

“Semantic Webs” and “The Semantic Web”: **Services, Resources and Technologies for Clinical Care and Biomedical Research**

Alan Rector

School of Computer Science / Northwest Institute of Bio-Health Informatics
rector@cs.man.ac.uk

www.co-ode.org

www.clinical-escience.org

www.opengalen.org

Semantic Web and Webs

▶ The Semantic Web

▶ A Global Information Resource

- ▶ Discoverable
- ▶ Collaborative
- ▶ Trust to be negotiated

▶ Semantic Webs

▶ Resources for Virtual Organisations

- ▶ Discoverable
- ▶ Collaborative
- ▶ Faithful and trusted
- ▶ Interworking

▶ BioMedicine is network of virtual organisations

- ▶ For care
- ▶ For Research

Semantic Web Technology

- ▶ **New ways to deliver information services**
 - ▶ **Service oriented computing**
 - ▶ Easy interworking of heterogeneous systems
 - ▶ *SOAP*
 - ▶ **Semantically rich computing**
 - ▶ **Workflows**
 - ▶ *“Macros on steroids”*
 - ▶ *Discovering appropriate services.*
 - ▶ **Knowledge representation**
 - ▶ *“Ontologies and metadata with everything!”*
 - ▶ *Data on its own means nothing*
- ▶ **New standards for things we have been doing**
 - ▶ **RDF(S), OWL, WSDL, xxML, SCUFL,**
- ▶ **New standard resources**
 - ▶ **Genes, proteins, pathways,**

**... Standards for everything
... and E-Science / E-Health
... and digital libraries
... and ... and**

- ▶ **RDF, RDFS, OWL, SWRL, WSDL, SOAP, ...**
- ▶ **W3C Healthcare and Life Sciences Special Interest Group**
- ▶ **ISO 11179**
- ▶ **Dublin Core**
- ▶ **SKOS**
- ▶ **...**

What about medical standards?

HL7? CEN? ISO? SNOMED? ...?

Do we have to do it on our own?

It's a big open world out there!

...and E-Science / Semantic Grid

▶ E-Science

- ▶ **Large scale collaborative science**
- ▶ **Collections based research**
 - ▶ Using information rather than gathering data
- ▶ **Often Uses Grids but not about Grids**
 - ▶ Image processing, Text mining, Neuro Computing
 - ▶ *Need Cycles and Petabytes*
 - ▶ Workflows, Information organisation, social computing
 - ▶ *Need connectivity & collaboration*

Three themes for this talk

- ▶ **Information discovery**
- ▶ **Joining up healthcare delivery and biomedical research**
- ▶ **Factoring huge problems into manageable chunks**
 - ▶ **Workflows & Service Oriented Architectures**
 - ▶ **Rich semantics, metadata and ontologies**

Theme 1: Discovering Information

- ▶ **Adding meaning**
 - ▶ That machines can process
 - ▶ That people can understand
 - ▶ From specifying “how to do it” to specifying “what to do”
- ▶ **To find it it must be described**
 - ▶ **Metadata & annotations**
- ▶ **To describe it you need a language;**
 - ▶ **RDF(S), OWL, SWRL, ...**
- ▶ **For the language you need words -**
 - ▶ **“Ontologies” and Terminologies**

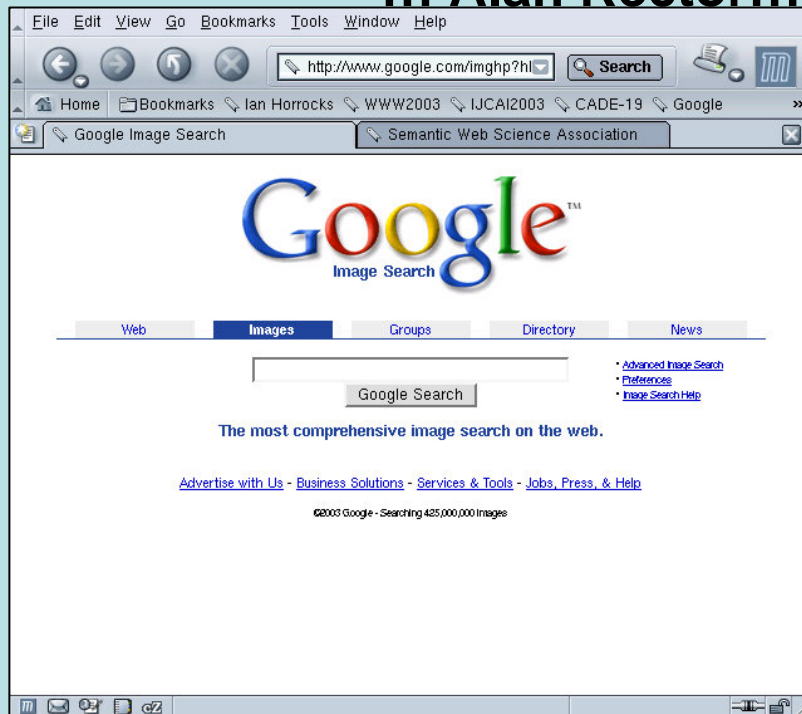
The promise of the *Semantic Web*

The *Syntactic Web* is easily confused...

Find images of Steve Furber
Carole Goble



... Alan Rector ...



Rev. Alan M. Gates, Associate Rector of the Church of the Holy Spirit, Lake Forest, Illinois

What information can we see...

WWW2002

The eleventh international world wide web conference

Sheraton waikiki hotel

Honolulu, hawaii, USA

7-11 may 2002

1 location 5 days learn interact

Registered participants coming from

australia, canada, chile denmark, france, germany, ghana, hong kong, india,
ireland, italy, japan, malta, new zealand, the netherlands, norway,
singapore, switzerland, the united kingdom, the united states, vietnam,
zaire

Register now

On the 7th May Honolulu will provide the backdrop of the eleventh
international world wide web conference. This prestigious event ...

Speakers confirmed

Tim berners-lee

Tim is the well known inventor of the Web, ...

Ian Foster

Ian is the pioneer of the Grid, the next generation internet ...

Solution: XML markup with “meaningful” tags?

