updates
• long-term transactions
• synchronisation

... and Apple still can’t get it right!!
internal and external consistency

• the exam board ....

is the world consistent anyway?

• departmental lists
a different approach

do not enforce consistency

but highlight inconsistency

• instead of views of central data, related yet different sources
• specify connections and automatically check inform of updates, highlight discrepancies but allow divergence
concept – workspaces

spreadsheet on colleague’s PC

‘workspace’

central institutional database

table in word doc on your own PC

fast forward ten years ...

• semantic web and RDF
  – open schema (but can be specified)
  – open world model
  – flexible and extensible (e.g. Volkswagen)

• individual data sets
  – ontology engineering – getting the model right

• linking open data
  – connecting web of data
  – shared vocabularies and URIs
linking open data

linking through:
• shared
• dereferencable
• URIs
fast forward ten years ...

- semantic web and RDF
  - open schema (but can be specified)
  - open world model
  - flexible and extensible (e.g. Volkswagen)

- individual data sets
  - ontology engineering – getting the model right

- linking open data
  - connecting web of data
  - shared vocabularies and URIs

the long tail

- a few very large data sets
  - e.g. Open Govt., OS, geonames, dbpedia

- the small data of ordinary life:
  - from local bus timetables
  - to squash club league tables
supporting small data?

- Google fusion tables
- Google refine
- Freebase
- Kasabi

• mostly for ‘middle’ sized data

really small?

personal, but also Govt.

often tables

describe semantics rather than ‘converting’

• explicit – simple description
• implicit – semantics through interaction
explicit – 3 levels

• the table as it is:
  – there are 5 columns
    col 1 is called “name”, col 2 is ‘population

• internal semantics of table (in its dataset)
  – each row is the properties of a country entity
    defined by the ‘name’ column

• external linkage to standard data/vocab
  – rules + exceptions
  – country is ‘sameAs’ geoname country by matching name
    except ‘Wales’ is geoname administrative region ...

implicit

action is specification:
• view a table and give it a name
• link items/columns from different data sources
• perform calculation

semantics are emergent through use
SO ...

long history of consistency
... but not always possible or desirable

do not enforce consistency
but highlight inconsistency

exploit the long tail of small data

plus ...

come to
Tiree Tech Wave
3-7 Nov 2011