The Challenges of Publishing Databases

Peter Buneman
University of Edinburgh
University of Pennsylvania

Thanks to: Susan Davidson, Wenfei Fan, James Frew and Gianmaria Silvello
Supported by National Science Foundation  IIS 130221
After 2000 a major change took place in how we disseminate knowledge.

The traditional function of 20th century libraries to disseminate knowledge has almost ceased to exist. (They still perform a useful archival function.)
And yet...

We (academics) still use traditional methods to publish information – methods that derive from a mode of publishing invented in China in the 3rd century and developed in Europe in the 14th century.

Some of the consequences:

- We publish huge amounts of data
- What we publish may be fluid (changes over time)
- We publish “data” as well as text
- Our publications have increasingly complicated structure
- Programs, as well as people, read our publications
- We publish in databases
We could “forget” academic publishing, but there’s an interesting middle ground: **curated databases**

- A digital reference work. Traditional dictionaries, gazetteers, encyclopedia have been replaced by curated databases.
- Value lies in the organization and annotation of data.
- Commonly constructed by copying parts of other (curated) databases.
- Rapidly increasing in scientific research. (> 1000 in molecular biology)
- Constantly checked/verified. Data quality and timeliness are important.
- Often group efforts. Produced by a dedicated organization or collaboration.
- Increasingly seen as “publications” by scientists. (You get kudos if someone uses your database – like a citation.)

Promoted by the Human Genome Project (1993-2003), they are precursors to much citizen science and “crowdsourced” data – Galaxy Zoo etc
We cannot separate the world of academic publishing from the much larger world of publishing data, so what are the problems?

Most conventional database issues show up: data transformation, data integration, query languages, etc. In addition:

- * Data provenance: where did this data come from.
- Database archiving: keep a record of old stuff.
- * Data citation: huge demand for this, but how?
- Data annotation: how many curated databases get built.
Provenance: 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 exploration, and sometimes for efficiency

Are these related?
Workflow provenance: what did I do to make these pictures?

Part of a workflow diagram
Data provenance

One of the first things one finds about curated databases is that much of their contents are obtained from other, possibly curated, databases. Knowing where data came from and how it was transformed is essential to its credibility and one of the main factors in data quality.

Some examples:
Swissprot (one of the first curated DBs)

Some of the fields in Swissprot are added by the curators (using the cited papers) others are copied from other databases – which ones.

Also note the use of the comment fields to store evolving knowledge!
Linked open data

A huge effort to link together and represent the data in hundreds of databases (about 30bn triples)

Almost entirely copied from other curated databases and, sadly, mostly stale.
<table>
<thead>
<tr>
<th><strong>Introduction</strong></th>
<th>AFGHANISTAN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Geography</strong></td>
<td>AFGHANISTAN</td>
</tr>
<tr>
<td><strong>People and Society</strong></td>
<td>AFGHANISTAN</td>
</tr>
</tbody>
</table>

**Nationality:**
- noun: Afghan(s)
- adjective: Afghan

**Ethnic groups:**
- Pashtun 42%, Tajik 27%, Hazara 9%, Uzbek 9%, Aimak 4%, Turkmen 3%, Baloch 2%, other 4%

**Languages:**
- Afghan Persian or Dari (official) 50%, Pashto (official) 35%, Turkic languages (primarily Uzbek and Turkmen) 11%, 30 minor languages (primarily Balochi and Pashai) 4%, much bilingualism, but Dari functions as the lingua franca

**note:** the Turkic languages Uzbek and Turkmen, as well as Balochi, Pashai, Nuristani, and Pamiri are the third official languages in areas where the majority speaks them

**Religions:**
- Sunni Muslim 80%, Shia Muslim 19%, other 1%

**Population:**
- 30,419,928 (July 2012 est.)

**note:** this is a significantly revised figure; the previous estimate of 33,609,937 was extrapolated from the last Afghan census held in 1979, which was never completed because of the Soviet invasion
The Population of Corfu