<name> 



 **</name>**


<location>  

 **</location>**


<date>   **</date>**

<slogan>   **</slogan>**

<participants>  



<introduction> 



 **</introduction>**

<speaker>  **</speaker>**

<bio>  **</bio>...**

Need to Add “Semantics”

- ▶ **Annotations**
 - ▶ **In languages that machines can process**
 - ▶ **Using terminology that people have agreed & machines can process**

Competitive/Complementary Technologies - machine learning & text mining



National Centre for Text Mining (NaCTeM)

<http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/>

Content viewer 1

Settings Help

Symbol GE NE sli-1

Name suppressor of LIneage defect SLI-1, homolog of mammalian oncogene c-cbl, adaptor-like E3 ubiquitin ligase (66.1 kD) (sli-1)

Organism Caenorhabditis elegans

Link DB

Synonym CELK03228 / M02A10.3a / M02A10.3b

Product

My Folder 1

Folder Help

Item	Memo
GE NE sur-5	+memo
GE NE sli-1	+memo
GE NE BRAF	+memo

Gene Dictionary

Search for genes or protein names

>> Organism >> Field >> Help

Search for **raf1** : 1 -- 14

You can drag symbol names into other windows.

Symbol	Name	Synonym	Product
GE NE lin-45	related to oncogene RAF, abnormal cell LIneage LIN-45 (95.0 kD) (lin-45)	raf-1	
GE NE Y73B6A.5		raf-1	
GE NE Dsor1	Downstream of raf1		
GE NE phi	pole hole	D- raf1 Raf1 raf-1 raf1: raf-like-oncogene-1	
GE NE BRAF	v-raf murine sarcoma viral oncogene homolog B1	B- raf 1 B- raf -1	

Content Viewer 2

Settings Help

Symbol GE NE sur-5

Name

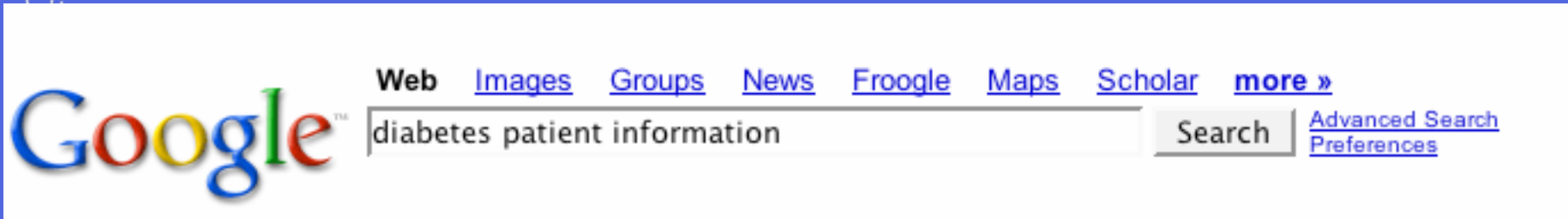
Organism Caenorhabditis elegans

Link DB

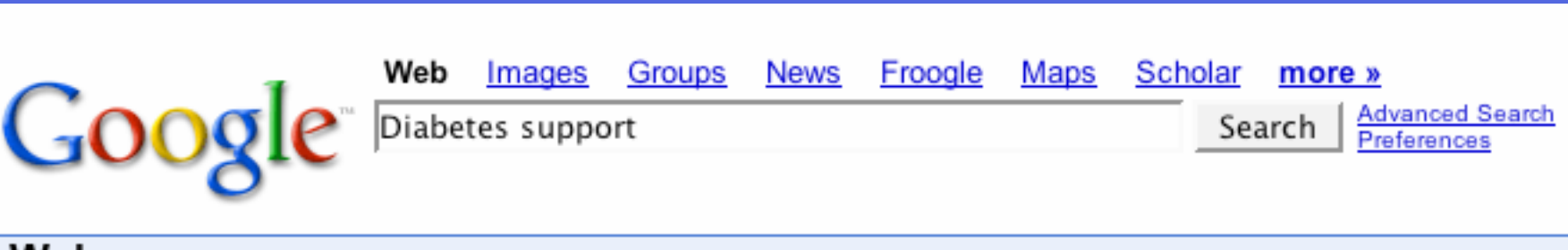
Synonym CELK01685 / K03A1.5

Product

Or web mining there's no lack of text out there



Results 1 - 10 of about 41,600,000 for [diabetes patient information](#). (0.49 seconds)



Results 1 - 10 of about 64,300,000 for [Diabetes support](#). (0.87 seconds)



Creative Commons Search

Full copyright applies to most stuff on the web. But this search helps you find photos, music, text, and other works whose authors want you to re-use it for some uses -- without having to pay or ask permission. ([More Info](#))

Diabetes support

Search with:

[\[More Info\]](#) [\[Forum\]](#) [\[Help\]](#)

optional

- Find me works I can use even for commercial purposes.
 Find me works I can modify, adapt, or build upon.

optional (nutch only)

Format:

Hits 1-10 (out of about 361 total matching pages):

[BioMed Central | Abstract | Systematic reviews of epidemiology in diabetes: finding the evidence](#)

... of epidemiology in diabetes: finding the evidence ...
 (v) <http://www.biomedcentral.com/1471-2288/5/2/abstract> ([more from www.biomedcentral.com](#))

[Lipids in Health and Disease | Full text | Relationship between Sialic acid and metabolic variables](#)

... the development of diabetes [2]. Diabetes is another risk factor for ... microvascular, and type-2 diabetes ...
 (v) <http://www.lipidworld.com/content/4/1/15>

[Cardiovascular Diabetology | Full text | Hypertension control: results from the Diabetes Care Progra](#)

... decade in patients with diabetes mellitus attending Diabetes Centres in the ... results from the ...
 (v) <http://www.cardiab.com/content/4/1/11>

[Health and Quality of Life Outcomes | Full text | Response shift and glyceimic control in children wi](#)

... he joins a diabetes support group and meets ... HbA1c, duration of diabetes, or ...
 (v) <http://www.hqlo.com/content/3/1/38>

Key:

- work is in the public domain
- must give attribution
- can't use commercially
- can't make derivatives
- must sharealike (use same license)
- sampling license
- sampling+ license

Web-Discovery of information

▶ Four competing technologies

▶ Semantic Web

- ▶ Or hand built ontologies
 - ▶ *OBO, FMA, SNOMED? , other ...*

▶ Social computing

- ▶ Open Directory, Wikipedia, FLIKR, FoF, ...

