Annotation and Citation

Peter Buneman
University of Edinburgh

What is the connection?

- Annotation - adding information to existing data
  - How is annotation different from any other data
  - How is it “attached” to data?
  - How does it propagate through queries?
- Citation – a form of annotation, but
  - Traditionally applied to papers/books etc., not general data
  - Not “attached” to data?
- But we want to apply citation to data

Annotation in Uniprot

```
ID   11SB_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;
```

```
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).
```

```
MURA I.;
```

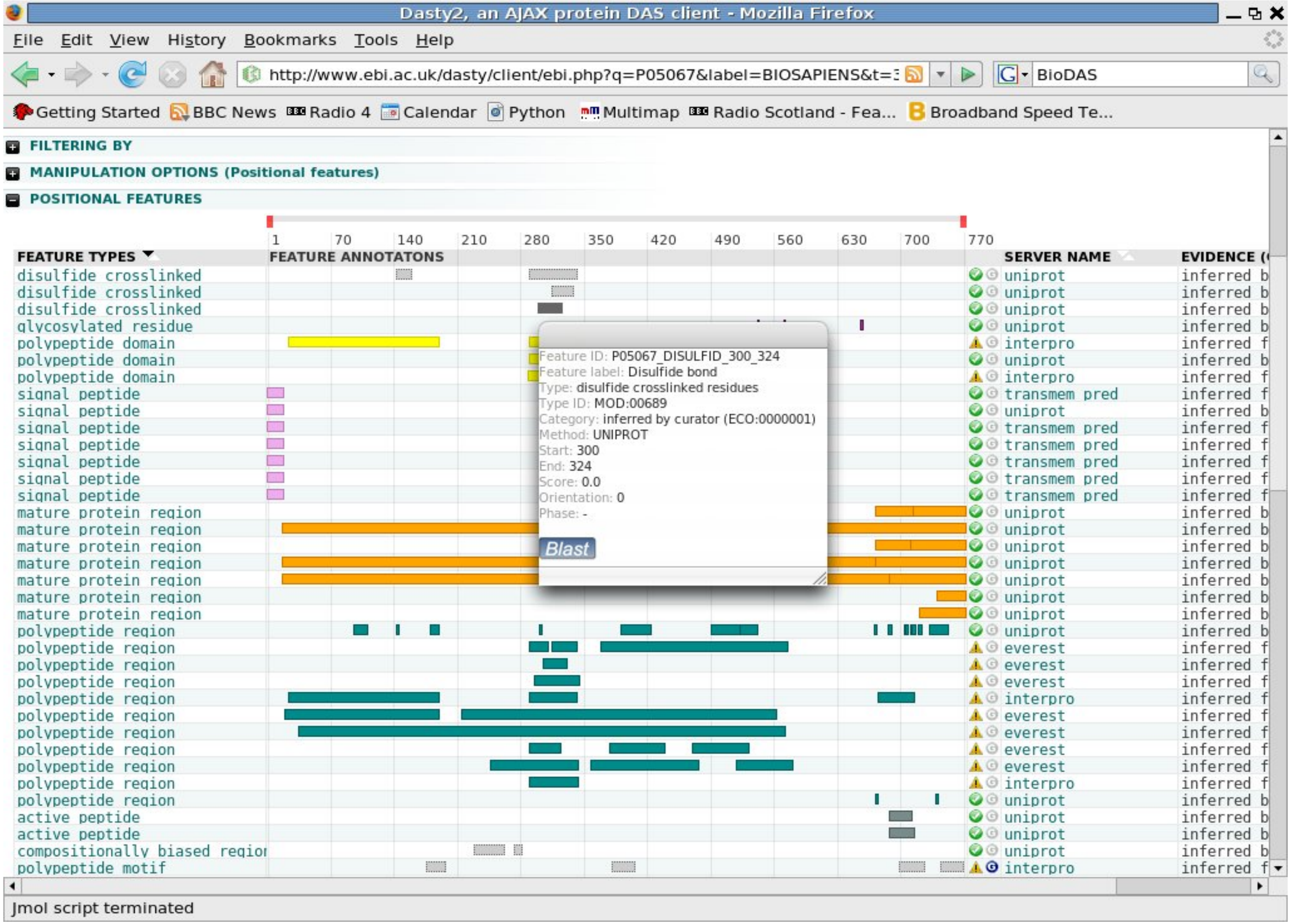
```
CC   -!- SUBUNIT: HEXAMER; EACH SUBUNIT IS COMPOSED OF AN ACIDIC AND A
CC       BASIC CHAIN DERIVED FROM A SINGLE PRECURSOR AND LINKED BY A
CC       DISULFIDE BOND.
```

```
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   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 FQGSEVWQQH RYQSPRACRL ENLRAQDPVR
     RAEAEAIFTE VWDQDNDEFQ CAGVNMIRHT IRPKGLLLPG FSNAPKLIFV AQGFGIRGIA
     IPGCAETYQT DLRRSQSAGS AFKDQHQKIR PFREGDLLVV PAGVSHWMYN RGQSDLVLIV
     FADTRNVANQ IDPYLRKFYL AGRPEQVERG VEEWERSSRK GSSGEKSGNI FSGFADEFLE
     EAFQIDGGLV RKLKGEDDER DRIVQVDEDF EVLLPEKDEE ERSRGRYIES ESESENGLEE
     TICTLRLKQN IGRSVRADVF NPRGGRISTA NYHTLPILRQ VRLSAERGVL YSNAMVAPHY
     TVNSHSVMYA TRGNARVQVV DNFGQSVFDG EVREGQVLMI PQNFVVIKRA SDRGFEWIAF
     KTNDNAITNL LAGRVSQMRM LPLGVLSNMY RISREEAQRL KYGQQEMRVL SPGRSQGRRE
//
```

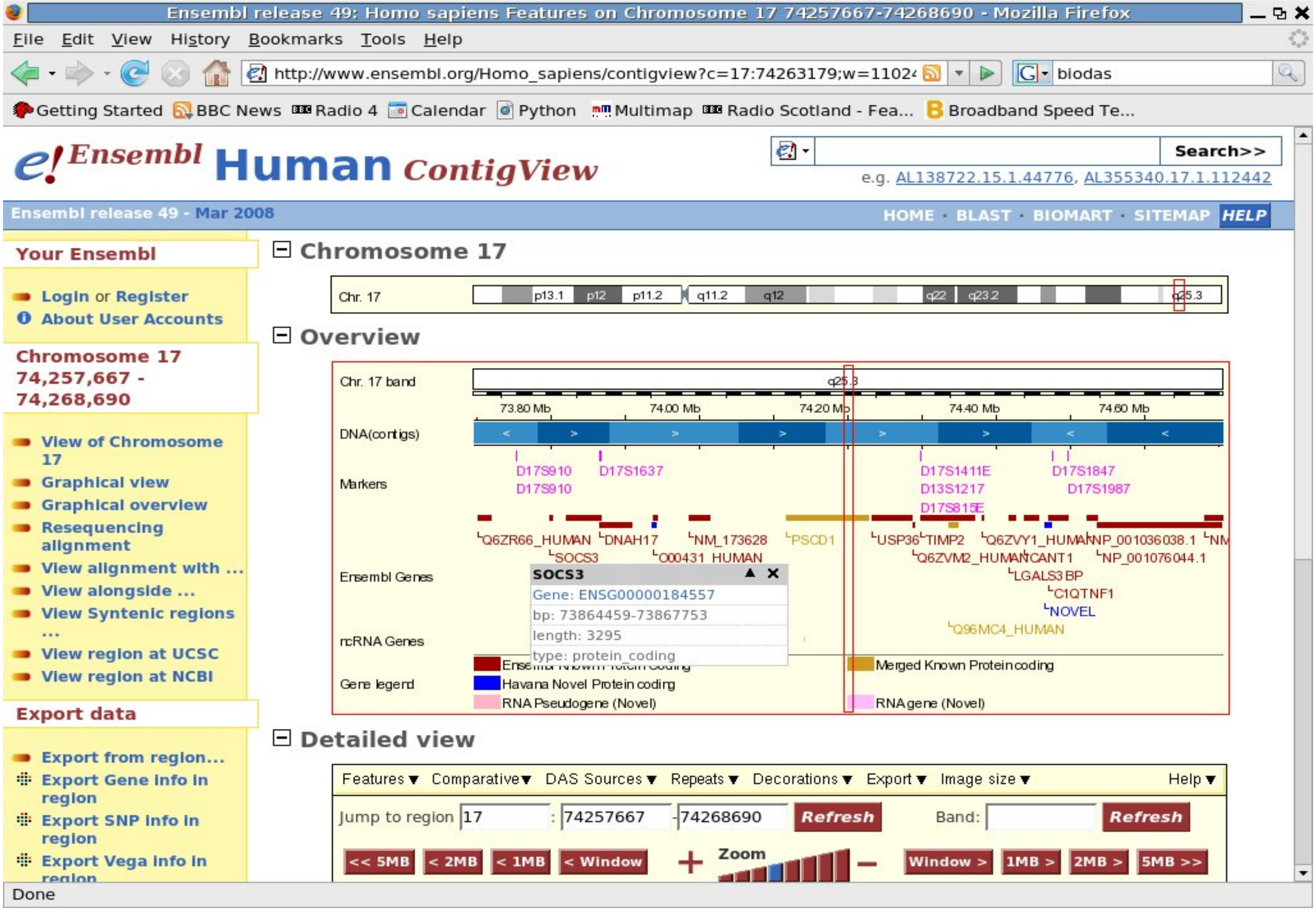
The Distributed Annotation Server (DAS)


Dasty2, an AJAX protein DAS client - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ebi.ac.uk/dasty/client/ebi.php?q=P05067&label=BIOSAPIENS&t=3
BioDAS
Getting Started BBC News Radio 4 Calendar Python Multimap Radio Scotland - Fea... Broadband Speed Te...
FILTERING BY
MANIPULATION OPTIONS (Positional features)
POSITIONAL FEATURES
1
70
140
210
280
350
420
490
560
630
700
770
Blast

