WP6:
Ontologies and Terminologies
Background, Findings & Recommendations

Alan Rector, U Manchester
with
Jean Marie Rodrigues, U Jean Monnet, Saint-Etienne
Anand Kumar, U Jean Monnet, Saint-Etienne
Dipak Kalra, UCL
Bedirhan Ustun, WHO
Pieter Zanstra
Sounds Easy but has proved Hard

- 150 years of effort has not produced a solution
  - 20 years of intensive work in IT has not provided a solution
    - 10 years work on SNOMED provides at most a start on a solution
      - May even have be an “anti-solution” to parts of the problem
        » Certainly resulting in building serious “pregacy” - pre-built legacy
          ... effort & expertise in circumventing flaws and problems of its own creation

- Too many requirements ... so priorities unclear
  - Intimately intertwined with
    - EHRs, Public health, Decision support, Clinical care...
      - “What’s it for?”

- Temptation to do more than is possible
  - The best is the enemy of the good

- Those who must pay do not benefit; those who benefit will not pay
  - The benefit is to the common; the cost is to the private
Basic constraints & Limitations on the Possible

• **Scaling and the combinatorial explosion**
  – You can’t provide a phrase book big enough to write a novel
    • Or an EHR

• **Real world variability**
  – Some things are just different
    • Standards will just mislead

• **Human variability**
  – People can’t always interoperate
    • Machines will never interoperate better than the people that use them

• **Poor match of problem space & solution space**
  – Poor definition of purpose
    • “What’s it for?”

• **Lack of Quality assurance**
The scaling problem: The combinatorial explosion

- It keeps happening!
  - "Simple" brute force solutions do not scale up!

$\text{Conditions } \times \text{ sites } \times \text{ modifiers } \times \text{ activity } \times \text{ context} \rightarrow$
  - Huge number of terms to author
  - CHAOS and unending projects
Too small and too big

• All enumerated pre-coordinated terminologies are
  – To big to be used easily
  – To small to contain what is needed

• How many noun phrases in medical-ese?
  – All languages follow Zipf’s law - an infinite tail
    • e.g. in GALEN we could form over $10^{10}$ legal combinations!
      – and it covered only limited parts of medicine and surgery
  – Any terminology will cover only a tiny fraction of all possible terms
    • Impossible to know which will be needed in advance

• Must create terminologies “just in time”
  – from the bottom up - as we need them
    • Around a modest (25K term) core
  – Compose most terms as “post-coordinated” expressions from a modest ontology
    • analogously to composing phrases from words
<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>
Accepting limitations: Human reproducibility

<table>
<thead>
<tr>
<th>Category</th>
<th>X</th>
<th>X</th>
<th>X</th>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headcloth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloth</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scarf</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Model Person</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woman</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Adults</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Standing</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Background</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brown</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Blue</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Chemise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Dress</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Tunics</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clothes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suitcase</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Luggage</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Attache case</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brass Instrument</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French Horn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Horn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Tuba</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Semantic Health

Information Society and Media
Accepting Limits:
Problem space & solution space

Problem space

Solution space

Tools
Vocabulary - Things sometimes called “Terminologies”

• Controlled vocabulary with identifiers (codes)
  – List of terms for some entities plus “codes” to identify them
    • Code - meaningless identifier independent of language

• Lexicon
  – The collection of linguistic entities - attached to a given controlled vocabulary, codes, or ontology
    • May include grammatical & other information; may be multilingual

• Classification
  – An organisation of entities into classes for a specific purpose, e.g. ICD and DRGs

• Thesaurus
  – A collection of entities/terms arranged for human navigation via broader-than/narrower-than and associative relations

• Ontology (sensu informatics)
  – A logical model of the meanings of the entities about which information is to be expressed for use in computers

• Knowledge Representation System
  – A model of the background knowledge assumed, expressed so as to be used in computer systems - including but not limited to the ontology
Findings
State of Play - Facts on the Ground

• Most data will be collected using local codes
  – SNOMED unlikely to supplant local codes in EU, US or elsewhere except
    • UK & ?Australia & ?Canada Infoway
      – Maps to SNOMED may be important
  – HL7 v2 + LOINC will dominate messaging standards for labs
    – except in UK which will use V3 + SNOMED + local schemes
      • except in primary care which will use Clinical Terms (Read) & Nordic countries

• ICD will continue to be the main form of international recording
  – ICD 11 might convergence with SNOMED

• Subsets of SNOMED will be used in many areas as a controlled vocabulary
  – Without major investments other aspects will remain valueless or worse
    • And developing subsets is proving costly and rarely re-usable
  – But an alternative international effort is unlikely

