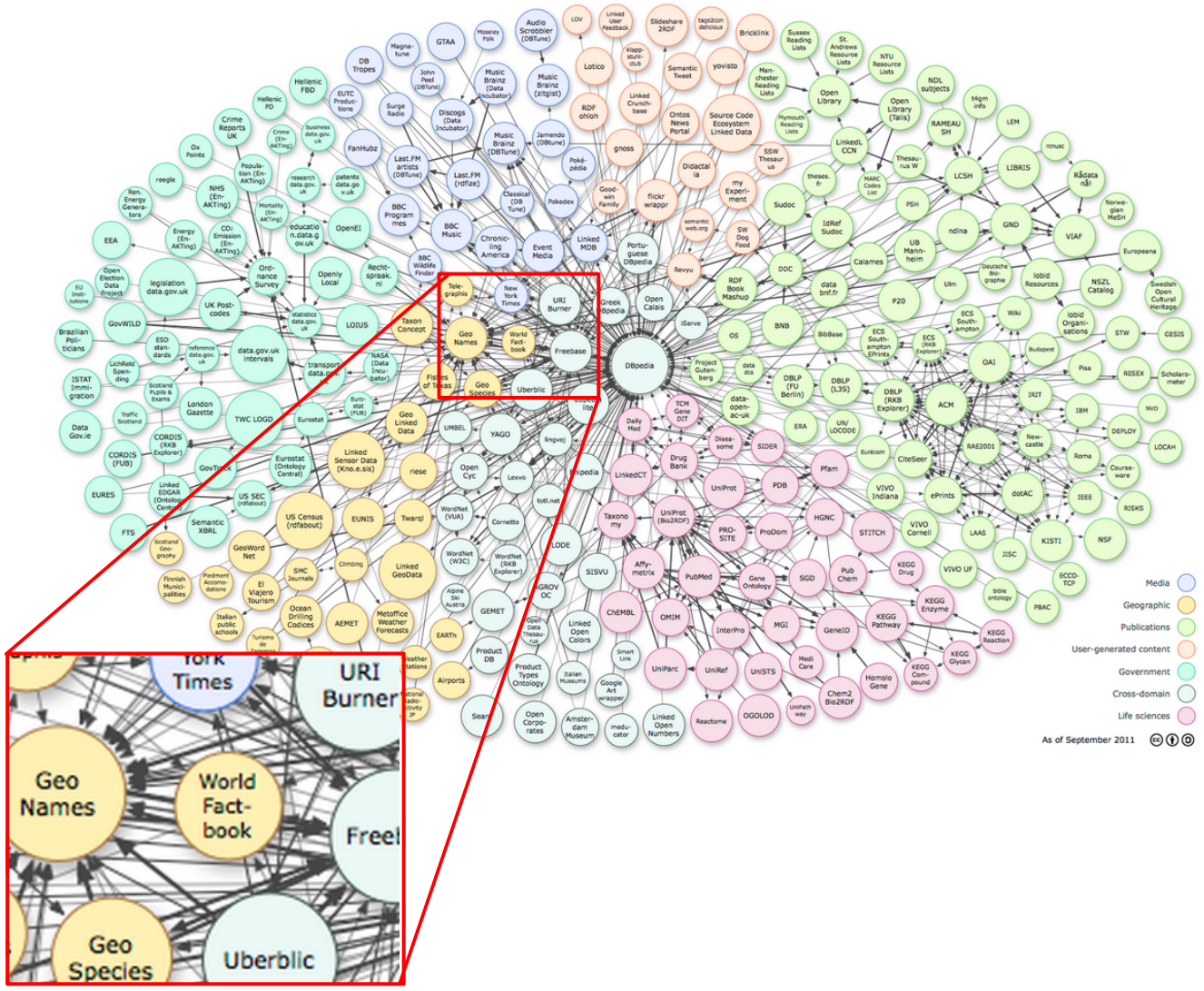
I Annotation

With Egor Kostylev and Stijn Vansummeren

Linked Open Data – the population of Afghanistan



factbook:population_total	31889923 (xsd:long)
factbook:populationbelowpovertyline	53.0E0 (xsd:double)
factbook:populationgrowthrate	2.625E0 (xsd:double)
factbook:portsandterminals	Kheyrabad, Shir Khan
factbook:radiobroadcaststations	AM 21, FM 5, shortwave 1 (broadcasts i

LOD

Introduction :: AFGHANISTAN

Geography :: AFGHANISTAN

People and Society :: AFGHANISTAN

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.)

country comparison to the world: 40

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

World Factbook
(March 2013)

Annotation has been lost. How do we preserve it? [DBNotes, Annotea, 3rd voice]...

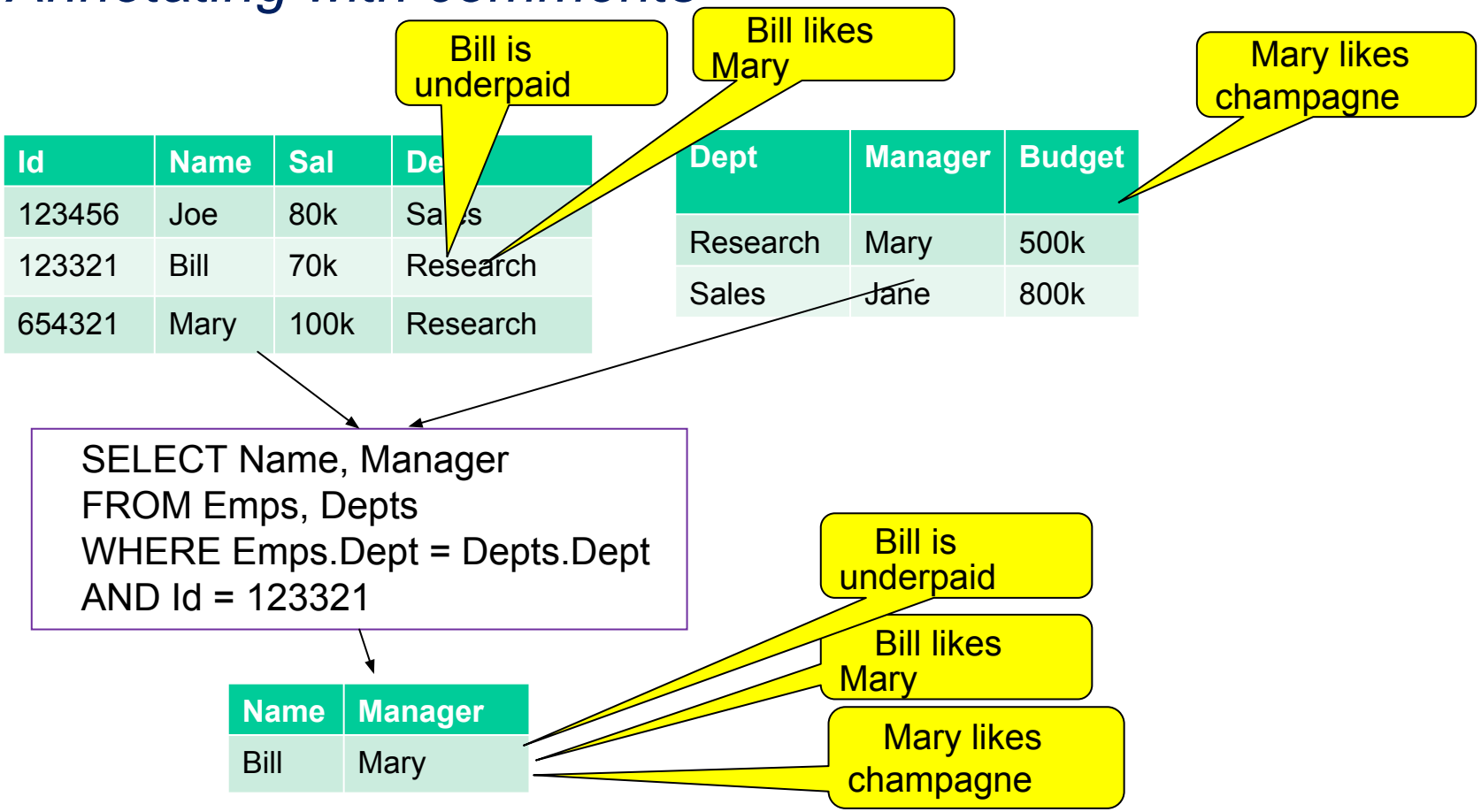
Name	Address	Job	*Time*
Jane	Edinburgh	Teacher	*1998-2001*
Jane	Edinburgh	Research	*2002-2005*
Jane	Paris	Research	*2006-2012*
Bill	New York	Actor	*1999-2001, 2004-2006*