FEATURE TYPES ▼	FEATURE ANNOTATONS	SERVER NAME	EVIDENCE (
disulfide crosslinked		✓ uniprot	inferred b
disulfide crosslinked		✓ uniprot	inferred b
disulfide crosslinked		✓ uniprot	inferred b
glycosylated residue		✓ uniprot	inferred b
polypeptide domain		⚠ interpro	inferred f
polypeptide domain		✓ uniprot	inferred b
polypeptide domain		⚠ interpro	inferred f
signal peptide		✓ transmem pred	inferred f
signal peptide		✓ uniprot	inferred b
signal peptide		✓ transmem pred	inferred f
signal peptide		✓ transmem pred	inferred f
signal peptide		✓ transmem pred	inferred f
signal peptide		✓ transmem pred	inferred f
signal peptide		✓ transmem pred	inferred f
mature protein region		✓ uniprot	inferred b
mature protein region		✓ uniprot	inferred b
mature protein region		✓ uniprot	inferred b
mature protein region		✓ uniprot	inferred b
mature protein region		✓ uniprot	inferred b
mature protein region		✓ uniprot	inferred b
mature protein region		✓ uniprot	inferred b
polypeptide region		✓ uniprot	inferred b
polypeptide region		⚠ everest	inferred f
polypeptide region		⚠ everest	inferred f
polypeptide region		⚠ everest	inferred f
polypeptide region		⚠ interpro	inferred f
polypeptide region		⚠ everest	inferred f
polypeptide region		⚠ everest	inferred f
polypeptide region		⚠ everest	inferred f
polypeptide region		⚠ everest	inferred f
polypeptide region		⚠ interpro	inferred f
polypeptide region		✓ uniprot	inferred b
active peptide		✓ uniprot	inferred b
active peptide		✓ uniprot	inferred b
compositionally biased regior		✓ uniprot	inferred b
polypeptide motif		⚠ interpro	inferred f

Feature ID: P05067_DISULFID_300_324
Feature label: Disulfide bond
Type: disulfide crosslinked residues
Type ID: MOD:00689
Category: inferred by curator (ECO:0000001)
Method: UNIPROT
Start: 300
End: 324
Score: 0.0
Orientation: 0
Phase: -

Jmol script terminated



Ensembl release 49: Homo sapiens Features on Chromosome 17 74257667-74268690 - Mozilla Firefox

http://www.ensembl.org/Homo_sapiens/contigview?c=17:74263179;w=1102

e! Ensembl Human ContigView

e.g. AL138722.15.1.44776, AL355340.17.1.112442

Ensembl release 49 - Mar 2008

Chromosome 17 74,257,667 - 74,268,690

Chromosome 17

Overview

SOCS3

Gene: ENSG00000184557

bp: 73864459-73867753

length: 3295

type: protein_coding

Detailed view

Jump to region 17 : 74257667 - 74268690

Done

Numerous attempts to define generic annotation systems:

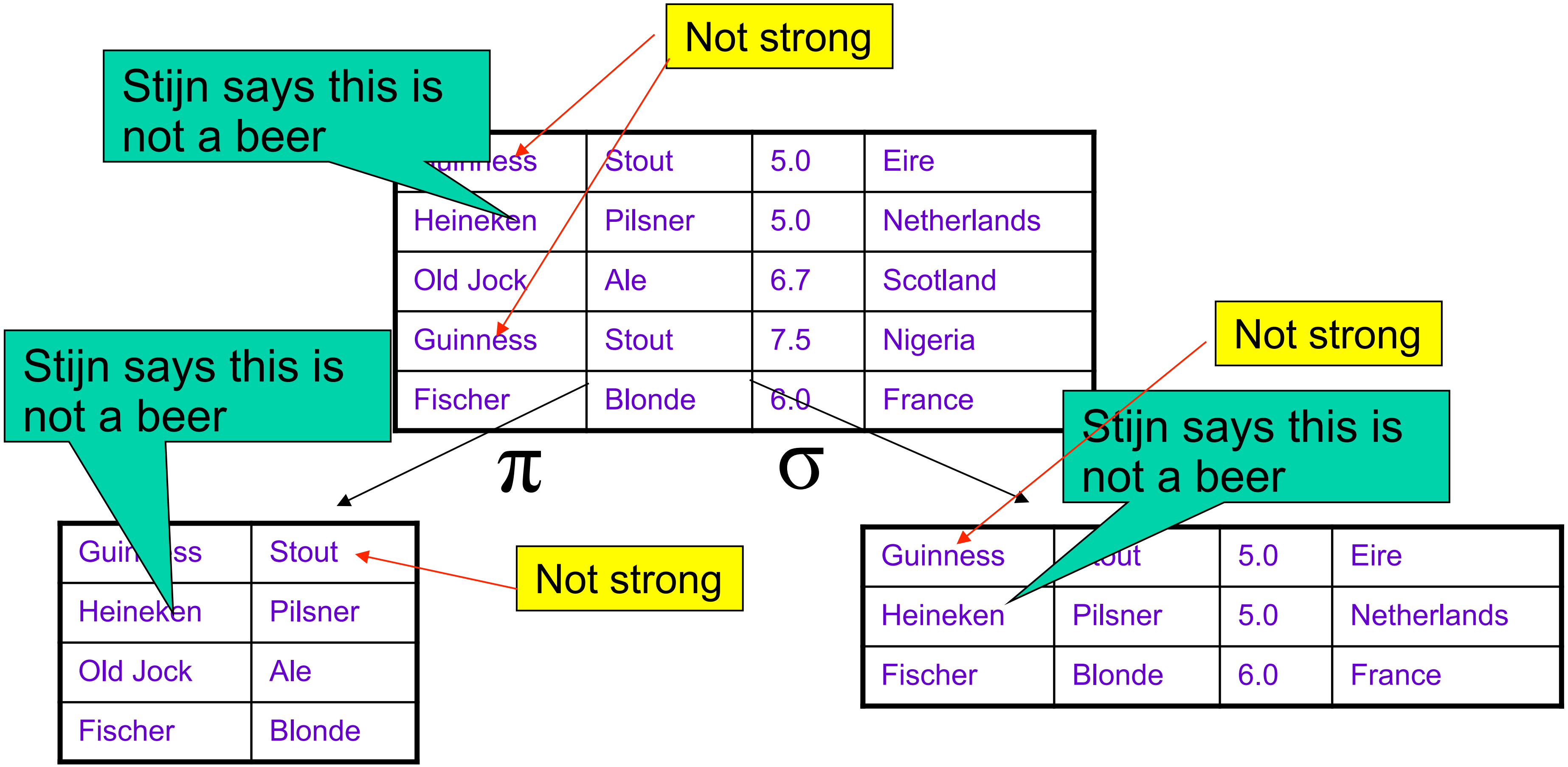
- Third voice (circa 1999) Web page annotation
- Annotea (2001) ditto
- DBNotes (Bhagwat *et al* 2005) Relational DB annotation
- Superimposed Information Systems (Murthy *et al* 2005) Documents and images
- Mondrian (Geerts et al 2007) More sophisticated RDB annotation
- DBWiki (B. *et al* 2011) Generic curated DB management

Highly successful annotation systems for specific structures:

- BioDAS
- Google Maps
- Other DAS's, e.g. AstroDAS

And isn't RDF about annotating the Web?

Annotating Databases



Guinness	Stout	5.0	Eire
Heineken	Pilsner	5.0	Netherlands
Old Jock	Ale	6.7	Scotland
Guinness	Stout	7.5	Nigeria
Fischer	Blonde	6.0	France

π σ

Guinness	Stout
Heineken	Pilsner
Old Jock	Ale
Fischer	Blonde

Guinness	Stout	5.0	Eire
Heineken	Pilsner	5.0	Netherlands
Fischer	Blonde	6.0	France

Not strong

Stijn says this is not a beer

Not strong

Stijn says this is not a beer

Not strong

Stijn says this is not a beer

Annotation propagation

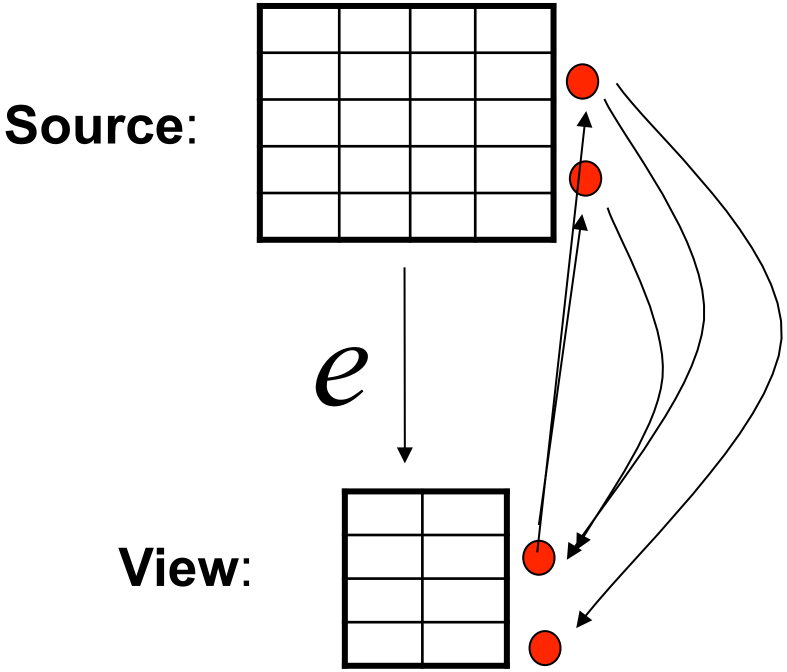
"Obvious" rules, e.g.:

π (t) is annotated in π (R) iff t is annotated in R

if $t \in \sigma$ (R) then t is annotated in R iff t is annotated in σ (R)

etc

Given a view annotation what source annotation causes least "spread"?
Is there a source annotation that causes no spread?



Results on annotation propagation

Suppose we have an annotation on a view. A source annotation is side-effect free if it causes exactly the view annotation to appear when propagated forward.

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

Similar results for minimising the “spread” of an annotation. [B., Khanna & Tan, PODS 2001]

View deletion problems are related

Side effect-free view deletion: given a tuple *t* in *Q*(*S*), find a subset *T* of of *S* whose removal causes precisely t to disappear ($\{t\} = Q(S) - Q(S - T)$). NP hard for

PJ queries (fixed query)

