Whose knowledge? Whose questions? Whose answers?

Conflicting Preconceptions for Knowledge Intensive Systems

Alan Rector
School of Computer Science, University of Manchester
Northwest institute of BioHealth Informatics

rector@cs.man.ac.uk
www.clinical-esciences.org
www.co-ode.org

© University of Manchester

with thanks to
Jeremy Rogers, Anthony Nowlan,
and the SWBP, OWL, CO-ODE & Protégé teams
This should be the age of knowledge acquisition

- The Web
  - Social computing & Web II
  - The Semantic Web/Grid

- Service oriented Architectures and discovery
  - Digital libraries, repositories, ...
  - Semantic computing

- “Ontologies with everything”

- ...

“... but couldn’t we just get back to 1985”
• **Problem space**
  - Answer questions
  - Advising on actions
  - Hazard monitoring
  - Creating forms
  - Discovering resources
  - Constraint actions
  - Assess risk
  - …

• **Solution space**
  - Logics
  - Rules
  - Frames
  - Planners
  - Logic programming
  - Bayes nets
  - Decision theory
  - Fuzzy sets
  - Open / closed world
  - …
• **Solution space**
  - Logics
  - Rules
  - Frames
  - Planners
  - Logic programming
  - Bayes nets
  - Decision theory
  - Fuzzy sets

• **Problem space**
  - Answer questions
  - Advising on actions
  - Hazard monitoring
  - Creating forms
  - Discovering resources
  - Constraint actions
  - Assess risk
  - ...
Problem space & solution space

Problem space

Guidelines, Patterns, Tools,...
Knowledge Acquisition

Solution space
The knowledge principle

• **A system exhibits intelligent understanding and action at a high level of competence primarily because of the specific knowledge that it contains about its domain of endeavor.**

• Knowledge comes from domain experts: *Derived from and delivered to the problem space*
We argue that logical soundness, completeness, and worst-case complexity are inadequate measures for evaluating the utility of representation services, and that this evaluation should employ the broader notions of utility and rationality …”

- Two theses of knowledge representation: Language restrictions, taxonomic classification and the utility of representation services.
Where I come from

Clinical Terminology

Data Entry

Clinical Record

Clinical research & Decision Support

Best Practice

GALEN Ontologies & Description logics

Data Entry

Electronic Health Records

Clinical research Decision Support & Knowledge Presentation

Best Practice

Mr. Ivor Bigun

Dun Roamin

Anytown

Any country

4431 3654 90273

Clinical research

Decision Support & Knowledge Presentation
By way of User Centred Design
Ontologies were a solution, not a goal
... and now to

- **BioOntologies, Anatomy, and Molecular Pathways**
  - Capturing biologists knowledge
  - Reconciling diverse views / Avoiding ontological dogmatism
  - Integrating ≥500 sources

- **Healthcare record standards: Data models and Terminology**
  - Improving the rigour of “UML” based message standards
  - Binding of terminology to data models

- **Medical terminologies / ontologies**
  - GALEN, SNOMED, NCI thesaurus

- **Forms generation for clinical trials, data entry & entry and decision support.**
  - PEN&PAD Revisited

- **User oriented Ontology Tools - CO-ODE/Protégé-OWL**

- **Semantic Web Best Practice WG & Healthcare & Life Sciences SIG**
  (& campaigner for OWL 1.1 and beyond)
Some Recent User Experiences

- Two companies building clinical systems
  - Intelligent forms creation and constraint management
    - Mastering a combinatorial explosion of $\geq 10^7$
  - Workshops, Training and subcontracting

- NHS National Programme for IT:
  - Additional constraints to UML and between UML and terminology/ontology
    - Different implementation of the standard don’t fit
  - Managing SNOMED
    - Large $\geq 450K$ ters DL based terminology

- Biologists research groups
  - Many clinics and workshops
    - Indefinitely many emails on HCLSig

- The Anatomy community
  - Workshops towards a common vertebrate anatomy

- “Can’t we just get back to 1985”
What are my prejudices?

