Models of Provenance

Peter Buneman

University of Edinburgh

Thanks to: Adriane Chapman, James Cheney, Susan Davidson, Egor Kostylev, Sam Lindley, Heiko Mueller, Wang-Chiew Tan, Val Tannen, Stijn Vansummeren and lots of others

The population of Corfu

113.479 (2001)	http://www.corfunext.com/corfu_geography.htm
107,879 (as of 2001)	http://en.wikipedia.org/wiki/Corfu ***
93,000	http://www.corfunet.com/corfu/
109,512	www.agni.gr/
110,000	www.corfuvisit.net
70,000	http://www.newadvent.org/cathen/04362a.htm
107,600	http://www.greek-hotels.com/
105,043	http://www.merriam-webster.com/dictionary/corfu
approximately 110,000	www.kassiopi.com/MenuContent.aspx?MenuId=6
approximately 120.000	http://www.gardeno-corfu.com/
115,200 (2003 est)	http://encyclopedia.farlex.com/Corfu
around 110,000	http://www.sunshinetravel.gr/CORFUGUIDE/CORFU_TRAVEL_GUIDE 0-1.htm
110.000	http://www.dialashop.com/travel/corfu.html
about 110,00	http://www.argobenitses.gr/greece.php
97,102 in 1981	http://geography.howstuffworks.com/europe/corfu.htm
107,880	http://catalogue.horse21.net/greece+hotels/corfu+hotels/hotels5/luxury
109,512	http://www.corfu-property.gr/content/view/14/38/lang,en/
about 100,000	http://members.virtualtourist.com/m/6ce90/67541/
110,000 approximately	http://www.corfu-island.org/features.htm
107,000	http://www.nytimes.com/2009/09/11/greathomesanddestinations/11iht-recorfu.html

*** The only site to give attribution/citation/reference!!!

Two schools of thought

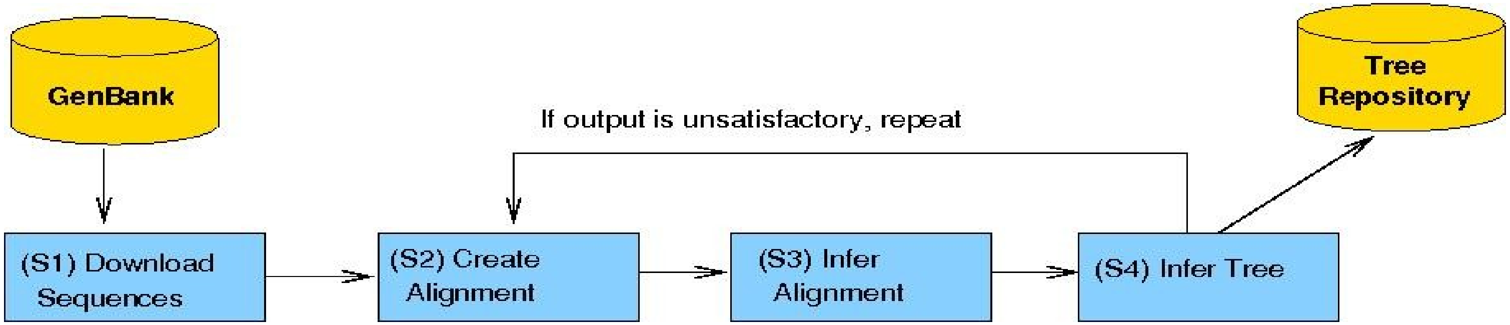
Data provenance – an explanation of

- *where* a piece of data came from,
- *how* it was created, or
- *why* it is there

Workflow provenance – record the execution/enactment of a workflow

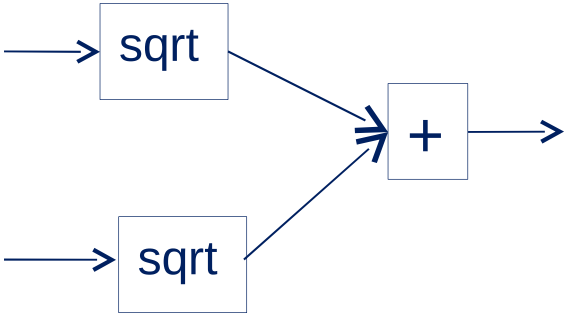
- In scientific workflows for repeatability, and
- sometimes for efficiency

Workflow provenance

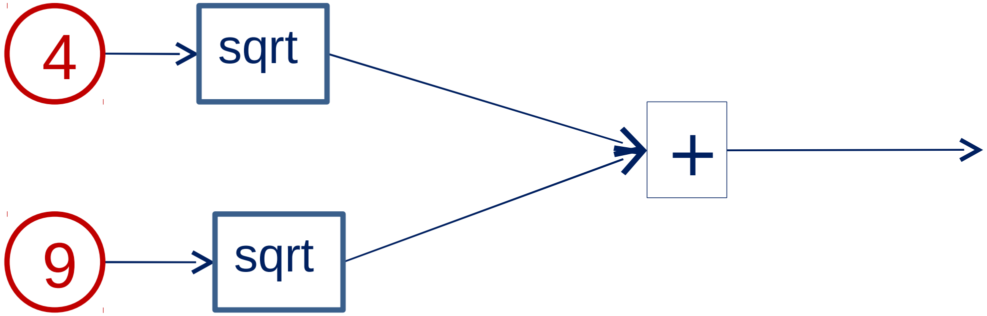


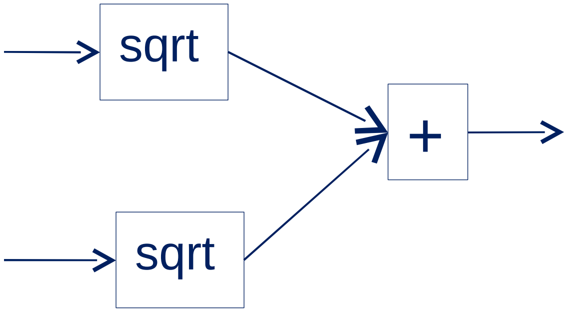
Taken from [Cohen, *et al* DILS 2006]