Name	Address	Job	*Believes*
Jane	Edinburgh	Teacher	{*Joe, Mary*}
Jane	Edinburgh	Research	{*Jane. Joe, Mary*}
Jane	Paris	Research	{*Jane. Joe, Mary*}
Bill	New York	Actor	{*Jane, Bill, Mary*}

Name	Address	Job	*Comments*
Jane	Edinburgh	Teacher	{"*I was a PhD student*"}
Jane	Edinburgh	Research	{ }
Jane	Paris	Research	{"*Lucky Jane*", "*It wasn't reseach*"}
Bill	New York	Actor	{"*He had bit parts*", "*He was a stage-hand*"}

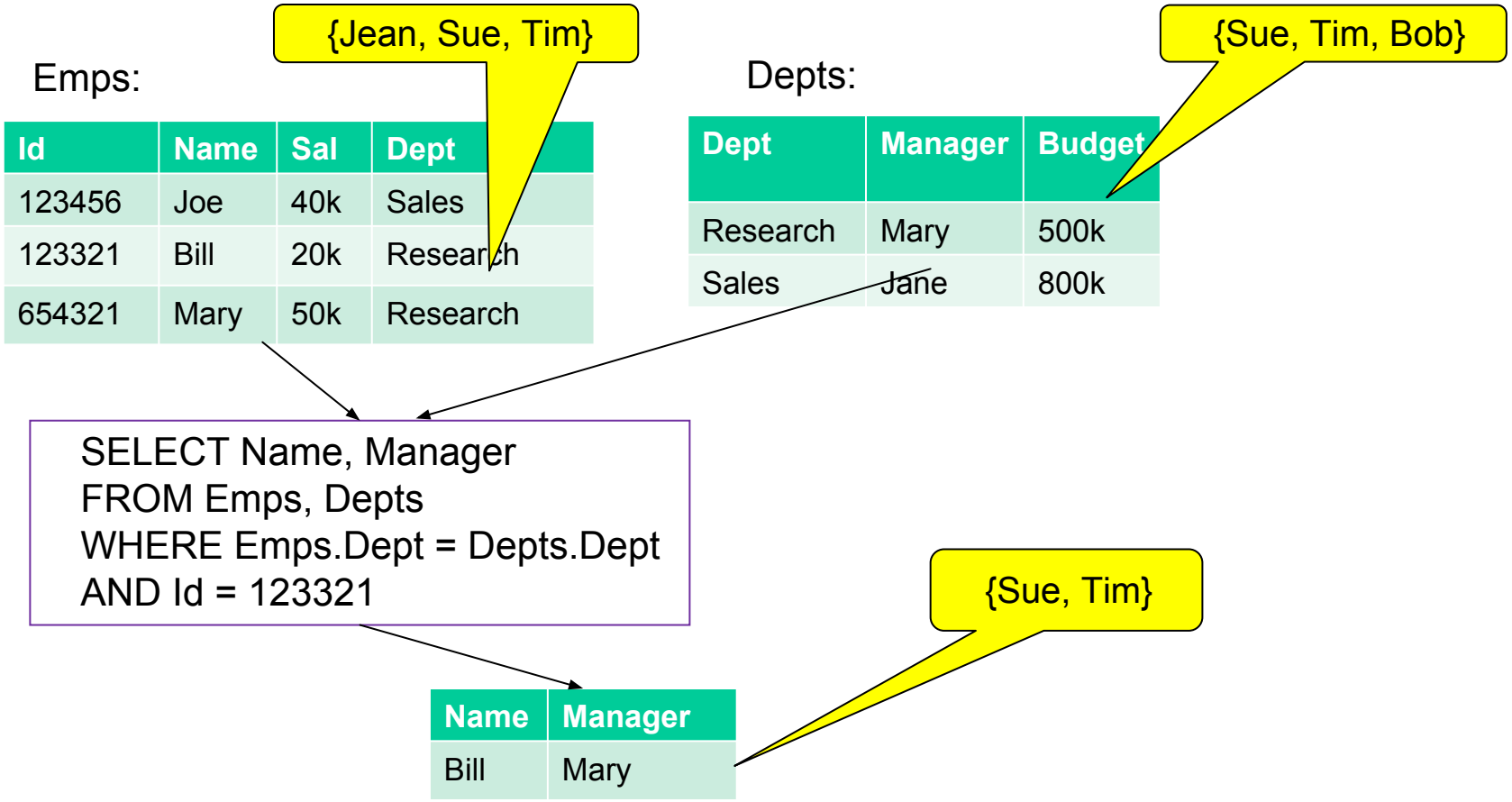
Annotation propagation

Annotating with comments



We probably want the *union* of the comments on the input

Annotating with beliefs: the people who believe a tuple to be true



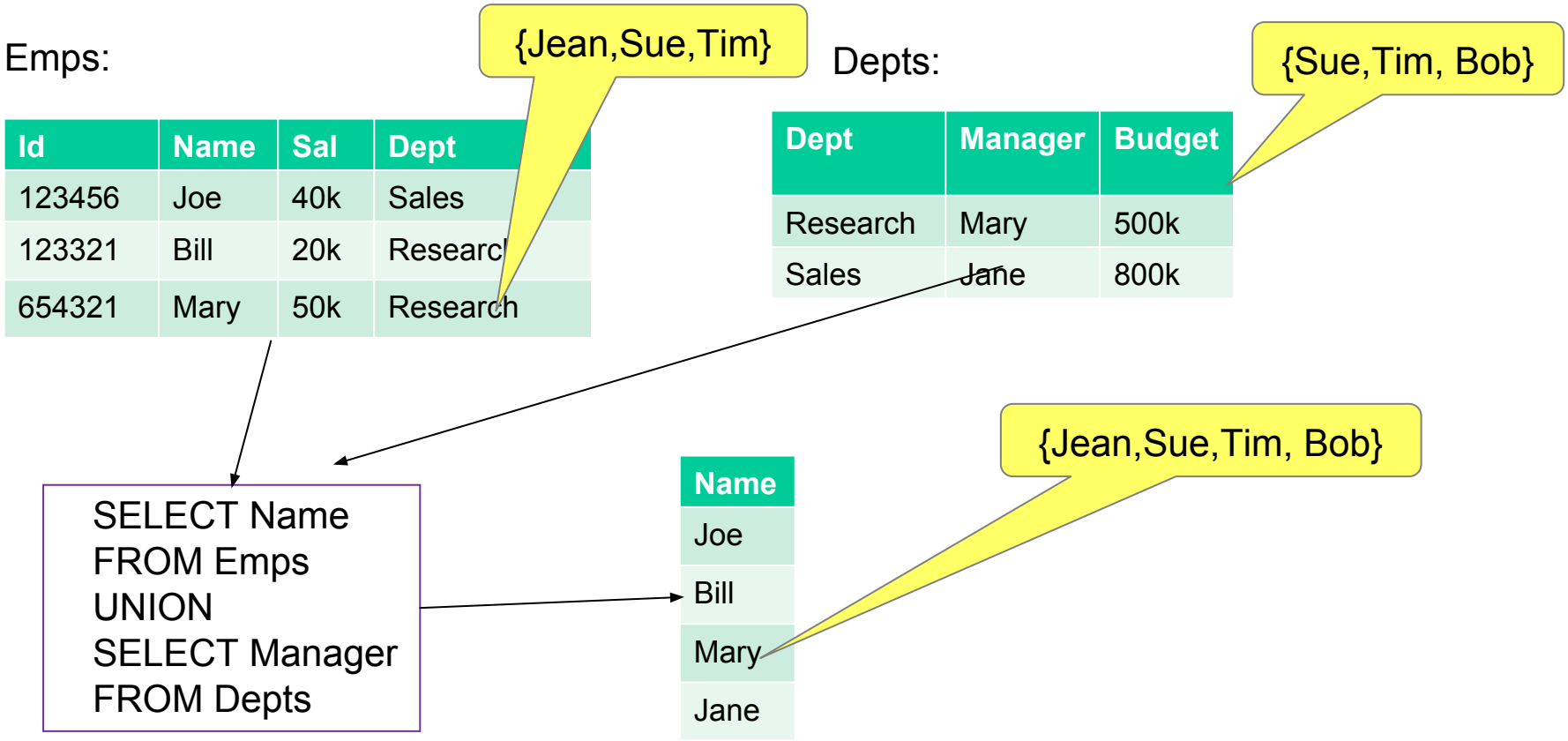
Id	Name	Sal	Dept
123456	Joe	40k	Sales
123321	Bill	20k	Research
654321	Mary	50k	Research