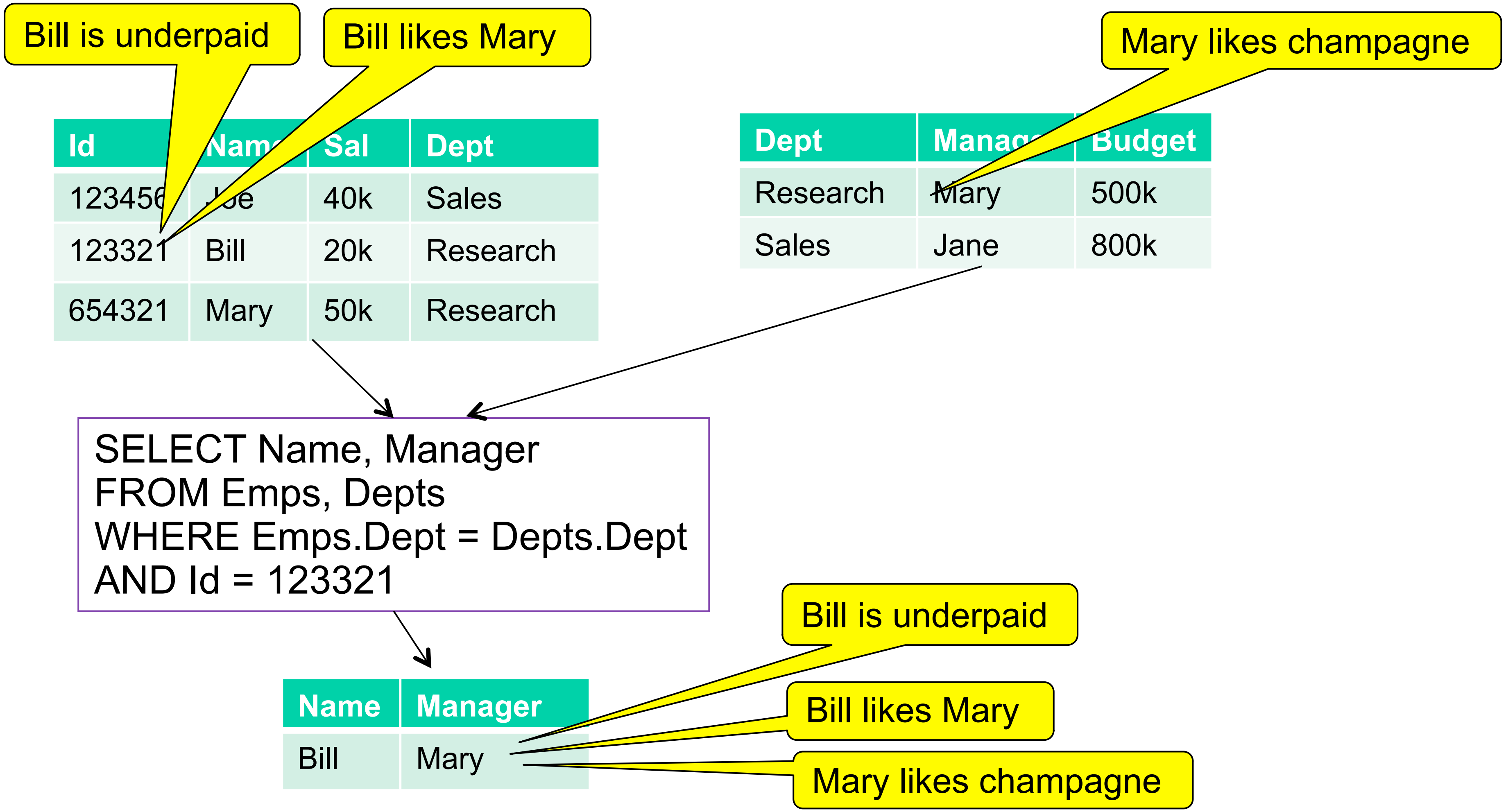
JU queries (not fixed)

All other cases have polynomial-time solutions.

“Key-preserving” transformations simplify annotation propagation, but the story for view deletion is mixed [Gao, Fan, Geerts, CIKM ’06]

More on annotation propagation

Annotating with comments



Bill is underpaid

Bill likes Mary

Id	Name	Sal	Dept
123456	Joe	40k	Sales
123321	Bill	20k	Research
654321	Mary	50k	Research

Mary likes champagne

Dept	Manager	Budget
Research	Mary	500k
Sales	Jane	800k

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

Name	Manager
Bill	Mary

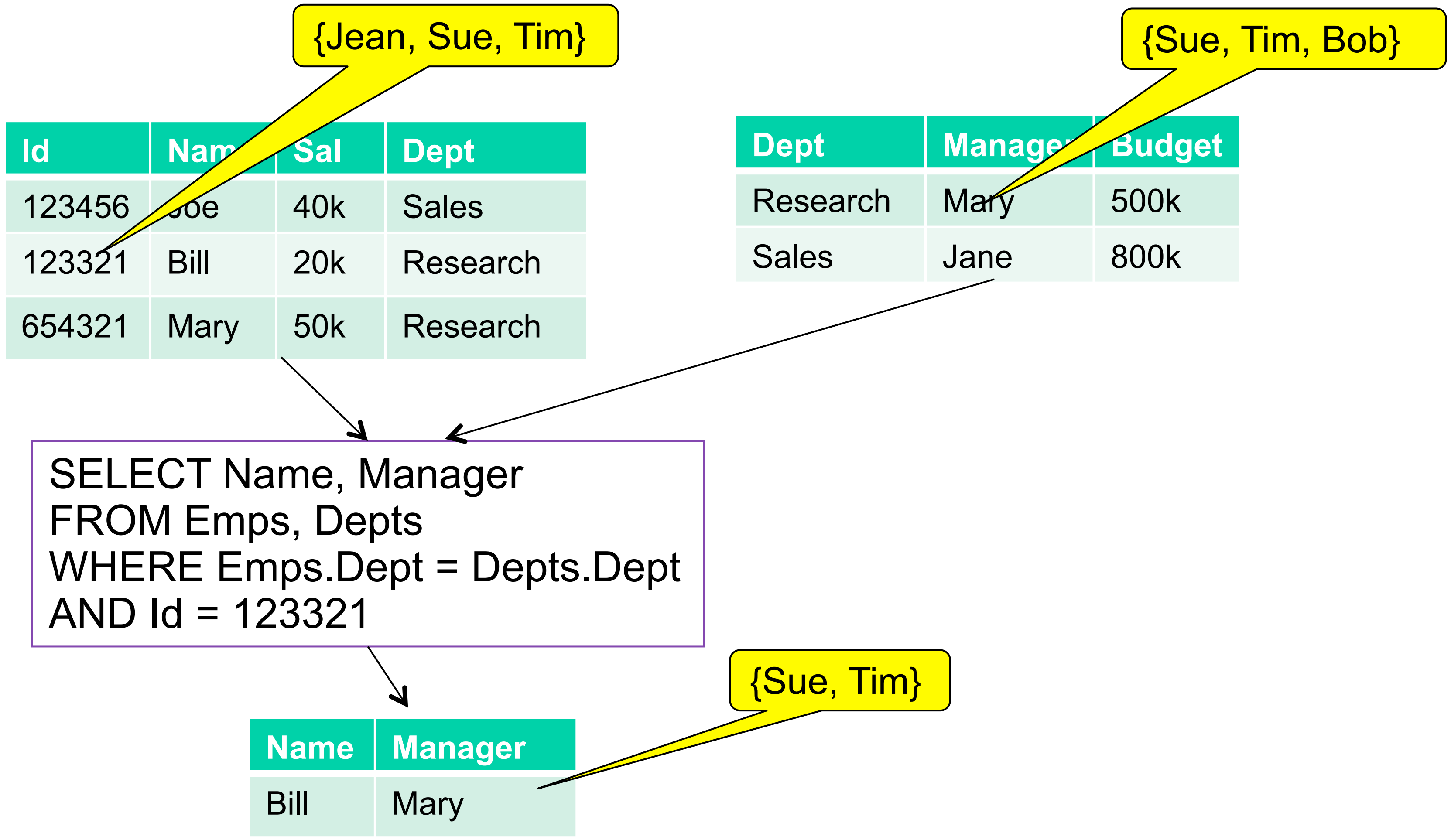
Bill is underpaid

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



{Jean, Sue, Tim}

Id	Name	Sal	Dept
123456	Joe	40k	Sales
123321	Bill	20k	Research
654321	Mary	50k	Research

{Sue, Tim, Bob}

Dept	Manager	Budget
Research	Mary	500k
Sales	Jane	800k

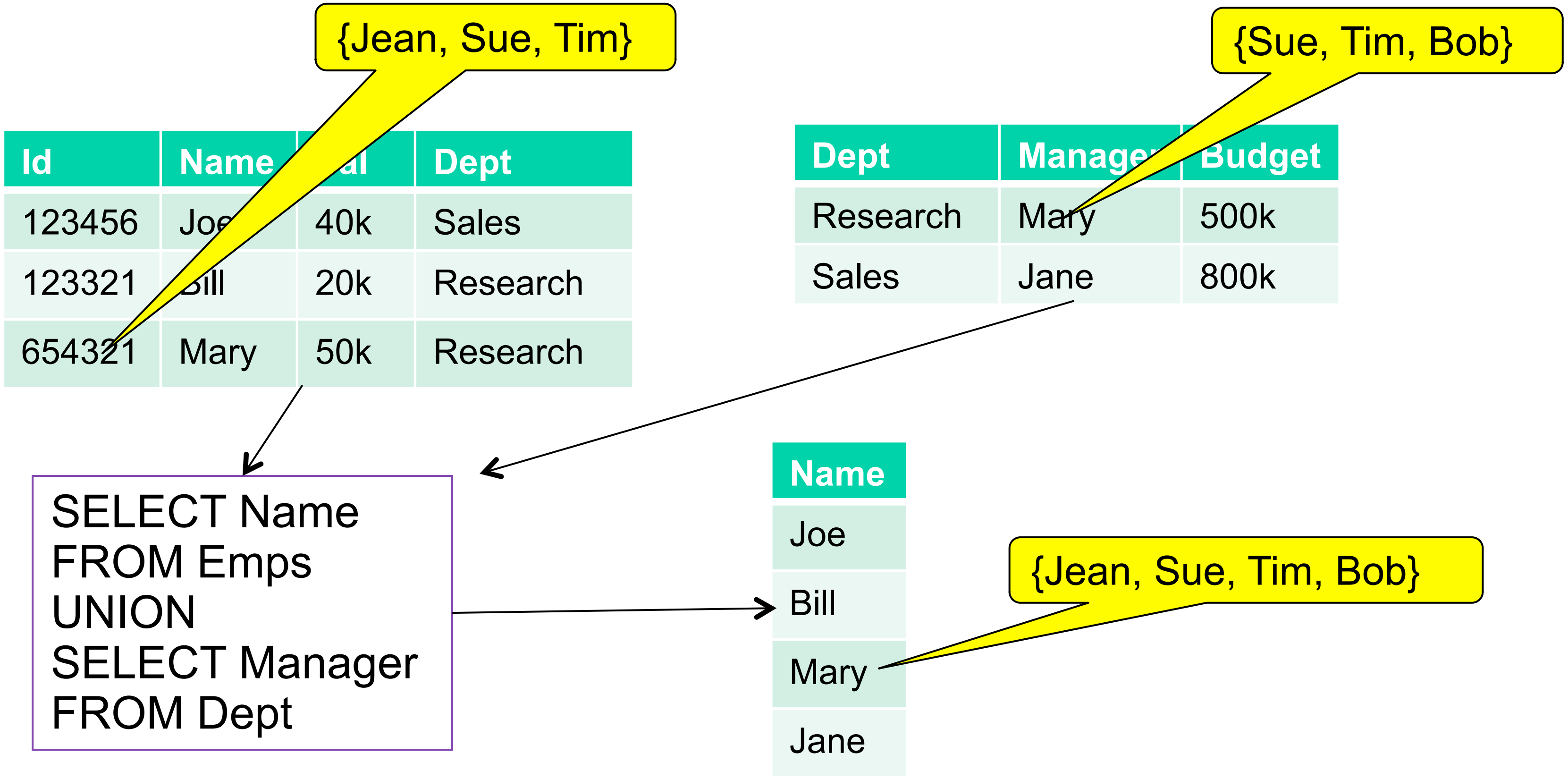
```
SELECT Name, Manager
FROM Emps, Depts
WHERE Emps.Dept = Depts.Dept
AND Id = 123321
```

{Sue, Tim}

Name	Manager
Bill	Mary

We want the *intersection* of the believers of the input tuple

Annotating with beliefs for another query:



{Jean, Sue, Tim}

Id	Name	Sal	Dept
123456	Joe	40k	Sales
123321	Bill	20k	Research
654321	Mary	50k	Research

{Sue, Tim, Bob}

Dept	Manager	Budget
Research	Mary	500k
Sales	Jane	800k

```
SELECT Name
FROM Emps
UNION
SELECT Manager
FROM Dept
```

Name
Joe
Bill
Mary
Jane

{Jean, Sue, Tim, Bob}

For UNION queries we want the *union* of the believers of the input tuples

Provenance/Annotation 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, +, · 0, 1) is a commutative semiring.
- Provenance is a polynomial over the abstract quantities *p,q,r,…*
- Comment semiring (STR, ∪, ∪, {}, {}) STR = set of strings
- Belief semiring (B, ∪, ∩, {}, B) B= 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 (**Z**,+,x,0,1)
  - Multiplicity of (d, e) in V is $r + (r \times r) + (r \times s)$

Two kinds of annotation?

