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**
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 experimentally
  - 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
- **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
Short Term

• 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
• 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
Key Recommendations for Terminology Research
Medium & Long Term

- 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 research into methods & metrics for Quality assurance

- Support research on foundations of terminologies
  - Representation and meaning
  - Practical issues in representation, scaling, and use in applications
  - Terminologies & software engineering

- Support studies of human factors and reproducibility -
  - Demand QA

- Support training & human capacity