• “First do the right thing; then do it right”
  – and hope that by the time you know what to do someone else can show you how

• “There is no one way!”
  – “Society for the Promotion of Secular Ontology”

• “Better” is better than “best”
  – The best is often the enemy of the good

• Unobtrusive assistance is better than advice
  – “Shedding light” rather than “uttering prophecies”

• “It doesn’t make the coffee!”
  – Nothing does everything - the secret is in factoring the solution space to fit the problem space
Who are my users?

- Intermediate users
  - Clinical terminologies developers
    - GO, SNOMED, GALEN, CCAM, FMA, …
  - IT staff configuring clinical systems
    - NHS, Healthcare information systems staff
  - Standards developers
    - BioPAX, HL7, ??ML,…,
- Near to end users
  - Bioinformaticians
  - Trialists and other specialist healthcare informaticians
- End users
  - Biolomedical scientists
  - Clinicians, carers and patients
The rest of the talk

• Our experiences
  – Our original application: Fractal tailoring of data input forms
  – Building large terminologies
    • Logic as the clips for “Conceptual Lego”
    • Untangling and normalisation
  – Working with biologists

• Observations on Knowledge Resources and “Ontologies”
  – Things we have found
  – Things we need
Where we started: A Vision

- Useful and Usable clinical systems
  - Information when and where you need it
    - to make better decisions
      - “Shedding light”
      - “Cuing”
    - In your own language
- Aggregated data as a byproduct of clinical
- …without getting caught in the combinatorial explosion
Clinergy/ Pen&PAD

Location None

- angina pectoris
- congestive cardiac failure
- left ventricular failure
- myocardial infarction
- pulmonary embolus

- anaphylactic shock
- anticoagulant therapy
- aortic aneurysm
- aortic incompetence
- aortic stenosis
- arrhythmia
- arterial thrombosis
- arteriosclerosis
- Buerger's disease
- cardiac failure
Forms

**Descriptors**
- **Presence**: present, possible, excluded
- **Duration**: 1 day
- **Severity**: mild, moderate, severe
- **Progress**: better, same, worse

**Associated Symptoms**
- Chest pain: absent, present, more...
- Arm pain: absent, present, more...

**Further History**
- FH, PMH, SH

**Examinations**
- **Examination**: normal
- **Cardiovascular system**: normal, more...
- **Pulse rate**: 70 per minute
- **Blood pressure**
  - Systolic BP: 120 mm Hg
  - Diastolic BP: 80 mm Hg

**Diagnoses**

**Interventions**
More on Chest pain

Descriptors

- **Chest pain**: absent, present
- **Sublocation**: left chest, right chest, clavicular region, sternal region, anterior ribs region, posterior ribs region, sternoclavicular joint, costochondral joint
- **Duration**: number of days
- **Character**: aching, piercing, dull, throbboning, pressing, burning, crushing
- **Severity**: mild, moderate, severe
- **Onset**: gradual, rapid, sudden

Additional forms
Generated Text

One of the great missed opportunities today

See Donia Scott’s work at OU

And automatic (Read) code generation
But there are combinatorially many possible forms

- The PROMIS system was discontinued in 1987 at 70K forms and accelerating
  - A major influence on our work
    - For us KR is a way to avoid combinatorial explosions
- A supplier recently expanded all XML forms for data capture in one clinic to $10^7$
Fractal tailoring of reusable resources: example of data collection forms for trials
Problems:
The exploding bicycle
(codes for injuries involving cyclists)

- 1972 ICD-9 (E826) 8
- READ-2 (T30..) 81
- READ-3 87
- 1999 ICD-10 ……

ICD = International Classification of diseases
• V31.22 Occupant of three-wheeled motor vehicle injured in collision with pedal cycle, person on outside of vehicle, nontraffic accident, while working for income
• W65.40 Drowning and submersion while in bath-tub, street and highway, while engaged in sports activity
• X35.44 Victim of volcanic eruption, street and highway, while resting, sleeping, eating or engaging in other vital activities
Building large Terminologies Logic 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 Membrane Transport of Chloride Ion causing Increase in Viscosity of Mucus in Cystic Fibrosis…”

“Hand which is anatomically normal”
Build complex representations from modularised primitives
Normalising (untangling) Ontologies