(A) Annotations that should be part of the data

Also called "Eilann a' Cheò"

Coords: 57.307N 6.23W

Eng. Name	Gaelic Name	Type	Pronunciation	. . .
Skye	An t-Eilann Sgitheanach	Island	<123.wav>	. . .

A problem for schema evolution?

(B) Annotations that are "higher order"

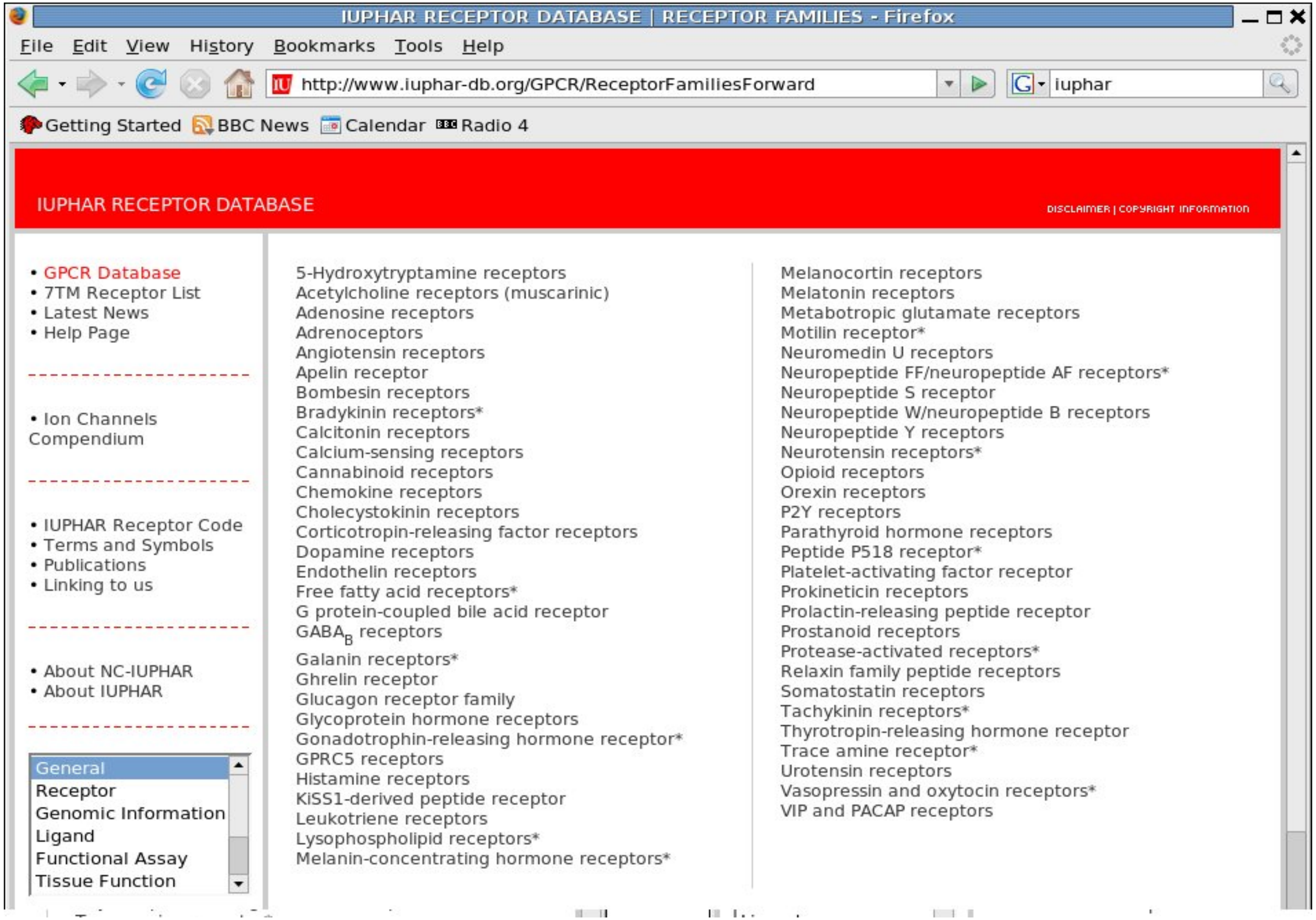
- "Jane believes this"
- "Created at time t"

How do we distinguish (A) and (B)?

Annotation and RDF

- Type (A) annotation presents no problems (just add new triples according to TBL)
- Type (B) is a real problem. How do we refer to a triple?
  - Reify?
  - Define the annotation target by a query?
  - Named graph?
- We'd like to reason about type B annotations *using RDF and some ontology language*:
  - If A trusts B and B believes T then A believes T
- Recent work by E. Kostylev and B. on annotation "semirings" for RDF and on combined annotations.

The IUPHAR database – an example of “brain-sourcing”


IUPHAR RECEPTOR DATABASE | RECEPTOR FAMILIES - Firefox
File Edit View History Bookmarks Tools Help
http://www.iuphar-db.org/GPCR/ReceptorFamiliesForward
iuphar
Getting Started
BBC News
Calendar
Radio 4
DISCLAIMER | COPYRIGHT INFORMATION

IUPHAR RECEPTOR DATABASE

- GPCR Database
- 7TM Receptor List
- Latest News
- Help Page

- Ion Channels Compendium

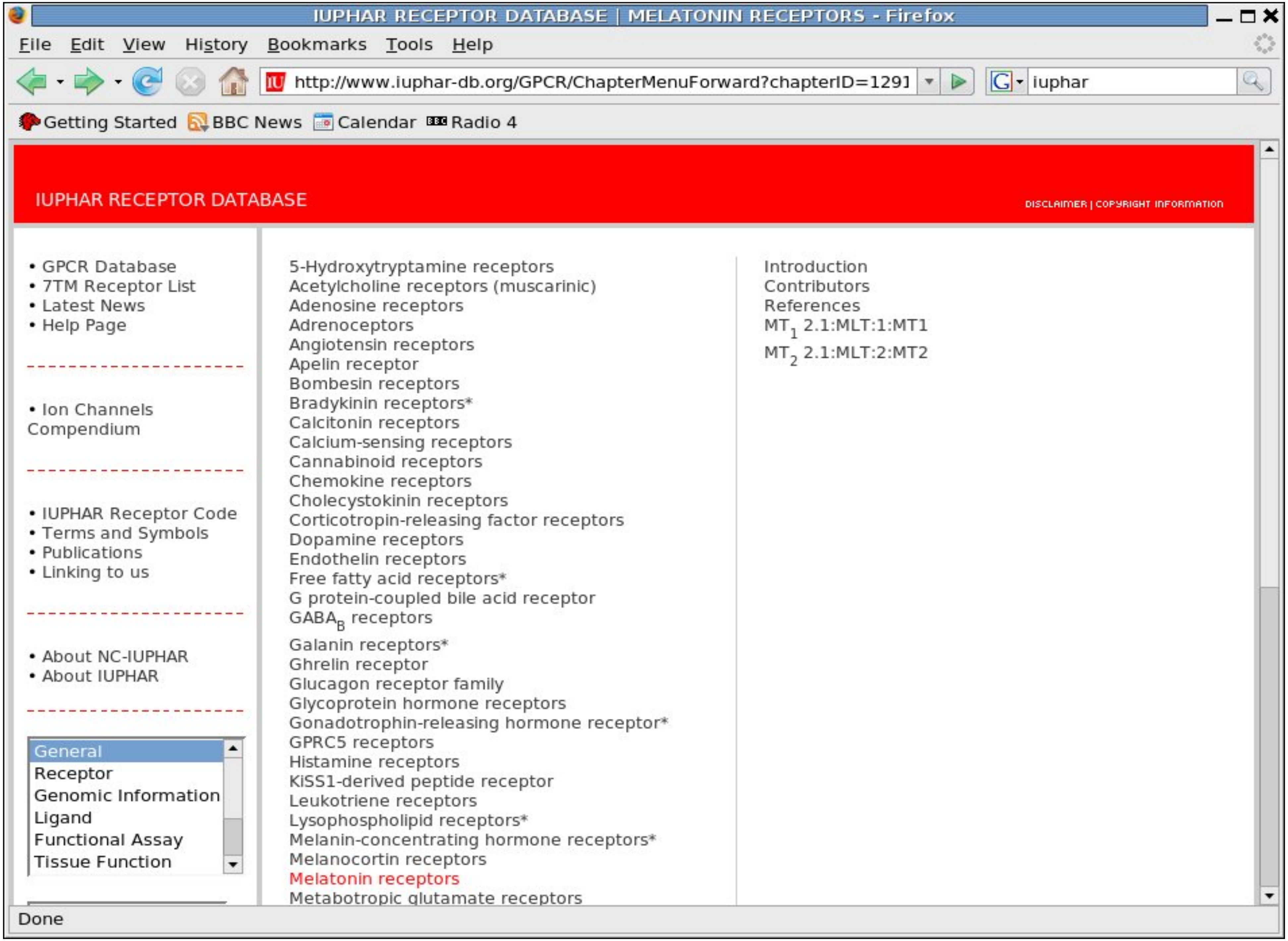
- IUPHAR Receptor Code
- Terms and Symbols
- Publications
- Linking to us