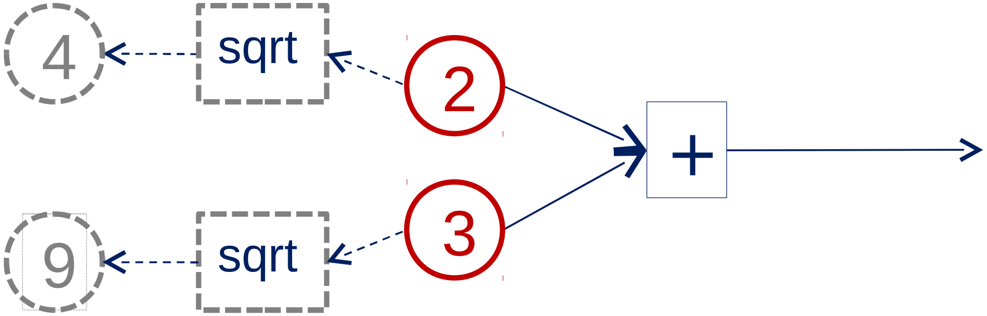
- Each step S1. . . S4 is itself a workflow.
- How does one record an “enactment” of the workflow?
- How much “context” does one record?
  - from people
  - from databases that change
- Recent attempts to produce a general model
  - Open Provenance Model [Moreau *et al.* IPAW 2007]
  - Petri Net + Complex Object [Hidders *et al.*Inf Syst 2008]

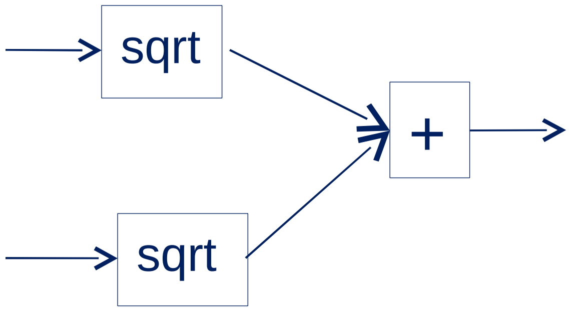


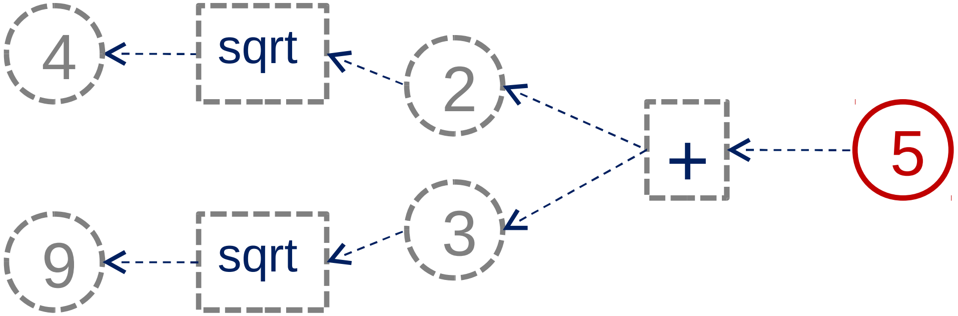
My imperfect understanding of OPM is that it is some kind of “unfolding” of the enactment of a workflow




