

Electronic Reporting in Pathology: Requirements and Limitations

A Paradigm for National Electronic Health Records Implementation
The ASPE Expert Panel on Cancer Reporting Information Technology

College of American Pathologists

and

The Altarum Institute

Washington, DC

September 2009

Table of Contents

Preface	3
Acknowledgements	4
Executive Summary	5
Section 1. Use Case	6
Section 2. Evidence of Value	11
Section 3: Common Identifiers	15
Section 4: Standard Terminologies	19
Section 5: Beyond Semantic Interoperability	26
Conclusion & Future Directions	31
Bibliography	33
Appendix 1: Unstructured (Traditional) Pathology Report	39
Appendix 2: Structured Pathology Report	41
Appendix 3: XML Pathology Report	43
Appendix 4: Semantic Web (RDF/N3) Pathology Report	45
Appendix 5: Relevant Tumor Registry organizations & projects	48

Preface

This white paper originated in the fall of 2008 from work carried out by the Pathology Electronic Reporting Taskforce (PERT) of the College of American Pathologists (CAP). Under contract support from the Centers for Disease Control (CDC), the PERT was tasked to propose electronic implementations of the CAP Cancer Committee's reporting templates. During discussions, it became obvious to participants that current frameworks for implementing structured reporting had distinct limitations, and that in particular problems of cross-terminology mapping and context representation posed serious hurdles for any possible implementation that aimed to support downstream use cases as complex as clinical decision support. These limitations seemed to be more than transient barriers that might be overcome by incremental technological enhancements to existing frameworks. Rather, they were in many cases inherent limitations of any computer-processable representation of complex medical reports. The Pathology use case, situated in the midrange of complexity of medical report types, provided a superb paradigmatic case for the investigation of these inherent limitations.

In order to understand these issues and limitations, a literature review was funded by the Assistant Secretary for Planning and Evaluation (ASPE), under project management and facilitation provided by the Altarum Institute. A panel of experts was convened to consider the Pathology reporting use case as a computational problem, and to offer perspectives on preconditions for its success. The panel convened monthly by teleconference and convened in Washington DC for two work sessions – October 14th, 2008 and February 6, 2009. The panel members are acknowledged below.

At the panel meetings, participants were introduced to the use case and asked to respond with commentary focusing on prerequisites for successful implementation based on considerations of technological feasibility, computability and inherent representation constraints. Each participant was asked to submit a selection of relevant publications from his or her field of expertise that would provide cogent summaries of the issues and challenges that he or she perceived to be most relevant.

The resulting discussion, the submitted literature references, and supplementary materials derived from PERT deliberations, have been collated and are presented here. We hope that this issues raised in this white paper will provide some perspective to assist in the planning and evaluation of implementation strategies for structured medical reporting in the U.S.

Regards,

Co-authors **John Madden, MD PhD**, Duke University and

Mary Kennedy, MPH, CT (ASCP), College of American Pathologists

Acknowledgements

We would like to acknowledge the following experts who participated in the panel meetings and contributed their valuable information and insight.

Dr. Monica de Baca, Physicians Laboratory & Avera McKennan Hospital

Mr. Christopher Carr, Radiologic Society of North America

Dr. Anni Coden, IBM Research

Ms Kate Flore, Altarum Institute

Dr. James Ostell, National Center for Biotechnology Information, National Institutes of Health

Mr. Eric Prud'hommeaux, World Wide Web Consortium (W3C)

Professor Munindar Singh, Department of Computer Science, North Carolina State University

Drs. John Sowa & Arun K. Majumdar, VivoMind Systems

Professor Jun Yang, Duke University

Executive Summary

Cancer is the leading cause of death in our nation; at any given time there are approximately 16 million patients with a cancer diagnosis in the United States¹ [Horner 2006]. The administration has set an ambitious goal for nationwide implementation of interoperable health records, with the aims of reducing medical error, controlling cost, streamlining communication among providers, and improving care quality. Enhancing management of cancer care ranks among the top challenges.

Patient records today—including existing electronic records—are often transmitted, and typically stored, as free text documents: the physician’s traditional “reports” and “notes”. But our national vision foresees complex data management, aggregation and processing tasks, for which unstructured free text is a poor starting point. For data-aware applications, physician observations, individual test results, diagnostic impressions, and the like must be represented as discrete, computer-processable values. Information must be available in “granular” form.

The challenges and effects of moving to granular data representation—referred to variously as “structured”, “synoptic” (from “synopsis”), or “templated” reporting—is best investigated on the basis of medical document types in which the underlying structure is already fairly predictable: medical reports in which the necessary and optional data content that constitutes the standard-of-reporting-practice have thoroughly been characterized by medical experts. Surprisingly, there are few such report types in medicine. At present, one of the best developed structured reporting domains is cancer diagnostic reporting in Pathology. This development is due to the efforts of the College of American Pathologists (CAP) and American College of Surgeons (ACoS). The efforts to create electronic versions of these checklists have been generously and consistently supported by the Centers for Disease Control (CDC).² In Section 1, we review the development of this structured reporting standard, its current implementation status in electronic reporting frameworks, and the “life-cycle” of documents that use these standards.

There is a widely held belief that adoption of records entry and storage systems that support discrete data entry improve provider communication and saves effort and expense. Yet published evidence for these supposed benefits is often difficult to adduce. In Section 2, we review the evidence for beneficial—or possibly detrimental—effects of adopting structured reporting in Pathology.

The computer-processable representation of free text source as discrete values is intimately related to the use of “medical vocabularies” or “controlled terminologies” to encode medical concepts. Any translation of a medical document into a fixed vocabulary necessarily ignores some contextual information and typically leaves some original intent of the physician’s meaning underspecified. The importance of such under specification varies depending upon the intended application. For search and retrieval of archived documents it may be unimportant. For medical decision support or automated quality assurance, it may be of critical importance. In Section 3, we discuss the status of standardized medical terminologies based on the example of Pathology, and comment on the interoperability problems raised by standardized terminology in document-oriented domains such as medicine.

Shared vocabularies are crucial for encoding the conceptual entities to which reports make reference; but equally important is a shared, easily deployed, and stable system of identifiers for the concrete entities to which medical reports refer: the particular patients, specimens, doctors, family members, and so on that are the tangible subjects of medical reports. In Section 4, we discuss shortcomings of existing universal identifier schemes, and suggest improvements and alternatives to address these shortcomings.

Finally, the effective dissemination of structured reports depends upon their availability within a linked network of healthcare providers who share certain quite specific expectations about data quality and reliability. In effect, the supporting network must provide a trust infrastructure that guarantees that structured data will be presented on the network in a fashion that does not adversely affect the semantic integrity of the data. Such trust can only exist on networks that are designed in awareness of the implicit rules governing data integrity that accompany data sharing between individuals behaving in certain roles within the network community. In Section 5, we discuss the computer professional’s emerging notion of trust-based services. We did not however conduct a detailed review of the larger issues related to the privacy and security of medical records in general.

Section 1. Use Case

In this white paper, we discuss an example use case in the domain of Pathology. The medical specialty of Pathology is concerned with the laboratory diagnosis of disease, and pathologists are physicians who render such diagnoses and oversee hospital-based or independent laboratories that carry out the technical procedures and tests that enable them. Non-physicians often associate the specialty of pathology with autopsies, but most pathologists function in a mixture of two work areas: “clinical pathology” and “surgical (or anatomic) pathology”.

Clinical pathology encompasses the management and interpretation of laboratory-based medicine, to include blood banking, clinical microbiology, clinical chemistry, hematology, and others. In the work of the HITSP under the aegis of the Office of the National Coordinator, the Laboratory Use Case focuses almost exclusively on this aspect of pathology practice. In this white paper, we will not address clinical pathology issues except in passing.