Structure

Part-whole

Function

Part-whole
Untangling and Enrichment
Using a classifier to make life easier

<table>
<thead>
<tr>
<th>Substance</th>
<th>Role</th>
<th>Values From</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein</td>
<td>- Hormone</td>
<td>- Insulin</td>
</tr>
<tr>
<td>Protein-Hormone</td>
<td>- Enzyme</td>
<td>- ATPase</td>
</tr>
<tr>
<td>Steroid</td>
<td>- Hormone</td>
<td>- Cortisol</td>
</tr>
<tr>
<td>Steroid-Hormone</td>
<td>- Enzyme</td>
<td>- ATPase</td>
</tr>
<tr>
<td>Insulin</td>
<td>- Enzyme</td>
<td>- ATPase</td>
</tr>
<tr>
<td>Cortisol</td>
<td>- Enzyme</td>
<td>- ATPase</td>
</tr>
</tbody>
</table>

Catalyst - Enzyme
- ATPase

^ some values from Catalyst Role
- Enzyme
- ATPase
Modularised into structure and function ontologies (all primitive)
Unified ontology after classification
Ontology normalisation: Enforced modularity

• All primitives form disjoint trees
• All polyhierarchies formed by inference from definitions
  – No arbitrary subclass axioms
• Therefore
  – All concepts have most specific primitive ancestor
  – No disjunctions in domain and range constraints
• Consequences
  – Changes occur in exactly one place
  – Defeasible inheritance can be constrained so that each property inherited only along one taxonomy
  • The same property cannot apply from two different sources that unless one subsumes the other
Led almost by accident to methods for Integrating multiple big vocabularies
And to real world problems
The Coding of Chocolate
An international conversion guide

<table>
<thead>
<tr>
<th>SNOMED-CT</th>
<th>Term</th>
<th>CTV3</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-F0811</td>
<td>Bounty bar</td>
<td>UbOVv</td>
</tr>
<tr>
<td>C-F0816</td>
<td>Crème egg</td>
<td>UbOW2</td>
</tr>
<tr>
<td>C-F0817</td>
<td>Kit Kat</td>
<td>UbOW3</td>
</tr>
<tr>
<td>C-F0819</td>
<td>Mars Bar</td>
<td>UbOW4</td>
</tr>
<tr>
<td>C-F081A</td>
<td>Milky Way</td>
<td>UbOW5</td>
</tr>
<tr>
<td>C-F081B</td>
<td>Smarties</td>
<td>UbOW6</td>
</tr>
<tr>
<td>C-F081C</td>
<td>Twix</td>
<td>UbOW7</td>
</tr>
<tr>
<td>C-F0058</td>
<td>Snicker</td>
<td>Ub1pT</td>
</tr>
</tbody>
</table>
...but logic & formal ontology scare users

- Require months to train even to a minimal level
- People make errors
  - Even experienced people
    - Logic inferences can be directly contrary to language maxims
- Make logic (& formal ontology) the “assembly language”
  - Give users high level languages
    - Ideally built with meta-authoring language construction kit
      - Use “Intermediate Representations”
Intermediate Representations
“Close to user forms”

Before:

(‘SurgicalProcess’ which
 isMainlyCharacterisedBy (performance which
 isEnactmentOf (‘SurgicalFixing’ which
 hasSpecificSubprocess (‘SurgicalAccessing’
 hasSurgicalOpenClosedness (SurgicalOpenClosedness which
 hasAbsoluteState surgicallyOpen))
 actsSpecificallyOn (PathologicalBodyStructure which <
 involves Bone
 hasUniqueAssociatedProcess FracturingProcess
 hasSpecificLocation (Collum which
 isSpecificSolidDivisionOf (Femur which
 hasLeftRightSelector leftSelection))>))))
Intermediate Representations
“Close to user forms”

After:
“Open fixation of a fracture of the neck of the left femur”

– MAIN fixing
ACTS_ON fracture
HAS_LOCATION neck of long bone
    IS_PART_OF femur
    HAS_LATERALITY left
HAS_APPROACH open