- About NC-IUPHAR
- About IUPHAR

General
Receptor
Genomic Information
Ligand
Functional Assay
Tissue Function

5-Hydroxytryptamine receptors
Acetylcholine receptors (muscarinic)
Adenosine receptors
Adrenoceptors
Angiotensin receptors
Apelin receptor
Bombesin receptors
Bradykinin receptors*
Calcitonin receptors
Calcium-sensing receptors
Cannabinoid receptors
Chemokine receptors
Cholecystokinin receptors
Corticotropin-releasing factor receptors
Dopamine receptors
Endothelin receptors
Free fatty acid receptors*
G protein-coupled bile acid receptor
$GABA_B$ receptors
Galanin receptors*
Ghrelin receptor
Glucagon receptor family
Glycoprotein hormone receptors
Gonadotrophin-releasing hormone receptor*
GPRC5 receptors
Histamine receptors
KiSS1-derived peptide receptor
Leukotriene receptors
Lysophospholipid receptors*
Melanin-concentrating hormone receptors*

Melanocortin receptors
Melatonin receptors
Metabotropic glutamate receptors
Motilin receptor*
Neuromedin U receptors
Neuropeptide FF/neuropeptide AF receptors*
Neuropeptide S receptor
Neuropeptide W/neuropeptide B receptors
Neuropeptide Y receptors
Neurotensin receptors*
Opioid receptors
Orexin receptors
P2Y receptors
Parathyroid hormone receptors
Peptide P518 receptor*
Platelet-activating factor receptor
Prokineticin receptors
Prolactin-releasing peptide receptor
Prostanoid receptors
Protease-activated receptors*
Relaxin family peptide receptors
Somatostatin receptors
Tachykinin receptors*
Thyrotropin-releasing hormone receptor
Trace amine receptor*
Urotensin receptors
Vasopressin and oxytocin receptors*
VIP and PACAP receptors


IUPHAR RECEPTOR DATABASE | MELATONIN RECEPTORS - Firefox
File Edit View History Bookmarks Tools Help
http://www.iuphar-db.org/GPCR/ChapterMenuForward?chapterID=1291
iuphar
Getting Started BBC News Calendar Radio 4
IUPHAR RECEPTOR DATABASE
DISCLAIMER | COPYRIGHT INFORMATION
Done

- GPCR Database
- 7TM Receptor List
- Latest News
- Help Page

- Ion Channels Compendium

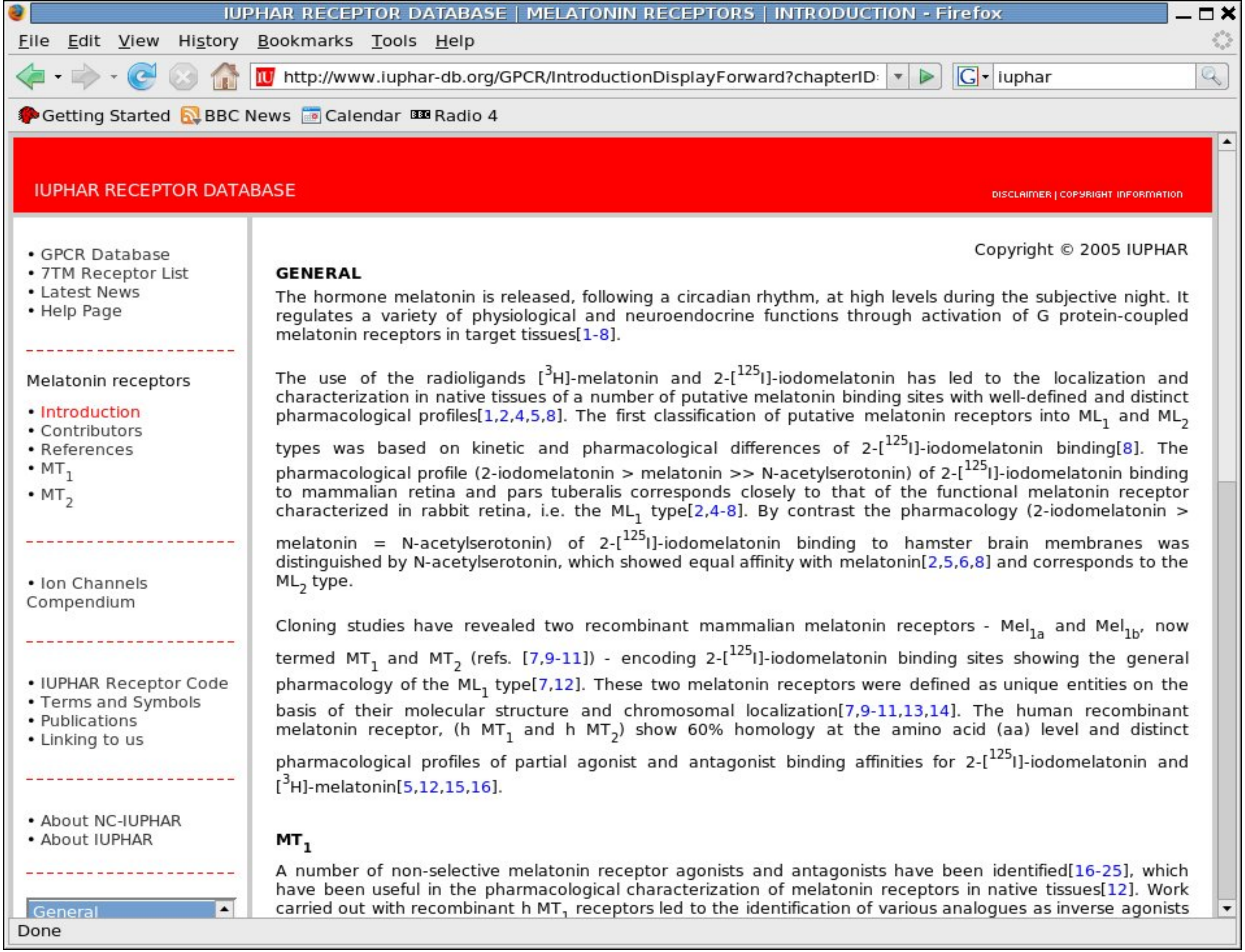
- IUPHAR Receptor Code
- Terms and Symbols
- Publications
- Linking to us

- About NC-IUPHAR
- About IUPHAR

General
Receptor
Genomic Information
Ligand
Functional Assay
Tissue Function

5-Hydroxytryptamine receptors
Acetylcholine receptors (muscarinic)
Adenosine receptors
Adrenoceptors
Angiotensin receptors
Apelin receptor
Bombesin receptors
Bradykinin receptors*
Calcitonin receptors
Calcium-sensing receptors
Cannabinoid receptors
Chemokine receptors
Cholecystokinin receptors
Corticotropin-releasing factor receptors
Dopamine receptors
Endothelin receptors
Free fatty acid receptors*
G protein-coupled bile acid receptor
$GABA_B$ receptors
Galanin receptors*
Ghrelin receptor
Glucagon receptor family
Glycoprotein hormone receptors
Gonadotrophin-releasing hormone receptor*
GPRC5 receptors
Histamine receptors
KiSS1-derived peptide receptor
Leukotriene receptors
Lysophospholipid receptors*
Melanin-concentrating hormone receptors*
Melanocortin receptors
Melatonin receptors
Metabotropic glutamate receptors

Introduction
Contributors
References
MT_1 2.1:MLT:1:MT1
MT_2 2.1:MLT:2:MT2



Copyright © 2005 IUPHAR

GENERAL

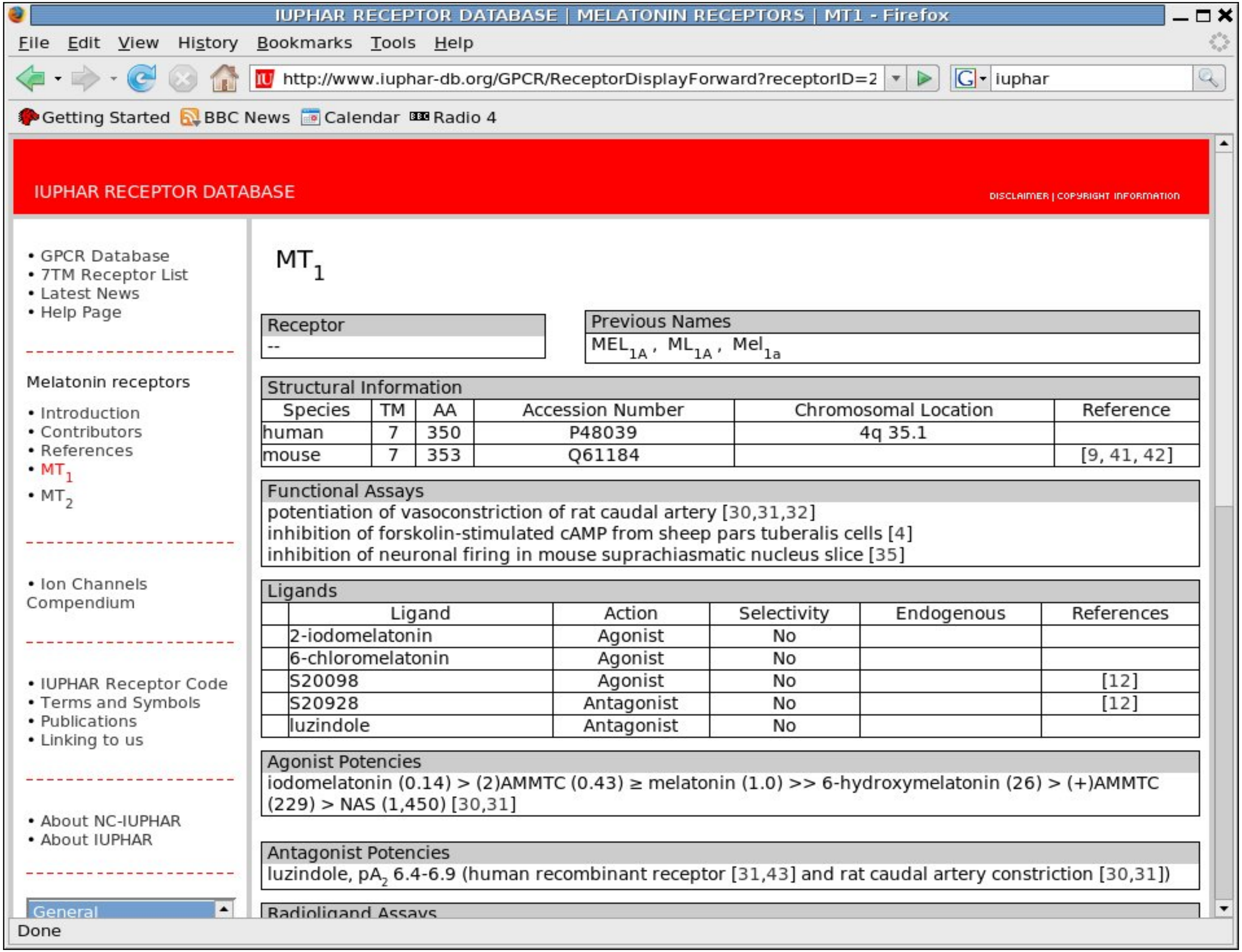
The hormone melatonin is released, following a circadian rhythm, at high levels during the subjective night. It regulates a variety of physiological and neuroendocrine functions through activation of G protein-coupled melatonin receptors in target tissues[1-8].

