Managing Information Models for E-health via Planned Evolutionary Pathways[1]

Helmut Duwe

Institute of Aerospace Medicine, German Aerospace Center (DLR),
Cologne, Germany
`Helmut.Duwe@dlr.de`

Abstract. E-Health is one of the many e-application areas promising to increase global connectivity and computational power of applied computer networks. Information modelling is an essential ingredient of developing e-health applications. The theoretical foundations of information modelling are growing. System theory, semiotics, semantics, and ontology have become allied partners to build powerful information models. Phrases like "ontology-based conceptual modelling" indicate this process of sophistication. When developing practical e-health systems the management of complexity is still a major challenge. Evolutionary pathways seem to be an appropriate approach to handle complexity. Significant gaps between existing pieces of avantgarde academic theory and broad practical success stories are seen. The idea of planning evolutionary pathways as mechanism to handle information model complexity is illustrated in one area, namely the development of a study and subject record manager for health-relevant physiology research.

1   Introduction

May be the 21st century will be known in the far future as the rise of the e-ages. E-business, e-commerce, e-government, e-science, last not least e-health are ringing in the decades which will hopefully see a tremendous increase in global connectivity and computational power of computer networks.

Herewith, a major issue of the ethical debate as well as the technical agenda is the handling of the right balance between simplicity versus complexity of representing data, information and knowledge concepts. Per se simplicity is not bad and complexity is not good, contextual adequacy is required.

2   Finding the Right Adequacy between Simplicity and Complexity

The handling of information has become a multidisciplinary scientific effort. The subject matter disciplines themselves have developed vast infrastructures to handle data, information and knowledge (e.g. Medline).

---

[1] The views and conclusions contained in this paper are those of the author and should not be interpreted as representing the official policies of any organization.

The following examples shall illustrate that existing approaches, useful as building blocks for e-health information models, indicate the need for paving the roads and bridges in-between.

A simple, but effective, mechanism to handle data for clinical trials has been set up by the CDISC (Clinical Data Interchange Standards Consortium) group. In the CDISC Operation Data Model (ODM)[2] the actual clinical data are handled as attribute-value pairs, as such a trivial technical approach.

The general approach behind this is called EAV (Entity-Attribute-Value) concept. It holds to a large extend the subject matter complexity within the data content and operates on the metadata level mainly with the very generic concepts of entity, attribute and value. It has been shown [see Brand et al] how this content internal complexity can be unfolded in Clinical Trial Data Management Systems, which are technically based on relational databases.

The unfolding of structural representations from internally complex data content to managed information objects with adequate metadata functionality is not only an issue of reliable "older" technologies like relation database systems, but also of object-oriented approaches. It is proven technique to avoid the explosion of classes via generic classes and attributes. Attributes work here as containers for typed content. The making of typologies is shifted in this case from a class typology (with explicit metadata entities) to the content structure (with more or less known internal metadata).

The examples to illustrate the potential and the handling difficulty with complex information and information models are taken from the LOINC (Logical Observation Identifier Name and Codes) and the openEHR (open Electronic Healthcare Record) project.[3]

The LOINC standardisation efforts of laboratory and clinical observations take care to find systematic names as universal identifiers. A LOINC long name has 6 parts: the component (analyte), the property measured, a time aspect, the sample type, the scale type, and where relevant the method used. A variant of an oral glucose tolerance test is written for example as `GLUCOSE^1H POST 100 G GLUCOSE PO`.[4] The precision of the naming needed can drive the appearance of the names towards what the creators of LOINC call themselves "awkward". Nevertheless, LOINC is not only a terminology listing, it is a very important starting point for more elaborated information models of all the identified and hence listed concepts.

The actors behind the openEHR (open Electronic Health Record) initiative have accumulated a rich documentation of information models, suitable to play a vital role in developing advanced healthcare record systems. Only one valuable central aspect should be mentioned here, namely the undertaking to clarify such an abstract concept as "data structure". The "Data Structures Reference Model" of openEHR contains sophisticated representations of a complex "structure" concept. Fundamental kinds of

---

[2] After all, the CDISC standards have their strength through the stringent way they require the explicit documentation of the metadata and by the richness in which they have captured the subject matter semantics of clinical trials.

[3] Further examples could have been taken from HL7 V3 RIM (Health Level 7, Version 3, Reference Information Model), due to size limitation of this paper this was not possible.

[4] H for hour, G for gramme, PO for *per ora*.

structure are identified, for example "single (values)"[5] (singularity), "list", "table", "tree" and temporal ones like "history", "event", and more. Further concepts like tree-tables and table-trees are thinkable and already implemented here and there. The challenge comes with the linking of complex subject matter issues and data structures.

