Terminological Systems: Bridging the Generation Gap

JE Rogers AL Rector
Medical Informatics Group, Department of Computer Science, University of Manchester, UK
http://www.cs.man.ac.uk/mig/giu/ or e-mail galen@cs.man.ac.uk

A rigorous formal description of the intended behaviour of a compositional terminology, a 'third generation' system, enables powerful semantic processing techniques to assist in the building of a large terminology. Use of an intermediate representation derived from such a formalism, but simplified to resemble a 'second generation' system, enables authors to work in an simpler and more familiar environment, avoiding many of the technical complications of the 'third generation' system.

INTRODUCTION

Developers of terminologies specifically designed for medical computer applications are increasingly exploring alternatives to the enumerative techniques embodied by traditional schemes such as ICD1 or READ version 1 or 2. The expressivity of such schemes is limited by whether appropriate, specific terms already exist. Existing terminologies such as SNOMED2, and many currently in development (e.g. DICOM SNOMED Microglossary, LOINC, ICNP 3, READ 3.14), have adopted compositional techniques: increased expressivity is achieved by fashioning descriptions from structured collections of more basic terms.

However, compositionality increases flexibility: a common clinical requirement is for sets of highly detailed terms in a particular specialised medical sub-domain - perhaps for research or audit purposes. Users of enumerative schemes must either wait for them to be included in the next major central revision or (more commonly) make *ad hoc* local additions. A compositional scheme enables principled local extension, by making new compositions.

European standardisation work reflects this move to compositional techniques. The European Committee for Standardisation (CEN) has produced several standards and pre-standards following ENV 12264[5], itself a pre-standard for representing terminologies as a semantic network.

Existing enumerative schemes are termed 'first generation' terminology systems by Rossi Mori[6]. In his study of compositional schemes in development he identifies four common components: a *categorial structure*, a *cross-thesaurus*, a *family of lists* and a *knowledge base of dissections*. Systems where all four components are well developed - Rossi Mori's 'second generation' - acquire new capabilities of semantic processing.

However, Rossi Mori notes that developing the four components and the resulting scheme must be an iterative process. Further, development of one component often complements, but may also depend upon, development of another. These dependencies may initially be expressed as a set of manually applied rules and checks. However, as the system and its dependencies become progressively more complex, it ceases to be possible to maintain integrity or coherence through human processing power alone.

Further progress requires formal specification of the system's intended behaviour and a software engine implementing this behaviour. Systems including such an engine - what Rossi-Mori terms 'third generation' systems - constrain and guide all user interaction according to the formalism. Further enhancements of semantic processing power are gained, but knowledge authoring becomes more demanding: the scheme, its terms and the formalism become so interdependent as to be inseparable and the whole becomes essentially a piece of software with unfamiliar added complexity for authors. This paper describes an approach which allows authors to use a simplified 'intermediate representation' resembling familiar 'second generation' systems while still retaining the full power of a 'third generation' system.

GALEN-IN-USE

GALEN-IN-USE is a European Union funded project to develop tools and methods to assist in the collaborative construction and maintenance of compositional surgical procedure classifications. The results from the previous GALEN project - the GRAIL formalism[7], GALEN Common Reference Model (CRM)[8,9,10], High Level Ontology[11] and Terminology Servers[12] - are providing 'third generation' system support for this task.

Taking part in the initial phase are four national coding and classification centres: WCC (Netherlands), SPRI (Sweden), CNR (Italy) and University of Ste. Etienne (France). During the project, conceptual representations of some 15,000 individual surgical procedures will be produced using the GRAIL formalism and integrated into the existing GALEN Common Reference Model[7,9,10,11].

GALEN and CEN ENV 1828

The relationship between GALEN and 'second generation' systems is illustrated by the GALEN approach to CEN ENV 1828[13], a pre-standard proposing a compositional structure for classifications of surgical procedures. The CEN schema (figure 1) reflects the way the terms are used in

1091—8280/97/$5.00 © 1997 AMIA, Inc.