sqrt
sqrt
+


4
sqrt
2
9
sqrt
3
+

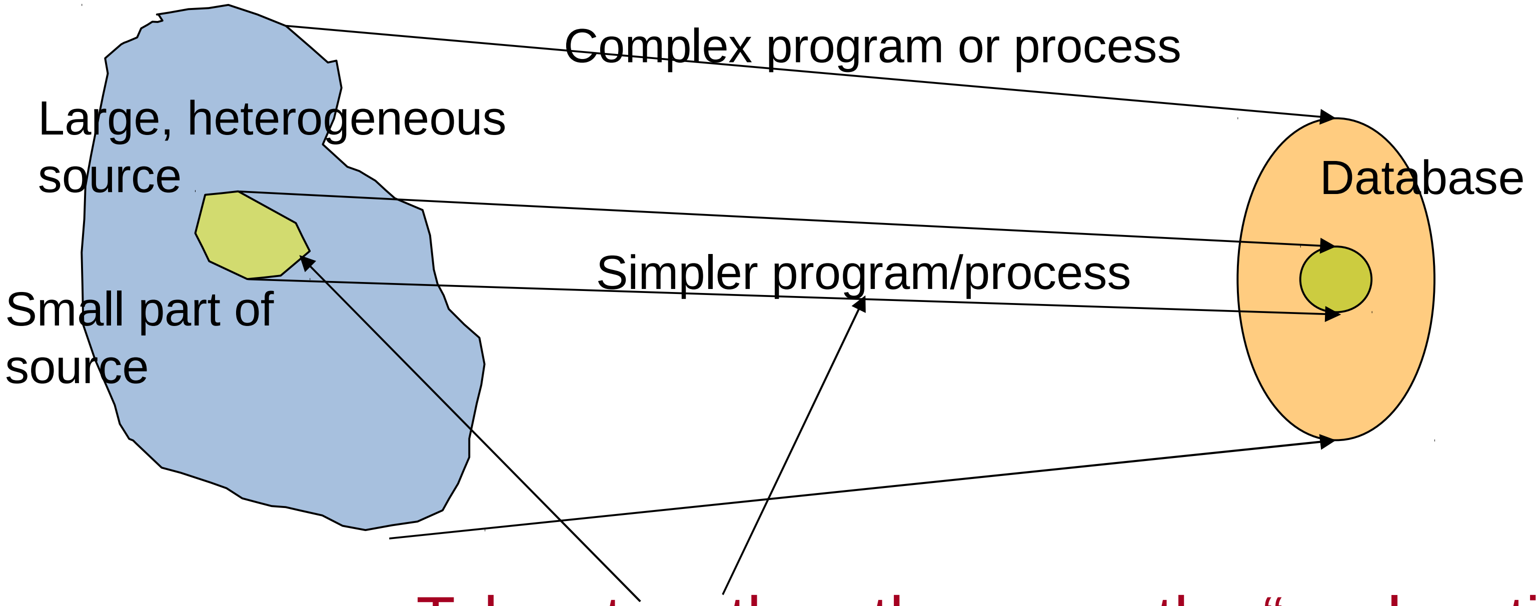

sqrt
sqrt
+


4
sqrt
2
9
sqrt
3
+
5

Data provenance

- Looks at elements of a (large) database
  - Where were they copied from
  - How were they constructed
  - Why are they in the output
  - What parts (tuples) of the input data “influenced” the output
- Relevant to:
  - Database archiving
  - Data annotation
  - Data citation

Workflow vs Data Provenance



Complex program or process

Large, heterogeneous source

Database

Simpler program/process

Small part of source

Taken together, these are the “explanation”.

Data provenance seeks to provide a simple explanation of *part* of the output, e.g.

- “This value was copied from that location”
- “This tuple was formed by combining those tuples”

Provenance Semirings or *How* provenance
(Tannen school: PODS ‘07 & ‘08 & '11)

R:

a	*b*	*c*	*p*
d	*b*	*e*	*r*
f	*b*	*e*	*s*

V:

a	*c*	$p + (p \cdot p)$
a	*e*	$p \cdot r$
d	*c*	$r \cdot p$
d	*e*	$r + (r \cdot r) + (r \cdot s)$
f	*e*	$s + (s \cdot s) + (s \cdot r)$

V(*X* ,*Z*) :– *R*(*X*, _, *Z*)
V(*X* ,*Z*) :– *R*(*X*, *Y*, _), *R*(_, *Y*, *Z*)

Tuples are created by :

- “joining” other tuples (join): $p \cdot r$
- “merging” other tuples (project and union): $p + r$

Both the “· ” and “+” are commutative and associative,
“· ” distributes over “+”: $p \cdot (r + s) = (p \cdot r) + (p \cdot s)$

Semirings

This structure $(K, +, \cdot, 0, 1)$ is a commutative semiring.

Provenance is a polynomial over the abstract quantities *p*,*q*,*r*, …

Various instantiations of this provide extensions to relational algebra:

- bag semantics
- C-tables
- probabilistic event tables
- various forms of why-provenance

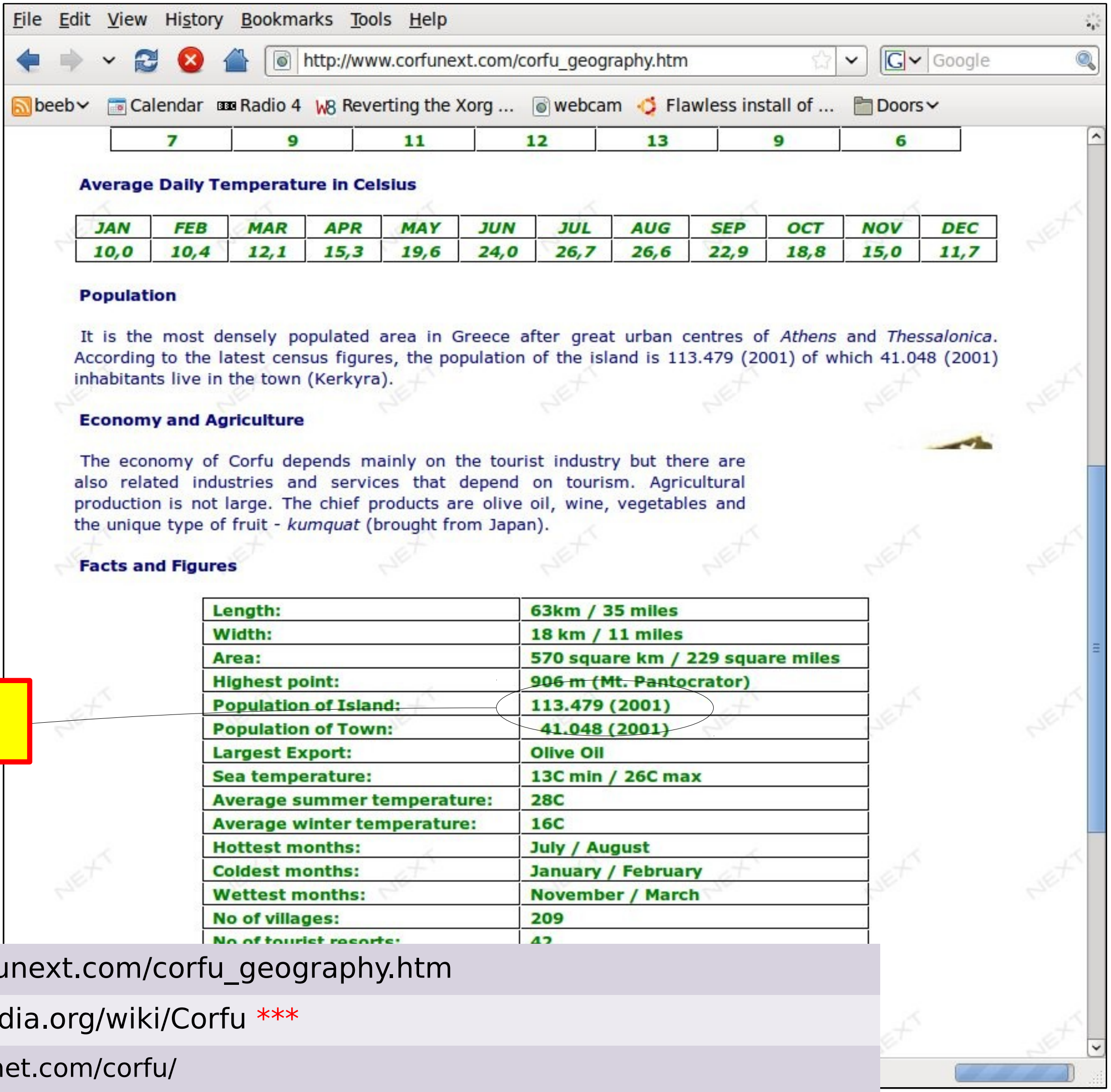
Note: the provenance polynomial is still only a partial description of the evaluation.

