

Curated Databases

Peter Buneman

University of Edinburgh

Susan Davidson	University of Pennsylvania
Alin Deutsch	UC San Diego
Wenfei Fan	Bell Labs -> University of Edinburgh
Carmem Hara	Universidade Federal do Parana
Sanjeev Khanna	University of Pennsylvania
Christoph Koch	University of Edinburgh/TU Wien
Keishi Tajima	University of Tokyo
Wang-Chiew Tan	UC Santa Cruz

<http://www.lfcs.inf.ed.ac.uk/research/database/dbs.html>

<http://db.cis.upenn.edu/Research/provenance.html>



Edinburgh has numerous research positions in databases, XML, web technology, fundamentals

Contact Peter Buneman, opb@inf.ed.ac.uk

Edinburgh is a great place to live!!!

Top-rated department. Excellent database group. Good connections with logical foundations, scientific DBs, distributed computation (Grid)

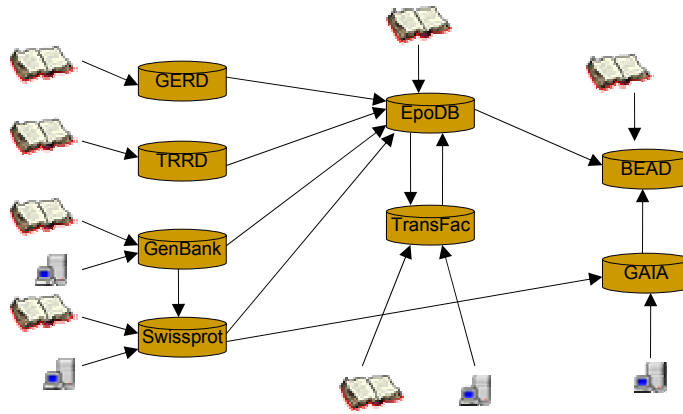
Scientific Databases

- Much modern science is dependent on databases
 - Prime example: molecular biology
- These databases are built and maintained with a great deal of human effort
- They often do not contain source experimental data.
- They borrow extensively from, and refer to, other databases
- They act as "dynamic" publications.
- They are **curated**

Much of the content of curated databases is **annotation**

Geography	France	Swissprot
<p>Location: <input type="checkbox"/> <input type="checkbox"/></p> <p>Western Europe, bordering the Bay of Biscay and English Channel, between Belgium and Spain, southeast of the UK, bordering the Mediterranean Sea, between Italy and Spain</p> <p>Geographic coordinates: <input type="checkbox"/> <input type="checkbox"/></p> <p>46 00 N, 2 00 E</p> <p>Map references: <input type="checkbox"/> <input type="checkbox"/></p> <p>EUROPE</p> <p>Area: <input type="checkbox"/> <input type="checkbox"/></p> <p>total: 547,030 sq km land: 545,620 sq km note: includes only metropolitan France; excludes the overseas administrative divisions water: 1,400 sq km</p> <p>Area - comparative: <input type="checkbox"/> <input type="checkbox"/></p> <p>slightly less than twice the size of Colorado</p> <p>Land boundaries: <input type="checkbox"/> <input type="checkbox"/></p> <p>total: 2,889 km border countries: Andorra 56.6 km, Belgium 620 km, Germany 451 km, Italy 488 km, Luxembourg 73 km, Monaco 4.4 km, Spain 623 km, Switzerland 573 km</p> <p>Coastline: <input type="checkbox"/> <input type="checkbox"/></p> <p>3,427 km</p> <p>Maritime claims: <input type="checkbox"/> <input type="checkbox"/></p> <p>contiguous zone: 24 NM territorial sea: 12 NM continental shelf: 200-m depth or to the depth of exploitation exclusive economic zone: 200 NM (does not apply to the Mediterranean)</p> <p>Climate: <input type="checkbox"/> <input type="checkbox"/></p> <p>generally cool winters and mild summers, but mild winters and hot summers along the Mediterranean; occasional strong, cold, dry, north-to-northwesterly wind known as mistral</p> <p>Terrain: <input type="checkbox"/> <input type="checkbox"/></p> <p>mostly flat plains or gently rolling hills in north and west; remainder is mountainous, especially Pyrenees in south, Alps in east</p> <p>Elevation extremes: <input type="checkbox"/> <input type="checkbox"/></p> <p>lowest point: Rhone River delta -2 m highest point: Mont Blanc 4,807 m</p>	<pre> ID 1158 CUCMA STANDARD; PRT; 480 AA. AC P13744; DT 01-JAN-1990 (REL. 13, CREATED) DT 01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE) DT 01-NOV-1990 (REL. 16, LAST ANNOTATION UPDATE) DE 115 GLOBULIN BETA SUBUNIT PRECURSOR. OS CUCURBITA MAXIMA (PUMPKIN) (WINTER SQUASH). OC EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE; OC VIOLALES; CUCURBITACEAE. RN 11; RP SEQUENCE FROM N.A. RC STRAIN=CV. KUBOKAWA AMARURI NANKIN; RK MEDLINE; 88166744. RA HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.; RL EUR. J. BIOCHEM. 172:627-632(1988). RN 12; RP SEQUENCE OF 22-30 ND 297-302. RA OHMIYA M., HARA I., MASTUHARA H.; RL PLANT CELL PHYSIOL. 21:157-167(1980). CC -1- FUNCTION: THIS IS A SEED STORAGE PROTEIN. CC -1- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A CC BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A CC DISULFIDE BOND. CC -1- SIMILARITY: TO OTHER 115 SEED STORAGE PROTEINS (GLOBULINS). DR EMBL; H3407; G167492; -. DR PIR; S00366; PNF018. DR PROSITE; PS00305; 115_SEED_STORAGE; 1. RW SEED STORAGE PROTEIN; SIGNAL. FT SIGNAL 1 21 FT CHAIN 22 480 115 GLOBULIN BETA SUBUNIT. FT CHAIN 22 296 GAMMA CHAIN (ACIDIC). FT CHAIN 297 80 DELTA CHAIN (BASIC). FT MOD_RES 22 22 PYRROLIDONE CARBOXYLIC ACID. FT DISULFID 124 303 INTERCHAIN (GAMMA-BETA) (POTENTIAL). FT CONFLICT 27 27 S -> E (IN REF. 2). FT CONFLICT 30 30 E -> S (IN REF. 2). SQ SEQUENCE 480 AA 54625 MW 55150506 CXC32; MARSELPFL CLAVFINCHQ SQIEQSPWE FQGEVYPOK RYQSPFRACL ENLRAQDPVR RAEAEALFTE VMDQNDPEQ CAGVNMIRH TRFGKLLPG FSNAPKLIFV AQGPGIGIA IFQCAFTYT ELKRSQSAQ AFKQQRIR PFRGDLIVV PAVSHMMH RQSDLVIV PDRYVANG IFTIIFPIL AGRFQVZG VEMERSRK ESKRSGKSH FSGFAEPL EAPQIDGLV RLKKGEDDR DRIVQVDEP EVLLEKDEE ERSRGRYIES ESEBNGLEE TICTRLKQK IGRSVKADVF NFRQGRVSTA NYHFLPLQK VRLSARQKV YSNRVAFPY YNSHSHWIA TRGNARQVQV ENFQGVFEG EYRQGVPLK PNFYVYRA SGRFEPIL KTDNATHL LAGRVSQMM LPLDVLNVY RISRKAQRL KYQGVKRVL SGRSQQRRE // </pre>	