Figure 1: CEN ENV 1828 schema for surgical procedures

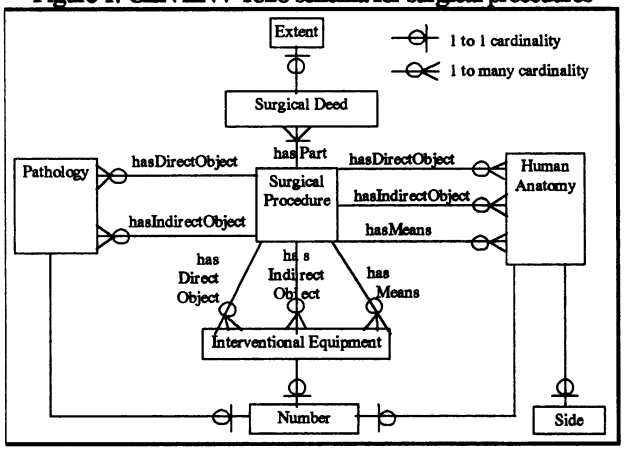
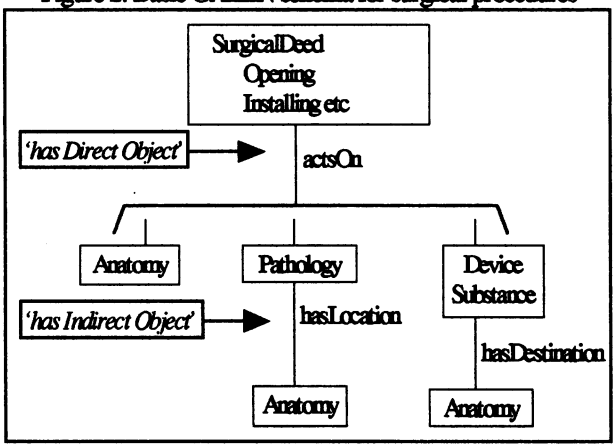


Figure 2: Basic GALEN schema for surgical procedures



language. Our experience has been that a conceptual model has slightly different requirements.[14] GALEN's schema must both support automatic classification and also integrate with an existing model which permits indefinite nesting of anatomical sublocation. These different treatments are illustrated by the GALEN interpretation of the section in the normative part of ENV 1828 which states: *'A surgical procedure must have anatomy either as a direct or an indirect object'*.

Within the GALEN Common Reference Model (CRM), neither the indirect nor the direct object is linked directly to the procedure. Instead, the direct object is always linked to the surgical deed itself:

(Removing which actsOn Kidney) name Nephrectomy.

A more significant difference in treatments concerns the indirect object. In the CEN scheme:

(SurgicalProcess:)*
– (hasPart) – (SurgicalDeed: Removal)
– (hasDirectObject) – (Pathology: Cyst)
– (hasIndirectObject) – (Anatomy: Kidney)

expresses the notion of 'excision of a kidney cyst'. However, in the Common Reference Model we are able to specialise [Cyst] according to its location:

(Cyst which hasLocation Kidney) name KidneyCyst.

If the CEN schema were followed, the constraining mechanisms in GRAIL could not prevent construction of obviously nonsense compositions such as 'removal of a renal cyst from the foot':