The use of the radioligands [^{3}H]-melatonin and 2-[^{125}I]-iodomelatonin has led to the localization and characterization in native tissues of a number of putative melatonin binding sites with well-defined and distinct pharmacological profiles[1,2,4,5,8]. The first classification of putative melatonin receptors into ML_1 and ML_2 types was based on kinetic and pharmacological differences of 2-[^{125}I]-iodomelatonin binding[8]. The pharmacological profile (2-iodomelatonin > melatonin >> N-acetylserotonin) of 2-[^{125}I]-iodomelatonin binding to mammalian retina and pars tuberalis corresponds closely to that of the functional melatonin receptor characterized in rabbit retina, i.e. the ML_1 type[2,4-8]. By contrast the pharmacology (2-iodomelatonin > melatonin = N-acetylserotonin) of 2-[^{125}I]-iodomelatonin binding to hamster brain membranes was distinguished by N-acetylserotonin, which showed equal affinity with melatonin[2,5,6,8] and corresponds to the ML_2 type.

Cloning studies have revealed two recombinant mammalian melatonin receptors - Mel_{1a} and Mel_{1b}, now termed MT_1 and MT_2 (refs. [7,9-11]) - encoding 2-[^{125}I]-iodomelatonin binding sites showing the general pharmacology of the ML_1 type[7,12]. These two melatonin receptors were defined as unique entities on the basis of their molecular structure and chromosomal localization[7,9-11,13,14]. The human recombinant melatonin receptor, (h MT_1 and h MT_2) show 60% homology at the amino acid (aa) level and distinct pharmacological profiles of partial agonist and antagonist binding affinities for 2-[^{125}I]-iodomelatonin and [^{3}H]-melatonin[5,12,15,16].

MT_1

A number of non-selective melatonin receptor agonists and antagonists have been identified[16-25], which have been useful in the pharmacological characterization of melatonin receptors in native tissues[12]. Work carried out with recombinant h MT_1 receptors led to the identification of various analogues as inverse agonists


IUPHAR RECEPTOR DATABASE | MELATONIN RECEPTORS | MT1 - Firefox
File
Edit
View
History
Bookmarks
Tools
Help
http://www.iuphar-db.org/GPCR/ReceptorDisplayForward?receptorID=2
iuphar
Getting Started
BBC News
Calendar
Radio 4
DISCLAIMER | COPYRIGHT INFORMATION
General
Done

IUPHAR RECEPTOR DATABASE

- GPCR Database
- 7TM Receptor List
- Latest News
- Help Page

Melatonin receptors

- Introduction
- Contributors
- References
- MT$_1$
- MT$_2$

- Ion Channels Compendium

- IUPHAR Receptor Code
- Terms and Symbols
- Publications
- Linking to us

- About NC-IUPHAR
- About IUPHAR

MT$_1$

Receptor
--

Previous Names
MEL$_{1A}$, ML$_{1A}$, Mel$_{1a}$

Structural Information

Species	TM	AA	Accession Number	Chromosomal Location	Reference
human	7	350	P48039	4q 35.1	
mouse	7	353	Q61184		[9, 41, 42]

Functional Assays

potentiation of vasoconstriction of rat caudal artery [30,31,32]
inhibition of forskolin-stimulated cAMP from sheep pars tuberalis cells [4]
inhibition of neuronal firing in mouse suprachiasmatic nucleus slice [35]

Ligands

	Ligand	Action	Selectivity	Endogenous	References
	2-iodomelatonin	Agonist	No		
	6-chloromelatonin	Agonist	No		
	S20098	Agonist	No		[12]
	S20928	Antagonist	No		[12]
	luzindole	Antagonist	No		

Agonist Potencies

iodomelatonin (0.14) > (2)AMMTC (0.43) ≥ melatonin (1.0) >> 6-hydroxymelatonin (26) > (+)AMMTC (229) > NAS (1,450) [30,31]

Antagonist Potencies

luzindole, pA$_2$ 6.4-6.9 (human recombinant receptor [31,43] and rat caudal artery constriction [30,31])

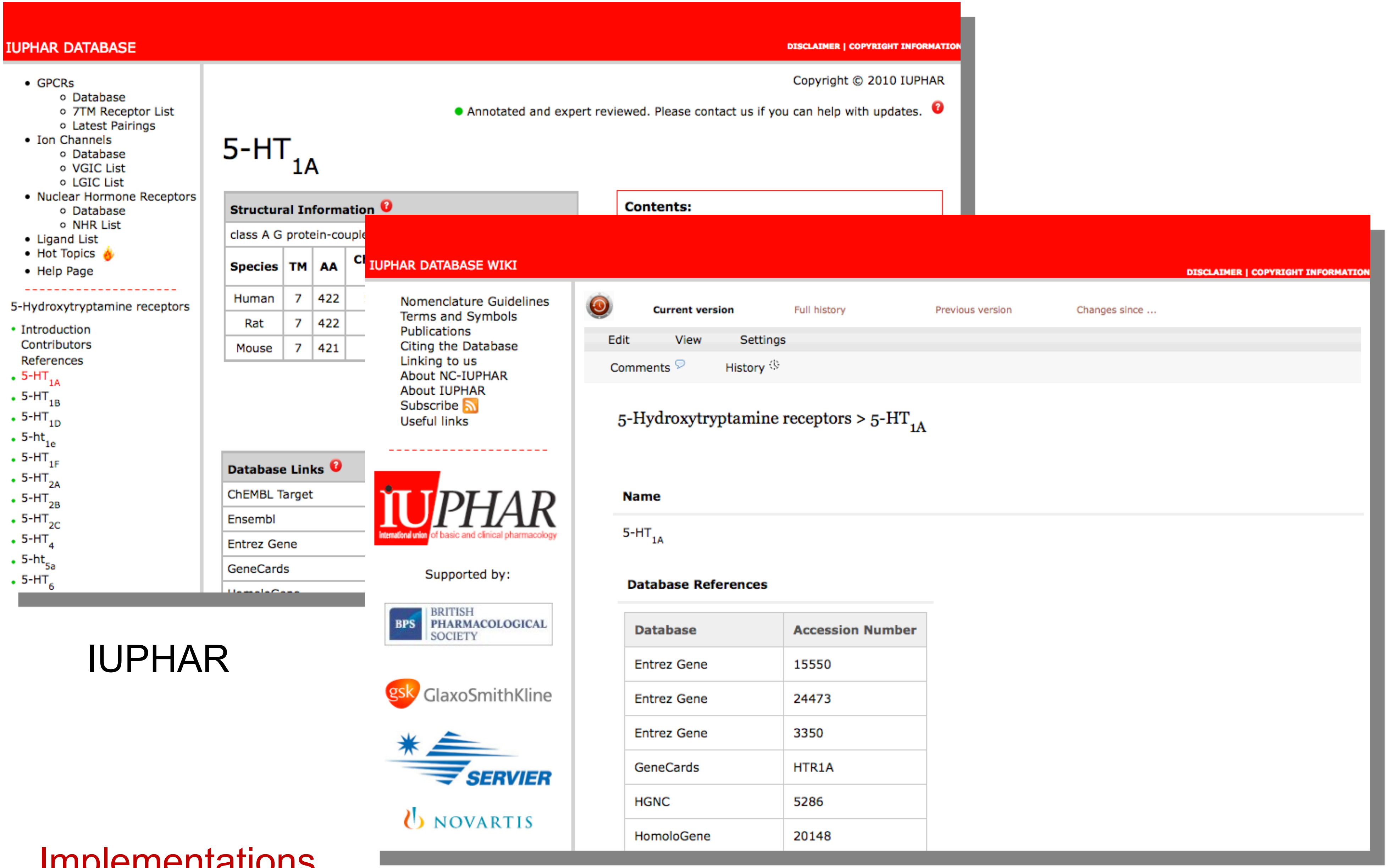
Radioligand Assays

DBWiki

A structured wiki for curated databases and collaborative data management

- Databases are great at storing and querying structured data, but hard to use.
- Wikis are easy to use, but bad at storing structured data.
- A *Database Wiki* is a system that combines the strengths of databases and wikis, to make it easier *collaboratively* to build valuable Web databases
    - In the same way "citizen science", brainsourcing or Wikipedia contributors already have built valuable Web sites *:*

A key feature is that any element can be annotated – including other annotations.
Annotations can be moved into *structure*



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

5-Hydroxytryptamine receptors

- Introduction
  Contributors
  References
- 5-HT_{1A}
- 5-HT_{1B}
- 5-HT_{1D}
- 5-ht_{1e}
- 5-HT_{1F}
- 5-HT_{2A}
- 5-HT_{2B}
- 5-HT_{2C}
- 5-HT_4
- 5-ht_{5a}
- 5-HT_6

Copyright © 2010 IUPHAR

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

5-HT_{1A}

Structural Information

class A G protein-couple

Species	TM	AA	Cl
Human	7	422	
Rat	7	422	
Mouse	7	421	

Contents:

Database Links

- ChEMBL Target
- Ensembl
- Entrez Gene
- GeneCards
- HomoloGene

IUPHAR DATABASE WIKI

DISCLAIMER | COPYRIGHT INFORMATION

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

Supported by:

Current version | Full history | Previous version | Changes since ...

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
Entrez Gene	3350
GeneCards	HTR1A
HGNC	5286
HomoloGene	20148

IUPHAR

Implementations by Heiko Müller and Sam Lindley

IUPHAR in DBWiki

Data(base) citation

- Scientists are increasingly publishing their data and expect credit for it.
- Scientific credit is measured by citations, so ...

How do we cite data in databases?