3   Some Fundamentals of Modelling Theory Exploited for E-health

It is stated here that system theory and semiotics shall not be underestimated when working in the field of information modelling [see Duwe et al].

Here, only the triadic sign relationship as core concept of semiotics shall be given as absolute basics. In the triadic sign relationship the nature of language is seen as a triangle of the reality object (thing), the representing sign or symbol (ink on paper or signs on a computer display), and the mental process of having a concrete thought based on sensory inputs (interpretant). The mapping of concepts into this triangle is still a major scientific challenge. For example, what is about the reality aspect of "visions" which bear "reality" years later ? What is about brain processes without sensory input from a reality source like it happens with synaestesia.

Within the scope of these considerations it is only important to keep all three corners of the triadic sign relationship in mind, when a term and concept is discussed.

The Renovated[6] Alliance Partners

An improved modelling theory for e-health applications should assimilate some of the major advancement made in conceptual modelling, in data semantics and ontologies.

The first cycle of UML (Unified Modelling Language) modelling during system analysis and early system design is basically centred around conceptual modelling. It has to be distinguish from the specification modelling in a second cycle, more or less the late phase of the system design phase of software development.

Handling the vocabulary of a knowledge domain is the key to semantic analysis. The primary issue of data semantics and hence semantic interoperability is the clarification of meaning. When reading the themes of conferences dealing with data semantics the overlap with conceptual modelling and data ontology becomes obvious.

Gruber calls ontologies an "explicit formal specification of a shared conceptualisation". [Gruber, cited after Heese, see Heese]. This illustrates once more the closeness of the term "ontology" and "concept/conceptual modelling", finally showing that (A) conceptual modelling, (B) data semantics and (C) data ontology are highly interwoven. Further topical constructions as for example "foundational ontologies" [see Schneider] represent the peak with regard to catching abstract complexity.

---

[5] resulting in class names like SINGLE_S, TREE_S, TABLE_S, etc.

[6] This word is used to indicate the picture of "high technology interiors (informatics) in ancient buildings (ontology)".

A   Conceptual Modelling   B   Data semantics   C   Data ontologies
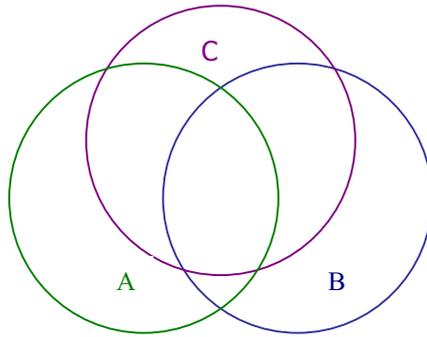
Fig. 1. Relationship between conceptual modelling, data semantics and ontologies

4   The Practical World of Information System Development in E-Health

Abstract modelling theory has to find more intensively its way into the real practical world of developing information systems for e-health.

A variety of textbooks of software engineering deliver good overviews of Software Engineering (SE) process models. The most well known kinds of SE process models are the so-called waterfall model, the V-model, the OO-models, and models with richness regarding model evolution like the incremental, the evolutionary and the spiral model. A model with increasing importance is the so-called Rational Unified Process (RUP) approach. RUP has been basically elaborated by the founding fathers of UML, which have pointed more than once to the importance of iterative and incremental development of software. There is a fruitful influence of these SE process models on each other noticeable. For example, the V-model, itself a highly structured waterfall model with integrated quality assurance, has incorporated incremental extension mechanism.

The health domain is huge. Some tentative figures from medical terminology systems shall give a first impression on the magnitude of complexity. The International Classification of Disease (ICD) contains in its four-level systematics more than 12.000 disease classes; the Medical Subject Heading (MeSH) covers around 20.000 main key words; the Logical Observation Identifier Names and Codes (LOINC) system lists more than 30.000 universal identifiers for laboratory and clinical observations, the Systematized Nomenclature of Medicine (SNOMED) covers a complex set of 200.000 terms, and the Unified Medical Language System (UMLS) project works with more than 600.000 concepts and more than a million terms. The estimation of concepts with relevance for medicine yields figures as high as 50 million.

The Institute of Aerospace Medicine (DLR-IAM) organizes and performs studies in the fields of human physiology, biology and medical technology. Scientific studies vary in the degree of complexity. An individual study is usually carried out by one institution, whereas a multi-centre study is performed by two or more institutions based on a common study protocol. Integrated studies comprise several related individual studies which are conducted by one or several institutions sharing test subjects, staff and resources at one study site. To support a wide range of study types the DLR-IAM is developing together with the Institute of Information processing and Computer supported new Media (IICM) at Technical University of Graz an open Electronic Study Record Manager (openESRM)[7]. The openESRM is based on an open Medicine Telematics Platform. This platform provides comprehensive security mechanisms like authentication, authorization, digital signature.