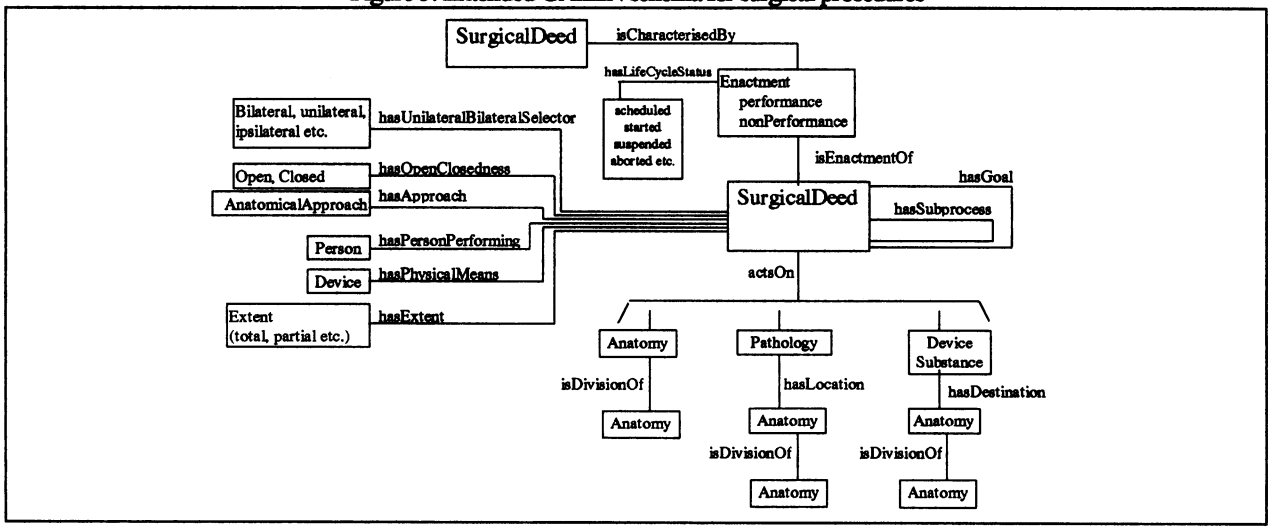
(Removing)
– (actsOn) – [(Cyst) – (hasLocation) – (Kidney)]
– (hasIndirectObject) – (Foot)

In the CRM, therefore, the indirect object is attached indirectly to the deed, via the direct object, thus:

(Removing which actsOn (Cyst which hasLocation Kidney)).

These changes result in a basic GALEN schema for surgical procedures (figure 2). This has subsequently been expanded to increase expressivity and to integrate it with the rest of the GALEN Common Reference Model (figure 3).

Figure 3: Extended GALEN schema for surgical procedures



AN INTERMEDIATE REPRESENTATION

GRAIL, the GALEN representation language, is necessarily complex - as would be any other 'third generation' representation. For the GALEN-IN-USE project, more than 20 clinicians were recruited across four countries to perform the analysis of original surgical procedure code rubrics into conceptual representations. However, few had any prior experience of GRAIL or the Common Reference Model.

To circumvent this problem we devised an intermediate representation[15] nearer to a 'second generation' system. It is structurally simpler than GRAIL, but may subsequently be automatically expanded into GRAIL. This expansion is possible because the design of the intermediate representation deliberately echoes that of the Common Reference Model. For example, the schema for surgical procedures in the intermediate representation is a systematic simplification of the corresponding extended GALEN schema. This enables automatic 'de-simplification' to occur when the dissections are expanded into GRAIL. Rogers has described this expansion process[16] and the GALEN software tools (TIGGER and SPET) which support it.

The intermediate representation is broadly similar to those used by the CANON group or the MED.[17,18,19,20] It is characterised by:

- a grammar defining a layout, or 'template', for well-formed representations.
- a relatively small set of semantic links (ACTS_ON, IS_PART_OF), compared to the GALEN CRM;
- a domain ontology specific to the surgical domain. The atomic terms (leg, excising, tumour etc.) are known as 'descriptors' and are explicitly typed by one of a small number of descriptor classes (e.g. anatomy, deed, lesion);
- a small set of constraints to control which links may be used with which descriptor classes .