Dept	Manager	Budget
Research	Mary	500k
Sales	Jane	800k

Name	Manager
Bill	Mary

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

Annotating with beliefs for another query



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

Semirings

These are two examples of *provenance semirings* (Tannen school)

- Belief semiring (B, $\cup$, $\cap$, {}, B) B= sets of believer -- Boolean algebra
- Comment semiring (STR, $\cup$, $\cup$, {}, {})* STR = set of strings -- not quite a semiring

Many well-known extensions to relational algebra are examples of semirings:, bag semantics, C-tables, probabilistic databases, various forms of provenance

Propagation of Boolean algebra and comment semirings can be expressed in relational algebra

Annotations on annotations

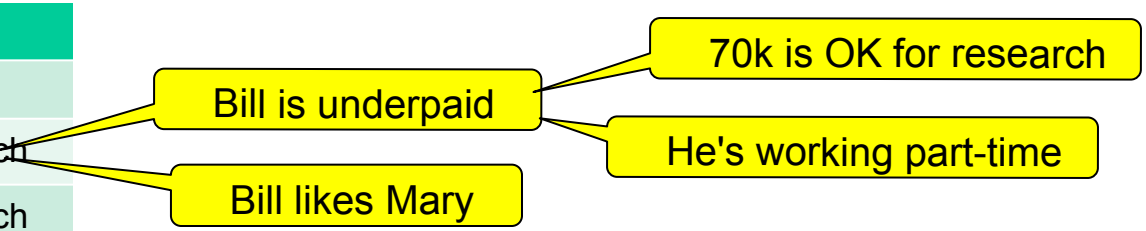
Overlooked (until recently): you can have annotations on annotations. In fact they are common in curated DBs

Examples:

At time t_1 A believed that at time t_2 ...

Also, chains of belief. Gatterbauer, *et al* Believe It or Not: Adding Belief Annotations to Databases. PVLDB 2(1): 1-12 (2009)

Id	Name	Sal	Dept
123456	Joe	80k	Sales
123321	Bill	70k	Research
654321	Mary	100k	Research



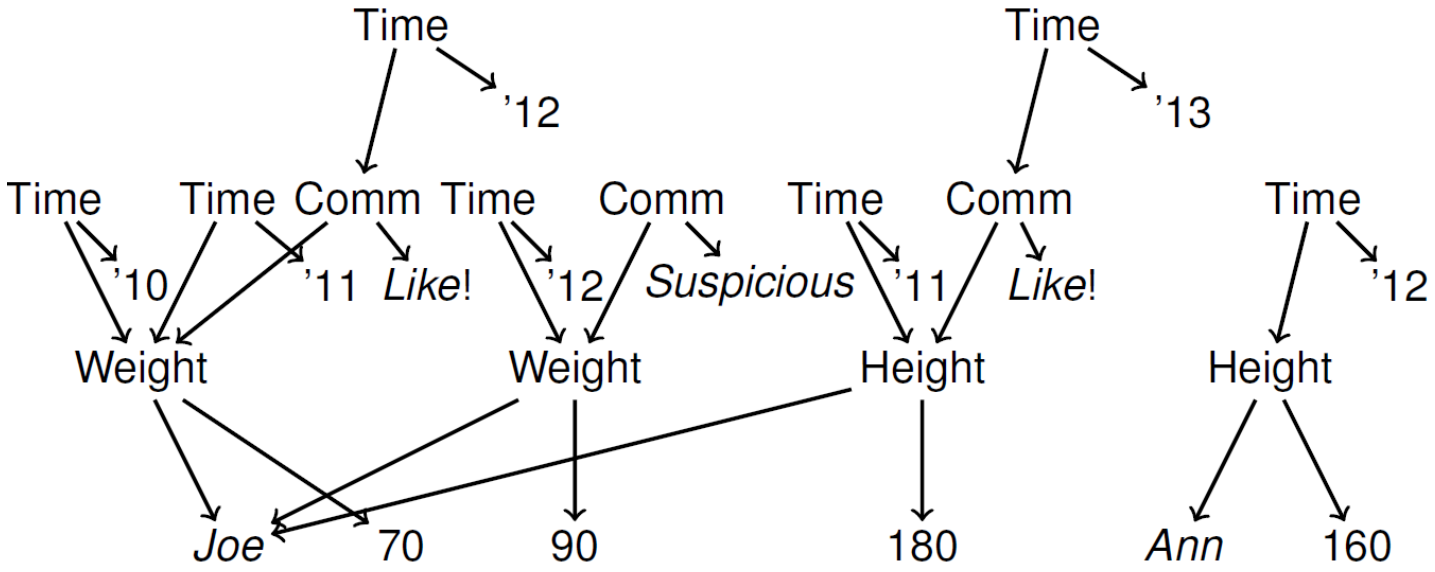
I.e., mail threads

Work in a term model

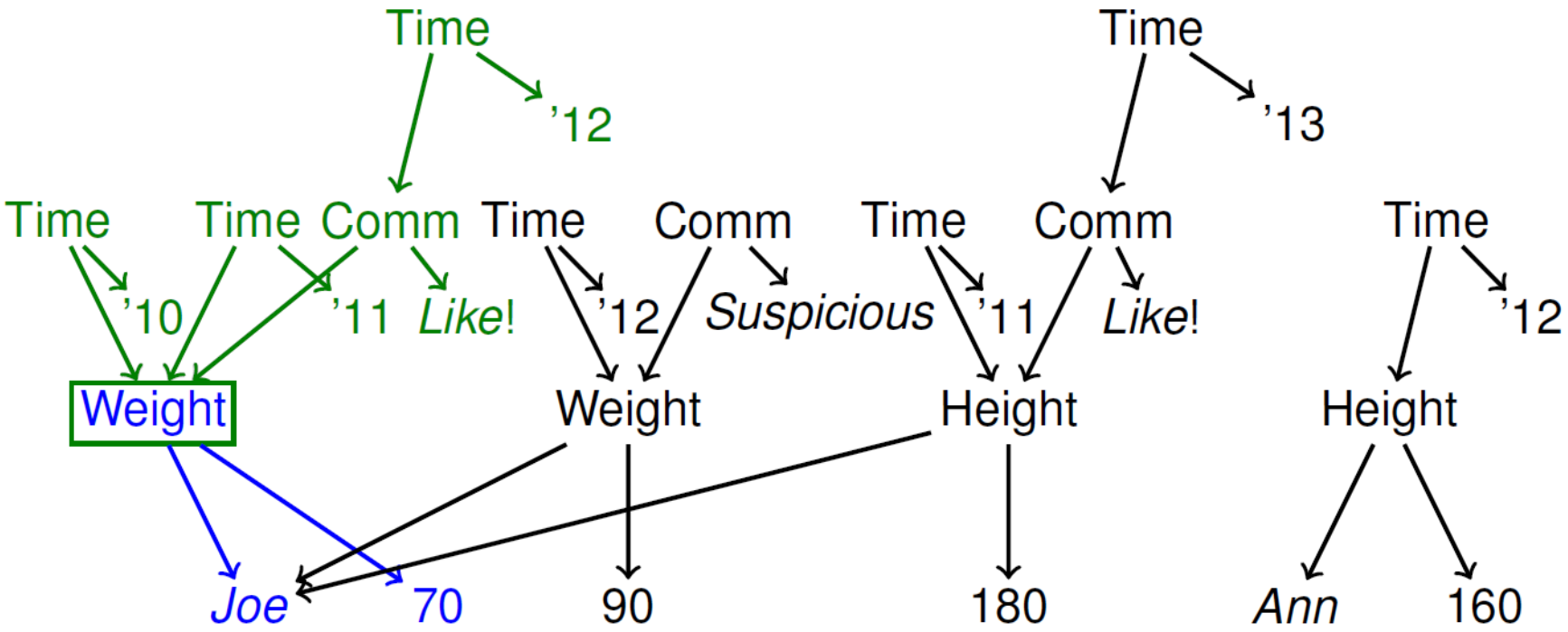
Weight(Joe,70)