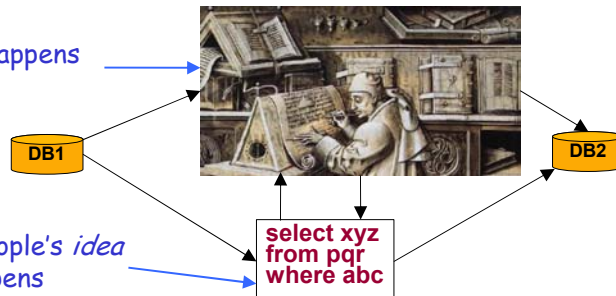
Curated databases “borrow”



Current database technology is mostly concerned with efficiency and robustness (high transaction rates, efficient queries)

It does not address many of the issues in curation

What *really* happens



On the menu for today

- Discussion of provenance
- Annotation - a related problem
- Models for semi-structured data
- Keys for XML
- Efficient Archiving
- A copy-and-paste model of curation

Provenance -- an old problem

W.V. Quine. "Words
enough..." *New York Review
of Books* XIII(10):3-4, 1969



A well-known encyclopedia, following tradition, incorrectly describes Monaco as having an area of "8 square miles."

A new edition using adds "...the length being 2 1/4 miles and the width varying from 165 to 1100 yards."

An editor, spotting the inconsistency, removes the *correct* information from the subsequent edition.

The area of Monaco today

Most sources	1.95 sq km
www.atlapedia.com	1.94 sq km
military.countrywatch.com	2 sq km

Most sources	1.95 sq km	0.75 sq m
www.state.gov	1.95 sq km	0.8 sq m
www.atlapedia.com	1.94 sq km	1 sq m

(1.95 sq km = 0.753 sq m)

The population of Monaco? (from a 2002 Web search)

"31,842 (July 2001 est.)"	CIA world factbook and 3 other sites
"31,693 (July 2000 est.)"	8 sites
"ESTIMATED 2000...32,500"	www.atlapedia.com
"32,000 (2000)"	biz.yahoo.com
"32,035(July 1998 est.)"	4 sites. <u>One of these gives attribution</u> -- to CIA world factbook (above)
"31,515 (July 1995 est.)"	www.immigration-usa.com
"(1995): 30,744"	2 sites
"29,972"	"according to the last official census in 1990" www.monaco.mc (the official web site of Monaco?)

Swissprot: a curated database

```
ID 11S_CUCMA STANDARD; PRT; 480 AA.
AC P13744;
DT 01-JAN-1990 (REL. 13, CREATED)
DT 01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE)
DT 01-NOV-1990 (REL. 16, LAST ANNOTATION UPDATE)
DE 11S GLOBULIN BETA SUBUNIT PRECURSOR.
OS CUCURBITA MAXIMA (PUMPKIN) (WINTER SQUASH).
OC EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE;
OC VIOLALES; CUCURBITACEAE.
RN [1]
RP SEQUENCE FROM N.A.
RC STRAIN=CV. KUROKAWA AMAKURI NANKIN;
RX MEDLINE; 88166744.
RA HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
RL EUR. J. BIOCHEM. 172:627-632(1988).
RN [2]
RP SEQUENCE OF 22-30 AND 297-302.
RA OHMIYA M., HARA I., MASTUBARA H.;
RL PLANT CELL PHYSIOL. 21:157-167(1980).
CC -1- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC -1- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC DISULFIDE BOND.
CC -1- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
DR EMBL; M36407; G167492; -.
DR PIR; S00366; FWPUB.
DR PROSITE; PS00305; 11S_SEED_STORAGE; 1.
KW SEED STORAGE PROTEIN; SIGNAL.
FT SIGNAL 1 21
FT CHAIN 22 480 11S GLOBULIN BETA SUBUNIT.
FT CHAIN 22 296 GAMMA CHAIN (ACIDIC).
FT CHAIN 297 480 DELTA CHAIN (BASIC).
FT MOD_RES 22 22 PYRROLIDONE CARBOXYLIC ACID.
FT DISULFID 124 303 INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT CONFLICT 27 27 S -> E (IN REF. 2).
FT CONFLICT 30 30 E -> S (IN REF. 2).
SQ SEQUENCE 480 AA; 54625 MW; D515DD6E CRC32;
MARSSLFTFL CLAVFINGCL SQIEQQSPWE FGQSEVWQOH RYQSPRACRL ENLRAQDPVR
RAEAEAFTE VWDQNDDEFQ CAGVNMIRHT IRPKGLLLFG FSNAPKLI FV A QGGFIRGIA
IPGCAEYQT DLRRSQSAGS AFKQHQKIR FFRGGDLLV FAGVSHWMYN RGQSDLVLV
FADTRVNAQ IDPYLRKFLV AGRFEQVERG VEVEPSSSK GSGSEKSGNI FSGFADEFLE
EAFQIDGGLV RKLKGEDDER DRIVQVDEF EVLPEKDEE ERSRGRVIES ESESENGLEE
TICTLRKQN IGRSVRADVF NFRGGRISTA NYHTLPILRQ VRLSAERGVL YSNAMVAPHY
TVNSHVMYA TRGNARQVV DNFQSVFDG EVREGQVLMV PQNFVVIKRA SDRGFEWIAP
KTNDNAINTL LAGRVSQMRM LPLGVLSNMY RISREEAQL KYGQEQMRVL SFGRSQGRRE
//
```