• Lack of tools and people will be major constraints

• Bio and Translational Medicine Terminologies will increase
  – e.g. Gene Ontology, NCI Thesaurus

• Web based initiatives will happen
  – Web 2.0 & Google-type approaches will be increasingly important
SNOMED CT: Current Assessment

- **Purpose**
  - remains ill defined ... but being used in UK and elsewhere for controlled vocabulary

- **Controlled vocabulary and identifiers**
  - *Well managed* but very slow response (months .. years)

- **Scale**
  - Overgrown - victim of combinatorial explosion and Zipf’s law
    - Too big to QA, manage and fix, find terms, use reproducibly
    - But still often ≤ 25% coverage for specific applications

- **Reproducibility**
  - *Poorly studied... often poor*

- **Hierarchies and relations**
  - *Unusable*
    - Too unreliable to depend on to behave as documented. Not QAed.
    - Systematically flawed in principle; Limited by State of Art circa 1990

- **Multilingual / cross-cultural support**
  - *Minimal - fundamentally an anglophone organisation*
    - Neither understood nor a priority of the IHTSDO
      - Spanish and Canadian French versions might appear
    - Separation of language and concepts still problematic; tools absent

- **Openness and accessibility for QA, contribution, & social computing**
  - *Unusable - remains effectively closed*
    - Not generally available on the Web
    - Opportunity cost of participation prohibitive; Influencing policy difficult
      - Remains the province of a small self-reinforcing clique

- **Could be fixed at modest cost relative to total cost of health it interoperability**
  - Priority is a feasibility study on ~20K concepts
Interaction of Terminology, EHRs & DSSs

- HL7-Terminfo provides a base
  - But no tools (yet) & proving difficult to implement consistently

- Archetype experiments provide a start
  - Terminology Query Languages promising
    - Common formalism a major challenge

- Better technologies exist and have been demonstrated experimentaly
  - Using logical tools, OWL, UML2, Model Driven Architectures...
    - But so far under-developed
      - Efforts at tools are under way in the UK
        » Outcome remains uncertain
  - More development urgently needed
Recommended Principles: Technical

• **Separate Language and Concepts**
  - Lexicon and Coding System/Ontology - more radically than in SNOMED today
    • Otherwise endless confusion
    • Develop multilingual cross-cultural systems
  - Otherwise endless confusion

• **Make it easy to participate -**
  - Hide complexity if not needed - separate levels
    • End users /Author-experts & configuration staff / Terminology experts

• **Develop binding to EHR and Decision support**
  - Requires new techniques and tools

• **Leverage modern tools for development, QA & deployment, especially from the Web community**
  - Web 2.0
  - Web Ontology Language OWL and modern logic formalisms
  - … but also what the EC has already paid for
    • See SNOB browser for SNOMED developed from GALEN tools

• **Build small core ontologies**
  - Use logic for post-coordination or provide just-in-time services for what people need as determined empirically

• **Develop QA methodologies**
  - and use them

• **Human factors and reproducibility matter as much as technical structure**
  - Success is seeing it get simpler - GALEN reduced training time from 3 mo to 3 days
Recommended Principles: Organisational

• **Process as well as product**
  - All actions must aim at long term institutions that can be sustained

• **Involve healthcare providers and systems vendors**
  - They are who must interoperate
  - Provide incentives; mitigate costs
    • Interoperability may be a *disbenefit* to them otherwise

• **The terminologies must be owned by their key end users**
  - Be responsive, cooperative and open
    • The Web gives us the tools - use it!
      IP worries are the enemy of interoperability

• **Terminology development must coordinate with EHR and Decision support Development**
  - All terminologies must have purposes
    • And be shown to be fit for purpose

• **Think global; act local**
  - Be multilingual and cross-cultural
    • Will only happen if EC intervenes
Key Recommendations for Terminologies

• Support WHO open collaborative development of ICD-11
  Try to open SNOMED to open collaborative development
  - Develop generic Web 2.0 / social computing
    • Seek mechanisms for opening the SNOMED process to social computing
• Develop open terminology tools that scale up to ICD & SNOMED
  - A European network of Terminology Servers & Web 2.0 Terminology sites
  - Tools based on SNOB, Protégé OWL, others
  - Cultivate open source communities - Empower users and specialist groups
  - Develop QA tools and techniques
• Develop language technologies
  - Text generation to present and QA
  - Text extraction to build and encode
• Support development of methodologies and tools for binding terminologies, EHRs and Decision support
  - Immediate investment in tools and medium term research
• Support feasibility study of reformulation of SNOMED on a small scale
  - 25K terms max
    • Show that SNOMED need not be a tax on medicine
      - ... or accept that it will be
• Support studies of human factors and reproducibility - Demand QA
• Support training & human capacity