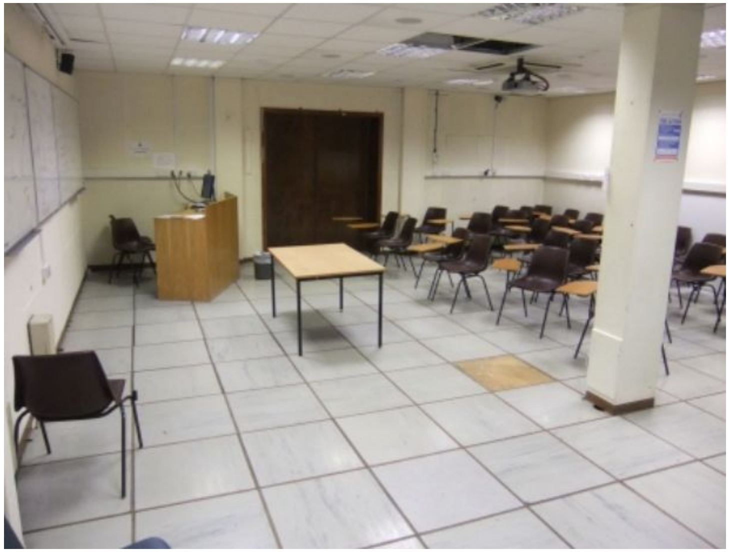
`charin.fntolist,erase.applist` and how not to do research

Peter Buneman

LFCS 30th anniversary

Once upon a time ...