WISE, Dec '03

11

```
ID 11S_CUCMA STANDARD; PRT; 480 AA.
AC P13744;
DT 01-JAN-1990 (REL. 13, CREATED)
DT 01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE)
DT 01-NOV-1990 (REL. 16, LAST ANNOTATION UPDATE)
DE 11S GLOBULIN BETA SUBUNIT PRECURSOR.
OS CUCURBITA MAXIMA (PUMPKIN) (WINTER SQUASH).
OC EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE;
OC VIOLALES; CUCURBITACEAE.
RN [1]
RP SEQUENCE FROM N.A.
RC STRAIN=CV. KUROKAWA AMAKURI NANKIN;
RX MEDLINE; 88166744.
RA HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
RL EUR. J. BIOCHEM. 172:627-632(1988).
RN [2]
RP SEQUENCE OF 22-30 AND 297-302.
RA OHMIYA M., HARA I., MASTUBARA H.;
RL EUR. J. BIOCHEM. 172:627-632(1988).
RN [2]
RP SEQUENCE OF 22-30 AND 297-302.
RA OHMIYA M., HARA I., MASTUBARA H.;
RL PLANT CELL PHYSIOL. 21:157-167(1980).
CC -1- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC -1- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC DISULFIDE BOND.
CC -1- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
DR EMBL; M36407; G167492; -.
DR PIR; S00366; FWPUB.
DR PROSITE; PS00305; 11S_SEED_STORAGE; 1.
KW SEED STORAGE PROTEIN; SIGNAL.
FT SIGNAL 1 21
FT CHAIN 22 480 11S GLOBULIN BETA SUBUNIT.
FT CHAIN 22 296 GAMMA CHAIN (ACIDIC).
FT CHAIN 297 480 DELTA CHAIN (BASIC).
FT MOD_RES 22 22 PYRROLIDONE CARBOXYLIC ACID.
FT DISULFID 124 303 INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT CONFLICT 27 27 S -> E (IN REF. 2).
FT CONFLICT 30 30 E -> S (IN REF. 2).
SQ SEQUENCE 480 AA; 54625 MW; D515DD6E CRC32;
MARSSLFTFL CLAVFINGCL SQIEQQSPWE FGQSEVWQOH RYQSPRACRL ENLRAQDPVR
RAEAEAFTE VWDQNDDEFQ CAGVNMIRHT IRPKGLLLFG FSNAPKLI FV A QGGFIRGIA
IPGCAEYQT DLRRSQSAGS AFKQHQKIR FFRGGDLLV FAGVSHWMYN RGQSDLVLV
FADTRVNAQ IDPYLRKFLV AGRFEQVERG VEVEPSSSK GSGSEKSGNI FSGFADEFLE
EAFQIDGGLV RKLKGEDDER DRIVQVDEF EVLPEKDEE ERSRGRVIES ESESENGLEE
TICTLRKQN IGRSVRADVF NFRGGRISTA NYHTLPILRQ VRLSAERGVL YSNAMVAPHY
TVNSHVMYA TRGNARQVV DNFQSVFDG EVREGQVLMV PQNFVVIKRA SDRGFEWIAP
KTNDNAINTL LAGRVSQMRM LPLGVLSNMY RISREEAQL KYGQEQMRVL SFGRSQGRRE
//
```

Citations for sequence
data

WISE, Dec '03

12

ID 11S_CUCMA STANDARD; PRT; 480 AA.
AC P13744;
DT 01-JAN-1990 (REL. 13, CREATED)
DT 01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE)

DE 11S GLOBULIN BETA SUBUNIT PRECURSOR.
OS CUCURBITA MAXIMA (PUMPKIN) (WINTER SQUASH).
OC EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE;
OC VIOLALES; CUCURBITACEAE.

RN [1]
RC STRAIN=CV. KUROKAWA AMAKURI NANKIN;
RX MEDLINE; 88166744.
RA HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
RL EUR. J. BIOCHEM. 172:627-632 (1988).

CC -!- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC -!- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC DISULFIDE BOND.
CC -!- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).

CC DISULFIDE BOND.
CC -!- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
DR EMBL; M36407; G167492; -.

FT CHAIN	22	480	11S GLOBULIN BETA SUBUNIT.
FT CHAIN	22	296	GAMMA CHAIN (ACIDIC).
FT CHAIN	297	480	DELTA CHAIN (BASIC).
FT MOD_RES	22	22	PYRROLIDONE CARBOXYLIC ACID.
FT DISULFID	124	303	INTERCHAIN (GAMMA-DELTA) (POTENTIAL).

FT MOD_RES 22 22 PYRROLIDONE CARBOXYLIC ACID.
FT DISULFID 124 303 INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT CONFLICT 27 27 S -> E (IN REF. 2).
FT CONFLICT 30 30 E -> S (IN REF. 2).
SQ SEQUENCE 480 AA; 54625 MW; D515DD6E CRC32;
MARSSLFTFL CLAVFINGCL SQIEQQSPWE FGSEVWQOH RYQSPRACL ENLRAQDPVR
RAEAEAFTE VWDQNDDEFQ CAGVNMRHT IRPKGLLLG FSNAPKLIV AQGFGIRGIA
IPGCAEYQT DLRRSQSAGS AFKQHQKIR PFRGGDLIV FVAGVSHWYN RGQSDLVIV
FADTRVNAQ IDPILRKFYL AGRFEQVERG VEWEERSKX GSGSEKSGNI FSGFADEFLE
EAFQIDGGLV RKLKGEDDER DRIVQVDEF EVLPEKDEE ERSRGRVIES ESENGLEE
TICTLRLKQN IGRSVRADVF NFRGGRISTA NYHTLPILRQ VRLSAERGLV YSNAMVAPHY
TVNSHVMYA TRGNARQVQV DNFQSQVFDG EVREQVLMV PQNFVVIKRA SDRGFEWIAP
KTNDNATNLT LAGRVSQMRM LPLGVLSNMY RISREEAQL KYGQEMRVL SPGRSQGRRE

Where does this information come from?
Which editor? Or was it the cited papers?

ID 11S_CUCMA STANDARD; PRT; 480 AA.
AC P13744;
DT 01-JAN-1990 (REL. 13, CREATED)

DT 01-JAN-1990 (REL. 13, CREATED)
DT 01-JAN-1990 (REL. 13, LAST SEQUENCE UPDATE)
DT 01-NOV-1990 (REL. 16, LAST ANNOTATION UPDATE)

RN [1]
RC STRAIN=CV. KUROKAWA AMAKURI NANKIN;
RX MEDLINE; 88166744.
RA HAYASHI M., MORI H., NISHIMURA M., AKAZAWA T., HARA-NISHIMURA I.;
RL EUR. J. BIOCHEM. 172:627-632 (1988).

RN [2]
RP SEQUENCE OF 22-30 AND 297-302.
RA OHMURA M., HARA I., MASTUBARA H.;
RL PLANT CELL PHYSIOL. 21:157-167 (1980).
CC -!- FUNCTION: THIS IS A SEED STORAGE PROTEIN.
CC -!- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC DISULFIDE BOND.

CC -!- SIMILARITY: TO OTHER 11S SEED STORAGE PROTEINS (GLOBULINS).
DR EMBL; M36407; G167492; -.
DR PIR; S00366; FWPUR

DR PROSITE; PS00305; 11S_SEED_STORAGE; 1.
KW SEED STORAGE PROTEIN; SIGNAL.
FT SIGNAL 1 21
FT CHAIN 22 480 11S GLOBULIN BETA SUBUNIT.
FT CHAIN 22 296 GAMMA CHAIN (ACIDIC).
FT CHAIN 297 480 DELTA CHAIN (BASIC).
FT MOD_RES 22 22 PYRROLIDONE CARBOXYLIC ACID.
FT DISULFID 124 303 INTERCHAIN (GAMMA-DELTA) (POTENTIAL).
FT CONFLICT 27 27 S -> E (IN REF. 2).
FT CONFLICT 30 30 E -> S (IN REF. 2).