▶ Web mining

- ▶ Google (& other web search)

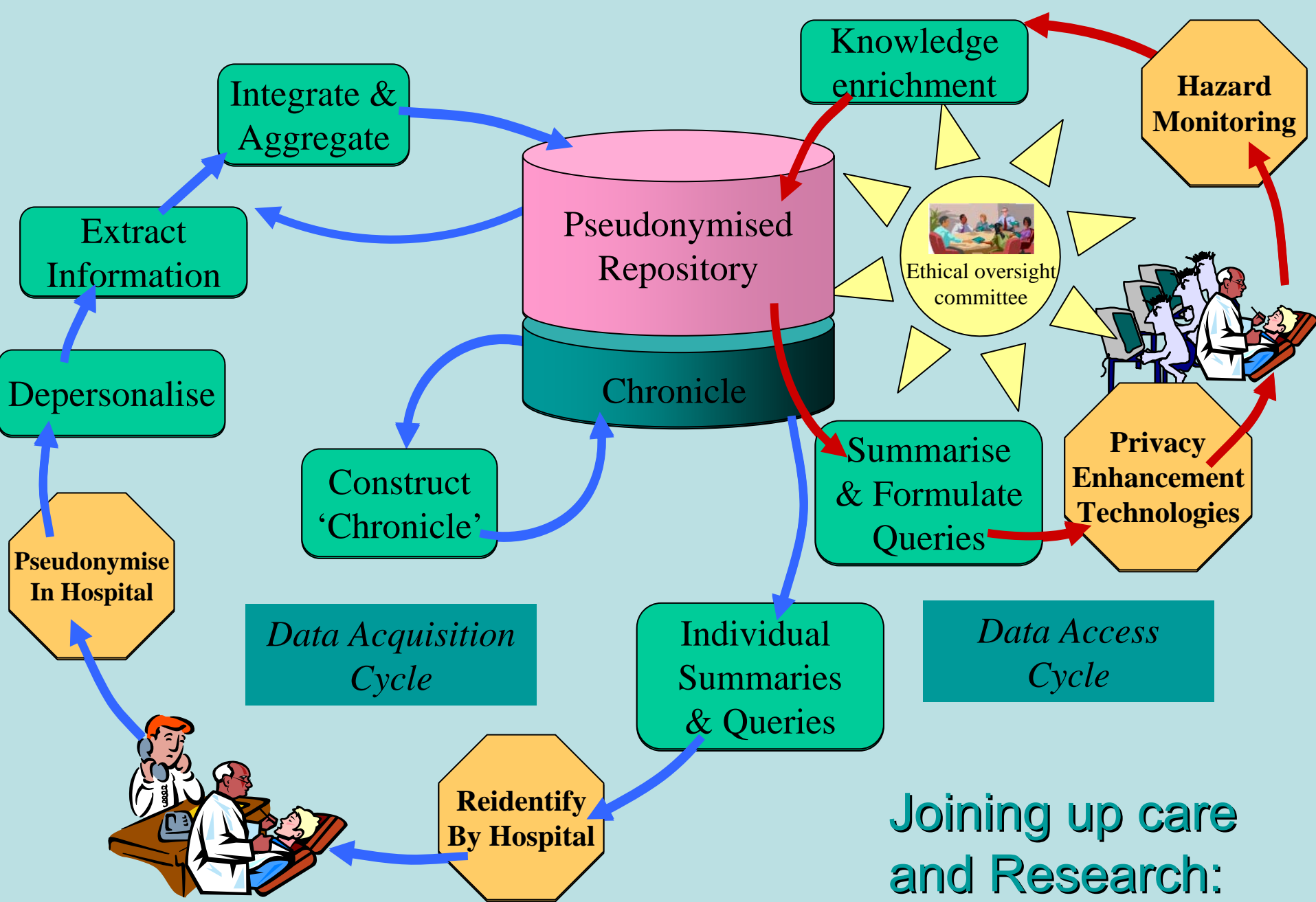
▶ Text mining

- ▶ Just becoming widely available, especially in biology
 - ▶ *All of pubmed abstracts about to be minable for relations*
 - ▶ *National Centre for Text Mining - NaCTeM*

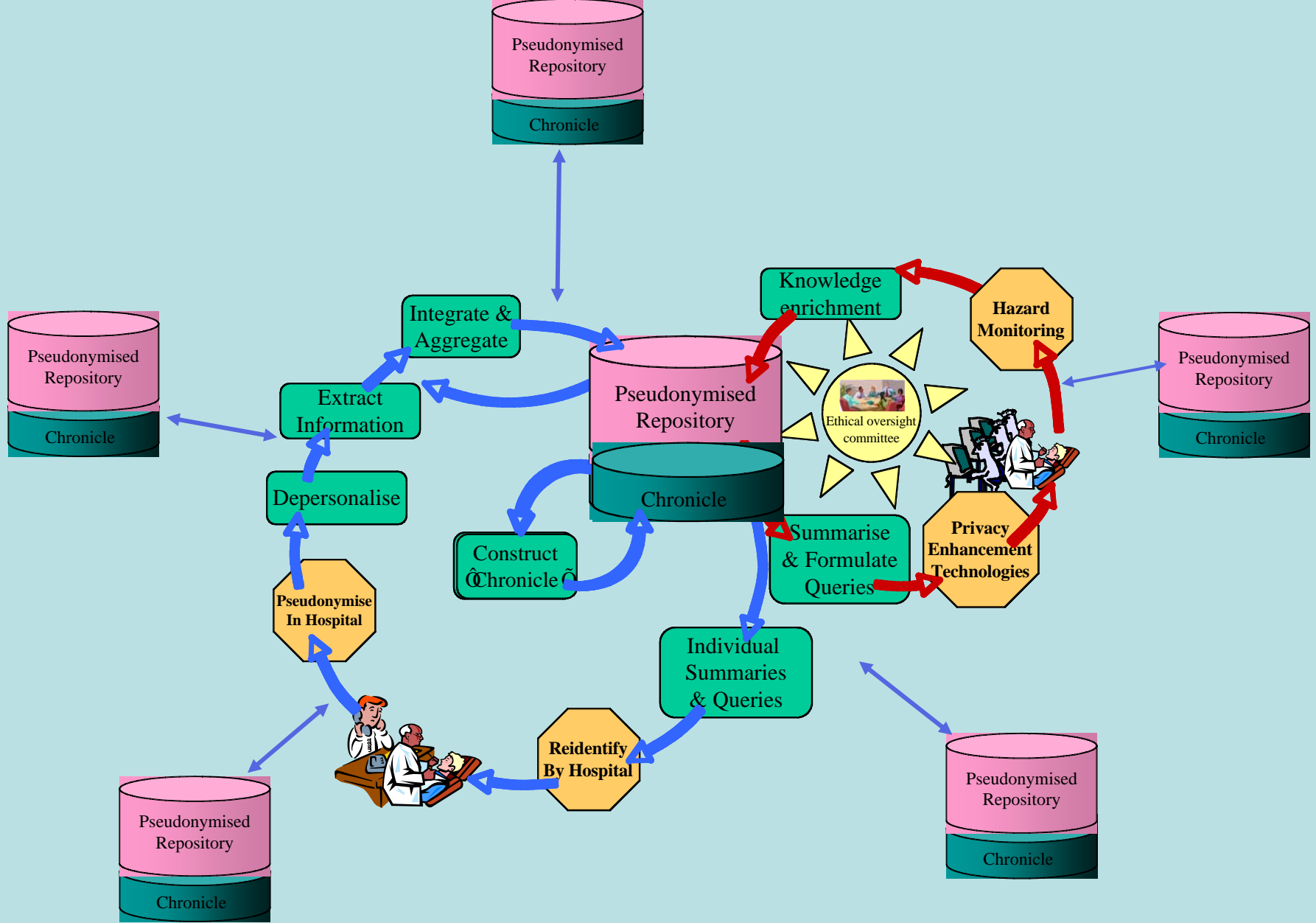
Theme II: Joining up healthcare delivery and Biomedical Research