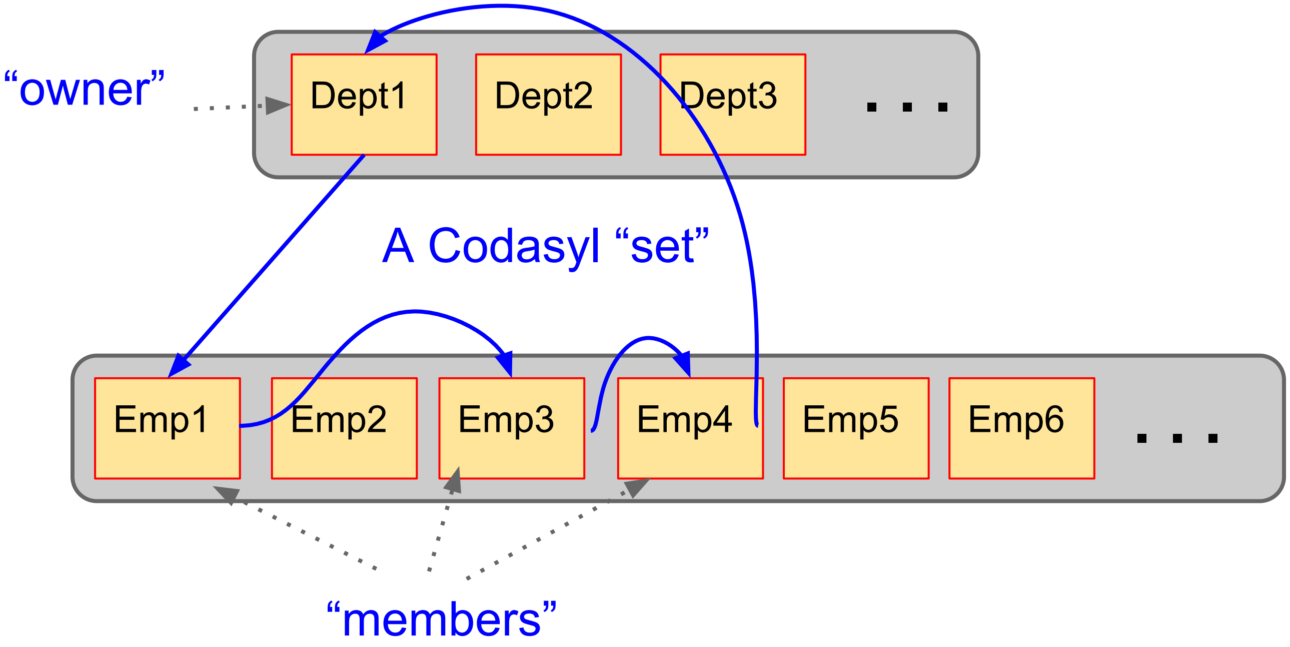
I spent a lot of time writing programs



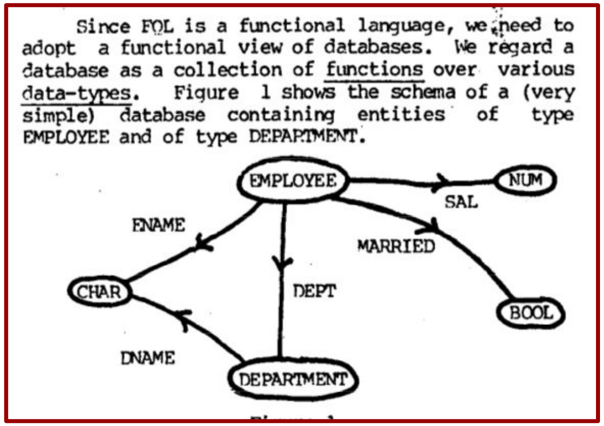
But I left the stimulating but tumultuous environment at Edinburgh to work in the US ...

… to work on databases

When relational databases were a theoretical nicety, we had Codasyl:

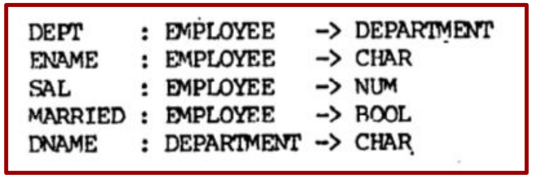


An embarrassingly long time ago, when LaTeX had not been invented.

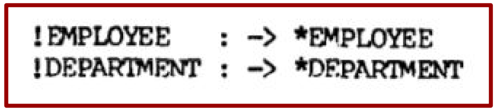


Since FQL is a functional language, we need to adopt a functional view of databases. We regard a database as a collection of functions over various data-types. Figure 1 shows the schema of a (very simple) database containing entities of type EMPLOYEE and of type DEPARTMENT.

EMPLOYEE
NUM
SAL
ENAME
MARRIED
CHAR
DEPT
BOOL
DNAME
DEPARTMENT



```
DEPT    : EMPLOYEE   -> DEPARTMENT
ENAME   : EMPLOYEE   -> CHAR
SAL     : EMPLOYEE   -> NUM
MARRIED : EMPLOYEE   -> BOOL
DNAME   : DEPARTMENT -> CHAR
```

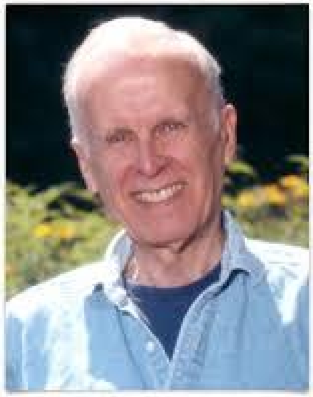


```
!EMPLOYEE   : -> *EMPLOYEE
!DEPARTMENT : -> *DEPARTMENT
```

Schema

* means “stream of”

Database instance is a set of functions



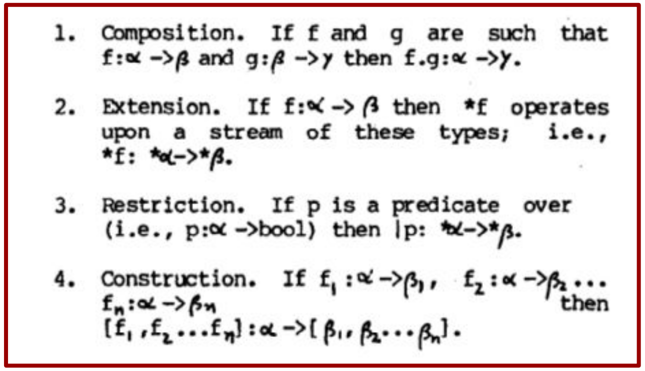
I had developed a taste of lazy and combinatory programming in POP-2. And Backus' FP had appeared.

```
charin.fntolist,erase.applist
```

Burstall, R.; Collins, J.; Popplestone, R. (1968). *Programming in POP-2*. Edinburgh: Edinburgh University Press.