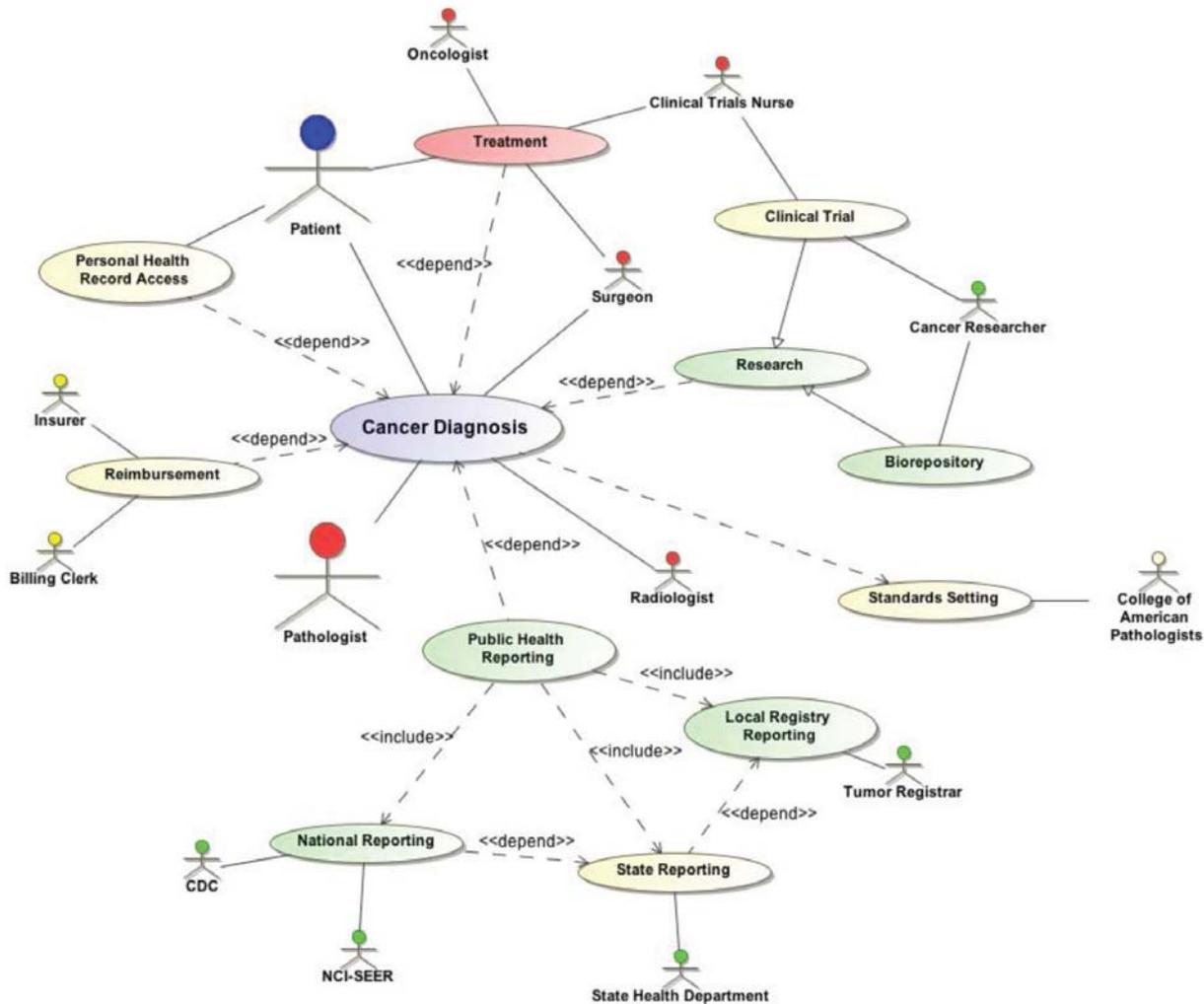
Anatomic pathology constitutes the bulk of the labor of most pathologists, and includes the diagnostic examination of tissues and fluids to establish the presence, nature and extent of disease. Such specimens include biopsies (e.g. breast biopsy, colonoscopy biopsy, etc.); resections (e.g. lung cancer surgery, throat cancer surgery, etc.); cytology (e.g. Pap smears, etc.). The physician or surgeon who performs such a procedure forwards the resultant specimen to the laboratory where a pathologist catalogs it, examines it and dictates a description, and submits some or all the specimen to laboratory technologists who prepare it for microscopic examination. The pathologist receives back representative portions of the prepared specimen as glass slides, which she examines under the microscope to determine the nature of the disease. The pathologist may order special laboratory procedures beyond the usual microscope examination if the case requires it. Finally, the pathologist dictates and signs a full report of her findings, which then becomes part of the patient’s medical record, and which is generally considered the definitive diagnosis of the patient’s illness.

The Pathology cancer diagnostic report is a paradigmatic example of a complex medical report that appears suitable for electronic rendering as a “structured document”.

We focus on this last step, the pathologist’s communication with other physicians and interested parties. We consider how it might be improved on a national basis by applying shared electronic information systems to the process. We focus on a particularly critical type of anatomic pathology specimen: a specimen containing cancer.

When a pathologist renders a diagnosis of cancer, her assessment includes a wealth of information that is critical for determining prognosis and optimal treatment of the patient. This information includes the type and aggressiveness (grade) of the cancer, how far it has invaded locally, whether the tumor was completely excised, the presence or absence of distant spread (metastasis), and special features that might influence treatment or prognosis. The information will be used by a variety of people who play a role in the patient’s subsequent care, in billing and reimbursement, by public health authorities, and (in some cases) by medical researchers (Figure 1).

Figure 1: Actors Involved in Pathology cancer diagnosis report and their dependencies (UML Use Case Diagram).



The pathology report document has a “life-cycle” that involves many downstream data consumers with a variety of interests.

The range of downstream data consumers for a pathologic diagnosis of cancer is a cross-section of the entire healthcare ecosystem, and as such, this use case epitomizes the full range of challenges in interoperable data sharing. As shown in Figure 1, the initial data consumer of the cancer report is the oncologist, whose context is treatment management; and the operating surgeon, whose interest is determination of the adequacy of the resection. Beyond this point, a large number of secondary people become consumers of the pathology diagnostic data. Examples include:

A cancer clinical trials nurse, will use the data to determine whether the patient qualifies for any experimental treatment protocols.

Reimbursement specialists, both in the hospital billing office and an insurer, will use the diagnostic information to determine the appropriate DRG’s for inpatient billing or appropriate outpatient billing.

Quality improvement managers and researchers (with patient agreement to use their information); will find added value in the diagnostic data that may, in part, become linked to a tissue banking or outcomes database.

The diagnosis becomes part of the quality assurance process of the hospital; and may be used in various ways to contribute information for quality metrics. The report itself may subsequently be subject to quality review for completeness and for correlation with other findings for the same patient.

Since cancer is a reportable disease in most state jurisdictions, the diagnosis becomes a data input for the hospital cancer registry. At the local registry the data undergoes further processing and correlation with other clinical records, and is then forwarded to a state or regional cancer registry, and from there to one or more national public health cancer registries [Wingo 2005].

Ultimately insofar as it is available for research, the diagnosis feeds back into the professional standards setting process, public health reports, outcomes studies and quality assurance.

Because of the importance of the pathologic cancer diagnosis to the patient's treatment and to myriad other healthcare functions, systematization of the content of the report has been a high priority of pathologists. The College of American Pathologists (CAP), the U.S. medical specialty organization of pathologists, has periodically convened a committee of cancer experts to assess the state of the scientific evidence regarding cancers of various sites and types, and to recommend requisite and optional data items for inclusion by pathologists into their reports [Anonymous 2009a] "Staging" is the method by which pathologists, oncologists, radiologist and public health workers communicate how advanced a patient's cancer has become, measured according to a standard scheme based on a combination of clinical and pathologic assessment of a variety of parameters. The CAP templates are also closely aligned with standards promulgated by the American Joint Commission on Cancer (AJCC), the North American Association of Central Cancer Registries (NAACCR), the CDC National Program of Cancer Registries (NPCR) and the National Cancer Institute's Surveillance Epidemiology and End Results (SEER) programs. These recommendations are published in periodically updated reports available to pathologists through the CAP website or as a printed document. These templates are one input to pathologists' best practices, and a particularly authoritative one. The sister specialty group, the American College of Surgeons (ACoS) has therefore adopted adherence to these templates as a criterion for assessing whether cancer treatment centers shall win accreditation as "Comprehensive Cancer Centers" under the ACoS's quality assessment program.

“Structured reporting” encompasses a range of document formats, whose suitability for automated processing varies widely.

The complexity of medical reports for future national health information infrastructure spans a spectrum. At one extreme, clinical laboratory reports (e.g. blood chemistries or cell counts) have a relatively predictable structure—making such tests the “low hanging fruit” of electronic reporting. But this predictability makes these inadequate models for highly complex and variable medical report types with substantial amounts of unstructured, free text content: reports such as progress notes and physical examinations.

Because the data elements appropriate to a cancer report are explicitly defined by the CAP and AJCC standards, these heavily structured, yet quite complex reports have long been seen as particularly suitable “mid-spectrum” model for investigating issues involved in translating complex medical reports into an electronically sharable and searchable formats.

To make this concrete, we exhibit here a fictionalized, but otherwise accurate, specimen of a traditional (pre-CAP template) cancer report (Appendix 1). Note that the report, including the diagnosis, is rendered in a free-text style throughout. The report has a sectional structure, to be sure, but the section headers merely adumbrate the content. The substance of the report is human-readable only; it is not machine-processable, though it may be stored as electronic text.