SQ SEQUENCE 480 AA; 54625 MW; D515DD6E CRC32;
MARSSLFTFL CLAVFINGCL SQIEQQSPWE FGSEVWQOH RYQSPRACL ENLRAQDPVR
RAEAEAFTE VWDQNDDEFQ CAGVNMRHT IRPKGLLLG FSNAPKLIV AQGFGIRGIA
IPGCAEYQT DLRRSQSAGS AFKQHQKIR PFRGGDLIV FVAGVSHWYN RGQSDLVIV
FADTRVNAQ IDPILRKFYL AGRFEQVERG VEWEERSKX GSGSEKSGNI FSGFADEFLE
EAFQIDGGLV RKLKGEDDER DRIVQVDEF EVLPEKDEE ERSRGRVIES ESENGLEE
TICTLRLKQN IGRSVRADVF NFRGGRISTA NYHTLPILRQ VRLSAERGLV YSNAMVAPHY
TVNSHVMYA TRGNARQVQV DNFQSQVFDG EVREQVLMV PQNFVVIKRA SDRGFEWIAP
KTNDNATNLT LAGRVSQMRM LPLGVLSNMY RISREEAQL KYGQEMRVL SPGRSQGRRE

History of the entry is incomplete (only version numbers of last updates are kept.)

How to describe provenance?

- Suppose a table is created by a query Q on a database D . How do we describe the provenance of some "piece" t of the output?
- Could just give Q and D
- Could try to identify those parts of D that "contributed to" t .
- What do we mean by "contributed to"?

Two kinds of provenance

name	born	period
J.S. Bach	1685	baroque
G.F. Handel	1685	baroque
W.A. Mozart	1756	classical

```
SELECT name, born  
FROM composer
```

```
SELECT name, born  
FROM composer  
WHERE born < SELECT AVERAGE born FROM composer
```

name	born
J.S. Bach	1685
...	...

Why is this element in the output?

Where does this element come from?

Why and Where

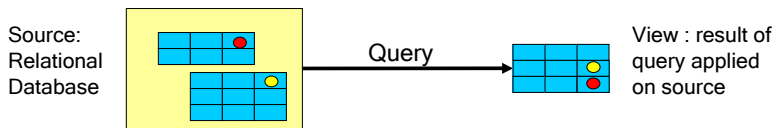
- **Where-provenance** of output data d
 - the set of all source locations whose contents are copied to d
- **Why-provenance** of an output tuple d
 - the set of all witnesses for d
 - a witness for d is a minimal set of source tuples which "proves" that d exists in the output
 - For positive queries -- a set of tuples in the source whose deletion causes d to disappear

A related topic -- annotation

- If I annotate part of the DB how should the annotation "spread" to other users.
- Annotations need to spread forward and backwards through queries.
- **Annotation Systems:**
 - Annotea ([W3C](#))
 - annotate web pages
 - location is defined with XPointer
 - BioDAS (Distributed Annotation Server) ([L.Stein et. al](#))
 - annotate on genome sequences

Annotation placement

(thanks to Wang-Chiew Tan for the next few slides)



- An annotation is placed in the view
 - where do we place the annotation on source?
- Annotation placement problem presented in relational setting
 - results carry over to fragments of XML (hierarchical model)

Example

Serves fine French Cuisine in elegant setting. Jackets required.

Extensive wine list!

Yummy chicken curry!!

NYRestaurants (Source Table)

Restaurant	Cost	Type	Zip
Peacock Alley	\$\$\$	French	10022
Bull & Bear	\$\$\$	Seafood	10022
Pacifica	\$	Chinese	10013
Soho Kitchen & Bar	\$	American	10022

All Restaurants (View 1)

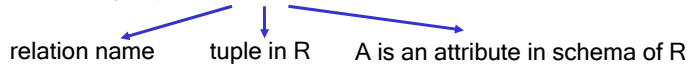
Restaurant	Cost	Type
Peacock Alley	\$\$\$	French
Bull & Bear	\$\$\$	Seafood
Pacifica	\$	Chinese
Soho Kitchen & Bar	\$	American

Cheap Restaurants (View 2)

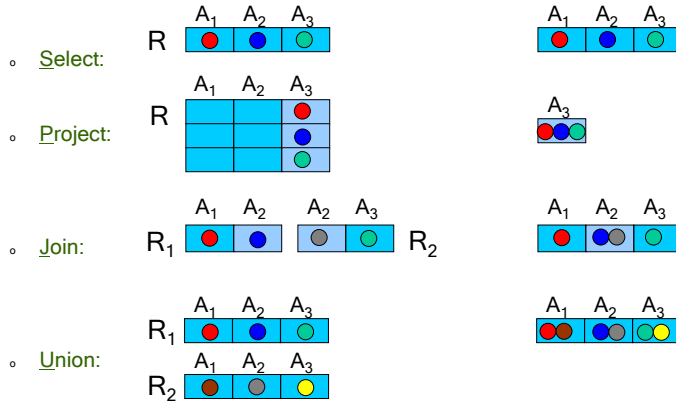
Restaurant	Cost	Type
Pacifica	\$	Chinese
Soho Kitchen & Bar	\$	American

Location and Propagation Rules

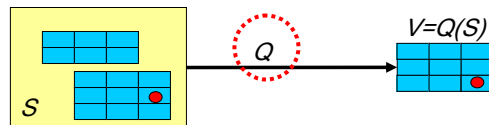
- A *location* is a triple: (R, t, A)



- Propagation Rules:



The Bad News



- It is NP-hard to decide if there is a side-effect-free annotation for a Project-Join (PJ) query.
- There is a polynomial time algorithm for SPJU queries that do not simultaneously contain a Project and a Join operation.

General results (Khanna, Tan)

- **View deletion problems.** Given a tuple t in $Q(S)$, find a subset T of S whose removal causes t to disappear from $Q(S)$ and
 - T is minimal (source minimisation) or
 - The effect on $Q(S)$ is minimal
- For both cases
 - PJ and JU: NP-hard
 - SPU and SJ: P
- This is a form of "why-provenance"

View Annotation Problem

- Given an annotation $Q(S)$ (i.e. a location in $Q(S)$), find a location in S to minimise the "spread" of annotations.
 - Only view minimisation is meaningful
- Queries involving PJ: NP-hard
- SPU, SJU: P

All the complexity problems arise because we can copy (PJU) two "locations" into one.

Are we going about this the right way?

- Should our queries be "annotation conscious"?

```
SELECT name, age
FROM employee
WHERE age = 50
```

```
SELECT name, 50 as age
FROM employee
WHERE age = 50
```

- Is the problem more complex?