Backus. *Can Programming Be Liberated from the von Neumann Style*? A Functional Style and Its Algebra of Programs. CACM August 1978

When Moggi had not spoken the Word, nor had Wadler preached it

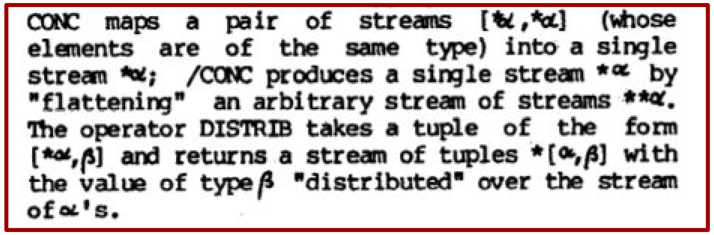


1. Composition. If f and g are such that $f{:}\alpha \rightarrow \beta$ and $g{:}\beta \rightarrow \gamma$ then $f.g{:}\alpha \rightarrow \gamma$.

2. Extension. If $f{:}\alpha \rightarrow \beta$ then $*f$ operates upon a stream of these types; i.e., $*f{:}\ {*\alpha} \rightarrow {*\beta}$.

3. Restriction. If p is a predicate over (i.e., $p{:}\alpha \rightarrow \text{bool}$) then $|p{:}\ {*\alpha} \rightarrow {*\beta}$.

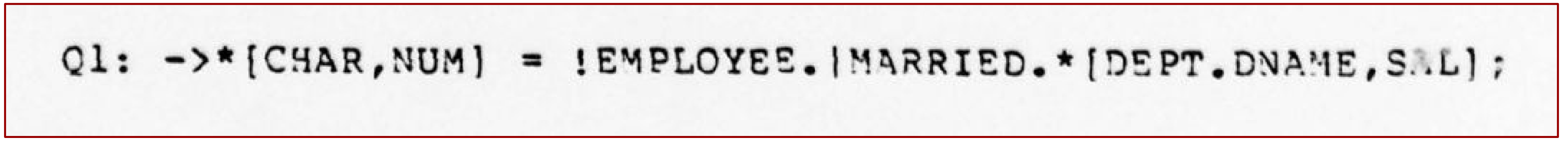
4. Construction. If $f_1{:}\alpha \rightarrow \beta_1$, $f_2{:}\alpha \rightarrow \beta_2 \ldots f_n{:}\alpha \rightarrow \beta_n$ then $[f_1, f_2 \ldots f_n]{:}\alpha \rightarrow [\beta_1, \beta_2 \ldots \beta_n]$.



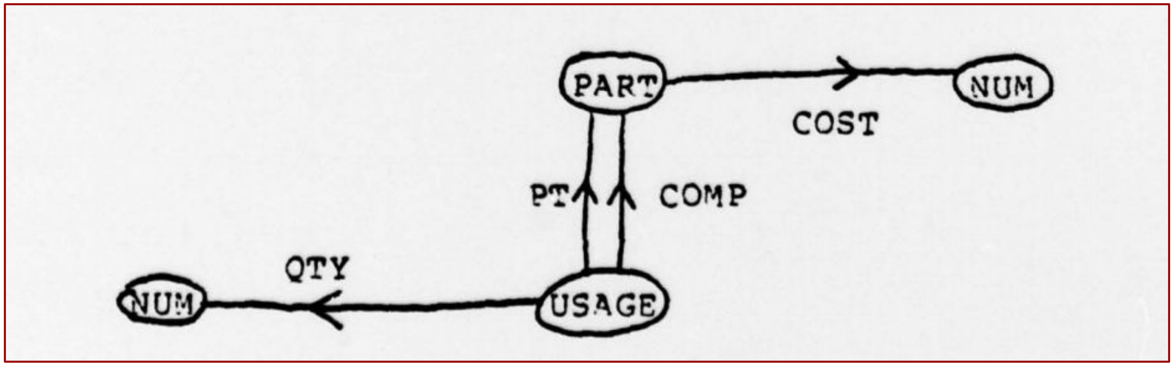
CONC maps a pair of streams $[*\alpha, *\alpha]$ (whose elements are of the same type) into a single stream $*\alpha$; /CONC produces a single stream $*\alpha$ by "flattening" an arbitrary stream of streams $**\alpha$. The operator DISTRIB takes a tuple of the form $[*\alpha, \beta]$ and returns a stream of tuples $*[\alpha, \beta]$ with the value of type β "distributed" over the stream of α's.