Time(Weight(Joe,70), 10)
Time(Weight(Joe,70), 11)
Comm(Weight(Joe,70),"Like")

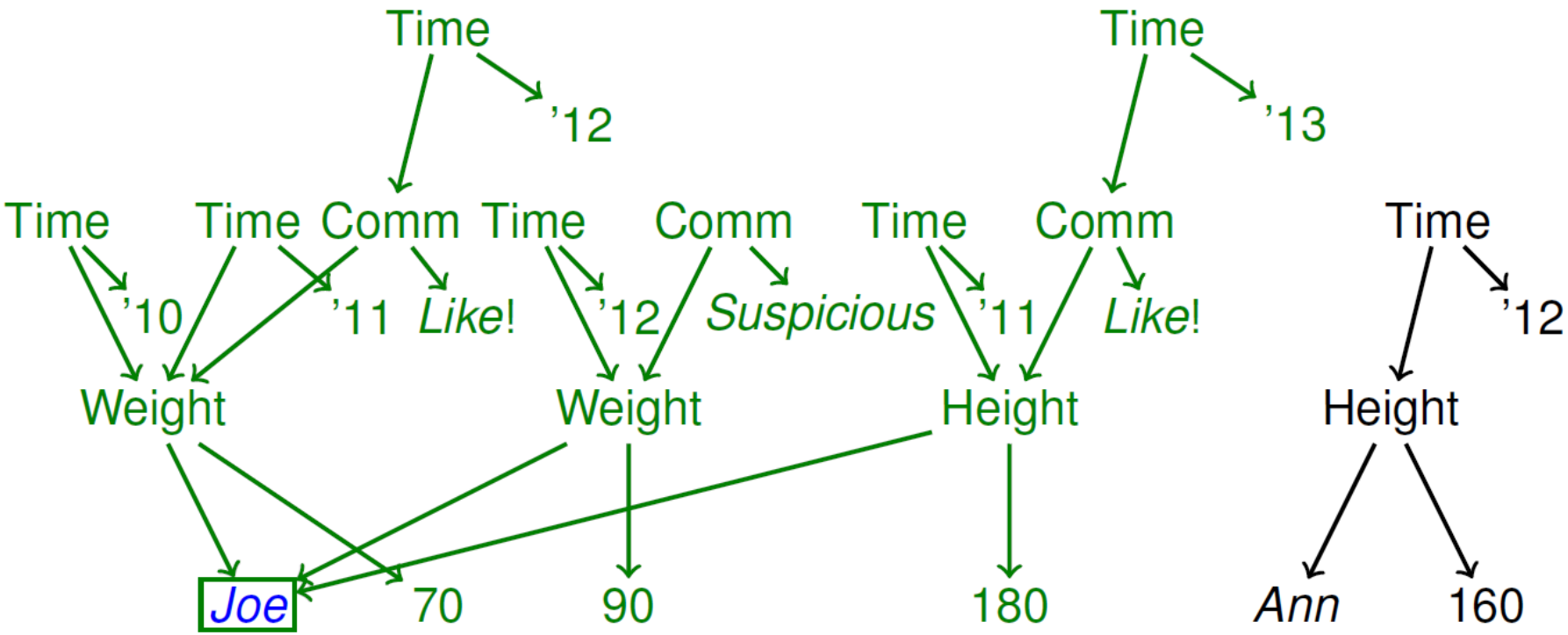
Time(Comm(Weight(Joe,70),"Like"), 12)



Annotation on Weight(Joe,70)



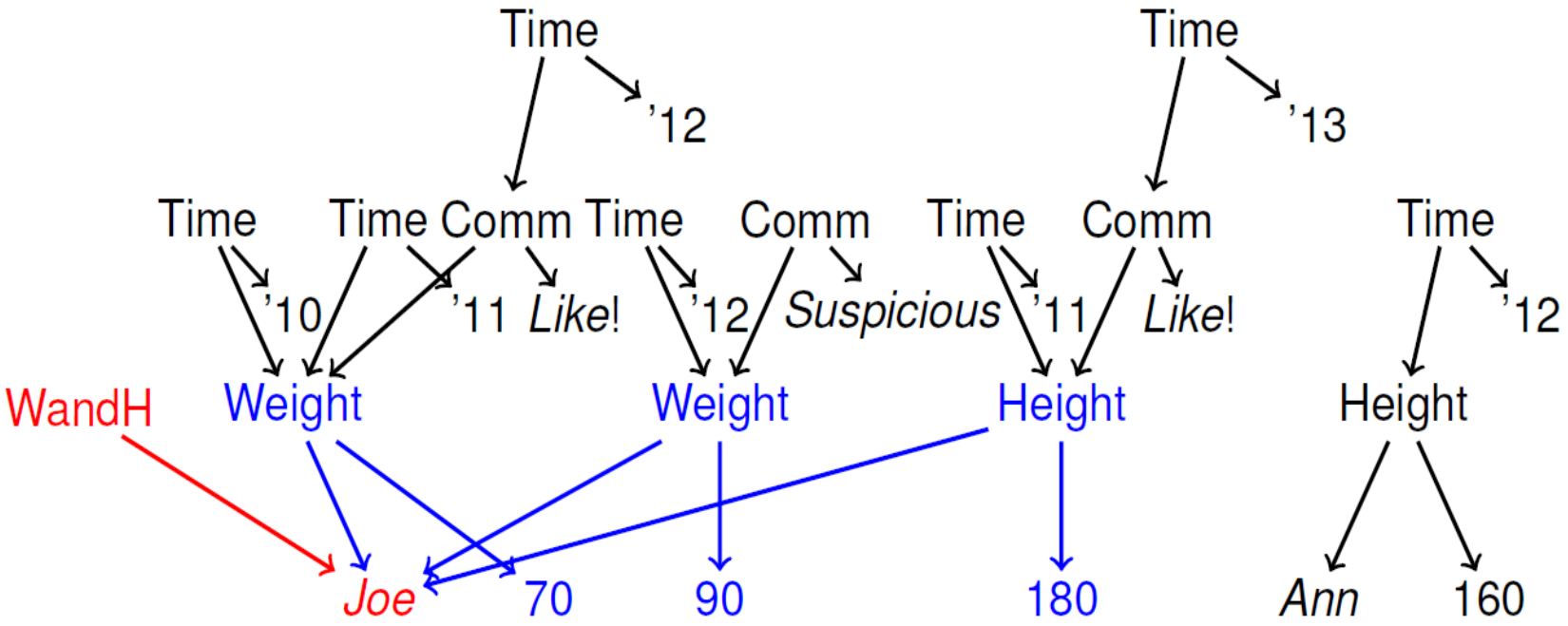
Annotation on Joe?