Name	Shoesize	Hatsize
Joe	8	47
...

Annotations: "47 is prime" and "47 is too low" with arrows pointing to the Hatsize value 47.

- Are we ignoring useful structure -- e.g. some notion of location?

Semistructured data and keys

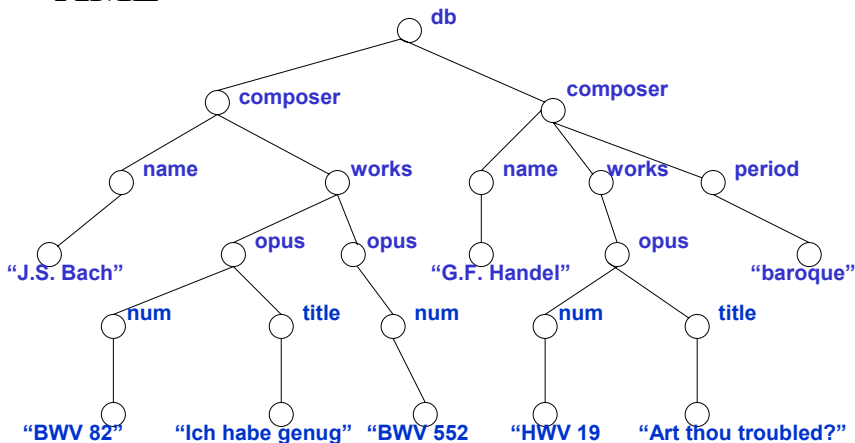
- Negative results on annotation are a consequence of mapping several locations to one.
- Much scientific data is hierarchical, if not semistructured, but has a well-defined notion of location.
- *Ad hoc* annotation requires some notion of location to make the annotation "stick"
- We get some interesting connections...

A *deterministic* model for semi-structured -- and structured-- data (Deutsch, Tan)

Synopsis:

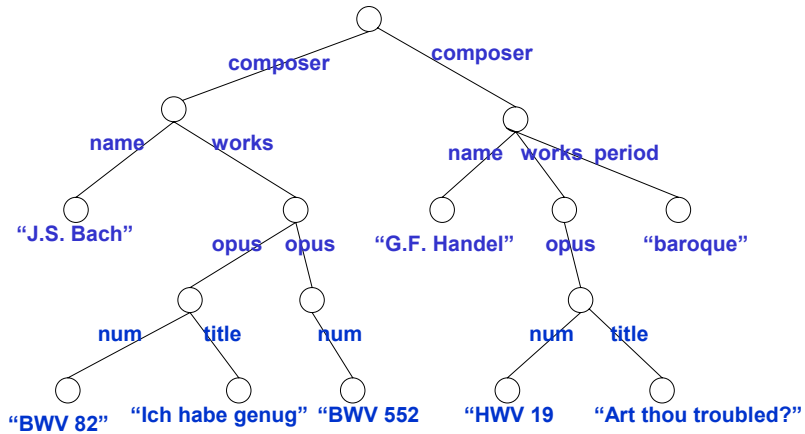
- Based on common model of semistructured data as an edge-labeled graph.
- Deterministic (out-edges have distinct labels) -- less general
- Consistent with well-designed relational databases (E-R diagrams)
- New connections between types and constraints
- Commonly used in practice (claim!)

Semistructured data: node-labeled as in XML



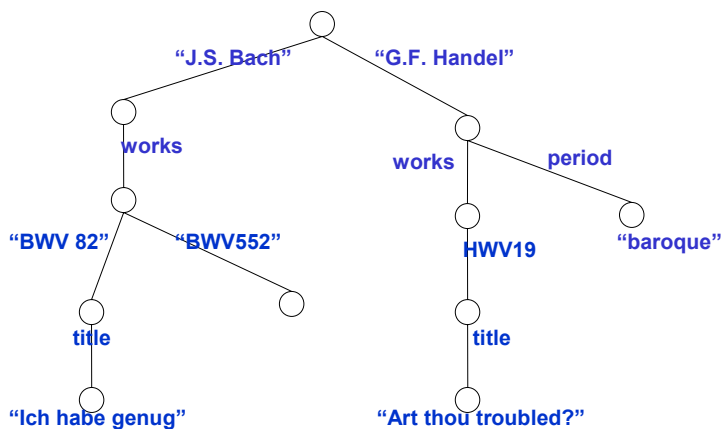
[order is important in XML]

Semistructured data: edge-labeled as in UnQL, XML-QL



[These systems mostly ignore horizontal order]

Semistructured data: deterministic model



Semistructured data -- syntax

```
{ composer: {  
  name: "J.S. Bach",  
  works: {  
    opus: {num: "BWV82, title: "Ich habe genug"},  
    opus: {num: "BWV552"}  
  } },  
  composer: {  
    name: "G.F.Handel",  
    . . .  
  }  
}}
```

non-deterministic

```
{"J.S. Bach":  
  {works: {"BWV82": {title: "Ich habe genug"},  
           "BWV552": {}  
  },  
  "G.F.Handel": { . . . }  
}
```

deterministic

Examples

- **A record:** {Name: "Bruce", Height: 6.2}}
- **An array:** {1: "a", 2: "b", 3: "c"}
- **A set:** {1: {}, 2: {}, 4: {}}
- **A multiset:** {"Guinness": 3, "Molson": 2}
- **A relation with an explicit key:**
{ 11: {Name: "Kim", Rate: 50},
 22: {Name: "Bob", Rate: 75} }

Precedents

- Relational databases from ER diagrams
- ACeDB, largely ignored by the DB community
- Most scientific data formats
- Model for file/data synchronisation by Pierce *et al*

- File and directory names that contain data

`/timit/train/dr1/fcjf0/sa1.wav`

corpus: timit
type: training
dialect-region:1
sex: f
speaker-id: cjf0
sentence-id: sa1
file-type: waveform

- Compound keys traditionally indicated location:

BL MS Cotton Nero A.ix

Manuscript in the British Library, which used to be in the library of a Mr. Cotton [which burnt down] under a statue of Nero, top shelf, nine books along from the left.