<table>
<thead>
<tr>
<th>Population</th>
<th>Source Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>107,879 (as of 2001)</td>
<td><a href="http://en.wikipedia.org/wiki/Corfu">http://en.wikipedia.org/wiki/Corfu</a></td>
</tr>
<tr>
<td>93,000</td>
<td><a href="http://www.corfunet.corry/corfu/">http://www.corfunet.corry/corfu/</a></td>
</tr>
<tr>
<td>109,512</td>
<td><a href="http://www.agni.gr/">www.agni.gr/</a></td>
</tr>
<tr>
<td>110,000</td>
<td><a href="http://www.corfuvisit.net">www.corfuvisit.net</a></td>
</tr>
<tr>
<td>70,000</td>
<td><a href="http://www.newadvent.org/cathan/04362a.htm">http://www.newadvent.org/cathan/04362a.htm</a></td>
</tr>
<tr>
<td>107,600</td>
<td><a href="http://www.greek-hotels.com/">http://www.greek-hotels.com/</a></td>
</tr>
<tr>
<td>105,043</td>
<td><a href="http://www.merriam-webster.com/dictionary/corfu">http://www.merriam-webster.com/dictionary/corfu</a></td>
</tr>
<tr>
<td>approximately 110,000</td>
<td><a href="http://www.kassiopi.com/MenuContent.aspx?MenuId=6">www.kassiopi.com/MenuContent.aspx?MenuId=6</a></td>
</tr>
<tr>
<td>approximately 120,000</td>
<td><a href="http://www.gardeno-corfu.com/">http://www.gardeno-corfu.com/</a></td>
</tr>
<tr>
<td>around 110,000</td>
<td><a href="http://www.sunshinetravel.gr/CORFGUIDE/CORFU">http://www.sunshinetravel.gr/CORFGUIDE/CORFU</a> TRAVEL GUIDE 0-1.htm</td>
</tr>
<tr>
<td>110,000</td>
<td><a href="http://www.dialashop.com/travel/corfu.html">http://www.dialashop.com/travel/corfu.html</a></td>
</tr>
<tr>
<td>about 110,000</td>
<td><a href="http://www.argobenises.gr/greece.php">http://www.argobenises.gr/greece.php</a></td>
</tr>
<tr>
<td>107,880</td>
<td><a href="http://catalogue.horse21.net/greece+hotels/corfu+hotels/hotels5j/luxury">http://catalogue.horse21.net/greece+hotels/corfu+hotels/hotels5j/luxury</a></td>
</tr>
<tr>
<td>109,512</td>
<td><a href="http://www.corfu-property.gr/content/view/14/38/ang,en/">http://www.corfu-property.gr/content/view/14/38/ang,en/</a></td>
</tr>
<tr>
<td>about 100,000</td>
<td><a href="http://members.virtualtourist.com/m/6ce90/67541/">http://members.virtualtourist.com/m/6ce90/67541/</a></td>
</tr>
<tr>
<td>110,000 approximately</td>
<td><a href="http://www.corfu-island.org/features.htm">http://www.corfu-island.org/features.htm</a></td>
</tr>
</tbody>
</table>

*** The only site to give attribution: http://www.statistics.gr/portal/page/portal/ESYE
Two big questions

- How do we describe data provenance, Several interesting ideas. But there is still much work to be done. Three loose classifications
  - Why is this piece of data in the database
  - How did it get there
  - Where did it come from

- Can we effectively record it?
  - Is keeping all this extra information going to make our databases impossibly large?
What is the difference?

Pedigree

Lineage

Provenance
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”
Annotation – a closely related issue

Annotating with comments

Bill is underpaid
Bill likes Mary

Emps:
<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Sal</th>
<th>Dept</th>
</tr>
</thead>
<tbody>
<tr>
<td>123456</td>
<td>Joe</td>
<td>40k</td>
<td>Sales</td>
</tr>
<tr>
<td>123321</td>
<td>Bill</td>
<td>20k</td>
<td>Research</td>
</tr>
<tr>
<td>654321</td>
<td>Mary</td>
<td>50k</td>
<td>Research</td>
</tr>
</tbody>
</table>

Depts:
<table>
<thead>
<tr>
<th>Dept</th>
<th>Manager</th>
<th>Budget</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>Mary</td>
<td>500k</td>
</tr>
<tr>
<td>Sales</td>
<td>Jane</td>
<td>800k</td>
</tr>
</tbody>
</table>

SELECT Name, Manager
FROM Emps, Depts
WHERE Emps.Dept = Depts.Dept
AND Id = 123321