- By a database, I mean anything that has internal structure or is subject to change

We (computer scientists) don't normally publish data, but …

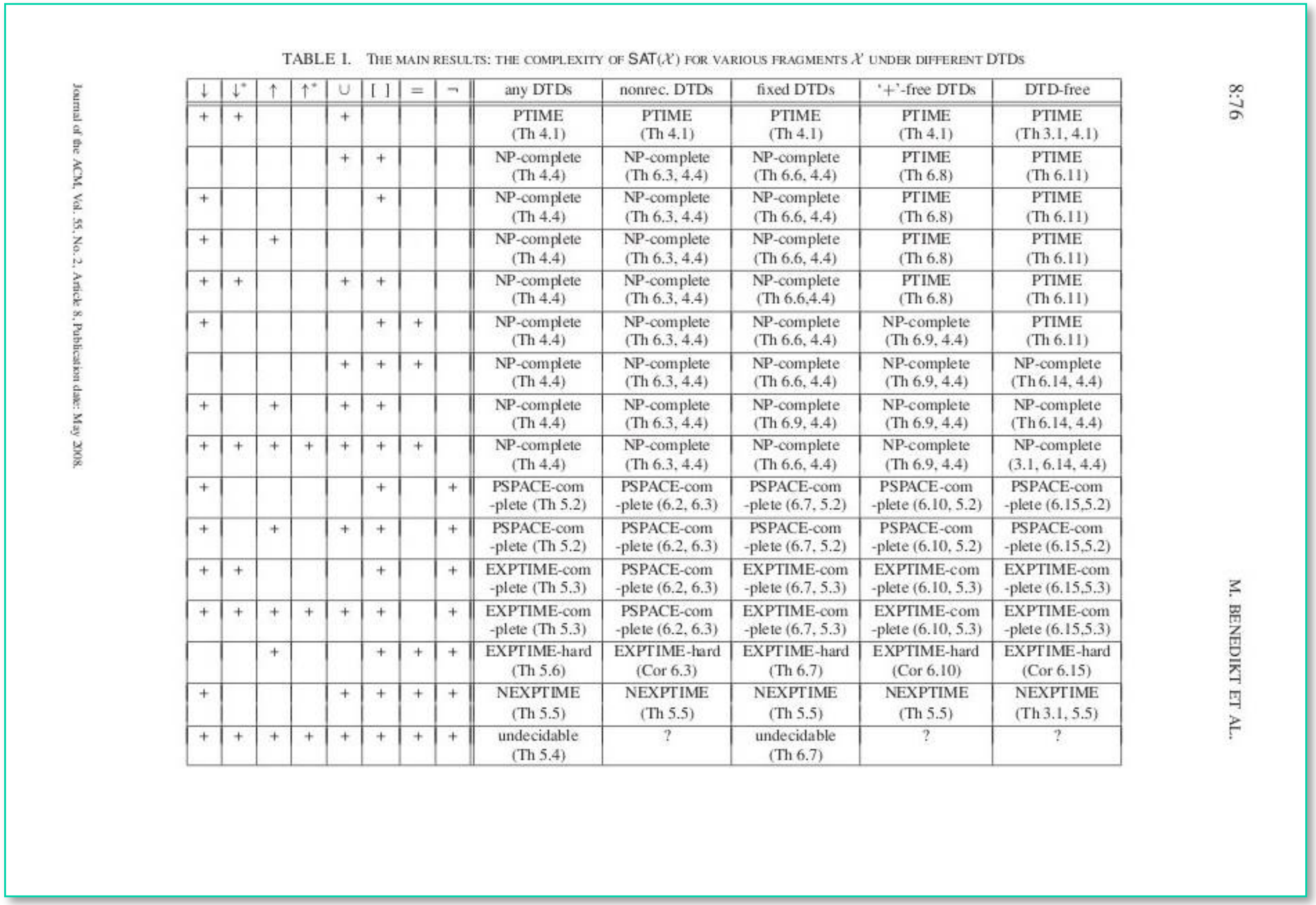


TABLE I. The main results: the complexity of $\mathsf{SAT}(\mathcal{X})$ for various fragments $\mathcal{X}$ under different DTDs

$\downarrow$	$\downarrow^*$	$\uparrow$	$\uparrow^*$	$\cup$	[ ]	=	$\neg$	any DTDs	nonrec. DTDs	fixed DTDs	'+'-free DTDs	DTD-free
+	+			+				PTIME (Th 4.1)	PTIME (Th 4.1)	PTIME (Th 4.1)	PTIME (Th 4.1)	PTIME (Th 3.1, 4.1)
				+	+			NP-complete (Th 4.4)	NP-complete (Th 6.3, 4.4)	NP-complete (Th 6.6, 4.4)	PTIME (Th 6.8)	PTIME (Th 6.11)
+					+			NP-complete (Th 4.4)	NP-complete (Th 6.3, 4.4)	NP-complete (Th 6.6, 4.4)	PTIME (Th 6.8)	PTIME (Th 6.11)
+		+						NP-complete (Th 4.4)	NP-complete (Th 6.3, 4.4)	NP-complete (Th 6.6, 4.4)	PTIME (Th 6.8)	PTIME (Th 6.11)
+	+			+	+			NP-complete (Th 4.4)	NP-complete (Th 6.3, 4.4)	NP-complete (Th 6.6,4.4)	PTIME (Th 6.8)	PTIME (Th 6.11)
+					+	+		NP-complete (Th 4.4)	NP-complete (Th 6.3, 4.4)	NP-complete (Th 6.6, 4.4)	NP-complete (Th 6.9, 4.4)	PTIME (Th 6.11)
				+	+	+		NP-complete (Th 4.4)	NP-complete (Th 6.3, 4.4)	NP-complete (Th 6.6, 4.4)	NP-complete (Th 6.9, 4.4)	NP-complete (Th 6.14, 4.4)
+		+		+	+			NP-complete (Th 4.4)	NP-complete (Th 6.3, 4.4)	NP-complete (Th 6.9, 4.4)	NP-complete (Th 6.9, 4.4)	NP-complete (Th 6.14, 4.4)
+	+	+	+	+	+	+		NP-complete (Th 4.4)	NP-complete (Th 6.3, 4.4)	NP-complete (Th 6.6, 4.4)	NP-complete (Th 6.9, 4.4)	NP-complete (3.1, 6.14, 4.4)
+					+		+	PSPACE-complete (Th 5.2)	PSPACE-complete (6.2, 6.3)	PSPACE-complete (6.7, 5.2)	PSPACE-complete (6.10, 5.2)	PSPACE-complete (6.15,5.2)
+		+		+	+		+	PSPACE-complete (Th 5.2)	PSPACE-complete (6.2, 6.3)	PSPACE-complete (6.7, 5.2)	PSPACE-complete (6.10, 5.2)	PSPACE-complete (6.15,5.2)
+	+				+		+	EXPTIME-complete (Th 5.3)	PSPACE-complete (6.2, 6.3)	EXPTIME-complete (6.7, 5.3)	EXPTIME-complete (6.10, 5.3)	EXPTIME-complete (6.15,5.3)
+	+	+	+	+	+		+	EXPTIME-complete (Th 5.3)	PSPACE-complete (6.2, 6.3)	EXPTIME-complete (6.7, 5.3)	EXPTIME-complete (6.10, 5.3)	EXPTIME-complete (6.15,5.3)
		+			+	+	+	EXPTIME-hard (Th 5.6)	EXPTIME-hard (Cor 6.3)	EXPTIME-hard (Th 6.7)	EXPTIME-hard (Cor 6.10)	EXPTIME-hard (Cor 6.15)
+				+	+	+	+	NEXPTIME (Th 5.5)	NEXPTIME (Th 5.5)	NEXPTIME (Th 5.5)	NEXPTIME (Th 5.5)	NEXPTIME (Th 3.1, 5.5)
+	+	+	+	+	+	+	+	undecidable (Th 5.4)	?	undecidable (Th 6.7)	?	?

(Thanks to Floris Geerts and Wenfei Fan)

Current practice

- Only very recently has the need to cite data in databases been recognized.
- Standards (e.g. Datacite) are being developed but they seem to be avoiding the problem of databases.
- Some DB publishers ask you to cite them but
  - don’t tell you how,
  - tell you to give the URL, or
  - tell you to cite some paper that they wrote about the database.

Nutrition Education for Diverse Audiences [Internet]. Urbana (IL): University of Illinois Cooperative Extension Service, Illinet Department; [updated 2000 Nov 28; cited 2001 Apr 25]. Diabetes mellitus lesson; [about 1 screen]. Available from `http://www.aces.uiuc.edu/~necd/inter2_search.cgi?ind=854148396`

NLM Recommended Formats for Bibliographic Citation.
Internet Supplement. NLM Technical report Bethesda, MD 20894, July 2001.

The structure of a citation

Bard JB and Davies JA. Development, Databases and the Internet. Bioessays. 1995 Nov; 17(11):999-1001

[Identifier and descriptive information]

Ann. Phys., Lpz 18 639-641

Nature, 171,737-738

[Identifier information alone]

Descriptive information is important, but is also somewhat arbitrary

Persistent identifiers

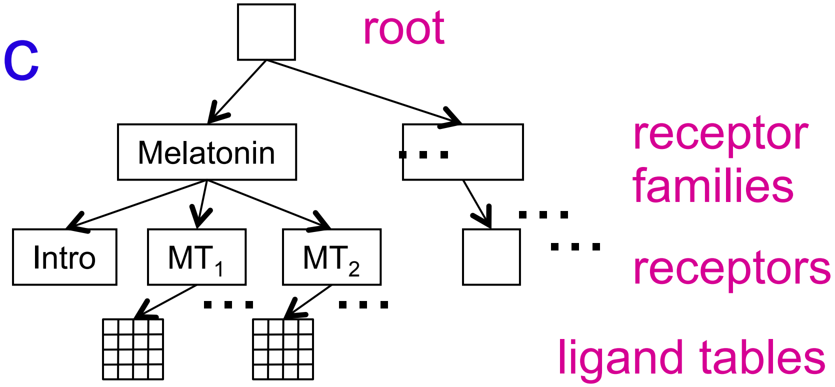
- The world seems to want to invent persistent identifiers for artefacts, digital or otherwise.
  - DOIs, URIs, ARKs, in addition to ISBNs and LOC#s
- Are they needed?
  - Do they confer any status on an object?
  - Do they ensure its persistence/longevity?
  - How do we use them with databases?