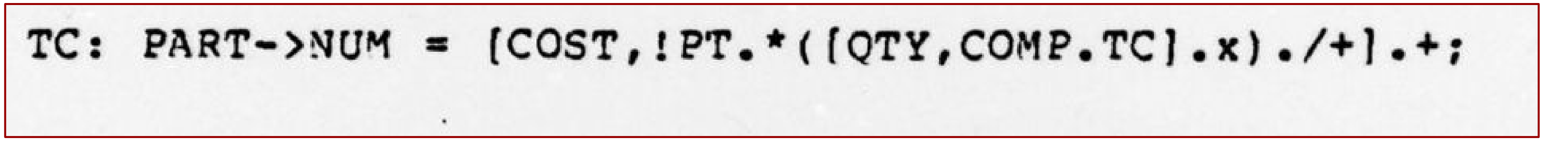
FQL got used by people building interfaces to Codasyl DBs.
Remember that most database queries are written by programs – not people.

[B., R Frankel, Sigmod 1979. *The Functional Data Model* D Shipman, Sigmod 1979]



```
Q1: ->*[CHAR,NUM] = !EMPLOYEE.|MARRIED.*[DEPT.DNAME,SAL];
```





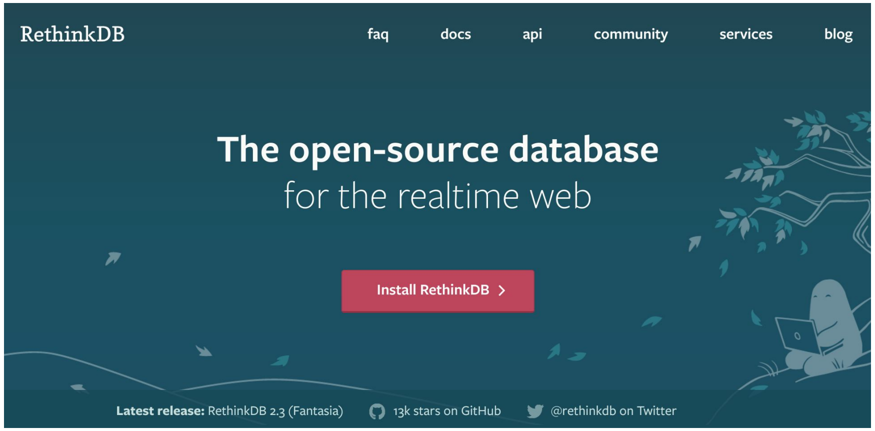
```
TC: PART->NUM = [COST,!PT.*([QTY,COMP.TC].x)./+].+;
```

Several years went by…..