Domain experts in the centres work from existing local coding schemes (WCC, NCSP etc.) to scope their task. Rubrics from these schemes are manually analysed to give, initially, a natural language paraphrase of what the expert believes the rubric means. A conceptual representation of each such paraphrase is then produced using the intermediate representation. The result of this two-step analysis is called a 'dissection' of the rubric. Each dissection has a header section which contains information about the original rubric and coding scheme. This is followed by the conceptual representation itself, introduced by the MAIN keyword. Semantic links are capitalised, descriptors are in lower case. Below is an example of a completed dissection:

```
RUBRIC "Insertion of intercostal catheter for drainage"
PARAPHRASE "Insertion of intercostal catheter in pleural space
for drainage"
SOURCE "ICD-9-CM" CODE "34.04"
MAIN inserting
    ACTS_ON catheter
    HAS_APPROACH intercostal route
    HAS_DESTINATION pleural space
MOTIVATED_OVERALL_BY draining
    ACTS_ON substance
        HAS_LOCATION pleural space
```

A GRAIL expansion from this dissection is automatically generated (below). The expansion algorithm requires that the primitive descriptors and links in the intermediate representation are given context dependent mappings to primitive or composed concepts and attributes in the Common Reference Model, as described by Rogers.[16]

```
[(SurgicalDeed whichG <
  isMainlyCharacterisedBy
    (performance whichG isEnactmentOf
      (Inserting which <
          hasSpecificSubprocess
(SurgicalApproaching whichG hasPhysicalMeans
  (Route which passesThrough IntercostalSpace))
          isActedOnSpecificallyBy
(Transport whichG hasSpecificConsequence
  (Displacement whichG hasBetaConnection PleuralCavity))
          playsClinicalRole SurgicalRole
          actsSpecificallyOn Catheter>))
  hasSpecificGoal (Draining which <
    playsClinicalRole SurgicalRole
    actsOn (Substance whichG hasLocation PleuralCavity)>) ]
  hasProjection
(('ICD-9-CM' schemeVersion 'default') code '34.04' 'code');
  extrinsically hasDissectionRubric
'ICD-9-CM 34.04 Insertion of intercostal catheter for drainage'.
```

ADDED VALUE OF GALEN

The GALEN intermediate representation is similar to a 'second generation' system. However, it results from a systematic simplification of a 'third generation' system rather than a gradual increase in sophistication of a 'first generation' enumerative system. It facilitates our knowledge authoring process whilst still allowing 'third generation' techniques to be exploited to build, maintain and validate the corpus and, ultimately, deliver it to end users. Four techniques, not applicable to 'second generation' systems or to the intermediate representation directly, are fundamental to our authoring process:

- Automated semantic normalisation and canonisation
- Automated and dynamic classification of compositions
- Automated maintenance of fixed knowledge database
- Automated generation of natural languages

Semantic Normalisation

Different authors, analysing the same rubrics, produce different dissections. These differences divide into those which are semantically equivalent, those semantically divergent and those which represent semantic errors. The expansion of dissections into GRAIL provides several different stages at which normalisation can occur. For example, differences of semantic equivalence such as varying encapsulation may be automatically normalised. A separate mechanism rejects many semantic errors:

Normalising varying encapsulation: in the rubric 'excision of lobe of lung', one author may determine that {lobe of lung} is an appropriate primitive descriptor, whilst another may choose to decompose it into {lobe IS_PART_OF lung}. The expansion into GRAIL normalises both into:

(Lobe which isSolidDivisionOf Lung).

because of the following previously declared mappings:

Descriptor / Link	GRAIL Mapping
lobe of lung	Lobe which isSolidDivisionOf Lung
lobe	Lobe
lung	Lung
IS_PART_OF	isSolidDivisionOf

Rejecting semantic error: The intermediate representation includes only limited constraints on which descriptors may be combined with which links. The CRM contains a richer set of constraints which are brought to bear when a dissection is expanded into GRAIL and when an expansion is presented to the engine for classification. Thus {fracturing ACTS_ON temperature} is permitted in the intermediate representation, but rejected at the GRAIL expansion stage.