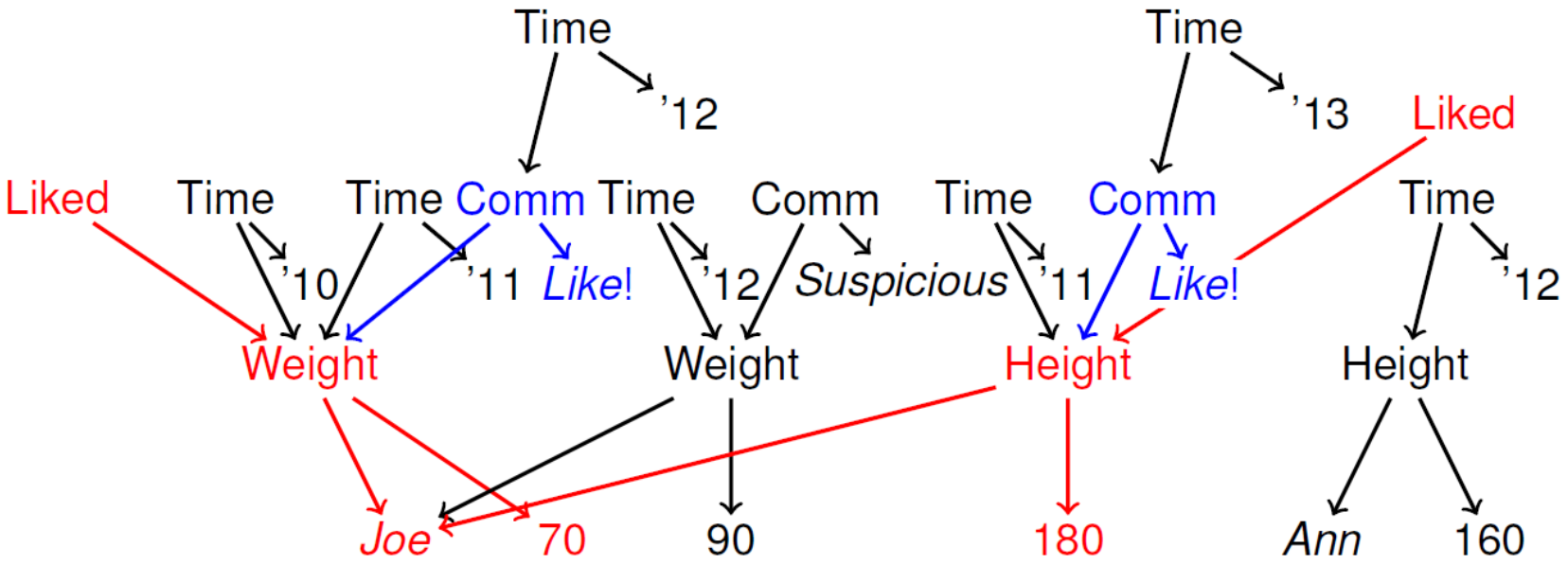
How does annotation differ from data?

The query defines what is data

WandH(X) <- Weight(X,V_1) & Height(X,V_2)



Liked(X) <- Comm(X, "Like!")



So how do annotations propagate through queries? Read [B., Kostylev and Vanssummeren. Annotations are Relative. ICDT 2013]

Or, better, go think about it!

Moving back to triple stores / RDF

At first sight this gives us a more "fine-grained" and accurate representation of annotation

Name	**Address**	**Job**	***Time***
Jane	Edinburgh	Teacher	*1998-2001*
Jane	Edinburgh	Research	*2002-2005*
Jane	Paris	Research	*2006-2012*
Bill	New York	Actor	*1999-2001, 2004-2006*

Job

Jane	Teacher	*1998-2001*
Jane	Research	*2002-2012*
Bill	Actor	*1999-2006*

Address

Jane	Edinburgh	*1998-2005*
Jane	Paris	*2006-2012*
Bill	New York	*1999-2001, 2004-2006*

How do we annotate RDF/LOD?

One of the unsung advantages of relational DBMSs -- especially column stores -- is that one can add a column *efficiently and without damaging the existing code base*]

What do we do with RDF?

- Add more columns. (We are already up to 4 or 5)
  - How many?
- Reify and
  - destroy the existing code base , and
  - re-reify each time you add an annotation?

What is done in practice?

- Google info boxes. Use a triple store and “something else” when needed.
- WikiData. Don’t yet have an adequate data model, but say they need one.
- Others. Reify anything that *might* be annotated.

"Annotation is the linking of a new commentary node to someone else's existing node. It is the essence of a collaborative hypertext." Berners-Lee 1992

Two solutions for annotating triple stores/RDF

Q: What do you do in a storm off a lee shore?

A: Don't be there in the first place.

In other words: don't use RDF

Major Heresy?

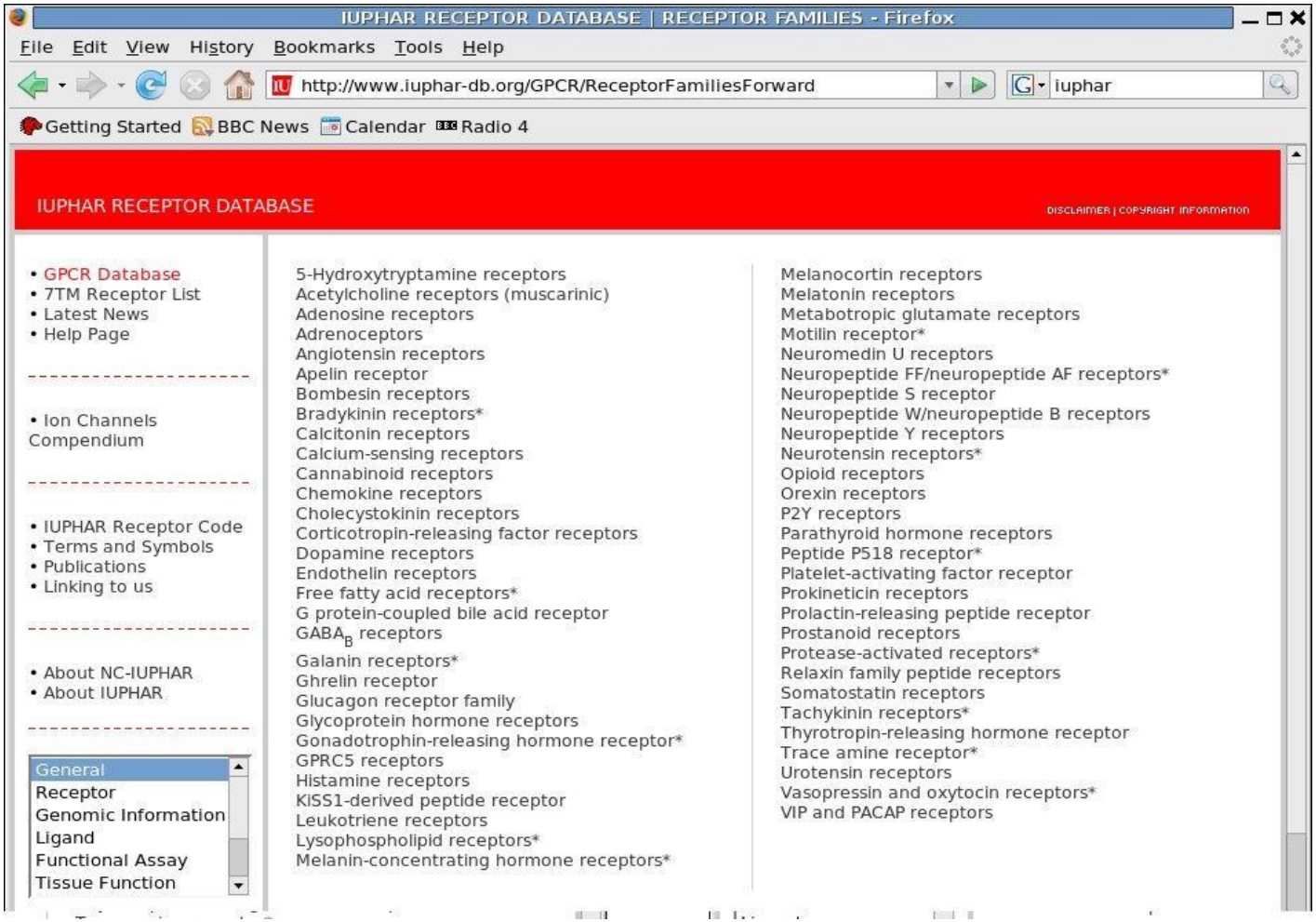
Use *Nested RDF* or (almost the same) use the 4th column as a triple identifier.

- Named graphs, blank nodes etc. all expressible *in RDF triples*
- Annotation is easy and can be compounded
- Ontological assertions like "If A trusts B and B believes X then A believes X" are easily represented.
- We've re-invented Lisp!