The CLEF Vision

www.clinical-escience.org

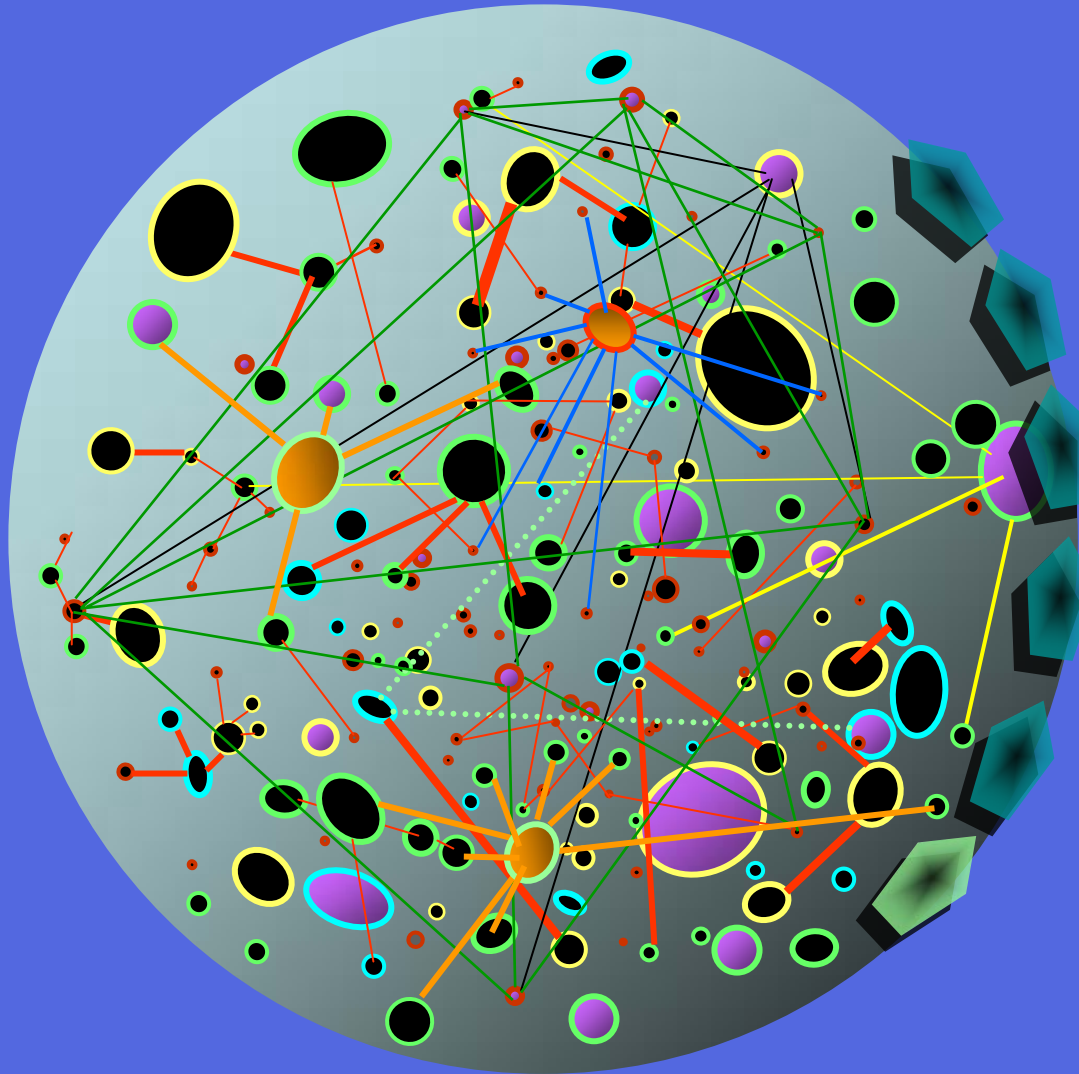


Joining up care and Research:
The CLEF Vision



The Chronicle

- ▶ A semantically rich summary of our best understanding of the patient
 - ▶ Inferred from data and metadata
 - ▶ Combined from many sources on semantic webs



(Increasing detail)