- Deterministic model has interesting implications for query languages and type systems for semistructured data.
- Suggests keys for XML
- Efficient archiving
- May give us better models of construction of curated DBs

Keys for XML (Davidson, Fan, Hara, Tan)

- Implicit keys are ubiquitous in scientific data formats (easily converted to XML)
- Some proposals for key specifications in XML work (DTD IDs, XML-Schema)
- "Deep citation" in digital libraries.
- Natural consequence of translating back from deterministic model to XML (node-labeled)

Keys for Relational DBs

Key attributes

Enrollment:	Student	Course	Grade	Project
<i>Target set</i>	Jones	Math2	95	B-
	Smith	Phil4	88	A
	Smith	Math2	77	C
	Rebus	Phil4	99	B+

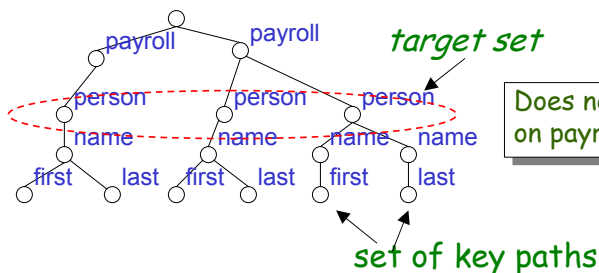
- Keys are critical in database design
- Keys are used to build indexes (optimization)
- Need to understand key inference

Key specification (node-labeled)

General form: $(Q \{P_1, \dots, P_n\})$

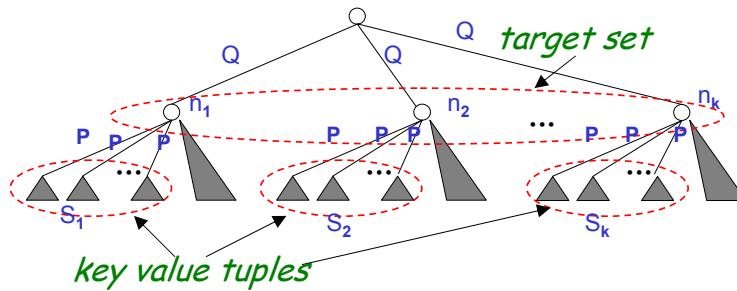
path expressions

Example: $(\text{payroll.person}\{\text{name.first}, \text{name.last}\})$



Does not impose uniqueness on payroll or person nodes

Meaning of a key spec. (single key path (Q{P}))



"Value" equality

nodes identical

- If $S_i \cap S_j$ nonempty then $n_i = n_j$
- ($|S_i| = 1$ ["strong" keys])

Key inference (absolute keys) [Davidson, Fan, Hara, Tan]

$\frac{(Q, S) \text{ } P \text{ any path expression (Superkey)}}{(Q, S \cup \{P\})}$	$\frac{(Q, Q', \{P\}) \text{ (Subnodes)}}{(Q, \{Q'.P\})}$
$\frac{(Q, S \cup \{P_i, P_j\}) \text{ } P_i \subseteq P_j \text{ (Path containment)}}{(Q, S \cup \{P_i\})}$	$\frac{(Q, S) \text{ } Q' \subseteq Q \text{ (Target containment)}}{(Q', S)}$
$\frac{(Q, S \cup \{P_i\}) \text{ } P_i \subseteq P \text{ (Key containment)}}{(Q, S \cup \{P\})}$	$\frac{S \text{ any set of path expressions (Root)}}{(\epsilon, S)}$

These are sound and complete !
(Path expressions may have .*)

Relative keys

General form: $Q\{P_1, \dots, P_n\}. Q'\{P'_1, \dots, P'_n\} \dots$

Example:

`book{name}.chapter{number}.verse{number}`

number specifies
chapter *only*
within book

number specifies
verse *only* within
chapter

Also:

`bible{}.book{name}.chapter{number}.verse{number}`

empty key: at most one bible node

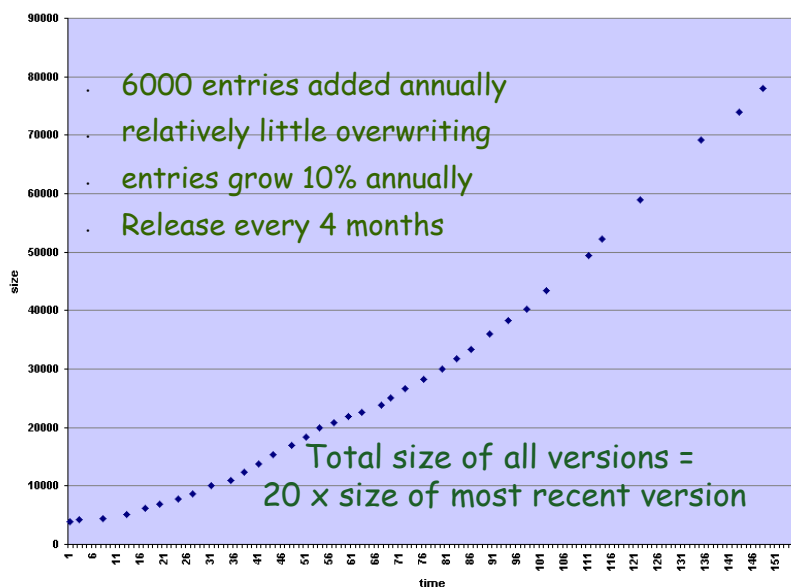
Notes on Keys

- Closely related to data models:
 - `payroll{}.employee{id}.[name{}, sal{}, ...]` (something like a "complex object"/nested relational model)
- Lots more is now known.
 - Interference with data model of XML Schema (Fan, Libkin)
 - Related to functional dependencies for XML (Arenas, Libkin)

How do we Build Archival Databases? [Khanna, Tajima, Tan]

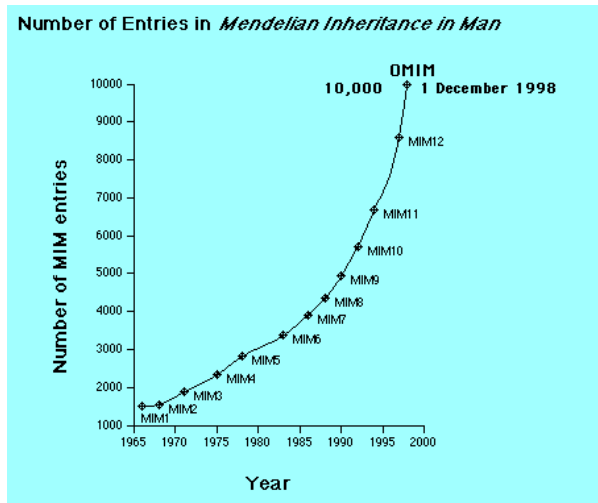
- Many scientific databases keep *archives*. It's important to preserve the state of knowledge as it was in the past
- Archive frequently: space consuming
- Archive infrequently: delay in getting recent information published.

Swissprot



Online Mendelian Inheritance in Man

- Printed editions stopped in 1998
- Updated daily!