We next exhibit a pathology cancer report formatted according to CAP template standards (Appendix 2). In this version of the report, data items are individually identified in a “question-and-answer” format, with one information item to a line. Two substantive differences from the narrative version are significant. The underlying content model is more precisely specified: atomic data items such as histologic type, tumor grade and so forth are explicitly specified in the model and a list of possible valid responses are given in the guidance document. Furthermore (not shown here), some items are explicitly stated to be required, while others are stated to be optional.

A limitation of the rendition described immediately above is that it remains a text document. Next, we exhibit a possible electronic extract of such a report, rendered in the XML format proposed by the Pathology Electronic Reporting Taskforce (PERT) of the CAP [Madden 2009] (Appendix 3). In this version, each of the elements is designated by a formally defined XML element associated with the CAP by means of its XML namespace, and [Madden 2009] in a defined syntax available in publicly accessible XML schemas. The XML markup renders the individual data items computer-identifiable, retrievable and processable.

Even this XML version has its limitations. Such a report would normally be stored in a local clinical database. While the retained information is locally meaningful, it is not immediately sharable with another hospital's database. In the second database, these data items would likely be stored in differently named database tables and columns, peculiar to the other institution's database logical design. The use of a

standard vocabulary does not of itself make the two databases transparent to each other's queries, since these would necessarily be formulated as queries against the logical design of the database tables, rather than against vocabulary markup.

A fourth rendition of the report is therefore presented in (Appendix 4). In this version, the semantic web language N3 [Berners-Lee 2006] translates the report into a series of assertions, or "triples" [Manola 2004]. It is such complete assertions, connecting a concrete person, specimen or act (designated by a unique identifier) to a conceptual framework (provided by a standardized terminology) that is the basis of all computer processing of data, whether database tables or web language. In the discussion of terminologies this representation will be further explored.

An anticipated benefit of structured reporting of pathology is improvement in the accuracy of coded renderings of medical data in standard vocabularies such as SNOMED CT.

The enhanced consistency and predictability of the structured style should simplify the rendering of pathology reports into computer-processable formats. In Section 2, we will review the evidence that manual encoding of reports is unsatisfactory, and that automated encoding of reports is suboptimal unless the reports are delivered in a structured format. In Section 3, we will discuss findings regarding the state of standard terminologies in the U.S. and their effect on reporting.

Because of its prominence as a standard terminology in the pathology domain, we will frequently refer to SNOMED CT as a terminology, beginning in the next section. For this reason, it may assist the reader to provide some background information about this particular controlled vocabulary.

SNOMED CT [Anonymous 2009b] is a comprehensive, multilingual clinical healthcare terminology. It contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into hierarchies. SNOMED CT also works to provide explicit links (cross maps) to health-related classifications and coding schemes such as ICD-9-CM, ICD-O3, as well as the OPCS-4 classification of interventions. Today, SNOMED CT is available in U.S. English, U.K. English, and Spanish. SNOMED CT is one of a suite of designated standards for use in U.S. Federal Government systems for the electronic exchange of clinical health information and is also a required standard in interoperability specifications of the U.S. Healthcare Information Technology Standards Panel (HITSP).

Section 2. Evidence of Value

The literature suggests that control of reporting variability reduces clinical errors. In pathology reporting of cancer, there is net evidence of such value in a synoptic report process.³ In most of the studies assessing the pathology reporting process, investigators have defined synoptic reporting as the presentation of information in a tabular rather than a descriptive format. The expected benefits of synoptic reporting are the guarantee that all information required for patient management is accurately included. The ease of extraction of data from synoptic reports is considered a secondary advantage.

Structured reporting enhances report quality.

Standardization of cancer reporting may reduce typographic errors as well as errors of omission. In one of the most referenced articles on the use of protocols to reduce medical practice variations [James 2000], the authors demonstrated improvement in reporting critical elements using a synoptic style. In anatomic pathology, the use of synoptic reporting for cancer cases reduced the amount of missing information with an increase in oncologist satisfaction and facilitation in patient care. The authors concluded that as a result, such reports offered a clearer path of communication between pathologists and other clinicians.

In [Karim 2008], the authors reviewed almost 1700 pathology reports with a diagnosis of primary cutaneous melanoma (PCM) to determine whether synoptic formatting increased the frequency with which pathologic features that influence prognosis and management were documented. They found that synoptic pathology reports were more often complete than non-synoptic reports with the same diagnosis, even in a specialized melanoma center.

[Cross 1998] audited the content of reports at a single hospital over a four year period after four different interventions. Initially reports were free text with no standardized guidelines for reporting. Subsequent audit periods introduced test guidelines, flow diagrams and synoptic reports. While each intervention increased completeness of feature reporting, only synoptic reports brought compliance to 100% for the data items audited.

Of note, the consistent recording of *negative* findings in both studies was deemed a significant outcome. By requiring an answer, whether negative or positive, the clinician has the added knowledge that a feature has been assessed. It was found that non-synoptic reports did not always document this distinction.

Optimum clinical communication requires supplementation of structured reports with narrative comments and direct pathologist-physician consultation.

Most authors recognized what the CAP standards also emphasize: that all synoptic reports must include the facility for free text to express degrees of uncertainty, among other reasons. If free text is an inevitable component of even well-designed structured reporting systems, then the problem of accurate textual encoding is unavoidable, and worrisome results regarding coding reproducibility (reviewed in Section 4) are relevant.

No statistically rigorous studies addressed differences of interpretation by clinicians of pathology cancer reports in the narrative versus the synoptic style. [Thompson 2004], in an editorial, stated that clear and free exchange of information was the most important aspect of interaction between pathologist and oncologist. This included a pathology report containing sufficient information to allow evidence-based patient management. The use of synoptic formats in pathology ensured all essential information was reported. However, pathologists must make clear if diagnostic uncertainty is present, and pathologists should refrain from making management recommendations. The use of synoptic report was viewed as a great value to the surgeon because it provided the necessary information to make management decisions and perform accurate staging. If this information was missing, the accuracy of staging and prognostic estimates was compromised.

Structured reporting results in initial workflow disruption. Productivity increases are delayed, but are ultimately achieved.

The literature recognizes fundamental incompatibilities between structured reporting and present-day

pathology workflow practices. A common workflow pattern in pathology practices was found to be a two-phase, paper-to-electronic data flow process [Qu 2007], in which *data collection* was performed on paper and “offline” whereas *data entry* was performed electronically in a second step while “online.” Thus substantial changes in workflow were necessary to implement standardization.

The literature reports a fair amount of resistance among pathologists changing from a totally narrative approach to a synoptic process, with or without accompanying narrative. One study [Mohanty 2007] suggested a drawback of synoptic reporting was the difficulty of persuading pathologists to change personal reporting preferences. These preferences may have little to do with report completeness or quality. In any event, workflow changes are likely to result, at least initially, in productivity decreases. [Mohanty 2007] recorded responses among pathologists to the introduction of a synoptic tool. Initially the synoptic tool was seen as cumbersome and time-consuming, and unsuitable for communicating nuanced diagnoses or findings. With increasing familiarity with the tool, resistance decreased.

There is some evidence that productivity may ultimately increase with the use of structured reporting, if only after a period of initial productivity decline and resistance to adoption. [Murari 2006] reported on an electronic synoptic reporting system for bone marrow cytology and histology reports. Formats used were based on the CAP checklists and incorporated SNOMED codes. Preformatted templates included remarks, gross description, microscopic features, final diagnosis and conclusion as well as sections for free text and numeric values. The study found an increase in productivity using this synoptic format. The time spent in entering and validating synoptic reports was almost 50% less than the time needed for free text reporting. Synoptic reports ensured inclusion of all relevant data and omission of unnecessary details. They were also found to be a suitable means to minimize typographical and transcription errors. Clinicians reportedly preferred the synoptic reports.

Given the changes in workflow that result from structured reporting, it is not surprising that successful implementation of structured reporting requires educational effort. [Chan 2008] studied pathologist compliance with a standardized colorectal cancer (CRC) synoptic reporting system based on the CAP cancer checklists before and after educational sessions. The authors instituted educational sessions covering gross and microscopic reporting protocols using diagrams and checklist templates as visual aids. After a six month period, the authors found statistically significant improvements in all previously underreported gross parameters as well as many microscopic features. A follow-up study 15 months later revealed that reporting of gross parameters and microscopic features continued to improve, the latter showing a 100% reporting of required parameters on the synoptic report. The authors concluded that workplace interventions to promote adoption must be considered if new guidelines are to be successful. In this study, supportive data presented in educational sessions and audits by Cancer Care Ontario may have further enhanced compliance.