Example (bag semantics).: Abstract quantities are multiplicities. Semiring is (**Z**,+,x,0,1)

Multiplicity of (d, e) in V is $r + (r \times r) + (r \times s)$

Where-provenance: how stuff gets copied

The evils of copy-paste!


File Edit View History Bookmarks Tools Help
http://www.corfunext.com/corfu_geography.htm
Google
beeb
Calendar
Radio 4
Reverting the Xorg ...
webcam
Flawless install of ...
Doors

7	9	11	12	13	9	6

Average Daily Temperature in Celsius

JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
10,0	10,4	12,1	15,3	19,6	24,0	26,7	26,6	22,9	18,8	15,0	11,7

Population

It is the most densely populated area in Greece after great urban centres of *Athens* and *Thessalonica*. According to the latest census figures, the population of the island is 113.479 (2001) of which 41.048 (2001) inhabitants live in the town (Kerkyra).

Economy and Agriculture

The economy of Corfu depends mainly on the tourist industry but there are also related industries and services that depend on tourism. Agricultural production is not large. The chief products are olive oil, wine, vegetables and the unique type of fruit - *kumquat* (brought from Japan).

Facts and Figures

Length:	63km / 35 miles
Width:	18 km / 11 miles
Area:	570 square km / 229 square miles
Highest point:	906 m (Mt. Pantocrator)
Population of Island:	113.479 (2001)
Population of Town:	41.048 (2001)
Largest Export:	Olive Oil
Sea temperature:	13C min / 26C max
Average summer temperature:	28C
Average winter temperature:	16C
Hottest months:	July / August
Coldest months:	January / February
Wettest months:	November / March
No of villages:	209
No of tourist resorts:	42

<cntrl>C <cntrl>V

113.479 (2001)	http://www.corfunext.com/corfu_geography.htm
107,879 (as of 2001)	http://en.wikipedia.org/wiki/Corfu ***
93,000	http://www.corfunet.com/corfu/
109,512	www.agni.gr/

Possible explanations of how something was copied

This data item was extracted from location L1 in document D1 and placed in location L2 in document D2

or

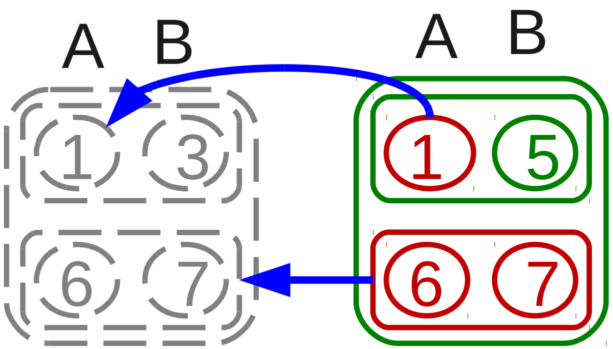
This data item was extracted from database D1 by query Q1 and placed in database D2 by update U2

(or some combination of the two)

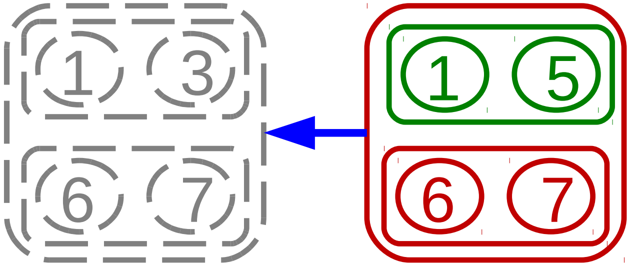
So we need to understand where-provenance for query and update languages

Query languages, update languages and where-provenance

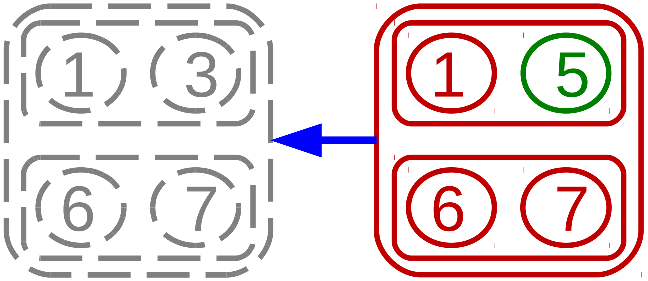
```
(select A, 5 as B from R where A = 1)
union
(select * from R where A <> 1)
```



```
delete from R where A = 1;
insert into R values (1,5)
```



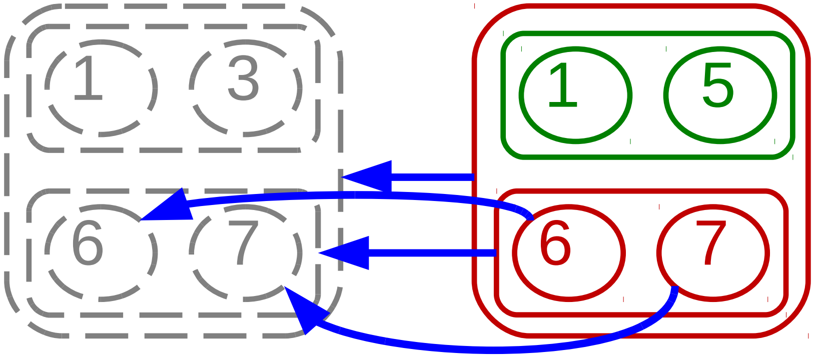
```
update R set B = 5 where A = 1
```



Update and Query Languages

- Database queries “rearrange” their input.
- The provenance graph shows where pieces of the output come from.
- Update languages are usually considered uninteresting because they are less expressive than query languages
- If the provenance graph is taken into account they become *more* expressive
- SQL (query fragment) is complete for “natural” value-preserving provenance
- SQL (update+query) is complete for “natural” kind-preserving provenance