BL MS Cotton Nero A X

- A manuscript (MS) in the British library (BL) formerly in the library of Joseph Cotton (which burnt down) under a bust of Nero shelf A ten (X) books along

Other ingredients in data citation

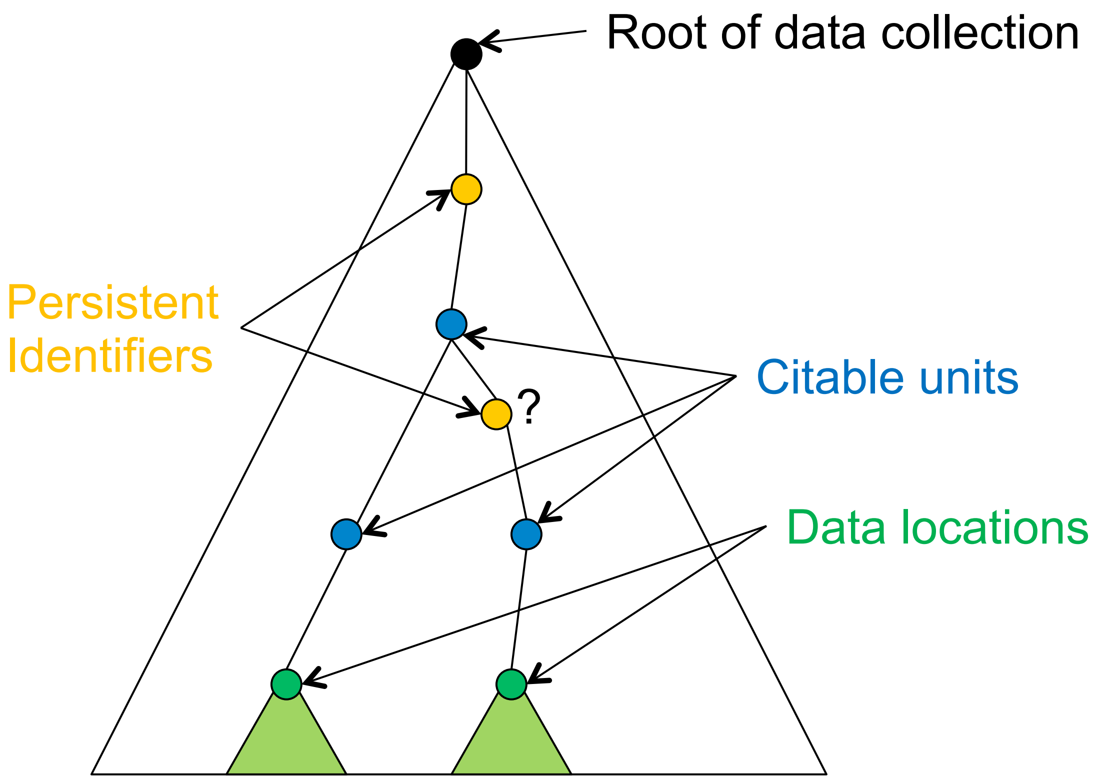
- The notion of a *citable unit*
  - An arbitrary piece/collection of data is not citable
  - (just as a page of a book is a not “the” citation”)
- The *location* of a piece of data within a citable unit
  - We need to be able to find the data of interest
  - (just as a page of a book is a useful location)
- It is often assumed that scientific databases/datasets are hierarchically organised



Some possible citations

1. The IUPHAR database (C1) contains no information about Ginandtonicin.
2. The IUPHAR database (C2) lists five ligands for Melatonin receptor MT_1.
3. The IUPHAR database (C3) asserts that luzindole is an antagonist ligand for receptor MT_1.

The Citation Hierarchy



Should PIDs be tied to citable units? Not clear.

Should we mint a new PID on each update to the database?

Bloggs, A.J. The Convolution of Reality. Elspringer (1977) p67 ISBN-00563744551

Citable unit

Data location

Persistent Identifier

We also need versioning

- Database archiving (Heiko Mueller' s archiver XARCH) provides:
  - A compressed archive successive versions of an XML document for stable citation
  - Also does naive archiving of relational data
- Why not assign version numbers to *parts* of the database?
  - We cannot query anything unless we know its state
- Versions should be recorded at the level of the highest citable (= queryable?) unit

Automatically generating citations

Why is this needed?

- Lots of citations may be required
- Evolving structure (e.g. authorship change)
- Accuracy
- Easy to change to agreed format (if there ever is one)
- Integrity check on the database

Requirement: a stable key/location structure

Idea: use a highly restricted version of Xpath to specify “patterns”

Example:

```
{DB=IUPHAR, Version=$v, Family=$f}

←

 /Root[]/Version[Number=$'v]/Data[]
   /Family[FamilyName=$'f]
```

generates, e.g.,

{DB=IUPHAR, Version=17, Family=Melatonin}

(identification and location information only)

Patterns and Constraints

- Patterns are expressed in the syntax of XPATH, but their function is to bind variables.
- Each step of the path must be qualified by a key variable (indicated by $′x)

```
/Root[]/Version[Number=$′v]/Data[]
  /Family[FamilyName=$′f]
```

FamilyName content uniquely specifies Family element (among all siblings with the same tag name)

Lack of a key variable means that there can only be one Data element (among all its siblings)

A rule that generates descriptive information

```
{ DB=IUPHAR, Version=$v, Family=$f Receptor=$r, Contributors= $a,
  Editor=$e, Date=$d, DOI=$i}

    ←

/Root[]
 /Version[Number=$′v, Editor=$?e, DOI=$.i, Date=$.d]
 /Data[]/Family[FamilyName=$′f]
 /Contributor-list/Contributor=$+a] /Receptor[ReceptorName=$′r]
```

What gets generated (example):

{ DB=IUPHAR, Version=11, Family=Calcitonin,
Receptor=CALCR, Contributors={Debbie Hay, David R. Poyner},
Editor=Tony Harmar, Date=Jan 2006, DOI=10.1234 }

Kinds of variables (non-key)

$.i	exactly one occurrence
$?e	at most one occurrence
$*a	arbitrary occurrences
$+a	one or more occurrences

[All these assume a given matching of key variables]

Efficiency: It is possible to generate and insert citations in linear time (one-pass under very mild constraints.)

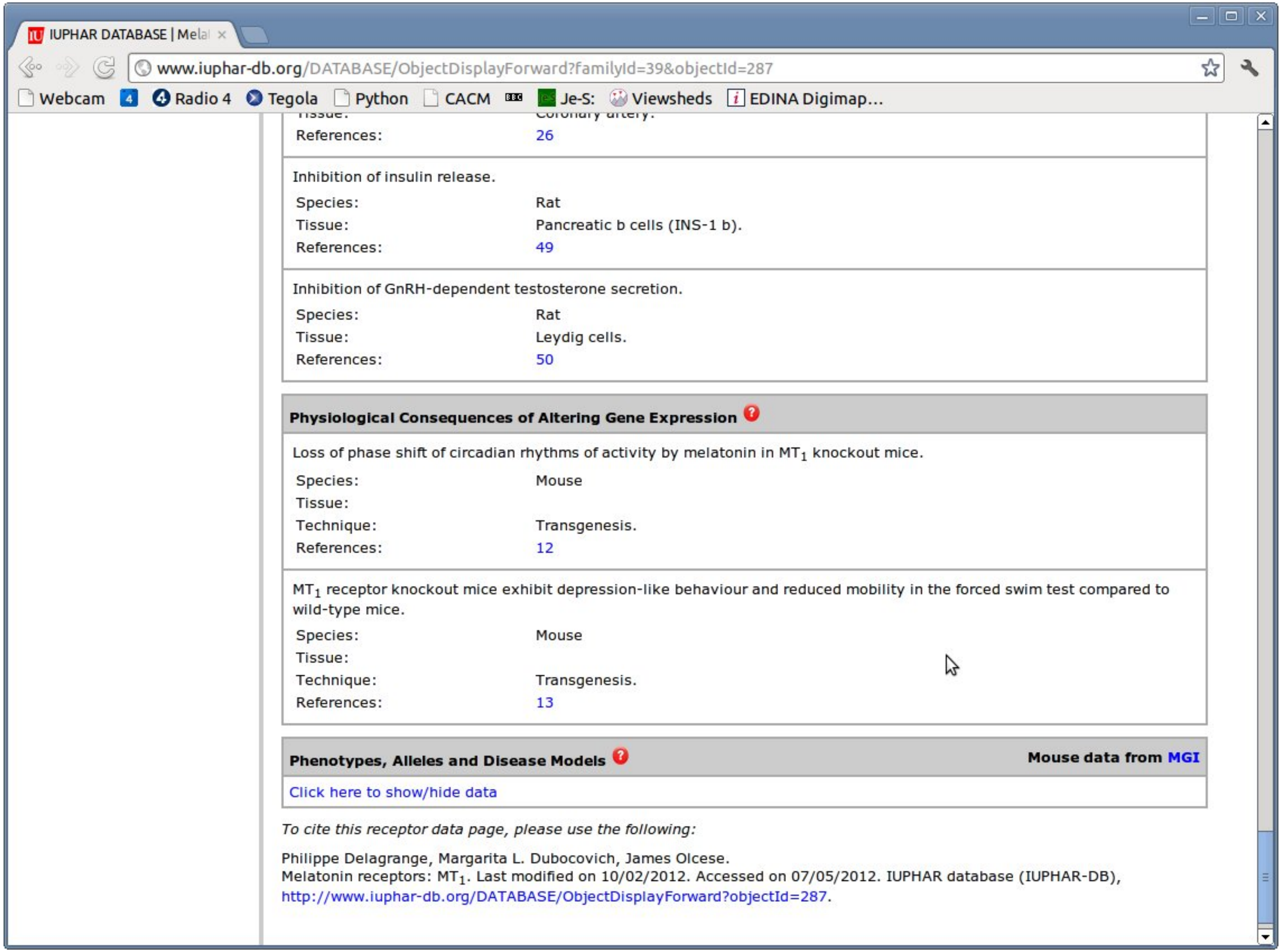
Implementation by Giammaria Silvello

Where we are

- Initial implementation by Gianmaria Silvello
- Citation abstract syntax: should be machine readable/mine-able and human readable.
  - JSON or XML Can we keep it human-readable?
- Concrete syntax a la BibTeX?
- Minimal required fields.
  - Location of the citable unit and/or
  - Persistent identifier
  - Location within the citable unit
- Partially implemented in IUPHAR-DB.

More (standard) database problems

- Source data usually conforms to some schema. The citation (e.g. Datacite) is required to conform to a schema. Can we guarantee this?
- How efficiently can we generate citations? What should be computed statically and what can be computed "on demand"?
- How much checking – or recomputation – needs to be done on update to the database or on schema modification?



Tissue: Coronary artery.

References: 26

Inhibition of insulin release.

Species:	Rat
Tissue:	Pancreatic b cells (INS-1 b).
References:	49

Inhibition of GnRH-dependent testosterone secretion.

Species:	Rat
Tissue:	Leydig cells.
References:	50

Physiological Consequences of Altering Gene Expression

Loss of phase shift of circadian rhythms of activity by melatonin in MT_1 knockout mice.

Species:	Mouse
Tissue:	
Technique:	Transgenesis.
References:	12

MT_1 receptor knockout mice exhibit depression-like behaviour and reduced mobility in the forced swim test compared to wild-type mice.

Species:	Mouse
Tissue:	
Technique:	Transgenesis.
References:	13

Phenotypes, Alleles and Disease Models **Mouse data from MGI**

Click here to show/hide data

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

Philippe Delagrange, Margarita L. Dubocovich, James Olcese.
Melatonin receptors: MT_1. Last modified on 10/02/2012. Accessed on 07/05/2012. IUPHAR database (IUPHAR-DB), http://www.iuphar-db.org/DATABASE/ObjectDisplayForward?objectId=287.

Not yet satisfactory because they don't publish past versions of the database

Citation and linked data?

- How does this work on an amorphous mass of RDF triples?
  - Where is the hierarchy (is there a hierarchy?)
  - What are the citable units?
- Problems similar to those for annotation
  - Define citable units by queries and use query containment to get the hierarchy?
  - Use named graphs? (How many columns do we need?)
- Should we express and link citations in RDF?
- And again there's efficiency...