Influences from PL theory and LFCS

- Impedance mismatch problem
- Domain theory and partial information in databases
- ML and record polymorphism
- Structural recursion, monads and nested relational algebra (FQL revisited)
- Partially static type systems for semi-structured data

Quite recently: Rethink DB



The open-source database for the realtime web

Latest release: RethinkDB 2.3 (Fantasia)

“It’s no secret that ReQL, the RethinkDB query language, is modeled after functional languages like Lisp and Haskell. The functional paradigm is particularly well suited to the needs of a distributed database while being more easily embeddable as a DSL than SQL’s ad hoc syntax. Key to functional programming’s power and simplicity is the anonymous (aka lambda) function.”

```
r.table('users').filter(r.row("age").eq(30)).map(r."
name").run();
```

```
charin.fntolist,erase.applist
```

Then I came back to Informatics and joined LFCS

Random thoughts on US vs UK research environment

- US more directed and less forgiving
  - Nothing like the intellectual ferment of the pre-LFCS years
- UK much more supportive of “interdisciplinary research”, but….
  - interdisciplinary research can (like writing programs) be a huge time-waster.
  - You spend most of your time doing boring/marginal stuff, but just occasionally something interesting turns up…
- And sometimes something whacky turns up, like semistructured data, provenance and (about 8 years ago) *Data Citation*

Now data citation is big business

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.”

Oxford University (on behalf of EPSRC) “Describe your data ... to enable other researchers to … cite them”

What is a (conventional) citation?

A collection of “snippets” of information: authors, title, date, etc. and some kind of access mechanism (DOI, URL, ISBN, shelf number etc.)

Not exactly provenance

Self contained, immutable (to within some choice of format)

Needed for a variety of reasons: kudos, currency, authority, recognition, access…

Especially important in curated databases – some kind of mixture of crowd- or expert-sourced data and conventional publication. (IUPHAR – hundreds of contributors, and they want to be acknowledged.)

So what’s the problem

Citations vary with what part of of the database is being cited. And the database changes over time.

There is a huge number of “parts” of a database

Web	URI/CGI
RDB	SQL
XML	XPath/XQuery
RDF	SPARQL
File system	set of paths

We cannot expect to put a citation for each “part” into DBLP. We are going to have to generate citations on the fly.

It gets worse

Start of Datacite 400 line XML schema specification for data citation

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Revision history
    2010-08-26   Complete revision according to new common specification by the metadata work
group after review. AJH, DTIC
                    2010-11-17 Revised to current state of kernel review, FZ, TIB
                    2011-01-17 Complete revsion after community review. FZ, TIB
                    2011-03-17 Release of v2.1: added a namespace; mandatory properties got minLength;
changes in the definitions of relationTypes
                    IsDocumentedBy/Documents and isCompiledBy/Compiles; changes type of property
"Date" from xs:date to xs:string. FZ, TIB
                    2011-06-27 v2.2: namespace: kernel-2.2, additions to controlled lists "resourceType",
"contributorType", "relatedIdentifierType", and "descriptionType". Removal of intermediate include-
files.
    2013-05 v3.0: namespace: kernel-3.0; delete LastMetadataUpdate & MetadateVersionNumber;
additions to controlled lists "contributorType", "dateType", "descriptionType", "relationType",
"relatedIdentifierType" & "resourceType"; deletion of "StartDate" & "EndDate" from list "dateType" and
"Film" from "resourceType";  allow arbitrary order of elements; allow optional wrapper elements to be
empty; include xml:lang attribute for title, subject & description; include attribute schemeURI for
nameIdentifier of creator, contributor & subject; added new attributes "relatedMetadataScheme",
"schemeURI" & "schemeType" to relatedIdentifier; included new property "geoLocation"
                    2014-08-20 v3.1: additions to controlled lists "relationType", contributorType" and
"relatedIdentifierType"; introduction of new child element "affiliation" to "creator" and "contributor"-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns="http://datacite.
org/schema/kernel-3" targetNamespace="http://datacite.org/schema/kernel-3" elementFormDefault="
qualified" xml:lang="EN">
                   <xs:import namespace="http://www.w3.org/XML/1998/namespace" schemaLocation="
http://www.w3.org/2009/01/xml.xsd"/>
                   <xs:include schemaLocation="include/datacite-titleType-v3.xsd"/>
                   <xs:include schemaLocation="include/datacite-contributorType-v3.1.xsd"/>
                   <xs:include schemaLocation="include/datacite-dateType-v3.xsd"/>
                   <xs:include schemaLocation="include/datacite-resourceType-v3.xsd"/>
                   <xs:include schemaLocation="include/datacite-relationType-v3.1.xsd"/>
                   <xs:include schemaLocation="include/datacite-relatedIdentifierType-v3.1.xsd"/>
                   <xs:include schemaLocation="include/datacite-descriptionType-v3.xsd"/>
                   <xs:element name="resource">
```

Start of a 700 line machine-generated SQL component of some OLAP API