Where-provenance graphs and OPM



- These graphs indicate that OPM (a very simple model) needs to be enriched with some form of nesting to deal with the structure of data
- The graphs for semi-ring provenance are even more complicated

Some practical outcomes

- Devised some techniques for recording and compressing where provenance in curated databases.
- Developed archiving techniques for databases – essential for provenance
  - Efficiently records all past states of a database
- Developed data citation techniques for curated databases
- A *Database* Wiki that – among other things – records and exports provenance

A working system for archiving evolving data sets

- Implemented by Heiko Müller
- For scale, we require external sorting of large XML files
    - Designed and implemented by Ioannis Koltsidas Heiko Müller and Stratis Viglas
- Has a simple temporal query language
- Experimented with recent (HTML) versions of CIA world factbook.
- Supports archiving of the IUPHAR database

The evolution of the World Factbook

Printed versions started around 1960

Released annually until 2008

On-line versions started around 1990

> Transcribed in a variety of data formats between 1990 and 1997. All simple “mark-up” of plain text – perhaps to enable machine readability

HTML (apparently hand generated) until 2007-2008

Some kind of Schema from 2008

Now updated weekly or more often

No publication (by CIA) of an archive!

What the archive looks like

```
<T t="2002-2007">
    <FACTBOOK>
        <COUNTRY>
            <NAME>Afghanistan</NAME>
            <CATEGORY>
                <NAME>Communications</NAME>
                <PROPERTY>
                    <NAME>Internet users</NAME>
                    <TEXT>
                        <T t="2004-2005">1,000 (2002)</T>
                        <T t="2006-2007">30,000 (2005)</T>
                        <T t="2002-2003">NA</T>
                    </TEXT>
                </PROPERTY>
                <PROPERTY>
                    <NAME>Radios</NAME>
                    <TEXT>167,000 (1999)</TEXT>
                </PROPERTY>
                <PROPERTY>
                    <NAME>Telephones - main lines in use</NAME>
                    <TEXT>
                        <T t="2006">100,000 (2005)</T>
                        <T t="2007">280,000 (2005)</T>
                        <T t="2002-2003">29,000 (1998)</T>
                        <T t="2004-2005">33,100 (2002)</T>
…
```

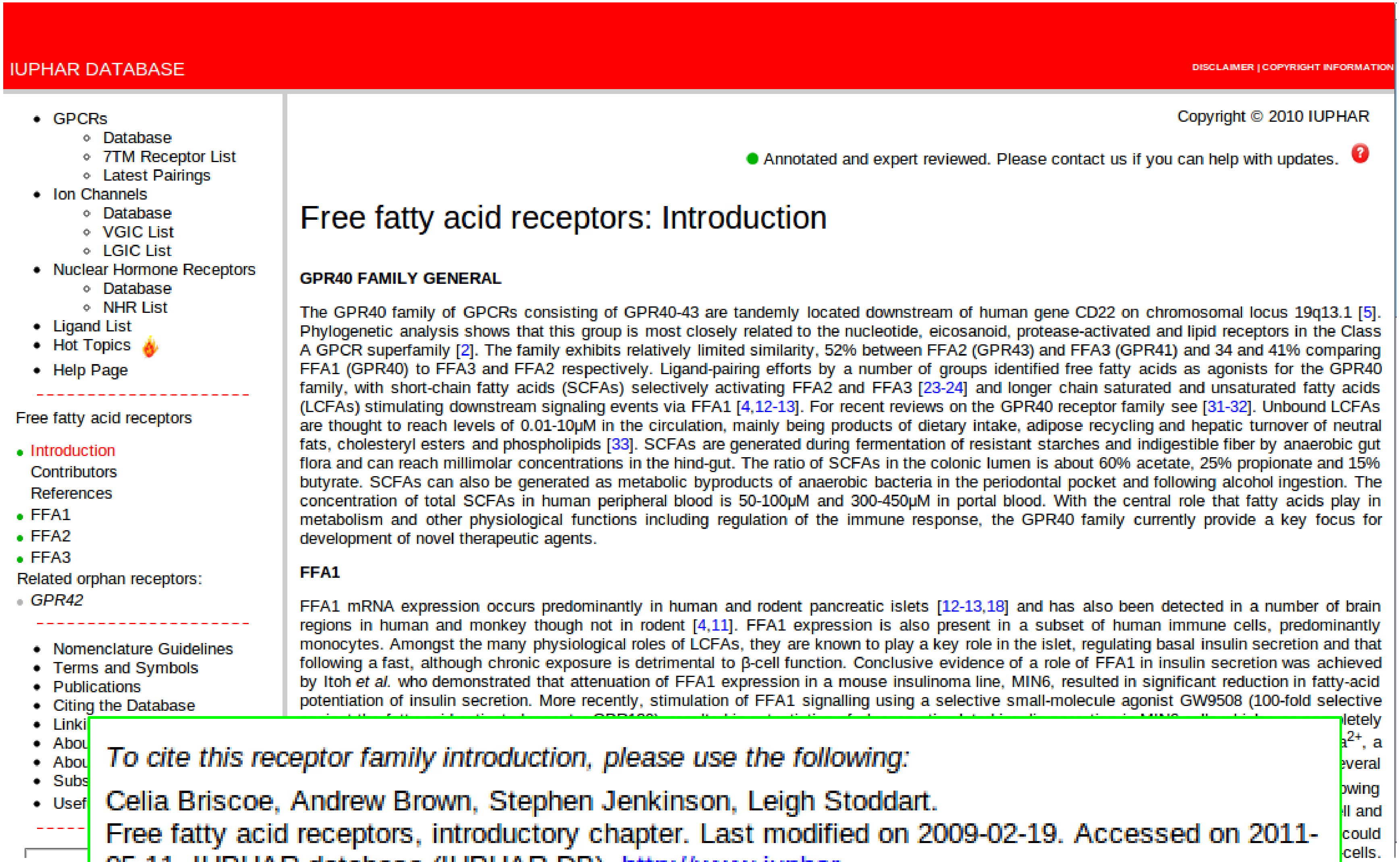
How did land area of countries change in 2002-2007?

```
<T t="2002-2007">
  <FACTBOOK KEY="">
    …
    <COUNTRY KEY="NAME Austria">
      <CATEGORY KEY="NAME Geography">
        <PROPERTY KEY="NAME Area">
          <SUBPROP>
            <NAME>land</NAME>
            <TEXT>
              <T t="2004-2007">82,444 sq km</T>
              <T t="2002-2003">82,738 sq km</T>
             </TEXT>
           </SUBPROP>
         </PROPERTY>
       </CATEGORY>
    </COUNTRY>
    …
    <COUNTRY KEY="NAME France">
      <CATEGORY KEY="NAME Geography">
        <PROPERTY KEY="NAME Area">
          <SUBPROP>
            <NAME>land</NAME>
            <TEXT>
              <T t="2002-2006">545,630 sq km</T>
              <T t="2007">640,053 sq km; 545,630 sq km (metropolitan France)</T>
             </TEXT>
…
```