Low haemoglobins
over a period =
anaemia

Coreferences

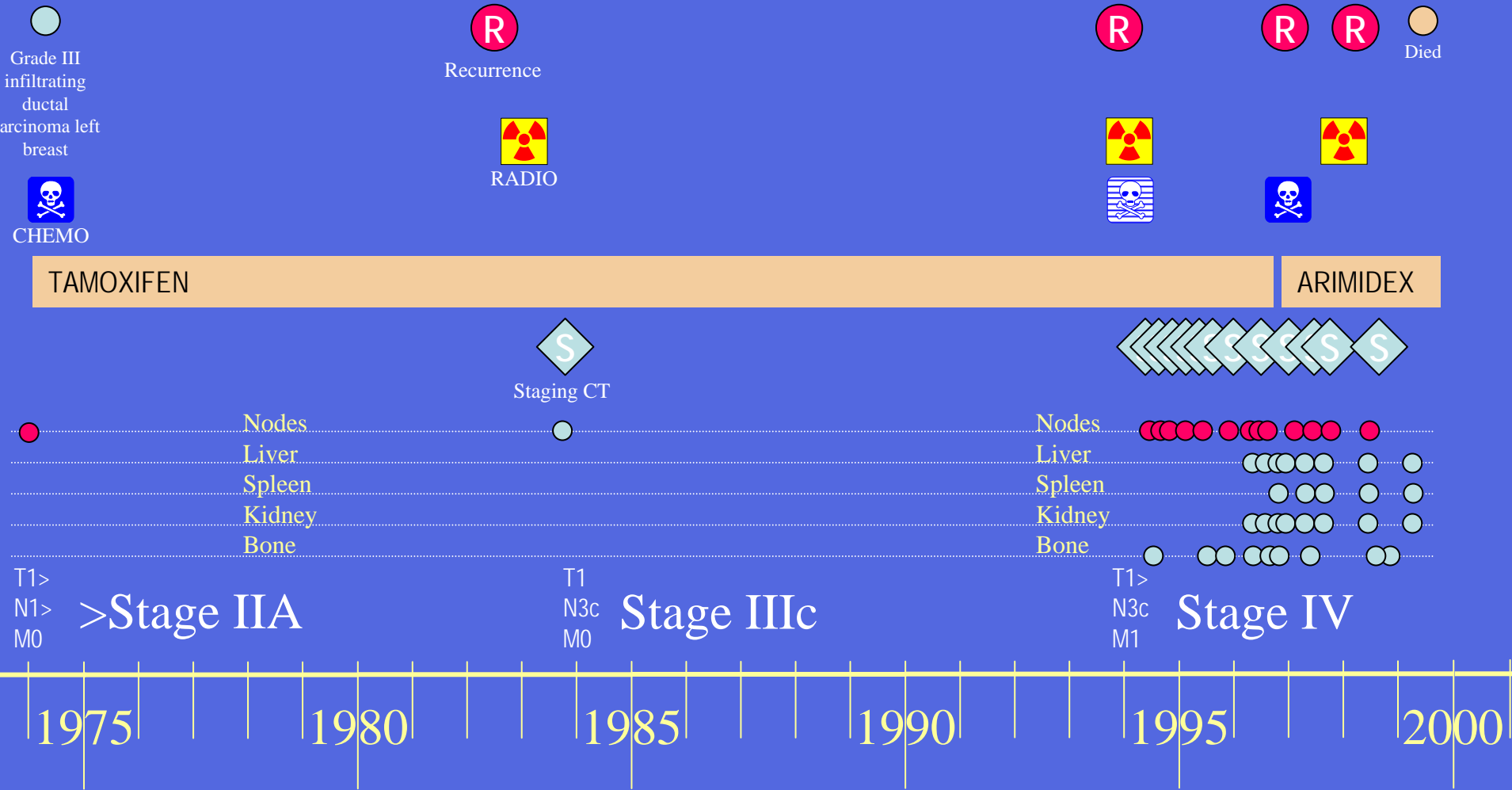
Time

Clinical pragmatics

Simplification

Abstraction

Inferred best view of the patient history - from whatever sources - the CLEF Chronicle



Privacy and Security

- ▶ The great barrier to clinical use
- ▶ Web/Grid security a key topic
 - ▶ For policy
 - ▶ How safe is safe?
 - ▶ What is the risk from medical information
 - ▶ *Your credit card company knows how much you drink!*
 - ▶ What counts as informed consent? Consent for what?
 - ▶ Benefits vs risks
 - ▶ Technology
 - ▶ Authentication - who are you?
 - ▶ Authorisation - what are you doing?
what are you allowed to do in that role?
 - ▶ Accounting - who pays? How much?

Theme III: Factoring huge problems

- ▶ **Medicine is big and complicated**
 - ▶ & full of niches
- ▶ **How to beat the combinatorial explosion**
- ▶ **Workflows**
 - ▶ *myGrid* & Taverna
- ▶ **Ontologies**
 - ▶ Protégé & CO-ODE
 - ▶ www.co-ode.org