The Study Protocol Information Model of RCRIM[8] was studied as input for an own study protocol information model. Putting all available parts of the RCRIM Protocol Model together indicates a little bit the growing complexity of modelling a subject matter like "Clinical Trials".

Displaying complete information models like RCRIM study information model or the HL7 V3 RIM tends to cover whole walls. Evolutionary extensions of one concept adds a further dimension of complexity to this "wall paper" phenomenon.

According to the idea of evolutionary pathways each identified "element" of such an information model will get a "genealogy vector". For each concept a multi-generation design shall be prepared that marks the way from simple to complex.

The staging shall be briefly illustrated with the OpenESRM concept of "Study", being here (besides "Study Record") the "root concept" (or root class). The prospective genealogy of the concept "Study" starts in the baseline generation with a simple associate construction, more or less a kind of mind mapping exercise.

The second generation has been designed with a data skeleton class, to delegate the basic data set just like it is done with the ProtocolSkeleton class of the RCRIM model. But, the StudySkeleton class is a data type candidate to be used throughout the entire life cycle, for example as super-class for the ProtocolSkeleton too (RCRIM has done the delegation so fare only for the Study Protocol). To reduce redundancy, generic sets of data are extracted and keep available for whatever other concept (class) is interesting to use it.

The third generation introduces the Collection concept as used within the "PlannedStudy" class of the RCRIM Protocol Model. The fourth generation of the Study class is designed as mix of singularity attributes and collection attributes. Furthermore, the problem of modelling "static" and "dynamic" features of a phenomenon is here reflected very intensively.

"Study" had been defined as a specific "process". The Workflow IT community has extensively reflected the term "process". But, a typical definition here just states that processes are chains of activities. The settlement of an adequate model of

---

[7] The description of the ESRM concepts are based on DLR-IAM internal papers, major contributions to the concepts have been mainly made by Doris Wilke and Erwin Bartels, especially the identification and description of the process activities is taken (with minor deviations) from the work of Doris Wilke.

[8] Regulated Clinical Research Information Model, the main player are here the CDISC group, the HL7 group, and FDA (the U.S. Federal Food and Drug Administration)

"process" is currently tried as part of a "cinematographic concept" of a "dynamic world (segment)". Here, a "process" should be modelled as "series of moving static still images". "Images" are then "situations", or "snapshots of reality" as McCarthy defines situations [see: McCarthy].

The author of this paper intends to elaborate the fifth generation of the "Study" class as illustration of advanced ontological concepts (going in the direction of Luc Schneider and utilizing the work of the new Leipzig Institute of Formal Ontology and Medical Information Science). The fifth generation of the concept "Study" is up to now thus a stage with an uncertain degree of feasibility. Nevertheless, the linking of realizable and experimental stages of an concept implementation is an important issue of technology assessment.

Last not least, the continuity of one and same concept is aligned with the explication of the subject matter, keeping the explication as close as possible to the use to normal natural language and domain language. The technical explication can and must use artificial constructs to remove ambiguity (with formal languages), for example use such term constructs for different aspects of the concept "Study" as StudyStructure (static), StudyProcess, StudyProcessStructure.

5   The Methodological Quintessence of Evolutionary Information Modelling

Already in the early nineties groups like Hahn et al [see Hahn] discussed the problem of concept versioning and concept drift. They proposed tracking methods to follow evolutionary concepts in dynamic settings and favoured an incremental concept build-up [see Klenner 1994, Hahn 1994]. Therefore, features like versioning, incremental development, iterative improvement have not only to be considered as elements of general software engineering, but are also important as explicit technique for conceptual information modelling in applied domains.

It becomes quite obvious that other techniques have to complement the methodological tool set of information modelling. The most important elements shall be an advanced definition approach [see Swartz] and an applied context theory of meaning.