Compensating workflow dislocation are the benefits cited above: error reduction, enhanced clinical relevance and inclusion of pertinent information. A particularly error-prone weak link exists at the seam between a paper-based data collection phase and electronic data entry. For example, the lack of a demographic section in the paper CAP checklists makes transfer of this entry of this data into an electronic system vulnerable to a patient mismatch error. To address these issues, [Qu 2007] created a template-based tumor reporting data entry system that was accessed on the web. Data was collected using web forms completed by pathologists, and automatically copied and pasted into text-editor buffers in the laboratory information system. Information in the web browser was purged following entry. The system allowed for free text comments. The authors stated that this system reduced typographic efforts and errors, simplified the reporting process, reduced the error-prone intermediate steps of going from paper copy to electronic format and counteracted the drawbacks of static checklists.

Existing laboratory and hospital software systems are not easily adapted to use structured reporting methods.

Incompatibility exists between present-day laboratory information software and the structured reporting data model, as noted by [Qu 2007]. Most commercial systems store data in a central (typically relational) database. The data schema for the central database tends to be fixed at the time the system is installed, and systems are not designed for frequent schema changes. Yet structured reports, if persisted in a database, typically require new data tables added to the existing database schema to accommodate the uniquely defined data elements that comprise the structured data model. Furthermore, since structured report standards evolve continually as medical content standards evolve frequent modifications to the underlying representations are required. This rapid evolution of the model is at odds with the design and maintenance routine of most commercial systems. In response to this incompatibility, [Qu 2007] offered a hybrid alternative model: a reporting system that can collect report data and disperse it into specific fields in an LIS. The authors warn that including the information into an existing LIS as a single field will make future retrieval difficult, but that even with this suboptimal solution the additional cost of implementation can be significant.

Structured reporting likely enhances the quality of cancer diagnostic data repurposed for research.

There is good evidence that structured pathology reporting facilitates reuse of data for biomedical research. Investigators have noted that lack of uniformity in pathology cancer reporting adversely affects research use of cancer diagnostic data. Inconsistencies result from non-uniform selection of reportable elements, diversity of terms used for common items, and variability in individual and institutional document styles. Standardized checklists have been advocated as a means to minimize these inconsistencies and improve granular and interoperable data collection in electronic data systems.

With the advent of molecular and translational medicine and desire to connect research to patient care, [Mohanty 2007] stated that synoptic reporting systems can be used to convey information to researchers in a consistent, concise way. Using the CAP cancer checklists for hematological and lymphoid neoplasms, supplemented by the WHO classification and additional custom data fields, the authors created a digital synoptic reporting module integrated into a commercial LIS. The system used the question-and-answer paradigm and stored results as discrete data fields. Terms were linked to SNOMED CT to support query activity. Validation logic was built into the system to ensure proper answer choices for all required fields. Synoptic reporting either replaced the traditional free text or supplemented free text in the final report. The authors concluded that this form of synoptic reporting resulted in more clear and consistent reports and reduced the need to re-review slides for missing information. By using discrete values, assessment of quality of care studies were expected to improve. Cancer surveillance was also expected to benefit from synoptic reporting by allowing the extraction of common data elements to populate the cancer registry environments.

[Harvey 2005] studied the changes in pathology reporting and histopathological prognostic indicators. During a ten year period, the quality of the pathology reports markedly improved in parallel with the adoption of a synoptic reporting process. The use of synoptic reports increased the effectiveness of data abstraction by a research scientist. Data in this study was extracted from case records by a research scientist and by a medically qualified author, recording whether each report used checklist format. Data reporting metrics improved during this ten year period in proportion to the number of reports using a synoptic format. In 1989, synoptic reporting was not used in the study area, but by 1999, 94.1% of breast

cancer reports were in this format. Overall, the authors attribute the marked improvement in completeness of reporting prognostic factors in breast cancer to: increased requests by clinicians for more information; introduction of synoptic reporting; standardized national approach to reporting and introduction of mammographic screening. The authors found limitations in the ability of non-pathologists to reliably extract data from text reports without pathologist assistance. When synoptic reporting was used, a research scientist was able to perform the task from most reports; in complex cases pathologist assistance was still needed. The authors concluded that pathologist involvement in data extraction remains necessary to reduce the potential for significant errors despite large improvements attributable to synoptic reporting.

[Tobias 2006] reported a case study using CAP cancer checklists within the Cancer Biomedical Informatics Grid (caBIG), an NCI-sponsored data network for cancer research. The authors stated a common data standard that permits interchange among clinical and research systems is urgently needed to advance tissue-based research. Citing the nearly ubiquitous use of the CAP cancer checklist in pathology practice, the researchers created caBIG-compliant data automated tissue annotation for cancer research using the checklists as a basis. The goal was to preserve the meaning of the CAP paper-based checklists in a way that supported semantic interoperability across diverse CaBIG systems. They reviewed CaBIG infrastructure and conceptualization of interoperability, and they carefully distinguished syntactic interoperability (ability to exchange information) from semantic interoperability (ability to understand and use information). Using caBIG methodology, they created UML models and semantic metadata for three CAP cancer checklists (invasive breast carcinoma, invasive prostate carcinoma and cutaneous melanoma), rendering the intent of the CAP cancer checklists as faithfully as possible. They encountered challenges to accurate rendering, including the need for new complex data types and relationships. In the end, three models were developed and form the foundation for future work to develop all CAP checklists as part of one common information model.

Structured reporting is perceived as desirable to enhance the quality of cancer diagnostic data reused for public health surveillance.

[Wingo 2005] reviewed the history of cancer registries in the United States and emphasized the value of interoperable cancer diagnostic data reporting at a granular level for the Public Health. (The relevant organizations are reviewed in Appendix 5.) The authors concluded that future cancer incidence surveillance should be accomplished within an integrated system that maximizes the use of data management information tools to maintain and enhance efficiency and quality and that allows flexibility in responding to new questions. Timeliness could be improved through the adoption of electronic messaging, standardized vocabularies, and interoperable systems.” They recommended improvements for cancer data collection at all phases of disease from pre-cancer to death. Other sources [Qu 2007] concur that a tumor reporting system should be standard and uniform while remaining institution-specific, and that the development of simple and accessible systems will improve cancer surveillance by cancer registries.

There is no current evidence that structured reporting in pathology makes care less expensive.

No published evidence was found to support the conclusion that structured reporting makes care cheaper. This is in contrast to other domains, for example, in aircraft maintenance⁴, where the end user cost is reduced since the turn-around time for repair and maintenance is faster (therefore, on an hourly billing basis, cheaper). In general, in any service-based industry that depends on skilled knowledge-based work (such as maintenance), if the workflow can be optimized then the end user costs can be reduced.

Section 3: Common Identifiers

A system of common identifiers for persons, events and things is a prerequisite for rendering synoptic-style reports fully computer processable.

During its life cycle the cancer diagnostic report touches people and information systems in many different roles. In the process it becomes semantically linked with many other documents. These links add to and modify the appropriate interpretation of the original document. If such semantic links are to be rendered machine-processable, a mechanism is necessary to uniformly identify other documents and those things to which they refer (including people and objects) in an unambiguous manner. A recent white paper from a leading software architecture consortium has posed the problem as follows:

How should an enterprise identify people and things to optimize its operation and facilitate collaboration with other enterprises? There are too many different ways of identifying people and things; and processes and systems relating to identity have grown up haphazardly and without linkage. This imposes a major overhead on the operation of enterprises today. There are many different ways of representing identities, and there is a proliferation of name forms across different computer systems. [Anonymous 2006]

The same group has distinguished *core identifiers* and *common core identifiers*:

A *core identifier* is an identifier that has the irreducible minimum of attributes, sufficient to distinguish its subject within the scope of a naming authority, and to identify that authority. A *common core identifier* is one that can be used between different organizations. [Anonymous 2006a]

In the realm of national healthcare, at issue are *common core identifiers*, since the goal is integration of services across multiple healthcare providers.

A system of *common core identifiers* is a prerequisite for effective cross-enterprise information technology integration. The *core* property in a distributed data model is analogous to the property of entity integrity in a traditional database, i.e. the property that no two rows in the same database table share the same primary key. The common property in a distributed data model is somewhat analogous to (although weaker than) the property of referential integrity in a traditional database, i.e. the property that every foreign key corresponds to a primary key. Large-scale distributed data storage systems such as Hadoop (Apache) [Anonymous 2007b], BigTable (Google) [Chang 2006], Pnuts (Yahoo) [Cooper 2008] and Dynamo (Amazon) DeCandia 2007 are the closest existing models for distributed datastores of the scale relevant to a national healthcare infrastructure. These stores are all distributed databases that use a simplified relational model to store key-value information. *Common core identifiers* for entities in a healthcare data system would have the same role as e.g. row keys in BigTable, whereas *common core identifiers* for vocabulary items would have roughly the same role as e.g. column keys in BigTable.