Semantic divergence: Differences of opinion between experts regarding what rubrics actually mean must remain problems for the experts to resolve. However, the other techniques discussed here combine to assist the domain experts in identifying when they do not agree.

Automatic classification

GRAIL expansions of the dissections are automatically classified according to the principles of the GRAIL formalism. Knowledge already present in the CRM affects this classification; for example, 'Operation on the Heart' subsumes 'Repair of Mitral Valve' because the anatomy model already knows the mitral valve is part of the heart.

Where a dissection has not been classified as expected, the task is to identify why. With the 'noise' of semantic equivalence removed through normalisation, the remaining causes are semantic divergence, and omissions or errors in the pre-existing knowledge base. Automated analysis, according to the formalism, of the relationships between expansions of dissections can answer questions such as 'why is this classified here?' and 'what should I change to have it classified there?'.

Automatic classification further ensures that the twin hierarchies of composed deeds and of the objects they act on must inevitably be exactly parallel, since one is derived formally from the other. Maintaining this 'parallelism' is presently commonly undertaken manually in other 'second generation' systems, (e.g. the READ 3.1 Thesaurus[4]).

Maintenance of the knowledge database

To hold a fixed form of the knowledge base, local implementations of compositional systems may need to instantiate 'artefact' concepts as well as the compositions originally provided by authors. This might be necessary to fit the knowledge base within a particular persistent data structure, (as occurs in the READ 3.1 Thesaurus[4]) or to optimise a classification or search algorithm.

In a GALEN system, knowledge authoring is decoupled from any particular implementation of the knowledge base. The local implementation determines for itself what it needs to instantiate, and is able to export the knowledge base to other implementations where the requirements for instantiated concepts may be different.

Automatic Generation of Natural Language

Early experiments provided the dissection authors with a display of their original scheme rubrics, ordered into a hierarchy according to the automatic classification of the GRAIL expansions. However, the original rubric is not always a satisfactory proxy for the dissection itself. The semantic information which directly determines the classification is hidden, and identifying the cause of an inappropriate classification from this presentation alone is not possible. Similarly, browsing the hierarchy of the GRAIL concepts themselves displays too much information, in too abstract a form, to be directly useful.

GALEN tools can generate from a GRAIL composition a natural language string which reflects the semantics of that composition.[21] Browsing hierarchies of these strings, in an editing environment which links them directly to their originating rubrics, dissections or GRAIL expansions, is expected to form a powerful Quality Assurance tool.

RESULTS AND FUTURE DEVELOPMENTS

More than 3000 original rubrics, in the fields of orthopaedics, urology, cardiology and gastroenterology have so far been dissected using the intermediate representation. These have subsequently been expanded into GRAIL and

classified within the Common Reference Model. Generation of Natural Language phrases for the results is now possible in four European languages, though the lexicons are not yet complete. Future experiments will examine delivering the corpus to the participating centres as either a first, second or third generation system according to local requirements.

CONCLUSION

'Third generation' systems, such as GALEN, offer advanced semantic processing techniques. We have shown the added value of using these to help build large and coherent terminologies. However, authoring compositional representations directly in a formalism such as GRAIL is time consuming and requires special skills.

An intermediate representation can bridge between the generations: 'third generation' system advantages can be gained whilst authoring effort remains closer to that required for 'second generation' systems. Existing standards can be extended or adapted to support this activity. A prerequisite is an automatic transformation between this representation and the formalism, and between the formalism and natural language.

Acknowledgements

With thanks to all in the GALEN-IN-USE consortium. GALEN-IN-USE is funded as part Framework IV of the EC Healthcare Telematics research program.