6   Outlook

Linking information modelling of health systems, basic life science and clinical practice to gain evidence based feedback loops should be a scientific subject topic put forward as issue on a e-health road map. Here, questions and answers concerning the need for and affordability of semantic grid computing [see De Roure] for e-health should have their place.
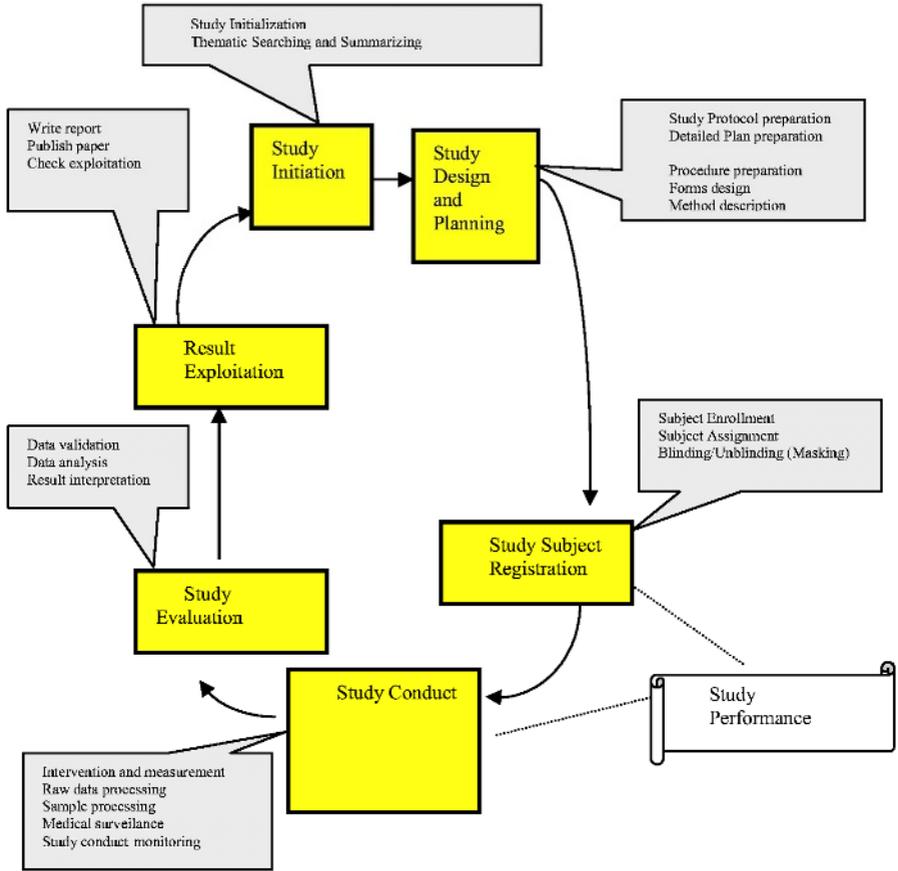
Fig. 2. The life cycle of a study and the main activities

To formulate an e-health road map for Common Europe is a science policy topic. An action item on this "road" could be the set up of a Model Library of European Generic Health System Components. Building an e-health system with a global generic backbone and integrating national as well as cultural variations is a global task to be put forward to gain benefits for all (e-health for all). After all, numerous groups around the globe may contribute to the effort that some day several millions of specific and inter-related health concepts will be available as computer-based re-usable model modules.

References

1. Duwe H, Schwartmann D, Bartels E: The Relevance of System Modelling and Semantic Analysis for the Development of E-Health Applications, International Conference on E-Health in Common Europe, Kraków, 2003
2. Clinical Data Interchange Standards Consortium (CDISC), http://www.cdisc.org
3. Logical Observation Identifier Names and Codes (LOINC). User Guide, edited by McDonald et al, http://www.regenstrief.org/loinc
4. openEHR (Electronic Health Record), http://www.openEHR.org
5. Schneider L: Designing Foundational Ontologies. The Object-Centered High Level Reference Ontology OCHRE as a Case Study, in: I.Y. Song et al. (Eds.), ER 2003, LNCS 2813, 2003. pp. 91-104
6. Regulated Clinical Research Information Model, http://www.hl7.org/Special/committees/rcrim/rcrim.htm
7. Friedman LM, Furberg CD, DeMets DL: Fundamentals of Clinical Trials, St. Louis, 1996
8. Brandt CA et al: Metadata-driven creation of data marts from an EAV-modeled clinical research, International Journal of Medical Informatics, 65, 2002, pp. 225-241
9. Hahn U, Klenner M: Tracking the evolution of concepts in dynamic worlds, in: D. Karagiannis (Ed.), DEXA '94 – Database and Expert System Applications, Proceedings of the 5th International Conference, Athens, Greece, September 7-9, 1994, Berlin, pp. 410-419
10. McCarthy: Actions and Other Events in Situation Calculus, http://www-formal.stanford.edu/jmc/
11. Swartz N: Definitions, Dictionaries, and Meanings, http://www.sfu.ca/philosophy/swartz/definitions.htm
12. De Roure D, Jennings NR, Shabolt: The Semantic Grid. A Future e-Science Infrastructure, http://www.semanticgrid.org