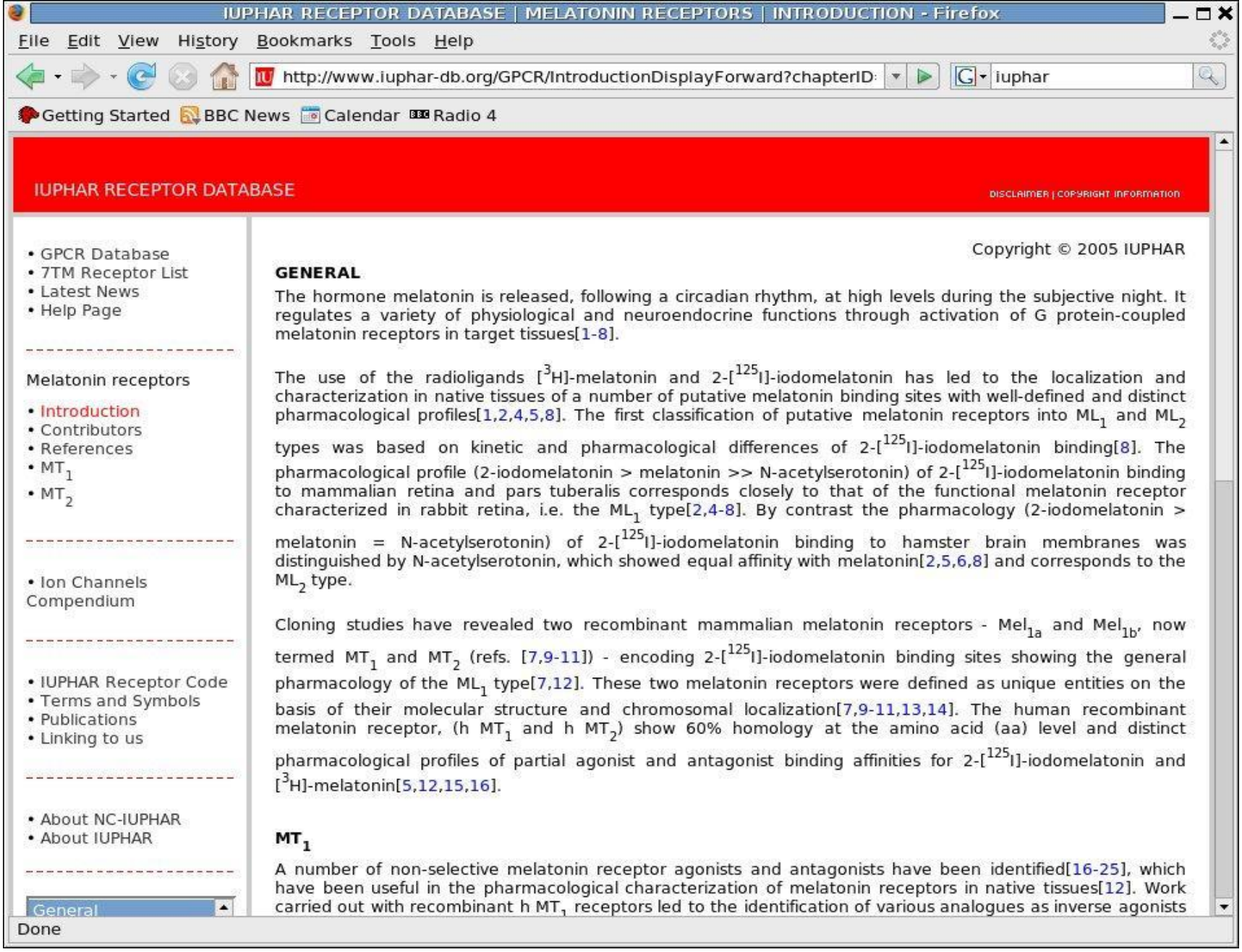
II Data Citation

With Susan Davidson, Wenfei Fan, James Frew, Tony Harmar and Val Tannen

The IUPHAR database – an example of "brain-sourcing"



*



GENERAL

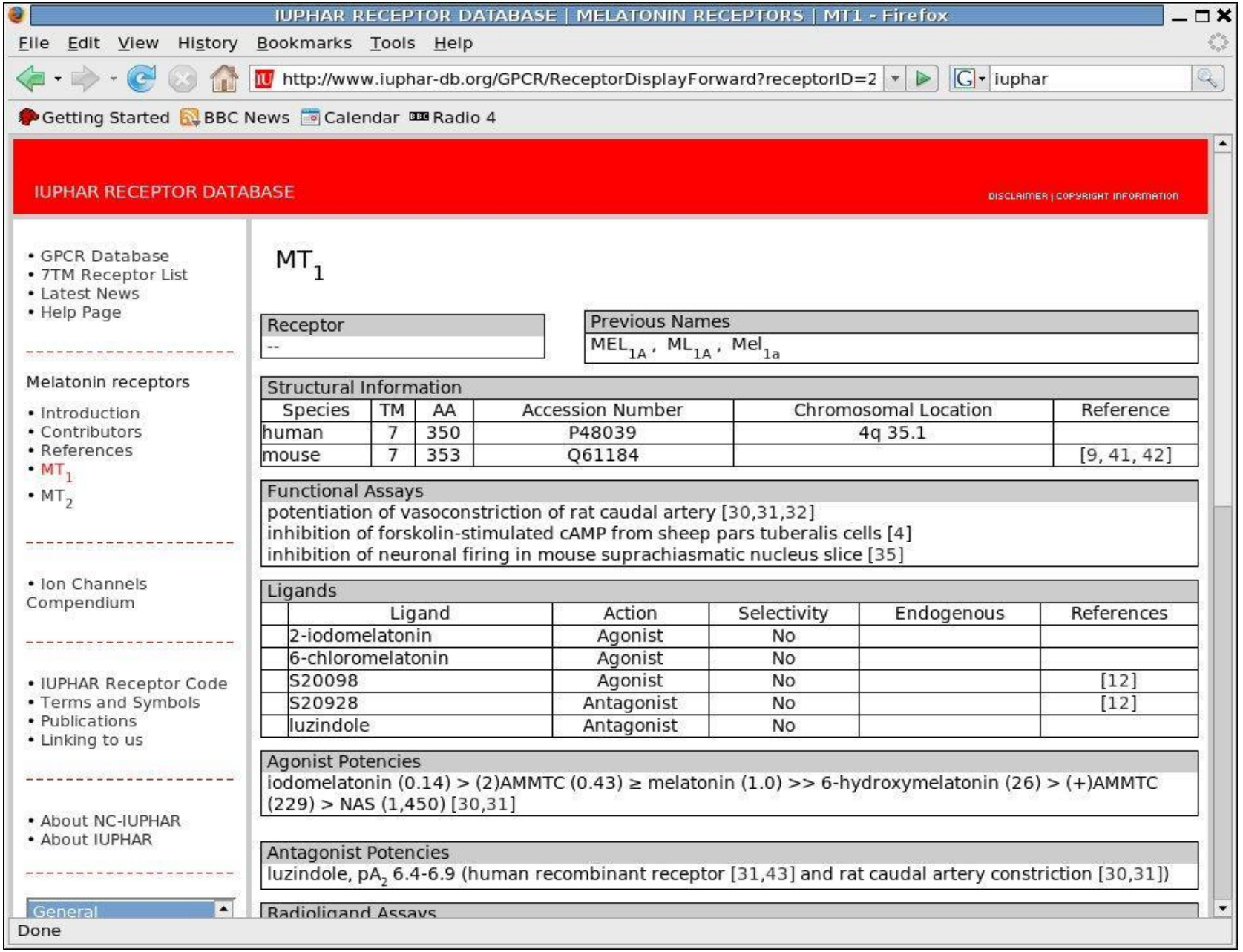
The hormone melatonin is released, following a circadian rhythm, at high levels during the subjective night. It regulates a variety of physiological and neuroendocrine functions through activation of G protein-coupled melatonin receptors in target tissues[1-8].

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
IUPHAR RECEPTOR DATABASE
DISCLAIMER | COPYRIGHT INFORMATION
• GPCR Database
• 7TM Receptor List
• Latest News
• Help Page
Melatonin receptors
• Introduction
• Contributors
• References
• MT1
• MT2
• Ion Channels Compendium
• IUPHAR Receptor Code
• Terms and Symbols
• Publications
• Linking to us
• About NC-IUPHAR
• About IUPHAR
General
MT1
Receptor
--
Previous Names
MEL1A, ML1A, Mel1a
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, pA2 6.4-6.9 (human recombinant receptor [31,43] and rat caudal artery constriction [30,31])
Radioligand Assays
Done

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

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

*

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)

Scientific databases/datasets are usually hierarchically organised, or are presented that way.