“Fixation of fracture of neck of left femur by open approach”

From 3-6 months training to 3-6 days training
Supports Loosely coupled distributed ontology development

From 80% central/global effort to 10% central/global effort

User effort cut by 75% compared with manual methods
Mostly in reduced committee meetings & arguments
A simple recent example: “Manchester OWL syntax”

- $\exists \text{owns}.\text{Dog}$ someValuesFrom(owns,Dog) owns SOME Dog
- $\forall \text{owns}.\text{Dog}$ allValuesFrom(owns,Dog) owns ONLY Dog

Or for one client just:

<table>
<thead>
<tr>
<th></th>
<th>property</th>
<th>min</th>
<th>max</th>
<th>filler</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>terms:has_a</td>
<td>0</td>
<td></td>
<td>Dosage</td>
</tr>
<tr>
<td></td>
<td>terms:has_a</td>
<td>0</td>
<td>1</td>
<td>Therapeutic Substance Unit Cl...</td>
</tr>
<tr>
<td>AND</td>
<td>terms:has_a</td>
<td>0</td>
<td></td>
<td>Route of administration</td>
</tr>
</tbody>
</table>
Users face a complex landscape
Inhabited by many tribes, Each tribe in its own teepee
and what feel like class divides
The chain of theorem envy

Pure Logician/Mathematician
Too neat
Too academic!
Doesn’t understand!

KR Researcher
Too scruffy
Too ad hoc
Doesn’t understand!

Knowledge Engineer
Too scruffy
Too ad hoc
Doesn’t understand!

Application Builder
Too scruffy
Too ad hoc
Doesn’t understand!

No one person can understand it all - must manage the chain
but logicians and ontologists are often seen as policemen

Incomplete! Undecidable! Higher order! No semantics! Not realistic! Not ontocleaned!

... and seem to insist on solving harder problems than the user actually has

... often without examples of why not
...or insist users understand the solution space

And don’t come back until you have the semantics clear

The opportunity & challenge of knowledge acquisition
Meet users where they are
Remember: logic is hard & most users are symbolophobes

• How many OWL experts here could get all of the following in the correct subsumption hierarchy first time?

- p some A
- p exactly 1 A
- p some ASub
- p exactly 1 ASub
- pSub some A
- pSub exactly 1 A
- pSub some ASub
- pSub exactly 1 ASub
- p only A
- pSub only A
- p only ASub
- pSub only ASub
- pSub only ASub

I didn’t: Why I need tools and a classifier
Matching problems and solutions
The science & craft of KA

- Patterns, guidelines and tools
- Reformulations of users’ “solutions”
- Collaborations between behavioural & domain scientists
- Challenges and demonstrations

Some observations…
Working with users:
The basics - I hope I am “teaching grandma to suck eggs”

• Do something useful - *First do the right thing!*

• Make it usable
  – Find a champion & early adopters
  – Embrace users’ attempts at solution
    • At least it is probably useful
      – an example; a working specification
  – Watch, ask, listen, and believe
    • There is no point in arguing with somebody that it is easy
    • There is no point in arguing with somebody that they understand
    • Get help -
      – Involve behavioural scientists
      – Use language and images - involve linguists

• Involve the implementers
• Build for feedback

• Accept that clever things may be useless
The interesting may not be useful

- Of medical decision support systems in the literature in 1995
  - 70% addressed diagnosis

- Success rate of evaluated medical decision support systems in the literature in 1995
  - Alarms 75%
  - Treatment Management 65%
  - Diagnosis 17%
And you can only tell if its useful if you know “What’s it for?”

- Tens of Millions spent on SNOMED’s 450K codes
  - Only 10K have ever been used
    - Many applications find 50%-75% missing

- Hard to predict users’ needs
  - So build it “just in time” if possible
  - And based on evidence
    - Mine the web
    - Mine corpora
    - Observe users
    - Build in feedback
Some arguments only have answers wrt problems

- **Open vs closed world reasoning**
  - Querying Rail Timetable: "Is there a train from Podebrady that gets me to Prague tonight?"
    - "No"
      - Possibly false, but useful
        » I can’t get better information
    - "Might be"
      - True but not useful & possibly misleading

- **Querying FoaF**
  - "Does Alan have any friends?"
    - "Don’t know"
      - At least not misleading
    - "No"
      - Wrong (I hope)
…which emphasizes again that: Metaknowledge is knowledge

- The railway timetable is presumed complete
  - But how do I know?
    - Do I trust the resource to tell me or depend on my experience?