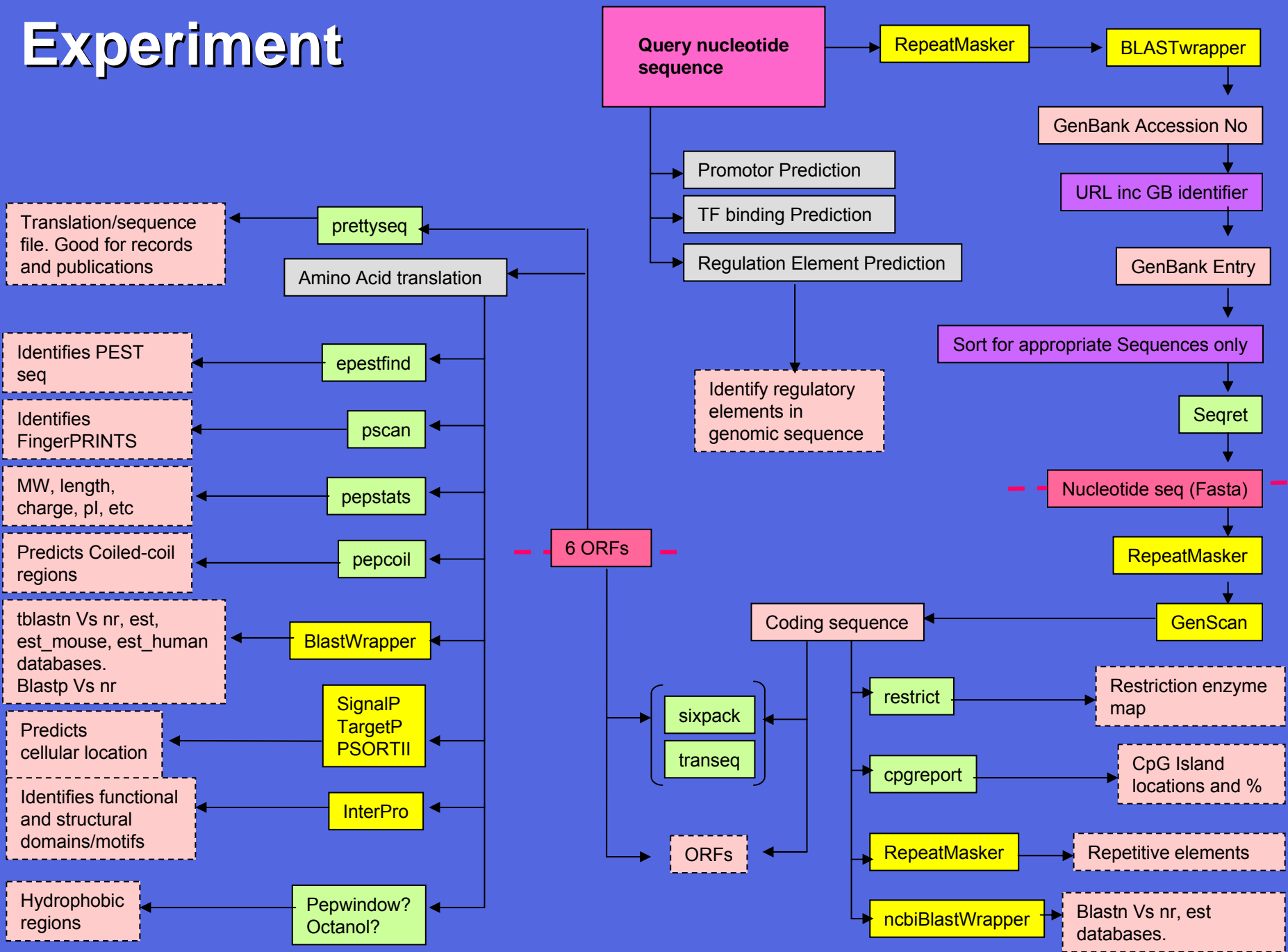
New ways of factoring problems

- ▶ Better ways to build from “Lego”
- ▶ Better ways of indexing and cataloguing
- ▶ Keys
 - ▶ Rich semantics
 - ▶ Discover rather than call
 - ▶ *Machine undersatndable*
 - ▶ Service oriented architectures
 - ▶ Workflows
 - ▶ Metadata and Provenance
 - ▶ Data on its own is meaningless
 - ▶ *What is in the repository?*
 - ▶ *What studies have used it?*
 - ▶ *What is known of its reliability?*
 - ▶ *???...???...???*
 - ▶ Terminology and ontology

Workflows in Biomedical Research

- ▶ **“Macros on steroids”**
 - ▶ **Specify what rather than how**
 - ▶ Describe the resources and tasks (RDF, WSDS, ...)
- ▶ **Break big problems down into little steps**
- ▶ **Reduce effort from days to hours for bioscientists**
 - ▶ **Can we move them to medical care**

Experiment



Analysis via 'Cut and Paste'

The image displays a workflow for genomic analysis using various web-based tools. Red arrows indicate the flow of data from one tool to another. The central DNA sequence is:

```

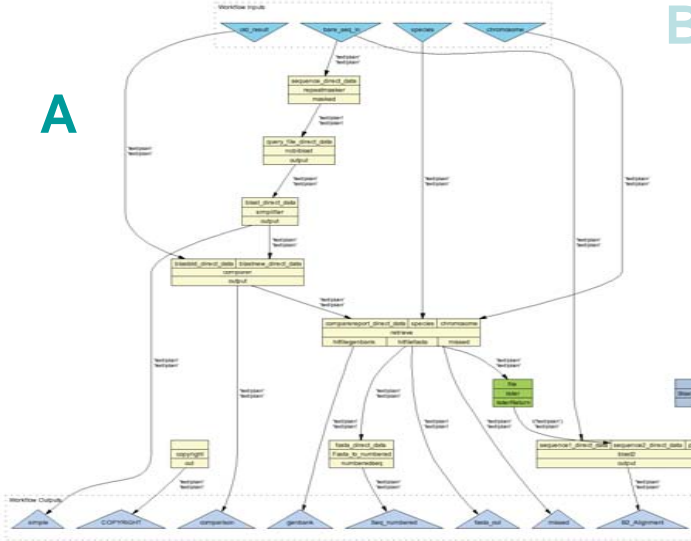
caattccac caacagtgga ttgagttgtt ggtctatggt caccacat
tgtt 12241 cagtccttca aatttaaac ttagagaga agcatacac
cot tttttagctt 12301 gaccatocca atagatacac agtgggtct
att ttaatttcca tttctcgt 12361 gactaattt gttgagcttg
tta gacaactca tttagagaagt gctaatatt 12421 tagtgactt
ttt ttttaattgg gatctcaatt tttttaatt attgattttg 12481
gagctatt tatatatct agatcaagt ttttatacag atacaagtt tggaa
541 tcttataag cctgtgggtt ttatattaat gttttattg tgaacttt
tacaattg 12601
  
```

The tools shown include:

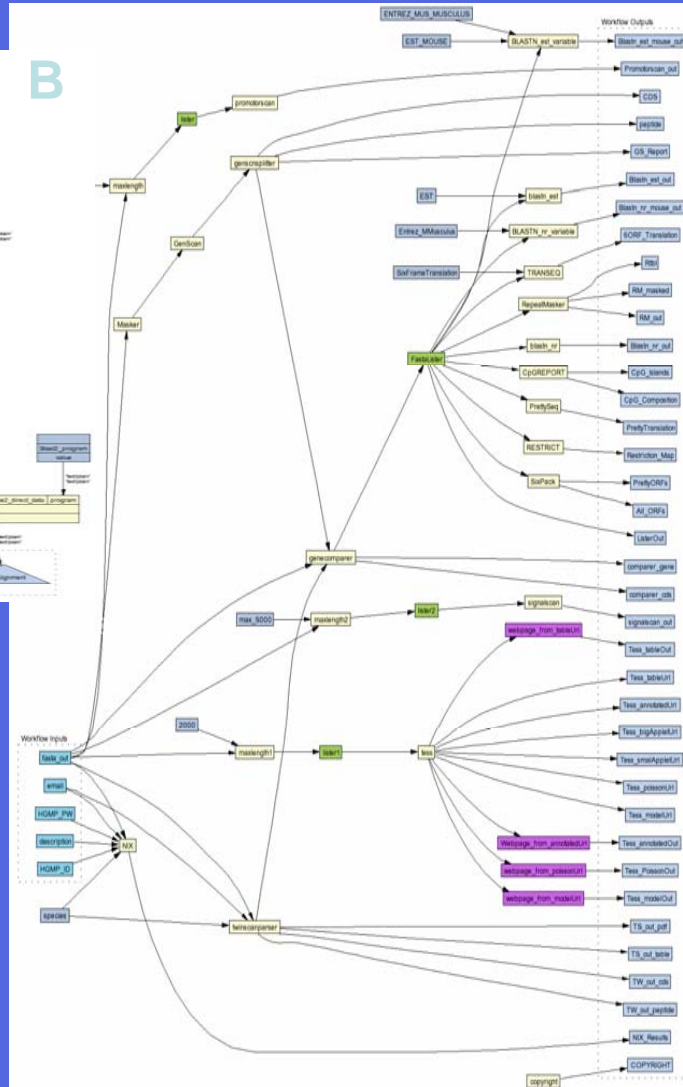
- AHBaba2.1**: A web server for identifying complete gene structures in genomic DNA.
- GENSCAN**: A web server for identifying complete gene structures in genomic DNA.
- TWINSKAN**: A web server for identifying complete gene structures in genomic DNA.
- BLAST**: A web server for searching nucleotide sequence databases.
- SignalP 3.0**: A web server for predicting signal peptides.
- iPSORT**: A web server for predicting protein sorting signals.
- RepeatMasker**: A web server for identifying and masking repetitive elements.
- EMBL-EBI**: The European Bioinformatics Institute.