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.

*

Automatically generating citations

Why is this needed?

- Lots of citations may be required at all levels of granularity
- The publisher of the data wants to specify the content of the citation
- Accuracy & conformance of the database to the citation
- Citations may have to conform to some complicated external specification
- New proposals for use of citations (micro-citations, executable papers, …)

Datacite Metadata Schema

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Revision history … 1-->
<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.xsd"/>
        <xs:include schemaLocation="include/datacite-dateType-v3.xsd"/>
        <xs:include schemaLocation="include/datacite-resourceType-v3.xsd"/>
        <xs:include schemaLocation="include/datacite-relationType-v3.xsd"/>
        <xs:include schemaLocation="include/datacite-relatedIdentifierType-v3.xsd"/>
        <xs:include schemaLocation="include/datacite-descriptionType-v3.xsd"/>
        <xs:element name="resource">
                <xs:annotation>
                        <xs:documentation>
        Root element of a single record. This wrapper element is for XML implementation only and is not defined in the DataCite DOI standard.
    Note: This is the case for all wrapper elements within this schema!</xs:documentation>
                        <xs:documentation>No content in this wrapper element.</xs:documentation>
                </xs:annotation>
                <xs:complexType>
                        <xs:all>
                                <!--REQUIRED FIELDS-->
                                <xs:element name="identifier">
                                        <xs:annotation>
                                                <xs:documentation>A persistent identifier that identifies a resource.</xs:documentation>
                                                <xs:documentation>Currently, only DOI is allowed.</xs:documentation>
                                        </xs:annotation>
                                        <xs:complexType>
                                                <xs:simpleContent>
                                                        <xs:extension base="doiType">
                                                                <xs:attribute name="identifierType" use="required" fixed="DOI"/>
                                                        </xs:extension>
                                                </xs:simpleContent>
                                        </xs:complexType>
                                </xs:element>
                                <xs:element name="creators">
                                        <xs:complexType>
                                                <xs:sequence>
                                                        <xs:element name="creator" maxOccurs="unbounded">
                                                                <xs:annotation>
```

...and 400 more lines of XMLSchema

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

```
{ 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 }

*

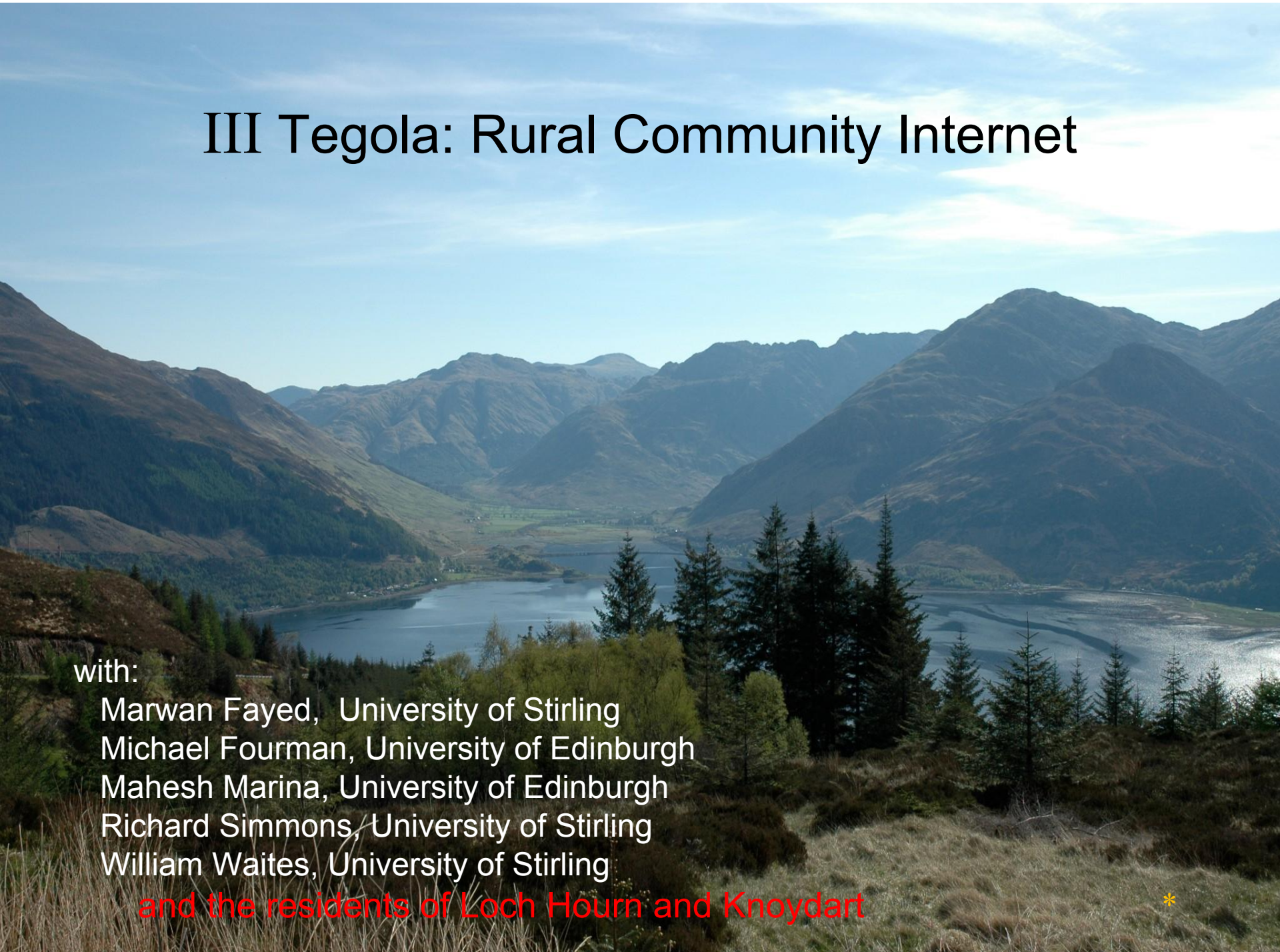
More (standard) database problems

- 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?

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...

III Tegola: Rural Community Internet



with:

Marwan Fayed, University of Stirling

Michael Fourman, University of Edinburgh

Mahesh Marina, University of Edinburgh

Richard Simmons, University of Stirling

William Waites, University of Stirling

and the residents of Loch Hourn and Knoydart *

A Deprived Community

No Workflows
No Active XML
No Probabilistic DBs
No Provenance Semirings
No RDF

*

Rural Internet

The need for fast Interet is *higher* in rural communities than urban ones, but they are the last to be connected.

Businesses: proportion of small businesses *higher* than in urban areas.

Online shopping: even for food.

Other communications often don't work: telephone, radio, TV, mobiles

Alternative to libraries, cinemas, etc.

Education: children on Eigg can lose 3 weeks school a year due to bad weather.

“Distance” communications: vital social importance, especially for the oldies!

Telemedicine & telehealth.

Willie Sandaig “teleconferences” with his grandchildren in New Zealand

Finlay, in Arnisdale, keeps in touch with his pal Sam in Worcester

The Tegola Testbed

Dedicated to research into high-speed, low cost rural Internet

Started 6 years ago with Bernardi and Marina to use commodity wireless hardware for long-range wireless distribution

Research issues:

- Network management **
- Propagation over water
- Power management
- Mast location planning
- Reliability
- Making things *simple*

And the residents of Loch Hourn were willing guinea-pigs

Practical experience

"Masts" need not be masts

"Green" power is neither green nor reliable

Communities and local business can deliver where centralised organisations cannot