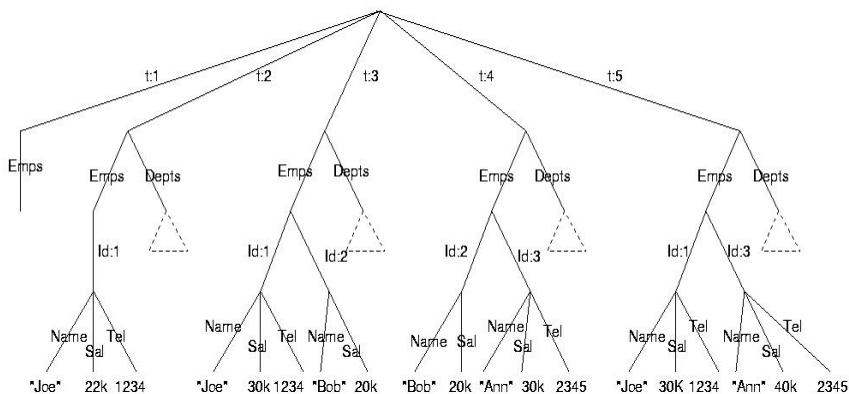
OMIM vs. Swissprot

- Both valuable curated databases
- Similar gross structure -- sequence of entries, each with internal structure
- Swissprot:
 - All past versions available
 - Slow release -- every 3-4 months
- OMIM
 - Past versions unavailable
 - Rapid release -- every day (or more often)

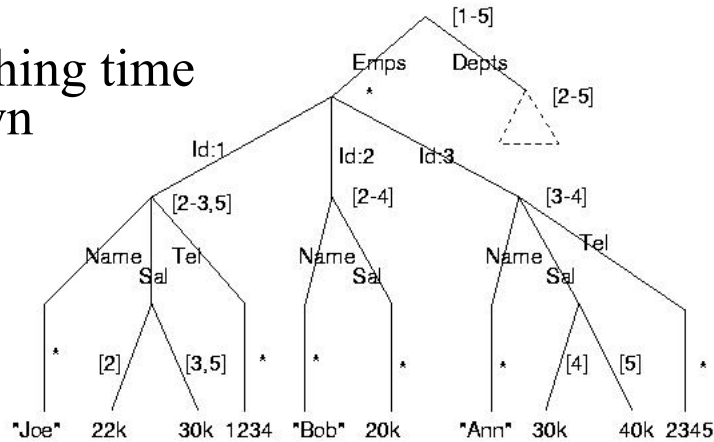
Why not use diff?

- Diff currently used for archival part of CVS
- Tree diffs have not yet come to market
 - Line diffs sometimes work well on formatted XML
- Diffs do not preserve "object-hood"
- Expensive to unwind 365 diffs

A Sequence of Versions



Pushing time down



This relies on a deterministic / keyed model

[Driscoll, Sarnak, Sleator, Tarjan: "Making Data Structures Persistent."]

An initial experiment

- Recorded all OMIM versions for about 14 weeks (100 of them)
- XML-ized all of them
- Combined into archive XML format file by pushing time down.
- Also recorded diffs between versions
- Did the same the same thing for the last 20 available versions of Swissprot

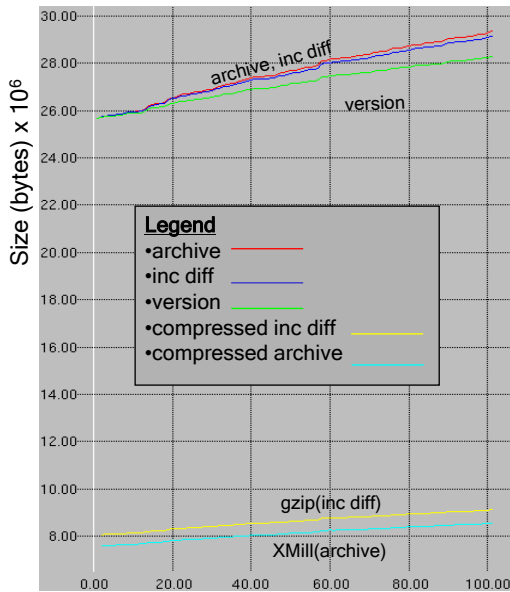
100 days of OMIM

Uncompressed

- Archive size is
 - ≤ 1.01 times diff repository size
 - ≤ 1.04 times size of largest version

Compressed

- archive size between 0.94 and 1 times compressed diff repository size
- gzip - unix compression tool
- XMill - XML compression tool



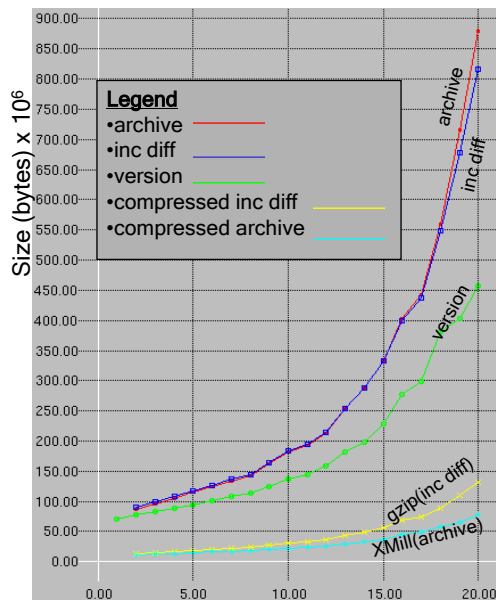
~ 5 years of Swissprot

Uncompressed

- Archive size is
 - ≤ 1.08 times diff repository size
 - ≤ 1.92 times size of largest version

Compressed

- archive between 0.59 and 1 times compressed diff repository size



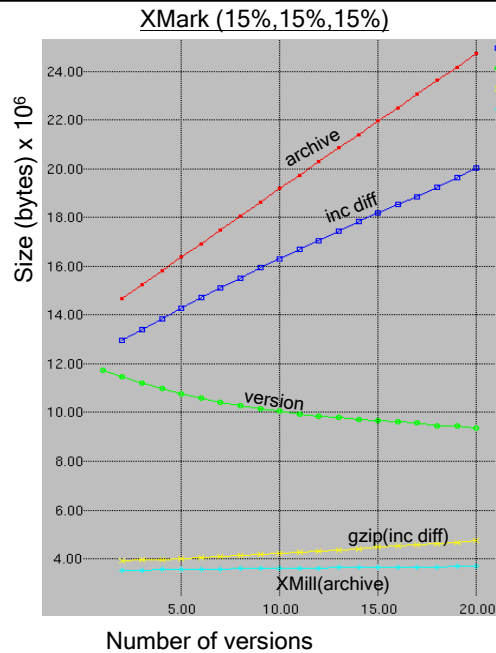
Synthetic XMark Data

Uncompressed

- Archive size is
 - ≤ 1.23 times diff repository size
 - ≤ 2.11 times size of largest version