What are the differences between the factbooks on 21/08/2007 and 10/09/2007?

```
<T t="21/08/2007-10/09/2007">
    <CIAWFB KEY="">
        <COUNTRY KEY="NAME Afghanistan">
            <CATEGORY KEY="NAME Communications">
                <PROPERTY KEY="NAME Internet users">
                    <T t="21/08/2007">
                        <TEXT>30,000 (2005)</TEXT>
                    </T>
                    <T t="10/09/2007">
                        <TEXT>535,000 (2006)</TEXT>
                    </T>
                </PROPERTY>
                <PROPERTY KEY="NAME Telephones - mobile cellular">
                    <T t="21/08/2007">
                        <TEXT>1.4 million (2005)</TEXT>
                    </T>
                    <T t="10/09/2007">
                        <TEXT>2.52 million (2006)</TEXT>
                    </T>
…
```

Citations in curated databases



IUPHAR DATABASE

DISCLAIMER | COPYRIGHT INFORMATION

- GPCRs
  - Database
  - 7TM Receptor List
  - Latest Pairings
- Ion Channels
  - Database
  - VGIC List
  - LGIC List
- Nuclear Hormone Receptors
  - Database
  - NHR List
- Ligand List
- Hot Topics
- Help Page

Free fatty acid receptors

- Introduction
- Contributors
- References
- FFA1
- FFA2
- FFA3

Related orphan receptors:

- *GPR42*

- Nomenclature Guidelines
- Terms and Symbols
- Publications
- Citing the Database
- Linki
- Abou
- Abou
- Subs
- Usef

Copyright © 2010 IUPHAR

● Annotated and expert reviewed. Please contact us if you can help with updates.

Free fatty acid receptors: Introduction

GPR40 FAMILY GENERAL

The GPR40 family of GPCRs consisting of GPR40-43 are tandemly located downstream of human gene CD22 on chromosomal locus 19q13.1 [5]. Phylogenetic analysis shows that this group is most closely related to the nucleotide, eicosanoid, protease-activated and lipid receptors in the Class A GPCR superfamily [2]. The family exhibits relatively limited similarity, 52% between FFA2 (GPR43) and FFA3 (GPR41) and 34 and 41% comparing FFA1 (GPR40) to FFA3 and FFA2 respectively. Ligand-pairing efforts by a number of groups identified free fatty acids as agonists for the GPR40 family, with short-chain fatty acids (SCFAs) selectively activating FFA2 and FFA3 [23-24] and longer chain saturated and unsaturated fatty acids (LCFAs) stimulating downstream signaling events via FFA1 [4,12-13]. For recent reviews on the GPR40 receptor family see [31-32]. Unbound LCFAs are thought to reach levels of 0.01-10μM in the circulation, mainly being products of dietary intake, adipose recycling and hepatic turnover of neutral fats, cholesteryl esters and phospholipids [33]. SCFAs are generated during fermentation of resistant starches and indigestible fiber by anaerobic gut flora and can reach millimolar concentrations in the hind-gut. The ratio of SCFAs in the colonic lumen is about 60% acetate, 25% propionate and 15% butyrate. SCFAs can also be generated as metabolic byproducts of anaerobic bacteria in the periodontal pocket and following alcohol ingestion. The concentration of total SCFAs in human peripheral blood is 50-100μM and 300-450μM in portal blood. With the central role that fatty acids play in metabolism and other physiological functions including regulation of the immune response, the GPR40 family currently provide a key focus for development of novel therapeutic agents.

FFA1

FFA1 mRNA expression occurs predominantly in human and rodent pancreatic islets [12-13,18] and has also been detected in a number of brain regions in human and monkey though not in rodent [4,11]. FFA1 expression is also present in a subset of human immune cells, predominantly monocytes. Amongst the many physiological roles of LCFAs, they are known to play a key role in the islet, regulating basal insulin secretion and that following a fast, although chronic exposure is detrimental to β-cell function. Conclusive evidence of a role of FFA1 in insulin secretion was achieved by Itoh *et al.* who demonstrated that attenuation of FFA1 expression in a mouse insulinoma line, MIN6, resulted in significant reduction in fatty-acid potentiation of insulin secretion. More recently, stimulation of FFA1 signalling using a selective small-molecule agonist GW9508 (100-fold selective