Large scale distributed datastores often relax the requirement for referential integrity. Instead of foreign keys, in effect a single, very large, sparse table is used. The absence of a true relational integrity property in these systems corresponds in a distributed data model to the possibility that a single entity may have multiple, unlinked identifiers. The presence of such unlinked aliases in an identifier system is common in the medical domain, where the same individual may be admitted under different names at different times. Thus the ability to tolerate such unlinked aliasing is an advantage, even though ultimately resolution of these links is desirable.

Desirable features of a common core identifier can be specified.

A common identifier has the following desirable properties [Prud'hommeaux 2008]:

Uniqueness: It clashes with previous or future identifier coinages must never occur.

Ownership: Common identifiers often carry some mark of ownership to indicate who created the identifier and has various administrative and/or usage rights.

Human readability: It is helpful if identifiers can be easily recognized and remembered by human beings.

Machine readability: Identifiers should be formatted in a way that is easily parsed, operated upon and stored in most computers.

Social infrastructure: A system of identifiers, to be useful, must have a community of committed users. This community can be small if the identifiers serve a specialized use case. But if the identifiers are intended to serve in a broad variety of use case contexts, then the user community must be correspondingly large.

Self-documenting: Identifiers stand in a relation to the thing(s) they denote. It may be helpful if identifiers are “self-documenting”, in the sense that a defined operation on the identifier can reliably exhibit the identifier’s object or a depiction of the object. This is a property, for example, of web addresses. When entered into a web browser, a web address causes the object of the identifier (a webpage document) to be exhibited on the screen.

Alias ability: Often, the same object may have multiple identifiers. In fact, freedom to coin a name “on-the-fly” for an object without first verifying that no one else has also applied a name to the same object is a practical requirement. But it is also necessary in many cases to be able to “resolve” identifiers, that is, to collect together all the identifiers that refer to a single object and combine them in some sense into a single identifier. Typically, this is done by selecting one identifier of a group of synonymous identifiers to take priority, and linking the others to the priority identifier as aliases.

Persistence: Identifiers, ideally, should never change or go out of existence. In particular, identifiers should not change their object. Two exceptions are of note: 1) Because objects themselves may change (e.g. a single person may become a married person, a living person may die) the definition of persistence itself may be problematic, and 2) In some cases, a single identifier may in the future point to multiple sub-concepts.

Unlimited supply: The supply of identifiers must be (for practical purposes) infinite.

Existing common core identifier frameworks for healthcare use a patchwork of alternative syntaxes.

Existing healthcare information frameworks provide virtually no support for core identification of particular things (medical specimens, pieces of medical equipment, parts of a particular individual’s body, geographical locations, individual actions and events, etc.); patchwork support for core identification of concepts (vocabularies); and some support for common core identification of healthcare providers.

The quest for common core identifiers for patients has been rendered difficult by complex issues of privacy and security [Appavu 1997], by resistance to government as the naming authority, and by the failure to identify any alternate naming authority. The publication of the National Provider Identifier (NPI) Final Rule in 2004 [Pickens 2005; Centers for Medicare & Medicaid Services, HHS 2004] resulted in a national system of identifiers for physicians and providers, but not one for patients. The equally critical need for common identifiers for entities *other than* patients, physicians, and healthcare organizations [Windley 2005] in a national health information architecture has received surprisingly scant coverage in the published literature.

One effort in the direction of a common identifier framework directed expressly at healthcare applications is the Referent Tracking System (RFS) of Ceusters & Smith [Ceusters 2005; Ceusters 2006; Ceusters 2006; Ceusters 2007]. This experimental central repository system provides a framework for tracking identity, defined according to ontological criteria, of entities including physical objects (medical equipment, specimens, etc.) through time. It is thus far more elaborate than simply a scheme for generating and publishing reusable identifiers. Whether practical healthcare contexts might require a system quite this elaborate is yet unknown.

Other than experimental systems such as RFS, what identifier systems are functioning or proposed in the healthcare domain? We consider here sharable patient and provider identification formats, but also systems intended to provide common core identifiers for physical and conceptual objects.

Health Level 7 version 3 (HL7v3) uses a subclass of the ASN.1 [Anonymous 2009c] Object Identifier (OID) type as its preferred identifier for coding schemes and identifier namespaces. HL7v3-compliant systems may also use OIDs for identification of individual information items (objects, concepts). OIDs can be rendered in a web context in a standardized way as URIs [Mealling 2001]. In contrast to other identifier systems in common use in healthcare, OIDs use a system of registries open to users outside the immediate HL7 community, even though this registry system is rudimentary. Therefore OIDs are genuine *common core identifiers*, in the sense defined above. Health Level 7 version 2 (HL7v2) uses a native identifier format based on two- or three-character alphanumeric characters. The identifiers are assigned centrally by

HL7, and hence while they are *core identifiers*, they are not *common core identifiers*.

Integrating the Healthcare Enterprise (IHE) does not propose its own system of identifiers, but instead relies on identifiers provided by other systems that are selected as part of its application profiles.

Medical vocabulary systems each use their own unique identifier systems and syntax, for which the vocabulary authority is the vocabulary publisher.

Identifiers based on Uniform Resource Identifiers (URI) stand out among existing identifier frameworks because of strong existing infrastructure, easy migration path from existing identifier systems, and the highest proven scalability.

Systems of unique identifiers rely upon two alternative mechanisms for avoiding name clashes: namespaces and randomization.

The namespace mechanism associates names with a graph structure that partitions names into non-overlapping families. In the common case that the partitioning structure is a tree, the namespaces are distributed in a hierarchy. This hierarchical structure assures absence of name-clashes. As the tree can be extended infinitely, the supply of names is assured.

There are several namespace-based identifier schemes in existence, including Digital Object Identifiers (DOIs) [International DOI Foundation 2009; Paskin 2005], forms of ASN.1 identifiers including Object Identifiers (OIDs) [Anonymous 2009c] and a variety of others [Anonymus 2007b]. The primary example of this identifier type is the Uniform Resource Identifier (URI) standard [Berners-Lee 2005], a family which includes among several dozen other subschemes, including the familiar web location identifiers of the form “<http://www.example.com>”. In what follows, several different URI schemes will be mentioned, but details of the various available schemes within the URI family will not be exhaustively treated.⁵ An alternate mechanism relies on maximizing randomness of names to guarantee uniqueness. The prototypical example is the UUID scheme [ITU 2004].

The technical landscape of identifiers has been thoroughly reviewed in [Anonymous 2007b]. This group systematically examined candidate identifiers for suitability as a common core identifier. All the standards that satisfy the technical requirements for common core identifiers and that are currently the subject of formal standardization belonged to the URI family. The XRI (Extensible Resource Identifier) scheme, developed in the XRI Technical Committee at OASIS (<http://www.oasis-open.org>), was the group’s recommendation for the global common core identifier form. Since the report was issued, discussions between OASIS and W3C have resulted in convergence between the XRI model and URI schemes in common use on the World Wide Web, in particular the [http:](http://) scheme [Barnhill 2008].

Members of the World Wide Web Consortium (W3C) Technical Architecture Group have consistently argued that novel URI schemes (i.e. schemes other than [http:](http://)) should rarely be used to name information resources on the Web, and that special registries for such identifiers (i.e. registries independent of the existing DNS infrastructure) should probably not be provided [Jacobs 2004; Thompson 2006]. [Sauer mann 2008] discusses additional requirements for the use of URIs as semantic identifiers, including content negotiation mechanisms to distinguish requests about an identified resource that seek computer-processable metadata from those that seek human-readable documentation.

[Tonkin 2008] reviewed a number of identifier schemes (including several also reviewed by [Anonymous 2007b]), with particular attention to the issue of managed persistence. She found, as expected, that persistence is an issue for URI-family identifiers, but she also noted several persistent URL family variants in successful use in the library community, including PURL, OpenURL, the Handle System, and the ARK system. Among the persistent identifier systems reviewed, DOI (a non-URI-family specification) was noted to have the widest dissemination in the library community. However, a specification exists for transmission of DOI information as a URI, and resolution of a DOI to a URI [Anonymous 2006b].

The OID is of special interest because it is the standard identifier for HL7 and DICOM. It was not reviewed in The Open Group report. However, since a well-defined mechanism exists for representing OIDs as URIs [Mealling 2001], the choice between these two representations may depend on the quality of the infrastructure available to support these two types of identifiers. Currently, there are fewer than ten registrars for OIDs worldwide. The largest OID repository (<http://www.oid-info.com>) has roughly 95,000 entries with new entries being added at an average rate of about 12 per day. By comparison, the [http:](http://) scheme in the URI system has hundreds of registrars worldwide, which are estimated [Anonymous 2009d] to have currently registered about 109,000,000 domain names, and to be adding domain names at a rate

of about 65,000/day, about 5,400 times the rate at which new OIDs are being registered. Furthermore, http: URIs are potentially dereferencable (i.e. URLs can be entered into a browser resulting in the retrieval of a webpage), and there exists a worldwide infrastructure supporting de-referencing, namely the Domain Name System), whereas there is no current mechanism for directly dereferencing OIDs.