- Biographies in FoaF are probably incomplete
  - But how do I know?
    - What about some other data somewhere on the Semantic Web?

- Need some new metadata standards urgently
Get the user’s to own it

• The Gene Ontology is the most successful “ontology”/ “thesaurus”
  – Not rigorous; many faults
  – But owned by the biologists and does their job

• The Gene Ontology Next Generation project accepts and delivers in their format
  – The OWL is just another quality assurance tool
QA with users:
The Logical answer isn’t always the right answer

“Cardiovascular Disorder” → “Disorder or injury of Heart or blood vessel”

“Slashed wrist” → “Self inflicted laceration of radial artery”

The user is right - fix the axioms:

“Cardiovascular Disorder” → “Disorder or injury of Heart or Great Vessels or arterial system as a whole”
...but may have to deal with differences in logic and language

• Logically:
  Four limbed animals do not subsume four legged animals
    – unless we represent explicitly our background knowledge that four legged animals have no other limbs
      • but we never say it
      • how to convey the intuitions to users?
Expressivity is more important than constraints

- Draw all the inference the user expects from “sensible” premises
  - Users won’t mind if wrong inferences could have been drawn from nonsense premises
    - Provided the nonsense premises can’t get in by accident
Find a way to do it: nonstandard reasoning: Commonalities and differences

• Classify all the proteins with phosphatase\textsuperscript{1} activities
  - Identify all the features in common
    • Simple subgraph algorithm
  - Create the generic abstraction
  - Classify again to find more phosphatases
    • 20 new phosphatases
      – Some probably implicated in disease
  - Classify by which tissue they come from
    • Repeat
      – Try to understand why some tissues suffer from disease and some don’t

\textsuperscript{1}”Phosphate-ase” - enzyme that breaks down phosphate bonds; the primary source of energy in cells; can drive super-oxidation; prime source of damage in diabetes, extreme inflammation in Intensive Care Units, atherosclerosis, and probably much else
Hide formalisms
Focus on meta-modelling

- Make the logic the assembly language
  - Build “compiler compilers” for close to user forms
    - Ought to be a major focus of our research
  - Transformations, import, export

- Support implementors and localisers
  - Tools to build user oriented special purpose tools, languages and interfaces

- Build quick tools for common patterns
  - Tabbed outlines for taxonomies
  - Easy visualisation and specification of disjoints
  - Wizards and Patterns for features & value sets
  - Special tools for lists
  - …
Help users to work together

• Science is Cooperative
E-Science is e-Cooperative

• Supporting cooperation requires new kinds of knowledge
  – Modularisation
  – Dependencies
  – Communication
  – User profiles
  – Views
  – Transforms
  – ...

Develop the whole chain of Knowledge Acquisition

Language → informal → Semi-formal → Formal

Language ← informal ← Semi-formal
Language Generation (an often forgotten technology)

• Even simple language generation helps
  – Even logicians may find it easier to work in pseudo-natural language

• If not natural language at least intuitive syntax
  – Syntax matters to understanding
Seek evidence

- Use corpora, NLP & the Web
  - Experts often wrong about what really happens
    - Beware of committee views
    - Beware of dogma
Find a sweet spot even if “trivial”
Prioritize issues to users’ requirements - thank you OWL 1.1

• Role assertions
  – “Fracture of neck of femur” is a kind of “Fracture of Femur”
    • hasLocus o isPartOf \(\rightarrow\) hasLocus
  – Accept repurposing
    • There are alternative solutions for anatomy but essential for other cases
      – “Uncle”
      – “Reified relations” (Ontological “qualities”)
Test at scale

• Many users’ problems are BIG
  – Biological examples ≥ 10K concepts
  – Medical examples 20K...500K concepts
Use the tricks of the trade
(especially to reduce arguments)