letely

a^{2+}, a

everal

owing

ll and

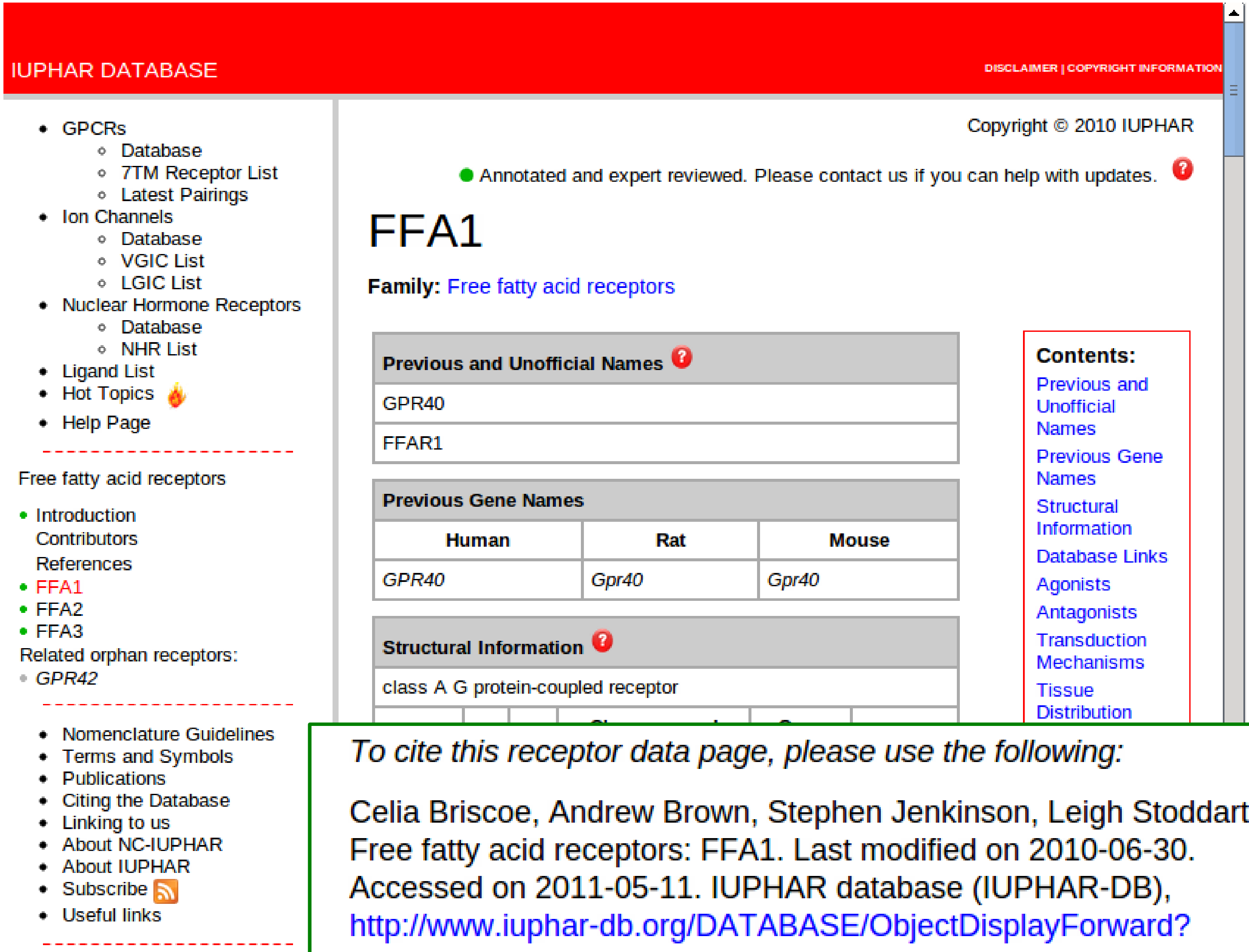
could

cells.

To cite this receptor family introduction, please use the following:

Celia Briscoe, Andrew Brown, Stephen Jenkinson, Leigh Stoddart. Free fatty acid receptors, introductory chapter. Last modified on 2009-02-19. Accessed on 2011-05-11. IUPHAR database (IUPHAR-DB), http://www.iuphar-db.org/DATABASE/FamilyIntroductionForward?familyId=24.

In another part of the same section...



IUPHAR DATABASE

- GPCRs
  - Database
  - 7TM Receptor List
  - Latest Pairings
- Ion Channels
  - Database
  - VGIC List
  - LGIC List
- Nuclear Hormone Receptors
  - Database
  - NHR List
- Ligand List
- Hot Topics
- Help Page

Free fatty acid receptors

- Introduction
  Contributors
  References
- FFA1
- FFA2
- FFA3

Related orphan receptors:

- *GPR42*

- Nomenclature Guidelines
- Terms and Symbols
- Publications
- Citing the Database
- Linking to us
- About NC-IUPHAR
- About IUPHAR
- Subscribe
- Useful links

● Annotated and expert reviewed. Please contact us if you can help with updates.