Workflows

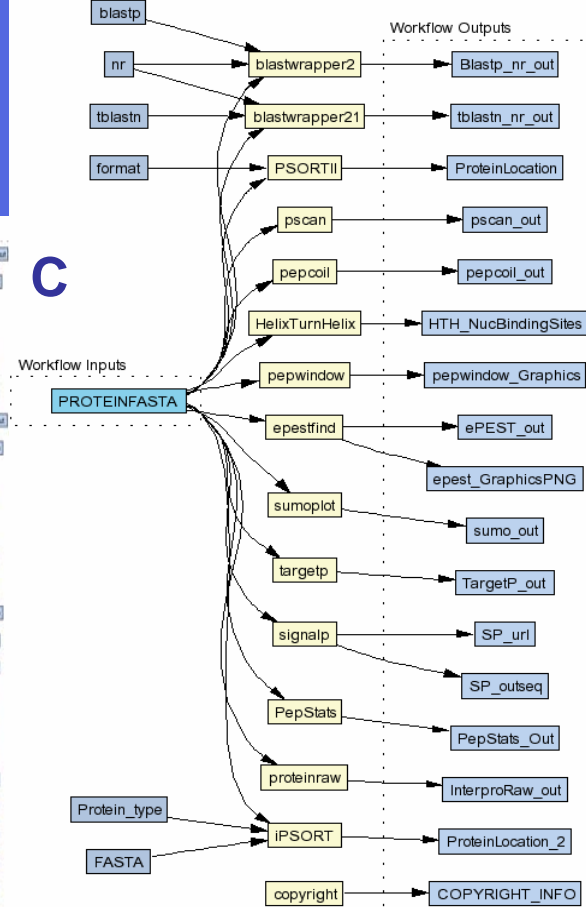
A



B



C



- A: Identification of overlapping sequence
- B: Characterisation of nucleotide sequence
- C: Characterisation of protein sequence

Description needs a language: Ontologies and Terminologies

- ▶ **Biologists manage quite well**
 - ▶ **Open Biological Ontologies**
 - ▶ The Gene Ontology, Micro-array / Gene Expression Database, etc.
 - ▶ **Little legacy**
 - ▶ It all started in 1980
 - ▶ **Fanatically open and collaborative**
- ▶ **Medicine has chaos and “the coding wars”**
 - ▶ **SNOMED (International,-RT, -CT), ICD, LOINC, DICOM, MEDDRA, NCI, ICPC, Read/CT (v1,v2, & v3), GALEN, NANDA,...**
 - ▶ It all started in 1880
 - ▶ **Closed and proprietary**

No longer a unique problem

New standards and interest

- ▶ Logicians and Computer Scientists from the mainstream
 - ▶ OWL, RDF, ...
- ▶ Ontologists from Philosophy
 - ▶ 3000 years of analysis
 - ▶ much of which is relevant
- ▶ ...but medicine is big and complicated
 - ... and combinatorially explosive
- ▶ A prime source of combinatorial explosions

Defusing the exploding bicycle: **500 codes in pieces**

- ▶ **10 things to hit...**
 - ▶ **Pedestrian / cycle / motorbike / car / HGV / train / unpowered vehicle / a tree / other**
- ▶ **5 roles for the injured...**
 - ▶ **Driving / passenger / cyclist / getting in / other**
- ▶ **5 activities when injured...**
 - ▶ **resting / at work / sporting / at leisure / other**
- ▶ **2 contexts...**
 - ▶ **In traffic / not in traffic**

V12.24 Pedal cyclist injured in collision with two- or three-wheeled motor vehicle, unspecified pedal cyclist, nontraffic accident, while resting, sleeping, eating or engaging in other vital activities

Conceptual Lego... it could be...

Goodbye to picking lists...

Structured Data Entry

File Edit Help

Cycling Accident

What you hit

Your Role

Activity

Location

The image shows a software window titled 'Structured Data Entry' with a menu bar (File, Edit, Help) and standard window controls. The main content area is titled 'Cycling Accident' and is organized into four rows, each with a category label on the left and a set of icons on the right. The 'What you hit' row contains icons for a bicycle, motorcycle, car, truck, train, tree, and person. The 'Your Role' row contains icons for a driver, a person in a car, a cyclist, and a taxi driver. The 'Activity' row contains icons for a person in bed, a person at a desk, a soccer player, and a painter. The 'Location' row contains icons for a traffic jam, a street intersection, a parking sign with a car, and a soccer field.

And generated language

Summary

Moderately severe angina pectoris for 1 day, getting worse
Rapid onset, moderately severe, pressing pain in left chest and sternal region present

On Examination

Cardiovascular system -

Slightly raised JVP

1st and 2nd heart sounds normal

No added heart sounds

Pulse rate 104 per minute

Blood pressure 138/90 mm Hg

Semantic Technology: Logic as the clips for “Conceptual Lego”