Bill is underpaid
Bill likes Mary
Mary likes champagne

We probably want the union of the comments on the input
Annotating with beliefs: the people who believe a tuple to be true

SELECT Name, Manager
FROM Emps, Depts
WHERE Emps.Dept = Depts.Dept
AND Id = 123321

We want the intersection of the believers of the input tuple
Annotating with beliefs for another query:

For UNION queries we want the \textit{union} of the believers of the input tuples.
Provenance/Annotation Semirings or *How* provenance
(Tannen atelier: PODS ’07, ’08 & ’11)

$$R:\begin{array}{cc|c}
  a & b & c & p \\
  d & b & e & r \\
  f & b & e & s \\
\end{array}$$  

$$V:\begin{array}{cc|c}
  a & c & p+ (p \cdot p) \\
  a & e & p \cdot r \\
  d & c & r \cdot p \\
  d & e & r + (r \cdot r) + (r \cdot s) \\
  f & e & s + (s \cdot s) + (s \cdot r) \\
\end{array}$$

$V(X,Z) := R(X, \_ , Z)$
$V(X,Z) := R(X, Y, \_ ), R( \_ , Y, Z)$

Tuples are created by:

“joining” other tuples (join): $p \cdot r$

“merging” other tuples (project and union): $p + r$

Both the “$\cdot$” and “$+$” are commutative and associative,

“$\cdot$” 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 \(p, q, r, \ldots\)
- Comment semiring \((\text{STR}, \cup, \cup, \emptyset, \emptyset)\) \(\text{STR} = \text{set of strings}\)
- Belief semiring \((B, \cup, \cap, \emptyset, B)\) \(B = \text{set of believers}\)
- Many well-known extensions to relational algebra are examples of semirings:
  - bag semantics
  - C-tables
  - probabilistic databases
  - various forms of why-provenance
- Example (bag semantics): Abstract quantities are multiplicities. Semiring is \((\mathbb{Z}, +, \times, 0, 1)\)
  - Multiplicity of \((d, e)\) in \(V\) is \(r + (r \times r) + (r \times s)\)
This approach also explained “why provenance” (one of the first attempts at a formulating data provenance)

$I$ is a database instance and $Q$ is a query.

If tuple $t \in Q(I)$, what subset $J$ of $I$ “influenced” the appearance of $t$? Presumably $t \in Q(J)$. Such a $J$ is a witness for $t$.

- Lineage – defined for relational algebra syntactically [Cui & Widom, TODS 2000]. Not obviously compositional.
- Set of minimal witnesses, [B. Khanna, Tan, PODS 2002]. But what is composition? This work told us.

But it didn’t help us with where provenance...
Practical “where provenance”: stuff gets copied

The evils of copy-paste!

| Length: | 63 km / 39 miles |
| Area: | 370 square km / 229 square miles |
| Highest point: | 296 m (971 feet) |
| Population of Town: | 113,479 (2001) |
| Exports: | Olive Oil |
| Sea temperatures: | 12°C min / 28°C max |
| Average summer temperatures: | 28°C |
| Average winter temperatures: | 18°C |
| Hotest months: | July / August |
| Coolest months: | January / February |
| Wettest months: | November / March |
| No. of villages: | 309 |
| No. of tourist resorts: | 42 |


93,000 [http://www.corfunet.com/corfu/](http://www.corfunet.com/corfu/)

109,512 [www.agni.gr/](http://www.agni.gr/)
DE  11S GLOBULIN BETA SUBUNIT PRECURSOR.
OS  CUCURBITA MAXIMA (PUMPKIN) (WINTER SQUASH).
OC  EUKARYOTA; PLANTA; EMBRYOPHYTA; ANGIOSPERMAE; DICOTYLEDONEAE; 
OC  VIOLALES; CUCURBITACEAE.

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).

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).