FFA1

Family: Free fatty acid receptors

Previous and Unofficial Names
GPR40
FFAR1

Previous Gene Names		
Human	**Rat**	**Mouse**
GPR40	*Gpr40*	*Gpr40*

Structural Information
class A G protein-coupled receptor

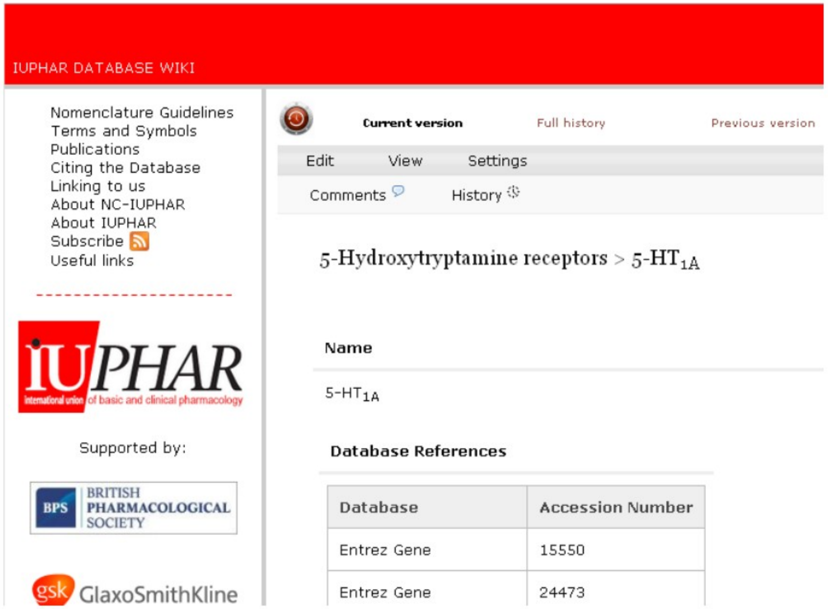
To cite this receptor data page, please use the following:

Celia Briscoe, Andrew Brown, Stephen Jenkinson, Leigh Stoddart. Free fatty acid receptors: FFA1. Last modified on 2010-06-30. Accessed on 2011-05-11. IUPHAR database (IUPHAR-DB), http://www.iuphar-db.org/DATABASE/ObjectDisplayForward?objectId=225.

A Database Wiki

A Wiki for developing structured data

- Free annotation of all data elements
- Automatic exporting and importing of provenance (currently only useful if moving data between other database wikis)



IUPHAR DATABASE WIKI

Nomenclature Guidelines
Terms and Symbols
Publications
Citing the Database
Linking to us
About NC-IUPHAR
About IUPHAR
Subscribe
Useful links

IUPHAR International union of basic and clinical pharmacology

Supported by:

BPS BRITISH PHARMACOLOGICAL SOCIETY

gsk GlaxoSmithKline

Current version Full history Previous version

Edit View Settings

Comments History

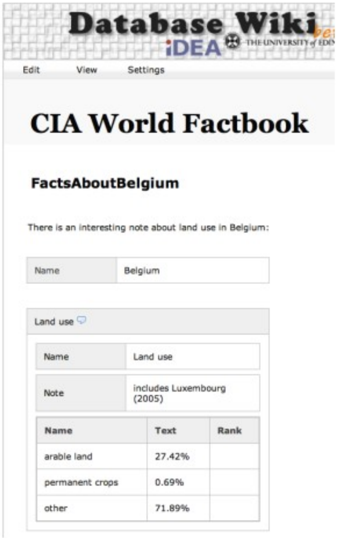
5-Hydroxytryptamine receptors > 5-HT_{1A}

Name

5-HT_{1A}

Database References

Database	Accession Number
Entrez Gene	15550
Entrez Gene	24473



Database Wiki beta iDEA THE UNIVERSITY of EDIN

Edit View Settings

CIA World Factbook

FactsAboutBelgium

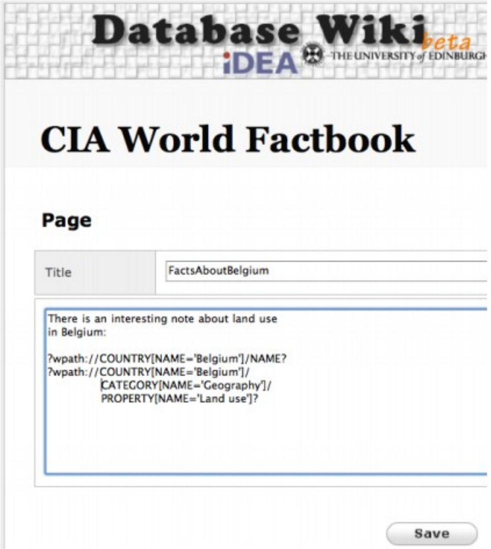
There is an interesting note about land use in Belgium:

Name	Belgium

Land use

Name	Land use
Note	includes Luxembourg (2005)

Name	Text	Rank
arable land	27.42%	
permanent crops	0.69%	
other	71.89%	



Database Wiki beta iDEA THE UNIVERSITY of EDINBURGH

CIA World Factbook

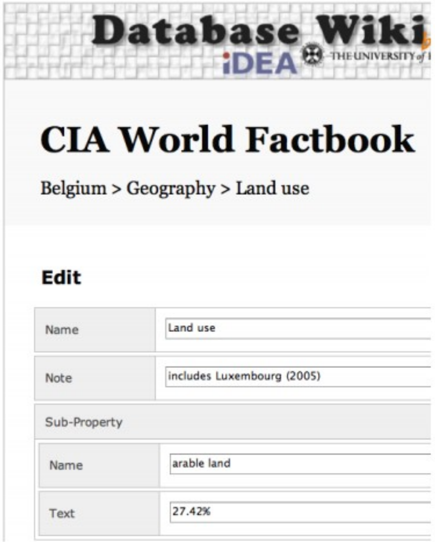
Page

Title: FactsAboutBelgium

```
There is an interesting note about land use
in Belgium:

?wpath://COUNTRY[NAME='Belgium']/NAME?
?wpath://COUNTRY[NAME='Belgium']/
         CATEGORY[NAME='Geography']/
         PROPERTY[NAME='Land use']?
```

Save



Database Wiki iDEA THE UNIVERSITY of E

CIA World Factbook

Belgium > Geography > Land use

Edit

Name	Land use
Note	includes Luxembourg (2005)

Sub-Property

Name	arable land
Text	27.42%

Provenance has lots of connections

- Intrinsic part of data quality. Fundamental to almost everything we do with data: quality, IP, copyright, etc.
- It is bringing together several areas of CS:
  - Semantics of update languages.
  - Probabilistic databases
  - Data integration
  - Debugging schema transformations
  - File/data synchronization
  - Program debugging (program slicing)
  - Program instrumentation
  - Security
  - Linked data and ontologies
- **The fundamental problem is finding the right model/models**

Questions?

Pedigree

Lineage

Provenance