In summary, there is wide recognition of the exceptional value of URIs as identifiers for real-world objects. In the healthcare domain this would include doctors, medical events, instruments, specimens, documents and any other entity in the healthcare domain. Among URI schemes, there is reason to prefer the http: scheme, although other URI-family schemes have adherents. Identifier schemes currently in use in healthcare applications, including OIDs, lack the infrastructure of http: URIs, but can easily be re-expressed as http: URIs. We do not examine the issue of retooling that might be required in existing software systems to accommodate the URI format, because such retooling is an issue for any of the identifier systems mentioned. Finally, an identifier system for healthcare must interoperate with frameworks for enforcing security, privacy, and trust, since the latter are inescapable requirements for sharing health information. In preparing this white paper, we did not directly review issues of security, privacy, and trust; these are complex issues that go well beyond our use case. It is however relevant to note that in the Web environment, which uses the URI as its indigenous identifier, security is the subject of a strong suite of existing implementation standards [Roessler 2009]; and privacy and trust frameworks are current topics of intensive research [Agarwal 2009; Bizer 2009; Dividino 2009; Liu 2009; Rao 2009].

Section 4: Standard Terminologies

Today's electronic medical records systems are not designed with granular, structured and interoperable-storage of clinical records data as a key functionality, making encoding of synoptic reports challenging.

There is wide agreement that one key to interoperable data in the medical domain will be the availability of standardized terminologies for sharing data that facilitate the use of strictly defined data elements over the free text approach. In the current landscape, patient data is stored by the provider, for example in a hospital or physician's office filing system. In the majority of cases today's systems rely on paper-based document storage, but the obvious hope is for increasing use of electronic storage.

The U.S. medical terminology landscape today is dominated by a moderate number of large, highly evolved terminologies that are centrally curated. Most terminologies are adapted for some particular business-process context. For example, for outpatient billing, the Current Procedural Terminology (CPT) system, curated by the American Medical Association, dominates, and is required for this use by CMS and insurers. For inpatient billing, the International Classification of Diseases (ICD-9) curated by the World Health Organization is used. Many other examples could be cited.⁶

Manual encoding of pathology reports using standardized vocabularies has poor reproducibility.

The literature did offer several studies that attempted to reproduce coding results from previously coded clinical concepts. Studies were also reviewed that attempted mappings between terminologies. The quality and integrity of coded medical data was reviewed in studies where the focus was on intra- and inter-observations using SNOMED CT as the coding vocabulary. While study numbers are small, there is evidence that variability in coding may be an issue for data integrity and reusability. More studies are needed that address issues of variation in coding results. We did not find any systematic studies of the intra-rater or inter-rater reliability of semantic annotation. The literature has little information on consistency and reliability of SNOMED CT coding across institutions. Few studies addressed the differences in coding accuracy and reproducibility for patient records versus coding for research use.

[Andrews 2007] compared the consistency of SNOMED CT encoding of clinical research concepts by three professional coding services. A random sample sent for coding consisted of question-and-answer pairs from Case Report forms. Coders were asked to select SNOMED CT concepts (including both pre-coordinated concepts and post-coordinated expressions) that would capture the meaning of each data item. All three agreed on the same core concept 33% of the time; two of three coders selected the same core concept 44% of the time; and, there was no agreement among all three services 23% of the time. The authors sought to determine the cause of this surprising lack of agreement. For example, was code choice inherently underdetermined? If so, coders should have expressed uncertainty about their selections. But this was not the case: each company evaluated their choice as an "exact match" for the vast majority of items they coded. Nor did companies agree about which items were "hard" to code. When asked to rate their own level of certainty for each item, all three reported same certainty level only 25% of the time; two of three companies reported same level of certainty in 55% of cases; and in 20% of cases there was complete disagreement about certainty. Considering the high variation in actual coding, this high level of certainty and the lack of agreement on which items are difficult to code are alarming. The authors conclude there is a need for efforts to make SNOMED CT more user-friendly. Yet the study did not change the authors' opinion of SNOMED CT as a viable and appropriate data standard for clinical research.

[Chiang 2006] investigated whether a controlled terminology can adequately support EHR systems by measuring coding agreement among three physicians using two SNOMED CT browsers. Both inter- and intra-coder variability was measured using ophthalmology case presentations. Pre-coordinated and post-coordinated concepts were acceptable. The study found that inter-coder agreement was imperfect and unequal and, intra-coder agreement was imperfect. Results obtained from exact code matching were different from those obtained by manual review to determine semantic equivalence. The authors raised concern about the reliability of coded medical information in real-world situations and also for retrospective clinical research. The specific browser used affected the way the reports were encoded, and it was suggested that improved browsing tools may improve reliability.

[Rothschild 2005] studied inter-rater agreement in physician-coded problem lists by having ten physicians review each of five cases. They concluded that inter-rater agreement in non-standardized problem lists is moderate, and that much variability may be attributable to differences in clinicians' style and the inherent fuzziness of medical diagnosis.

Finally, with regards to validation of the logical consistency of an ontology, [Wang 2008] developed a method for detecting "complex concepts" within SNOMED CT, which map upward to more than one hierarchy. Complex concepts were found to have an error rate (as defined in the paper) of 55%, while a control sample had an error rate of 29%. The study was limited to the Specimen hierarchy and did not examine the remainder of SNOMED CT. Issues raised in this article about the Specimen hierarchy have been addressed by the IHTSDO.

Automated encoding of reports will likely require a combination of techniques including full-text indexing, language processing and constrained entry.

To make free text machine-processable, rather than merely retrievable by loose criteria for strictly human perusal, several strategies are possible.

Full-text indexing of text fields in medical databases is the best-tested strategy [Hanauer 2006; Erinjeri 2008; Erdal 2006; Ding 2007; Wilcox 2003]. Modern text-indexing systems are extremely fast and thorough, and allow for very complex search criteria to be specified. Full text search indexes are not commonly provided in commercial medical records systems today, and advocates argue that they are underutilized. One obvious limitation is that purely lexical indexing fails to take account of the fact that natural language makes extensive use of synonyms and alternative spellings for words with similar or identical meaning. But this limitation can be remedied by supplementing the lexical index by synonym lists, and indeed such synonym lists are among the simplest types of "terminologies". The addition of "smart" retrieval algorithms to full-text indexing has proven a remarkably scalable and useful strategy for dealing with masses of free text (as for example in the domain of web search engines) [Moskovitch 2007].

A second strategy is to use *natural language processing* (NLP) techniques to parse free text into logical units based on algorithms that take grammatical, as well as lexical, criteria into account [Jagannathan 2009; Wang 2009; Chen 2006; Friedman 2004]. Some experts consider controlled natural languages as a necessary bridge to support human computer interactions. They point out that natural languages evolved to express and support human ways of thinking while computer languages enable IT professionals to think about the data and operations inside the computer system. Forcing clinical subject matter experts (SMEs) to think about their own subject in computer terms is counterproductive because:

- 1) Many of the SMEs simply become bad IT professionals;
- 2) Some of the SMEs become good IT professionals, but compromise and distort their intuitions about their own subject; and
- 3) Ultimately very few SMEs become good at both.⁷

The advantages of NLP are based on their ability to infer relationships among topics not merely based upon proximity of words but on linguistic structure. This sensitivity to structure allows, in theory, more accurate computable depictions of the intended meaning of the text. NLP techniques recognize that meaning consists not merely in juxtaposition of entities, but equally in the syntactic relationships among entities.⁸ This strategy has already been implemented in the field of cancer diagnostics. The National Cancer Institute's caTIES automatically extracts coded information from "free text surgical pathology reports" to support cancer research. However, no accuracy results have been reported. More recently [Codon 2009] has described an updatable knowledge representation model for cancer research (MedTAS/P) that uses open source NLP tools to parse free text pathology reports. Of note this study developed a methodology to validate the system against a set of colon cancer pathology reports. Results were highly satisfactory for histology and lymph nodes, with lower scores for metastatic tumors (possibly due to the limited number of samples in the training and test sets).