```
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
```

Another principle/recommendation

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.

NB. The citation depends on *both* Q and D

The database problem

Looks hard because any analysis of a query is likely to be hard, if not undecidable, but there’s hope.

Key concept is that of a database *view* – a function that when applied to a database in one schema produces a database in another schema (and model)

It is common for authors/publishers to supply citations for some parts of the database. These can be expressed as views $V_1 \ldots V_{n.}$.

So given a query *Q*, a database *D* and a schema *S*, can *Q* be factored through a view. That is, is there a Q_i such that

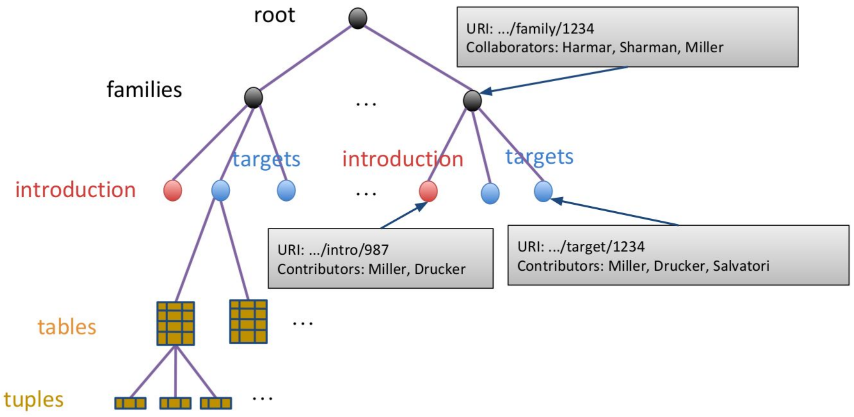
$$\forall D \in S.\ Q(S) = Q_i(V_i(D))$$

If so, the citation for V_i is the citation for Q.

This is a well-known database problem that comes from optimization. In fact our problem is a bit more subtle because the citation also depends on D, and we have to introduce the notion of a *parameterized* view. But the known machinery can be adapted.

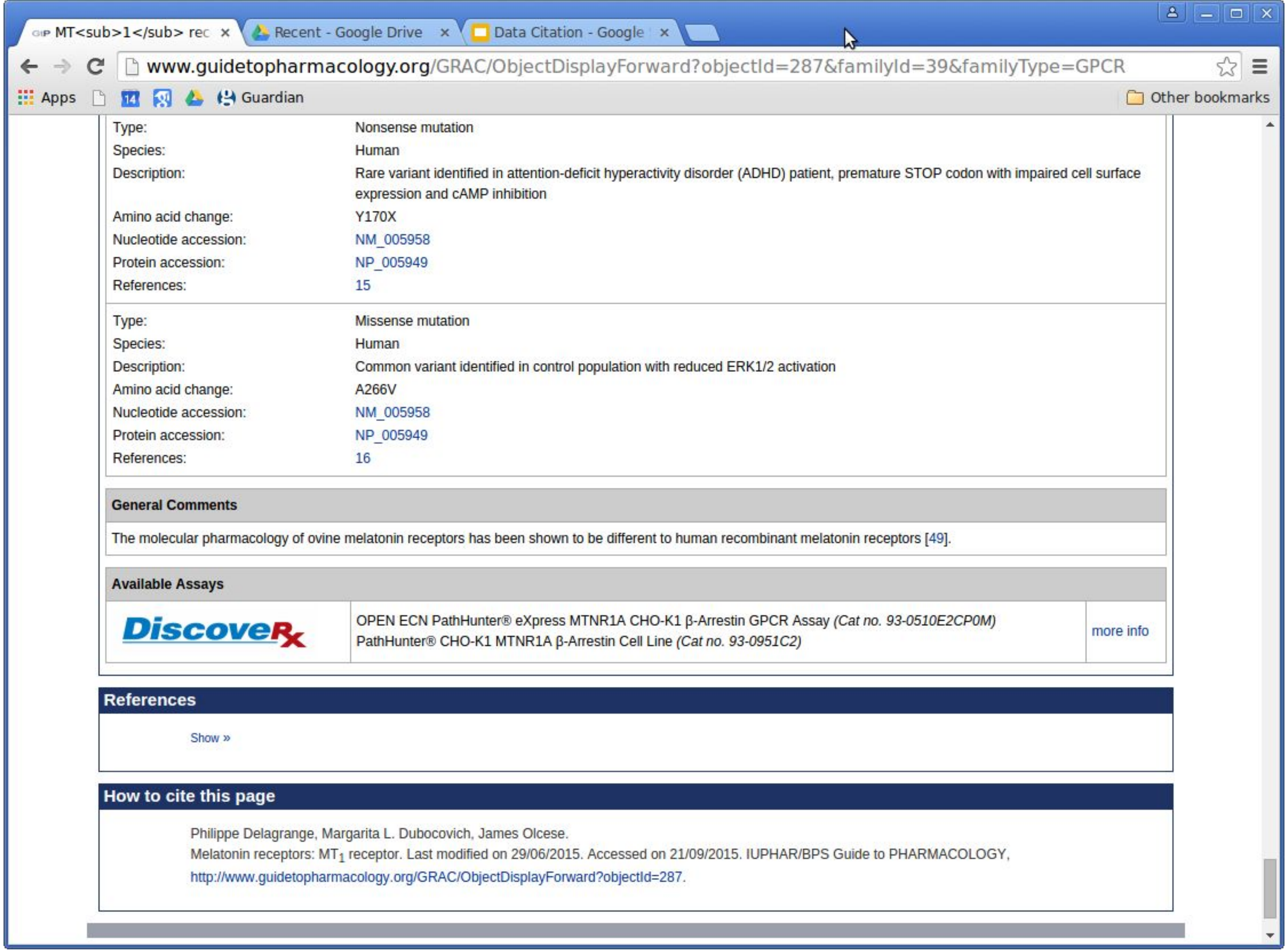
Hierarchical data (files, XPath, some URLs)