Where does this information come from? Which curator? Or was it the cited papers?
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 (copy-paste)

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 structured documents and for query and update languages
Copy-paste model of curated DBs

[B. Chapman, Cheney, Sigmod ’06]
(a) A biologist copies some SwissProt records into her DB.
(b) She fixes entries so that SwissProt PTMs are not confused with hers.
(c) She copies in some publication details from OMIM
(d) She corrects a mistake in a PubMed publication number.
A very simple copy-paste language (uses a “deterministic” tree model)

(1) delete c5 from T;
(2) copy S1/a1/y into T/c1/y;
(3) insert {c2 : {}} into T;
(4) copy S1/a2 into T/c2;
(5) insert {y : {}} into T/c2;
(6) copy S2/b3/y into T/c2/y;
(7) copy S1/a3 into T/c3;
(8) insert {c4 : {}} into T;
(9) copy S2/b2 into T/c4;
(10) insert {y : 12} into T/c4;

How costly is it to record all this?
How to reduce space

- Complete provenance: Record every update.
- Transactional provenance: Record the links at the end of some user-defined transaction (sequence of updates)
- Hierarchical (inferred) provenance. Only record a link if it cannot be inferred from the provenance of a higher node

Taken together these provide a substantial saving on storage. Overhead comparable with the size of the DB in some realistic simulations.
Publications have become databases and databases have become publications.

How do we cite them?

Generally you cite part of a database.

Databases are large and complex and they evolve over time.
People who think you should cite data

Large number of organizations: Datacite, DataONE, GEOSS, D-Lib Alliance, DCC, COPDES, Force-11, AGU, ESIP, DCMI, CODATA, ICSTI, IASSIST, ICSU…

Force 11: “Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.”

DataCite: “We believe that you should cite data in just the same way that you can cite other sources of information, such as articles and books.”

Amsterdam Manifesto: “Data should be considered citable products of research.”

Most important: the major research funders want you both to publish your data and to have it cited.
Our manifesto [B. Davidson & Frew, CACM Sept 2016]

The principles and standards for data citation that have been proposed are unlikely to be used unless the process of extracting information is coupled with that of providing a citation for it.

We need to generate citations *automatically* – as the data is extracted.

Data citation is a computational problem.
Start of a 700 line SQL component of some OLAP API

```sql
SELECT /*+ NOPARALLEL bypass_recursive_check */
    SP_ALIAS_190,
    (CASE SP_ALIAS_191
        WHEN 1 THEN 'PROVIDER::ALL_PROV::'
        WHEN 0 THEN 'PROVIDER::PROV:'
    ELSE NULL END) || SP_ALIAS_190 ALIAS_3553,
    SP_ALIAS_194,
    SP_ALIAS_191,
    SP_ALIAS_192,
    SP_ALIAS_193,
    SP_ALIAS_205,
    D4_AGE_GROUP_ET,
    (CASE D4_AGE_GROUP_GID
        WHEN 1 THEN 'AGE_GROUP::ALL_AGE_GRP:'
        WHEN 0
```
An Example: GtoPDB
a major curated database for pharmacologists

IUPHAR Guide to Pharmacology (GtoPDB) is a database of information about drug targets, and the prescription medicines and experimental drugs that act on them.

Information is presented to users through a hierarchy of web views, with an underlying relational implementation.

Contents of the database are generated by hundreds of experts who, in small groups, contribute to portions of the database. Thus the authorship depends on what part of the database is being cited.
GtoPDB is normally presented as web pages, but

- an abstract is published as a book, and
- one can also write queries against the underlying database
Citation to the IUPHAR database as a whole (the root) is a traditional paper written by the main curators of the database.

IUPHAR Family and Family Introduction pages have independent citations.
The hierarchy simplified

In some cases they would like to have citations for individual rows in a table
What we have seen:

Citations vary with what part of the database is being cited.
There may be a huge number of “parts” of a database.
We cannot expect to put a citation for each “part” into DBLP.
To repeat:

Unless we couple the process of generating a citation with the act of extracting the data, the advocacy of data citation is pointless.

The main problem:

*Given a database $D$ and a query $Q$, generate an appropriate citation.*

- Database owners need to be able to specify citations to parts of the database – schema level information.
- Database users need to have citations “served up” as they extract the data.
The idea:

Database owners can usually specify the citations for some parts of the database. They can for GtoPDB and several databases we have seen.

The person who wants to cite the database may have no idea what parts of the database were accessed by the query that got the data, and may not even know what the query was.

In any case what do we mean by a “part”?