A third strategy, epitomized by the CAP structured cancer reporting templates that are the subject of this white paper, is to restrict the freedom of the author by presenting a *predetermined form*, rather than a blank page, as a slate for content entry. The limitations of free text are overcome by making the text less free. When appropriate, these forms require the pathologist to select from a predetermined set of coded responses. The codes may be derived from or mapped to other logical structures (such as ontologies) to

enable complex analysis of the information. Free text inputs are limited to fields of narrowly defined scope. As we note below there are several methods of implementing this third strategy of constraining free text, including logic-based strategies that recognize that assertions consist of a subject and a predicate, the latter in turn consisting of a property and its value. More rudimentary terminological artifacts like synonym lists provide limited recognition of this fact.

A fourth strategy is to integrate all three prior approaches by merging database and information retrieval systems [Chaudhuri 2005]. As noted above, there are advantages and disadvantages to both free text and constrained data entry. Researchers have begun to study approaches that balance database functionality (structured data or DB), with information-retrieval (or IR as typified by NLP driven internet search engines) and have noted that the use cases that originally drove these two functions were very different. Traditional business applications such as payroll and inventory control drove database applications, and article or abstract indexing drove information retrieval. Today “virtually all advanced applications need both structured data and text documents, and information fusion is a central issue.” The authors reviewed currently available commercial technologies in this field, discuss architectural issues and review solutions and challenges in query optimization and evaluation.

More recently, [Weikum 2009] discussed barriers and opportunities in this rapidly growing field and give relevant examples of the utility of IR-styled queries across structured databases, and database styled queries on originally unstructured data. The article discussed several projects that attempt to bridge these fields as well as use of a query language based on extensions to the current SPARQL protocol and RDF query language. The authors conclude that while deep DB/IR integration “may still be wishful thinking” there is considerable nearer term opportunity to providing increased functionality to end users by integrating information from both methodologies.

No single existing terminology provides fully adequate coverage for all important report contexts, yet mapping among terminologies to extend coverage is technically cumbersome and yields poor quality results.

Early work by Cimino’s group [Patel 2007a;Patel 2007b] identified the variety of challenges that may arise in mapping among existing controlled terminologies. To take an example, Is SNOMED CT adequate for coding clinical research? Are there SNOMED CT structure and implementation issues that are unique to clinical research applications? [Richesson 2006] discusses the difference between standards in clinical medicine and the use of standards for clinical research and state there is little known about whether data standards for clinical medicine are adequate for clinical research. A study was conducted to determine if concepts on selected case report forms could be represented in SNOMED CT, not how they would be coded. A sample of data concepts from several vasculitis research studies was used to estimate the coverage provided in SNOMED CT for clinical research as well as the semantic integrity relevant to post-coordinated concepts.

Results showed that most clinical concepts needed for clinical research were covered by SNOMED CT. The key finding was that a majority of the concepts to be coded could only be partially represented by existing SNOMED CT concepts but rather had to be post-coordinated to either clarify context or to better capture complex clinical concepts. Either way, the authors conclude that both domain expertise and an intimate understanding of SNOMED CT are important requirements for effective coding. This study also suggest the need for improved understanding of how best to capture context in clinical research. The authors offer that the clinical research community must clearly define the purpose of data standards, whether it be to share data items (e.g. from a CRF), to share concepts, to represent context or to share data sets.

[Patel 2007a] describes a large case study that explored the applicability of medical ontologies (i.e. SNOMED CT), used primarily for terminology services, for use in automating common clinical tasks (e.g. cohort selection of patients for clinical trials). The assumption was that there is a need to bridge the semantic gulf between raw patient data, such as laboratory tests, and the way a clinician interprets the data. The technical challenges of knowledge engineering (mappings from MED to SNOMED CT), scalability (reasoning) and noisy data (cleansing) are the basis for the solutions proposed by the authors and are discussed in detail. The events used in the case study were laboratory test results, radiology findings and drug treatment.

An interesting problem concerning clinical data and clinical trials queries is that of open versus closed

world reasoning. While the former works for reasoning in radiology and laboratory data, the latter is used in pharmacy data. The authors assert that integrating the open and closed world reasoning is a key issue to resolve.

To be useful, terminologies must be expressive enough to cover a domain, yield reproducible results when implemented by a wide range of end users, and be maintainable through the expenditure of reasonable resources. It may be necessary to map terminologies so that comparisons can be made between specific terms of differing terminologies. Given this background, the literature was reviewed to determine the utility of these terminologies to the cancer checklist use case, including the broader ability to link this use case with similarly structured documents from other domains. No references were found that specifically addressed SNOMED CT's coverage of the Electronic Cancer Checklist use case. However SNOMED has a long history of use in the field of anatomical pathology, and is currently being reviewed for this coverage as part of an ongoing contract between the CDC and CAP. It should be noted that several studies have analyzed SNOMED CT's coverage of problem list coding for the general medical use case. For example in [Brown 2006] 1573 terms from the Veteran Affairs (VA) general medical evaluation template were studied. Sensitivity of SNOMED CT as a reference terminology was 63.8%, ranging from 29.3% for history items to 92.4% for exam items. SNOMED CT's sensitivity as an "interface terminology" was 55.0%.

[Elkin 2006] reviewed SNOMED CT's coverage of the 5,000 most frequent Mayo Master Sheet Terms. The investigators used both a NLP interface and browser. SNOMED CT had a sensitivity of 92.3%.

[Wasserman 2003] evaluated the scope of SNOMED CT through coding diagnosis and problem lists within a computerized physician order entry (CPOE) system. The study involved the mapping of local synonyms to existing SNOMED CT concepts. The system involved a caregiver selecting a diagnosis either from a list of frequently chosen diagnosis terms or through a free-text query of a subset of SNOMED CT database. In the latter, the user has the ability to view SNOMED CT concepts at their less granular or more granular levels.

The results showed that the majority of terms from these lists were found in SNOMED CT. The authors reported that the concept coverage overall was 98.5% Missing descriptions were considered simple to remedy. While there were relatively few missing concepts, their absence was considered significant and required the generation of a new concept with both single and/or multiple parents. Issues arose when caregivers searched by abbreviations (e.g. BPH) which resulted in the inclusion of new local descriptions to speed the free text query.

The authors state the benefits of computerized problem lists, such as used in this study, include more readily accessible information as compared to a paper chart and the codified terms can lead to clinical decision support features. A limitation of the study may have included the clinician's acceptance of terms that did not exactly match the intended meaning.

[Vikström 2007] focused on efforts to map the Swedish version of ICD-10 (ICD-10/KSH97-P) to SNOMED CT. The study involved the exploration and development of mapping rules; evaluation of this inter-coder reliability in the mapping (two coders); and the characteristics in the two coding systems that may hinder high quality mapping. The CLUE browser was used. The paper notes that differences between mapping involving terminologies and classifications are designed based upon the intended use of the map and may be dissimilar for different use cases.

Results of the study found that mapping rules were important and evolved over the course of the project. New mapping rules implemented after the first sequence significantly affected the results obtained from the second sequence although there was no significant improvement between the second and third sequences. The inter-coder reliability reached 83% by the end. Most obstacles to high quality mapping were differences between coders due to structural and content factors in both SNOMED CT and ICD-10/KSH97-P. The authors concluded those obstacles which prohibited higher quality mapping included similarity between concepts in SNOMED CT; structure of the exclude rule in ICD-10 as well as inconsistencies in the ICD-10 axes of classification; and the decision not to use post-coordination of terms. Limitations of the study include the use of an entirely manual mapping process and the lack of a Swedish translation of SNOMED CT.

[Wade 2009] explored the impact revisions to SNOMED CT had on terminology interfaces at a large medical center. Using the 2005 version of SNOMED CT 1570 of the initial 2002 terms were expressed as SNOMED CT concepts. Of these, 1118 initial mappings (71%) had to be revised to achieve consistency with the 2006 release.

[Bodenreider 2008] focused on issues in mapping LOINC laboratory tests to SNOMED CT. This proved challenging as LOINC consists of 6-axes (component, property, time, system, scale, and method), and each of these tends to be expressed at a greater level of detail than the corresponding SNOMED CT expressions. Thus LOINC is finer-grained than SNOMED CT and a large number of LOINC codes could not be mapped. For example out of 9511 LOINC codes there was no match for 1697, and only the component matched for 5686.

These findings emphasize that mapping itself is use-case specific and vocabularies can only be successfully mapped if:

- 1) Coding is accurate and reproducible (browser dependent).
- 2) Origin of the code system is critical to accuracy (context specificity).
- 3) Use of codes for intensive reasoning tasks (e.g. clinical support) requires a logically expressive ontology, and most existing medical ontologies lack the requisite expressivity.
- 4) Selection of code system should depend on the intended application and should differ for research, clinical decision support, search & retrieval, billing, etc.
- 5) Technology for mapping among coding systems is inadequate to support the study's use cases

These challenges have precedents. Prior experience in the field of artificial intelligence [Anonymous 2008] has indicated that similar projects become exponential as their scope increases, and that small projects with a hierarchical structure are most likely to succeed. The CAP electronic Cancer Checklists eCC fit both these criteria as they have limited scope (categorizing the type of malignancy and its anatomical spread), and are organized by organ system. These limitations may become more relevant as these reports are combined with a broader set of information.