References

1. World Health Organisation. International Classification of Diseases, 9th Revision. Geneva: WHO, 1977
2. Cote RA, Rothwell DJ, Palotay JL, Beckett RS, Brochu L (eds), The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International, College of American Pathologists , Northfield, IL:, 1993, 3rd edition
3. Mortensen RA (ed) The International Classification for Nursing Practice ICNP with TELENURSE introduction. Copenhagen: The Danish Institute for Health and Nursing Research, 1996
4. Price C, et al. Anatomical Characterisation of Surgical Procedures in the Read Thesaurus. JAMIA 1996; symp. Suppl.;110-114
5. CEN ENV 12264:1995. Medical Informatics - Categorial structure of systems of concepts - Model for representation of semantics. Brussels: CEN, 1995
6. Rossi Mori A, Consorti F, Galeazzi E, (1997) Standards to support development of terminological systems for healthcare telematics. (proceedings of IMIA Working Group 6 meeting, Jacksonville, Florida)
7. Rector A, and Nowlan WA (1993). The GALEN Representation and Integration Language (GRAIL) Kernel, Version 1. The GALEN Consortium for the EC AIM Programme. (Available from Medical Informatics Group, University of Manchester).
8. Rector A (1994). Compositional models of medical concepts: towards re-usable application-independent medical terminologies. Knowledge and Decisions in Health Telematics P. Barahona and J. Christensen (ed). IOS Press. 133-142.
9. Rector A, Gangemi A, Galeazzi E, Glowinski A and Rossi-Mori A (1994). The GALEN CORE Model Schemata for Anatomy: Towards a re-usable application-independent model of medical concepts. Twelfth International Congress of the European Federation for Medical Informatics, MIE-94, Lisbon, Portugal, 229-233
10. Rector A (1995). Coordinating taxonomies: Key to re-usable concept representations. Fifth conference on Artificial Intelligence in Medicine Europe (AIME '95), Pavia, Italy, Springer. 17-28.
11. Rector A, Rogers JE ,Pole P (1996) The GALEN High Level Ontology. Fourteenth International Congress of the European Federation for Medical Informatics, MIE-96, Copenhagen, Denmark
12. Rector A, Solomon WD, Nowlan WA and Rush T (1995). A Terminology Server for Medical Language and Medical Information Systems. Methods of Information in Medicine, Vol. 34, 147-157
13. CEN ENV 1828:1995 Health care informatics - Structure for classification and coding of surgical procedures. Brussels: CEN, 1995
14. Ceuster W, Beukens F, De Moor G, Waagmeester A (1997). The Distinction between Linguistic and Conceptual Semantics in Medical terminology and its Implications for NLP-Based Knowledge Acquisition. (proceedings of IMIA Working Group 6 meeting, Jacksonville, Florida)
15. Gaines BR, Shaw ML and Woodward JB (1993). Modelling as Framework for Knowledge Acquisition Methodologies and Tools. International Journal of Intelligent Systems 8(2): 155-168.
16. Rogers JE, Solomon WD et al. (1997) Rubrics to Dissections to GRAIL to Classifications. Fifteenth International Congress of the European Federation for Medical Informatics, MIE-97 Thessaloniki, Greece
17. Campbell KE, Das AK and Musen MA (1994). A logical foundation for representation of clinical data. JAMIA 1(3): 218-232.
18. Cimino J (1994). Controlled Medical Vocabulary Construction: Methods from the Canon Group. Journal of the American Medical Informatics Association 1(3): 296-197.
19. Evans D (1988). Pragmatically-structured, lexical-semantic knowledge bases for unified medical language systems. Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care, Washington DC, IEEE Computer Society Press: 169-173.
20. Huff S and Warner H (1990). A comparison of Meta-1 and HELP terms: implications for clinical data. Fourteenth Annual Symposium on Computer Applications in Medical Care (SCAMC-90), Washington DC, iEEE Computer Society Press: 166-169.
21. Baud RH, Rassinoux A-M, Lovis C, Wagner J et al. Knowledge Sources for Natural Language Processing. In: Cimino JJ (ed) Proceedings of the 1996 AMIA Annual Fall Symposium Philadelphia: Hanley & Belfus, Inc. 1996: 70-84