Compressed

- archive size is between 0.78 and 1 times compressed diff repository size



The Bottom Line

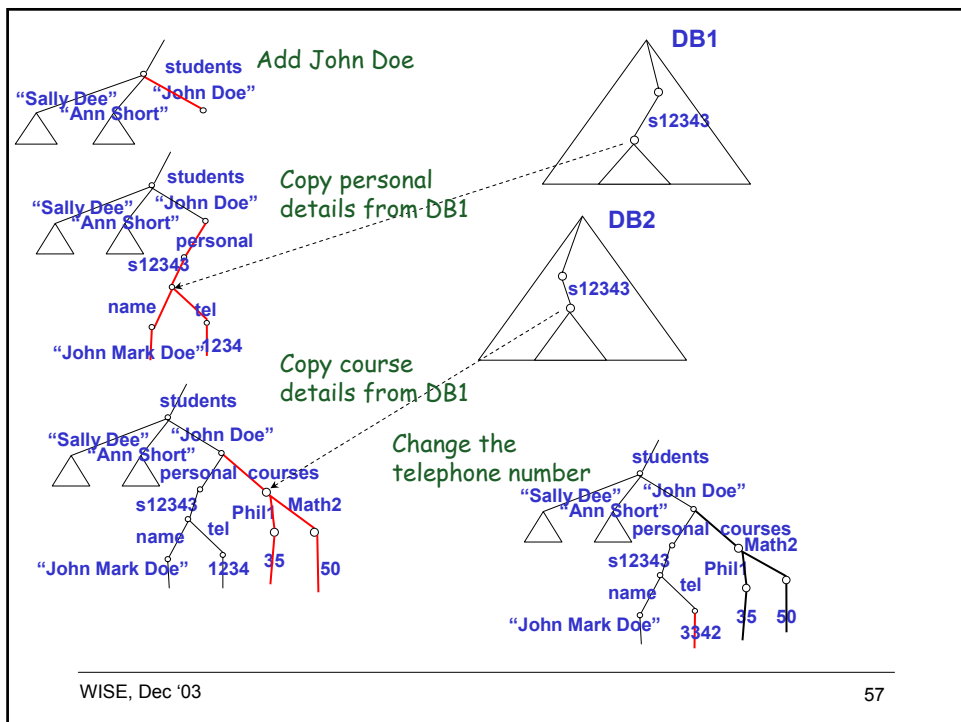
- Can archive a whole year of Swissprot or OMIM with < 15% overhead (size of current file)
- Retrieval is a linear scan
- Works well with compression to less than 30% of current file.
- Archive as often as you like! (Almost)
- Permits temporal queries on objects

A copy-and paste model of curation

- Manually curated database are *not* views (constructed by single queries over other database)
- They are built by a long sequence of manual insertions and transfers from other databases.
- The underlying transactions are expressed as a sequence of SQL updates, but this may not be a helpful model.

Example: my “advising” database

- Student John Doe is assigned to me. I create a database entry under the key [John Doe](#)
- I obtain his personal details from the applications database and his academic record from another database.
- When he comes to see me, I correct his telephone number.



In existing systems

- Insertion is done through a forms interface or an XML editor
- Copying whole "chunks" of a database can be tricky. Even using SQL to "copy and paste" a value from another DB is unwieldy.
- Reasoning about provenance is correspondingly hard.

Simple copy-and-paste commands

MyDB.Students."John Doe" ← {}

Create edge

MyDB.Students."John Doe".personal.s12343 ← DB1.people.s12343

Create edges and transfer subtree

MyDB.Students."John Doe".courses ← DB2.enrollment.s12343

Create edge and transfer subtree

MyDB.Students."John Doe".personal.s12343.tel ← 3323

Overwrite subtree

These would/could normally be performed by a visual editor

Implemented with symbolic links or "cp -r ..." ?

What are the advantages?

- Easy to attach (where) provenance information to nodes.
- If everyone uses the same model, one can produce a "source" tree.
- Archiving after each transaction is efficient.
- Transaction log can be recovered from archive (with provenance annotations)

But we also need to do bulk transfers, e.g.

Import *all* my students' course information from DB2

This starts to look like a database query!

Bulk Transfers

Incorporate e-mail addresses from my address book

```
MyDB.Students.$n.personal.email:$e ← Addrbook.$n.mail:$e
```

(Explicitly naming the subtree to be moved)

Incorporate home addresses from DB3:

```
MyDB.Students.$n.personal.$i.home:$e ← DB3.people.$i.addr:$e
```

(Note \$n only occurs on LHS - this is OK)

Import course information for *all* my students

```
MyDB.Students.$n.courses:$e ← DB2.enrollment.$i:$e,
```

```
MyDB.Students.$n.personal.$i
```

(Includes a pattern/condition. RHS occurrence of \$i is problematic.)

Copy-and-paste language (CPL)

- We want our rules *not* to put two subtrees in the same place (unique provenance)
- Every variable must either occur in the LHS or be functionally determined by a variable on the LHS (key/uniqueness information)
- Also need to check that two rules do not conflict.
- CPL *cannot* perform "aggregate" queries such as (set theoretic) projection.

Bag Languages

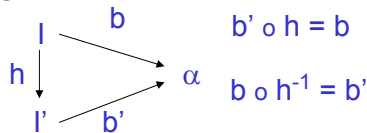
SQL is often called a bag language - doesn't eliminate duplicates:

`SELECT age FROM employee.`

Comprehensions and their connection with monads generalise SQL-like query languages to nested structures and to bags and lists [Wadler, UPenn].

Usual definition of a bag of α , $b: \alpha \rightarrow \text{nat}$ which is zero on all but a finite set

Alternative definition: equivalence class of finite index sets:



Relationship with Bag Languages

Arbitrary construction of index functions in comprehensions doesn't work:

$$\{(x,x) \mid (i,x) \in B\}$$

But it's easy to say how the monad operations should construct explicit index functions

map: $(\alpha \rightarrow \beta) \rightarrow (I \multimap \alpha) \rightarrow (I \multimap \beta)$

flatten: $(I \multimap (I' \multimap \alpha)) \rightarrow (I \times I') \multimap \alpha$ (lossy - *not* currying)

singleton: $\alpha \rightarrow (\text{unit} \multimap \alpha)$

pairwith: $(I \multimap \alpha) \times \beta \rightarrow I \multimap (\alpha \times \beta)$

union: $(I \multimap \alpha) \times (I' \multimap \alpha) \rightarrow (I + I') \multimap \alpha$


These are both index-preserving languages

- Fact. CPL can simulate this bag language.
- Is some form of the reverse true? Need to describe appropriate tree types and null values.
- What can we say about CPL if we generalise the constraints/conditions?
- What constraints are preserved under CPL transformations?
- An SQL-like language that preserves keys?
- How do we add aggregating functions to CPL?

Summary

- Useful results:
 - Keys for XML
 - DB archiving
- Provenance/annotation:
 - A BIG problem. We have only scratched the surface. Ideas needed!
- A model for curated DB construction.
 - Is copy-and-paste a good model?





Edinburgh has numerous research positions in
databases, XML, web technology, fundamentals

Contact
Peter Buneman,
opb@inf.ed.ac.uk

Edinburgh is
a great place
to live!!!

Top-rated department. Excellent database group. Good connections with
logical foundations, scientific DBs, distributed computation (Grid)