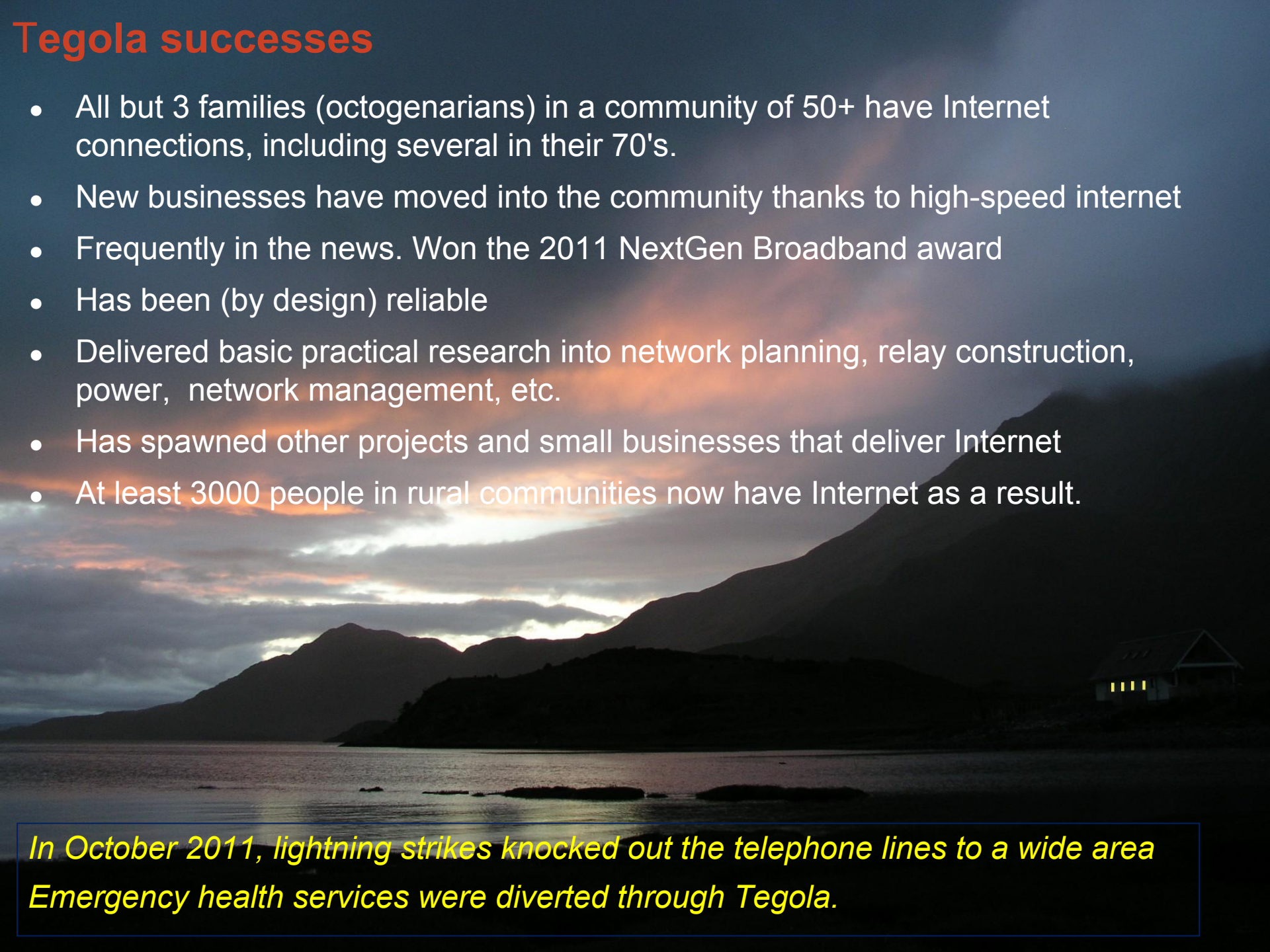
- Rural communities are resourceful
- Travel costs are minimal
- Relay sites can be negotiated by local agreement

Finlay (now age 14)
our on-site engineer

Various transportation systems

Tegola successes

- All but 3 families (octogenarians) in a community of 50+ have Internet connections, including several in their 70's.
- New businesses have moved into the community thanks to high-speed internet
- Frequently in the news. Won the 2011 NextGen Broadband award
- Has been (by design) reliable
- Delivered basic practical research into network planning, relay construction, power, network management, etc.
- Has spawned other projects and small businesses that deliver Internet
- At least 3000 people in rural communities now have Internet as a result.



In October 2011, lightning strikes knocked out the telephone lines to a wide area Emergency health services were diverted through Tegola.

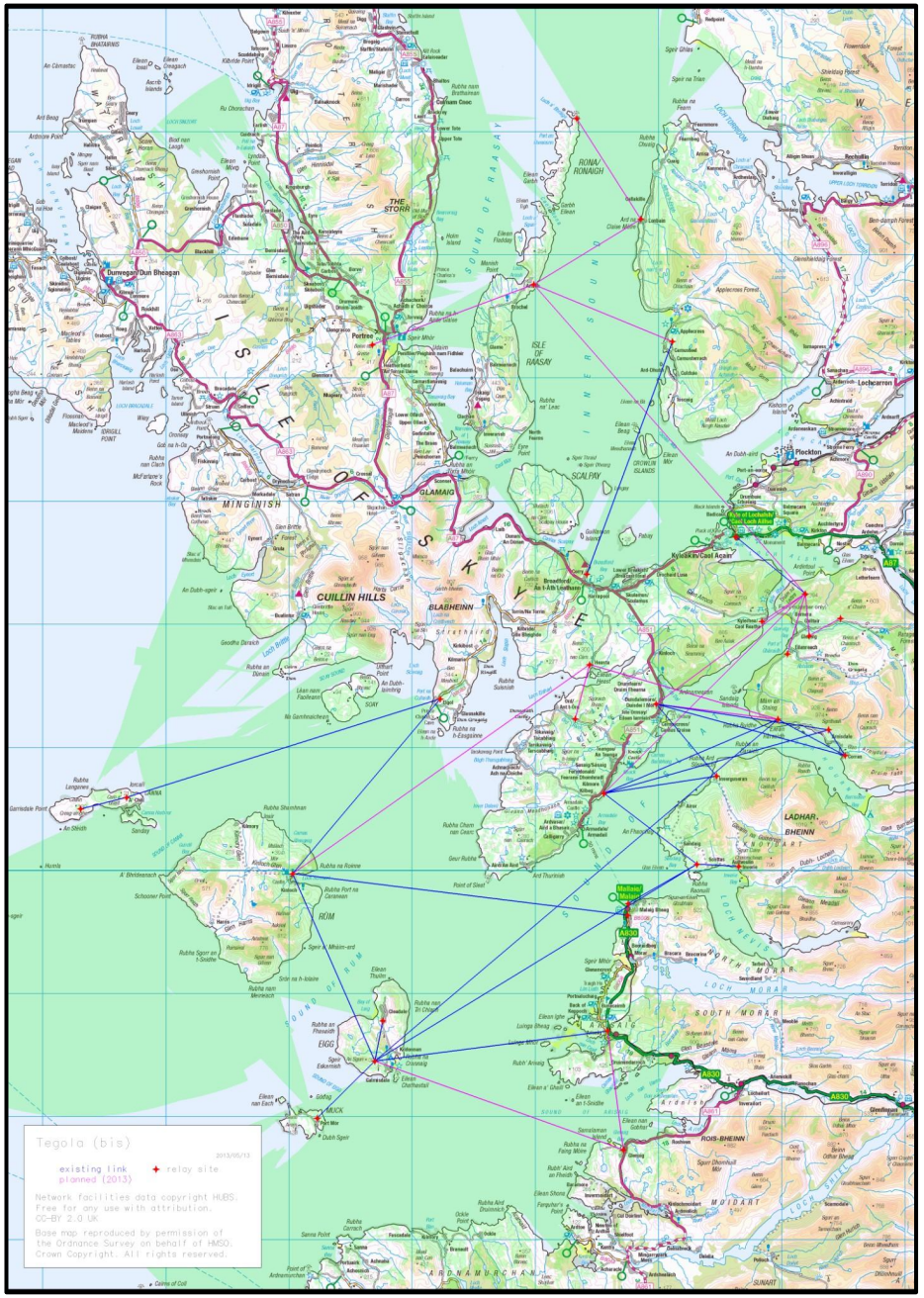
The extended Tegola network (by autumn 2013)

A confederation of seven interconnected community networks created by HUBS

Sharing infrastructure & services for economies of scale and reliability.

Delivering speeds above 30Mb/s .

One of the largest remote community networks in Europe.



What is needed

- An open access fibre backbone
- A light-touch and agile approach to community funding
- Open source, open access community tools for network management -- especially database expertise!

www.tegola.org.uk