- "Pericardium" (covering of the heart) & "Heart":
  - To anatomists: they are two different organs
  - To the clinicians: they are parts of the same functional unit
- The role hierarchy can let us support both views

  is_clinical_part_of
  ← is_structural_part_of
...but hide them from (most) users

- Neither the anatomist nor the surgeon need to know about the other
  - In GALEN we had several different variants of the intermediate representation compiler to deal with intractable arguments
    - QA always passed by all
Make representations maintainable & evolvable (basic software engineering)

• Don’t be too clever (Avoid “Write once read never”)
  – “Shortest S-Expression for a function”
    “Maximum code in a C ‘for’ statement”
  – “Clever” inference in an OWL KB
    • Avoid
      – Making inferences from domains/range
        » May better to check before inference
        * Or do inference in multiple step
      – Complex disjoint axioms and disjunctions

• Design so all changes made in only one place
  – Normalise ontologies

• Be systematic - follow patterns

• Provide regression testing from the beginning
  – Preferably before building

• Support debugging
Remember: Your part is only one source of error

It will require empirical testing

- Errors come from
  - Errors & low inter-rater reliability in input
    - Often under 80%
  - Implementation errors in applications
  - The perversity of the universe - the Rumsfeld effect

- It is going to require empirical testing
  - The ONLY test that will be accepted for most applications is empirical
    - Whether or not the algorithms are complete
My top next priority for OWL: “What is it sensible...

- **To ask**
  - Everything is permitted unless it is forbidden
    - Natural open world logical view
      - Sets of constraints

- **To tell (sanctioning)**
  - Everything is forbidden unless it is permitted
    - Natural frames/OO/UML view
      - Sets of templates
- It is sensible to ask about “Green entities”
  - There are entities that are green
    - It is not sensible to ask about “Green dreams”
      - No: dreams have no colour
    - *The more I know the less it is sensible to ask*

- Is it sensible to tell that balls are green?
  - If it might sensibly be green
    - *The more I know, the more it is sensible to tell*
      - A hard/impossible notion to capture in first order theories

But the basis for intelligent user interfaces

Needed by every user we have encountered
Related: “Optional”

- What are the semantics of:
  - \(0..n\) in UML?
  - or \textit{canonical graphs}\ in Conceptual Graphs?
  - or \textit{sensible sanctions}\ in GRAIL?
  - or \textit{slot attachment}\ in Frames?
  - or \textit{minCardinality}(0)\ in OWL?

- Supported by every formalism any user will ever have used
  - There must be something to it
    - Even if it isn’t first order

- Needed by every data application we have encountered
Related but different: “May”

• **May**
  - “What *may* cause pneumonia?”
    What classes of bacteria *may* cause pneumonia?
  • Hard to express in DLs/OWL
    - Need richer relations amongst properties
      » causes → may_cause
        Almost captures it
      » “Anything that acts has the potential to act”
        “Anything that does not have the potential to act, cannot act”
        …is there additional semantics?
      » Very messy without special tools and operators in user language
... others need for a better match to biomedical knowledge

- Prototypical knowledge
  *The heart is generally on the left but sometimes on the right*
  - Normative solutions OK for some things
    - “The normal hand has five digits”
    - “The normal heart is on the left”
  - Simple explicit exceptions OK for some things
    - “All human cells have nuclei except red blood cells”
  - But otherwise need prototypical knowledge & exceptions
    - “Typically birds fly”
    - “Typically gram positive bacteria are penicillin sensitive”

- Uncertainty - another step? a different formalism?
  - 70% of appendices are in right lower abdomen
    - Can fuzziness/probability/… be contained in the same framework or do we need a true hybrid system?
      - Needs to link to utilities and decision theory
Summary: Help fill the gap

Problem space

Guidelines, Patterns, Tools, Modularity, Views, Environments, metadata, workarounds, hybrid systems, ...

Solution space

Watch, ask, listen to, & believe users

Design for modularity, maintenance & evolution

Support meta users & application builders

Fit solutions to problems

embrace users’ partial solutions as requirements

Look to user requirements for implementation priorities