hand

extremity

body

chronic

acute

abnormal

normal

ischaemic

deletion

polymorphism

mucus

gene

protein

polysaccharide

cell

expression

Lung

infection

inflammation

bacterium

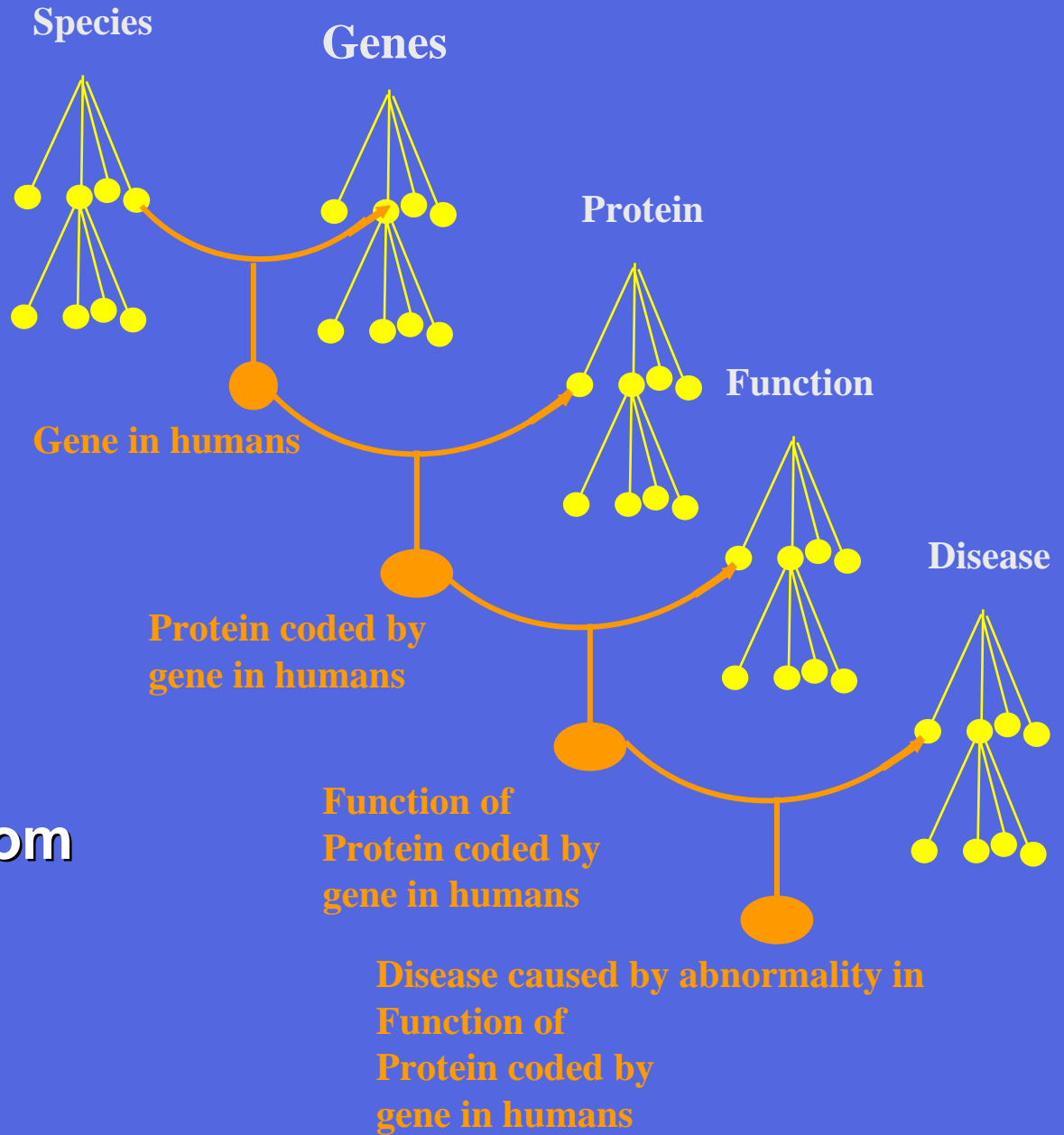
virus

Logic as the clips for “Conceptual Lego”

“*SNPolymorphism* of *CFTRGene* causing *Defect in MembraneTransport* of *Chloride Ion* causing *Increase* in *Viscosity* of *Mucus* in *CysticFibrosis*...”



“*Hand* which is
anatomically
normal”



Build complex representations from modularised primitives

**...but whatever the technology,
how will people interpret it?**

Inter-rater variability



ART & ARCHITECTURE THESAURUS (AAT)

Domain: art, architecture, decorative arts, material culture

Content: 125,000 terms

Structure: 7 facets, 33 polyhierarchies

Associated concepts (*beauty, freedom, socialism*)

Physical attributes (*red, round, waterlogged*)

Style/Period (*French, impressionist, surrealist*)

Agents: (*printmaker, architect, jockey*)

Activities: (*analysing, running, painting*)

Materials (*iron, clay, emulsifier*)

Objects: (*gun, house, painting, statue, arm*)

Synonyms

Links to 'associated' terms

Access: lexical string match;
hierarchical view

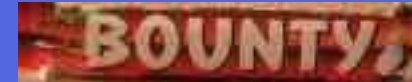
The “coding wars”: UMLS helps

- ▶ US National Library of Medicine
- ▶ *De facto* common registry for vocabularies
- ▶ Metathesaurus
 - ▶ 1.8 million concepts
 - ▶ categorised by semantic net types
- ▶ Semantic Net
 - ▶ 135 Types
 - ▶ 54 Links
- ▶ Specialist Lexicon
- ▶ Now a key web resource
 - ▶ Source of reference IDs
 - ▶ CUIs and LUIs
 - ▶ *LSIDs elsewhere in biology*

...but cultural differences can still catch you out:

An international conversion guide

SNOMED-CT	Term	CTV3
C-F0811	Bounty bar	UbOVv
C-F0816	Crème egg	UbOW2
C-F0817	Kit Kat	UbOW3
C-F0819	Mars Bar	UbOW4
C-F081A	Milky Way	UbOW5
C-F081B	Smarties	UbOW6
C-F081C	Twix	UbOW7
C-F0058	Snickers	Ub1pT



Creating open distributed communities

Open 'Just-in-time Development using Semantic Webs

- ▶ **Open just-in-time development**
 - ▶ For professionals
 - ▶ For patients
 - ▶ For public
 - ▶ By health informaticians
- ▶ **Social development**
 - ▶ By & for professionals
 - ▶ By & for patients
 - ▶ By & for public
 - ▶ By & for health informaticians

Critical for everything: Human Factors

Helping with a humanly impossible task

- ▶ **Doing the right thing**
 - ▶ As well as doing it right
- ▶ **Useful and usable applications**
 - ▶ Useless cleverness is easy & fun

*Requires **serious** investment and
Commitment*

Summary: The Semantic Web & Semantic Web/Grid Technology

- ▶ **Web or Webs**
 - ▶ **New methods**
 - ▶ *Discovery*
 - ▶ *Cooperation*
 - ▶ For the world
 - ▶ For virtual organisations
 - ▶ **Scaling up to medicine**
 - ▶ *Better ways to factor problems*
 - ▶ *Services rather than programs and data*
- ▶ **Depends on shared meaning & semantics**
 - ▶ **RDF, RDFS, OWL, WSDL, SWRL,**
- ▶ **Joining up care & research**
- ▶ **Human factors**