A simple pattern-matching language for generating citations in a hierarchy



{ DB: IUPHAR, Version: $v, Family: $$f, Contributors: $a,
URI: ”www.iuphar.org”, DOI: 10.3.14159}
←
/Root[VersionNumber: $v]/Family[FamilyName: $$f]
/Introduction[Contributor-list: $a]

{ DB: IUPHAR, Version: 26, Family: ”Calcitonin”, Contributors: [”Debbie Hay”, ”David R. Poyner”], URI: ”www.iuphar.org”, DOI: 10.3.14159}



MT₁</sub> rec
Recent - Google Drive
Data Citation - Google
www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=287&familyId=39&familyType=GPCR
Apps
Guardian
Other bookmarks

Type:	Nonsense mutation
Species:	Human
Description:	Rare variant identified in attention-deficit hyperactivity disorder (ADHD) patient, premature STOP codon with impaired cell surface expression and cAMP inhibition
Amino acid change:	Y170X
Nucleotide accession:	NM_005958
Protein accession:	NP_005949
References:	15

Type:	Missense mutation
Species:	Human
Description:	Common variant identified in control population with reduced ERK1/2 activation
Amino acid change:	A266V
Nucleotide accession:	NM_005958
Protein accession:	NP_005949
References:	16

General Comments

The molecular pharmacology of ovine melatonin receptors has been shown to be different to human recombinant melatonin receptors [49].