Could the user’s data have been obtained by another query Q on the part? If so we could use the citation for the part.
More precisely

What we call a “part” is what database people call *views* which are functions (expressed as queries – usually rather simple ones) that map the database into some other “smaller” database.

So owners supply citations for some views of the database, $V_1 \ldots V_n$.

The problem becomes: Given a schema $S$ and a query $Q$, can it be factored through a view? That is, is there a $Q_i$ such that

$$\forall D \in S. \ Q(D) = Q_i(V_i(D))$$

If so, the citation for $V_i$ is a *candidate citation* for $Q$. 
Database theory to the rescue!!

Questions involving the analysis of programs are generally undecidable, but views are rather simple programs and rewriting through views is a well-understood problem and for the cases we are dealing with there is often an efficient solution.*

But there’s a twist to the problem: we have may have lots of views. For GtoPDB we have (at least) three classes of views $C_{\text{Family}}$, $C_{\text{Intro}}$, and $C_{\text{Target}}$, each of which contains hundreds of views. The $C_{\text{Family}}$ class is defined by a single query with a parameter that can be instantiated by any Family name in the database.

How does this increase the difficulty? Recently answered by Ting Deng

Hierarchies make life simpler

In the web presentation of GtoPDB there is a natural hierarchy of views, e.g. each member of $C_{\text{Target}}$ is a subview of a member of $C_{\text{Family}}$. To specify these views we can use simple XPATH-like language:

**Family view:** /Root/Family[FamilyId=\$f]

**Family Introduction view:** /Root/Family[FamilyId=\$f]/Introduction

**Target view:** /Root/Family[FamilyId=\$f]/Target[TargetId=\$t]

A query, generated by traversing web interface, is another path, e.g.:

/Root/Family[FamilyID="F123"]/Target[TargetID="MT1"]/LigandTable

This can be answered using the Family and Target views above (both are candidate views), but the Target view is more specific (smaller) so we would use its citation.
We can use this to create a simple citation generation rule

The (parameterized) rule for introductions is:

/Root[VersionNumber: $v]/Family[FamilyName: $$f]/Introduction[Contributor-list: $a]```

And a sample result is (in some random syntax, but could be BibTeX or any other style):


But these XPath-like patterns also correspond to relational queries on the underlying database.
This works, but there are lots of unanswered questions

- What happens if the query can be answered using the union of two incomparable views – give both citations?
- What happens if the query can be answered using either of two incomparable views?
- What happens if the query can be answered using 100 views – give 100 citations or cite a super-view – if one exists?
- How do we find whether the query can be answered using the union of views?

How applicable is this to other databases?
Other databases

In biology:

• eagle i: A “resource discovery” tool built to facilitate translational science research. Allows researchers to collect and share information.
• Reactome: A free, open source, curated and peer-reviewed pathway database.

Not quite as complicated as GtoPDB. Several others have the problem that the citation data is not available – not in the database or in any machine-readable form.

And another interesting database in Earth Sciences...
MODIS (MODerate-resolution Imaging Spectrometer)

Data products include properties like snow cover, ocean color, reflectance, etc, and are distributed as granules.

A granule is a fixed size subset representing an interval of the satellite orbit, or tile within a standard map projection of all or part of the earth.

Highlighted tiles of spatial extent for California.
This is rather trivial except for the fact that the views form a lattice, and this is needed because they want citations for sets of queries.
Conclusions

Provenance:

- A huge number of open questions (especially about combining data and workflow provenance)

Citation:

- We have taken a conventional view of citations but there are all sorts of other ideas about citation: transitive citation, citation ontologies, etc.
- Do we want papers with 10,000 citations?
- Can we make something that is reasonably generic and easy to use?
- As is often the case, database theory has useful answers
And there are a large number of related topics for computer scientists:

- Annotation of data (everyone wants it)
- Preservation of data
- Databases that keep their history (needed for citation and provenance)
- Understanding social and legal issues with data. E.g. copyright: a bad law based on a flawed concept.

Thank you
谢谢
Some of my favorite citations:

BL Cotton Nero A. X

Cotton Otho A. XII


Nature, 171,737-738

Peter Buneman

wget -qO - http://mirror.hmc.edu/ctan/FILES.byname | grep ".bst$" | sed 's/.*\(.*\)/\1/' | sort -u | wc -l

Executed on 18 November 2014