Available Assays

DiscoveRx	OPEN ECN PathHunter® eXpress MTNR1A CHO-K1 β-Arrestin GPCR Assay *(Cat no. 93-0510E2CP0M)* PathHunter® CHO-K1 MTNR1A β-Arrestin Cell Line *(Cat no. 93-0951C2)*	more info

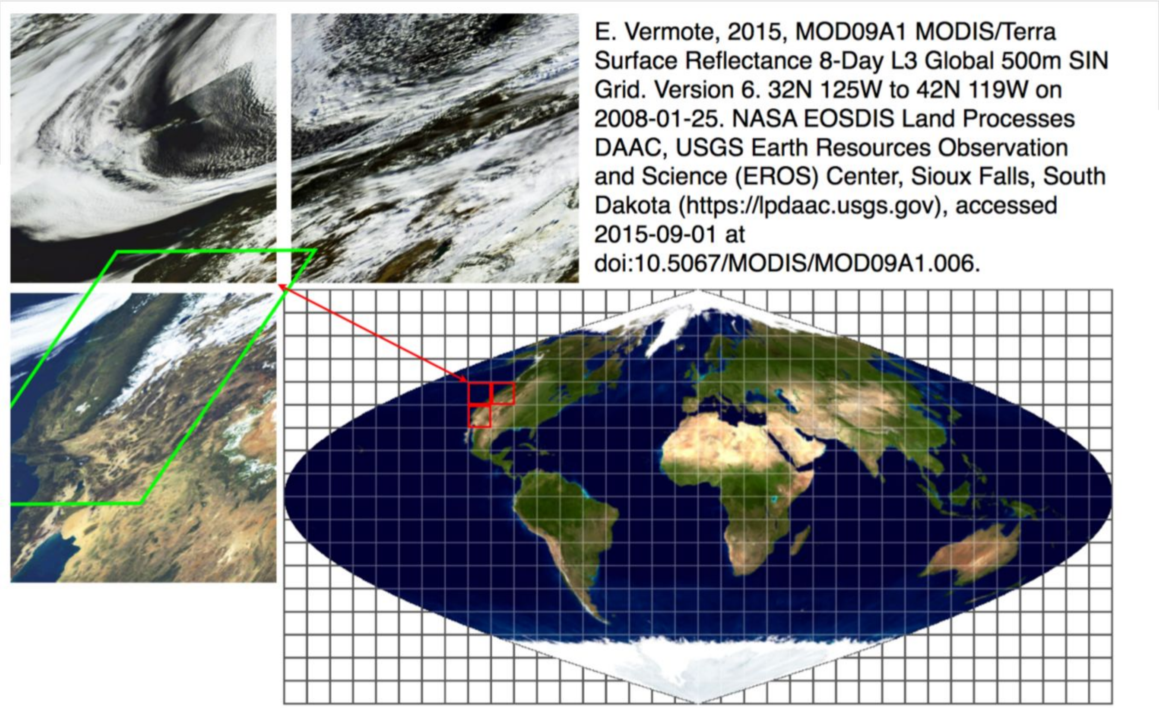
References

Show »

How to cite this page

Philippe Delagrange, Margarita L. Dubocovich, James Olcese.
Melatonin receptors: MT$_1$ receptor. Last modified on 29/06/2015. Accessed on 21/09/2015. IUPHAR/BPS Guide to PHARMACOLOGY,
http://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=287.

Also works for other kinds of data



E. Vermote, 2015, MOD09A1 MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid. Version 6. 32N 125W to 42N 119W on 2008-01-25. NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (https://lpdaac.usgs.gov), accessed 2015-09-01 at doi:10.5067/MODIS/MOD09A1.006.

The views are spatio-temporal bounding boxes

B., Davidson & Frew *Why Data CItation is a Computational Problem*. CACM. to appear

Thank You, LFCS!

Some of my favorite citations:

BL Cotton Nero A. X

Cotton Otho A. XII

Ann. Phys., Lpz 18 639-641

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 2011

Aad, G. *et al*. (ATLAS Collaboration, CMS Collaboration) *Phys. Rev. Lett.* **114**, 191803 (2015).