Reformulation of existing terminologies into more expressive logically-based and standardized modeling languages may provide superior utility and interoperability for encoded report information.

[Rector 1999] proposes that there are ten reasons why the development of a reusable clinical terminology for patient-centered systems has proven to be so difficult. The paper begins with three assumptions about the context in which terminologies are used:

1. The purpose of clinical terminology is to support clinical software;
2. All terminologies will have to support conversion to existing reporting coding schemes; and
3. All terminologies will need to be multilingual.

In addition, the author states that the literature offers little into what the specific tasks a terminology needs to perform or facilitate. The author goes on to define the types of information that needs to be collected, the tasks to be performed with that information, and the potential end-users of the information.

The ten reasons for difficulty in the development of clinical terminology are discussed in length. They include the vast scale that terminologies must serve; conflicts between user needs and software requirements; complexity of clinical pragmatics; separation of language and concept representation; clinical conventions versus logical/linguistic paradigms; underestimation of the difficulty of defining formalisms and populating them with clinical knowledge; need for terminologies to allow local tailoring; issues with existing classification systems; coordination between the medical record and messaging models; and the need to manage change.

The author proposes that for a clinical terminology to be relevant it must solve problems associated with clinical linguistics, clinical pragmatics and format concept representation. He continues with four possible approaches to solving these issues. They include the notion of simplifying the problem to achieve the highest priority as well as applying more effort, combining these efforts and obtaining better tools.

Lastly, the author addresses the conflict between the needs of software and the needs of human users. As part of the resolution there is a need to validate the use of clinical terminologies in actual implementations. Successful terminologies will be those that routinely share the same information across independent systems.

[Rector 2008] discusses several problem areas of SNOMED CT but focus on three particular issues:

SNOMED CT's "context model"; the representation of part-whole relations; and, the problems of determining semantic equivalence between findings and observables. The authors argue for a schema that integrates context with other concepts; one for concepts themselves; and one for concepts occurring in situations. Such a schema would require SNOMED CT logical formalism to include negation, disjunction and general concept inclusion axioms. The authors state the obvious choice for such formalism is W3C standard Web Ontology Language (OWL).

The authors discuss at length the major advantages to using a more expressive language that include: a uniform, clear, and understandable schema for all concepts used in clinical records; eliminating the need for special mechanisms to deal with context, parsimony and role groups; leveraging of the logical representation to organize and quality assure the hierarchies; improving the ability to recognize semantic equivalences between pre and post-coordinated expressions; improving the ability to modularize SNOMED CT for specific purposes; and, allowing access to techniques developed by the Semantic Web.

Reformulation of SNOMED CT would involve a threefold process: the syntax would need to be transformed to OWL 1.1 using entirely automated processes currently available; the explicit content of SNOMED CT's definitions would need to be reviewed and extended; and different tools and classifiers would need to be used. The authors argue that the result would be a SNOMED CT that is more regular, uniform, and has better defined and consistent semantics allowing for easier querying ability. The authors also admit that these assertions need to be tested by a feasibility study on a limited subset of SNOMED CT. For example in the United Kingdom 1000 (or 1.1% of the codes) of READ codes accounted for 81% of coded data of primary care practitioners while 10,000 codes accounted for 99%.

The semantic web approach to terminology differs in a number of ways from the existing state of affairs in medical terminology in the U.S. Characteristics of the semantic web approach include:

- 1) a common format for identifiers,
- 2) a common syntax, and,
- 3) a common formal semantics for vocabularies precedes the dissemination of content.

Given a common syntax and formal semantics, the ability to create, cross reference, merge or extend vocabularies is available to all users. In the Semantic Web model, the common identifier format is the Uniform Resource Identifier (URI) [Berners-Lee 2005]; the common syntax is the RDF triples syntax [Klyne 2004b]; and the common formal semantics is the model theoretic semantics of RDF/OWL [Hayes 2004], that defines the criteria for determining the truth of an assertions, and allows determination of the computational complexity of the task of computing the truth of a statement.

Therefore, in the semantic web environment, vocabularies are intended to be inherently mixable. Vocabularies tend to be smaller, more intensively modeled and even more highly contextualized than in the existing U.S. terminology landscape.

How can the two worlds of the existing terminology landscape and the emerging web-enabled terminological landscape be bridged?

Of central importance is the development of a common system of identifiers that can facilitate access to terminological knowledge by standard software and facilitate mixing of terminologies. The UMLS Metathesaurus curators developed one approach to this requirement [Anonymous 2009e] when they created the system of concept unique identifiers (CUI), lexical unique identifiers (LUI), string unique identifiers (SUI) and atom unique identifiers (AUI) that give an interoperable format to all conceptual and lexical entities in the source vocabularies. This system has a somewhat different intent than what is described here, since the unique concept identifiers of the source vocabularies, while preserved as informational items in this system, do not have guaranteed one-to-one mappings to any one of the four UMLS identifier types. A similar system is the Mayo LexGrid system [Pathak 2009], which assigns to each mapped identifier. Further discussion of common identifiers is elsewhere in this report.

A second key concern is the need to maintain a faithful semantic rendering of terms that have already been established in earlier systems through specification and use, while at the same time translating these terms into new formats and use cases in the web environment. Re-rendering existing vocabularies in new syntax necessarily puts pressure upon—and may do violence to—the established semantics of the term as represented in the legacy vocabulary. The alternative is to completely redesign vocabularies for a linked, cross-enterprise environment, but the cost of retooling makes this alternative undesirable.

A third concern is the choice of appropriate “helper” technologies to ease the transition from legacy document representation styles and semantic representation styles to new environments. An important component of this task is appropriate choice of common core identifiers for concepts across ontologies. The concepts themselves need not be identical, but the identifier format should be compatible across systems so that mapping can be performed using a common technology. This issue is discussed in detail in a subsequent section on Universal Identifiers.

Current terminology schemes used in medicine typically annotate medical documents by appending the term to the document in the manner of a label. For example, a list of terms names or identifiers might be listed at the end of the document, or in one or more special (possibly hidden) sections within the document. But it is rare to see structure applied to the terms beyond listing them. In a few cases (the HL7 CDA is an example), it may be possible, albeit with some difficulty, to record the precise location in the text of the document that caused the label.

Several examples of highly interoperable information representation frameworks currently exist that offer a semantic model and standard syntax within which arbitrary vocabularies can be expressed. The more expressive of these, including Common Logic (CL) [International Organization for Standardization 2007] and OWL [Bechhofer 2004] are logic-based formalisms that allow representation of complex interrelationships among terms and can serve as inputs to computer inference engines that are capable of using the information for tasks like medical decision support. Less expressive formalisms are also available, that allow simple hierarchical or list-like relationships among concepts to be captured. These are unable to capture the more advanced relationships expressible in logic-based formalisms, but are unquestionably simpler to understand and adequate for certain kinds of tasks, including many types of search and retrieval.

Note that in this categorization, the distinguishing feature of ontology is that terms in the vocabulary are understood as “classes”, or sets of instances, rather than as abstract ideas. This view of terms in an ontology as describing sets of actual data in a computable way allows us to use ontology terms not merely as “labels” for data, but rather as the computable schema for real-world data itself. Thus the transition from taxonomy to ontology “closes the loop” between data and vocabulary by allowing us to talk about real-world things using the identical syntax that we use to model our concepts. (In some circles, this is referred to as closing the gap between the “terminology model” and the “information model”). Viewed in another way, any vocabulary scheme at the level of a taxonomy or below lacks a sufficiently well defined computability relationship to the data to serve as the full-blown underlying schema for an actual database of patient data; whereas an ontology has a sufficiently well-defined and computable relationship to the data to be used, if desired, as a formal specification for a database schema.⁹

Ontologies are invariably founded upon the model of a well-formed expression consisting of a subject, predicate and object (i.e. is a triple) [Klyne 2004a]. Whereas legacy terminological markup systems tend to use term symbols as a labeling device, logic-based languages use vocabularies to construct complete sentences, or assertions, each of which may have a truth value. Rector has forcefully argued [Rector 2008] that the OWL language is the most appropriate language for expressing knowledge in the SNOMED CT domain.

Section 5: Beyond Semantic Interoperability

The ability to share data syntactically and semantically is necessary but not sufficient to support pragmatic (functional) interoperability among systems.

The IEEE¹⁰ defines interoperability as “The ability of two or more systems or components to exchange information and to use the information that has been exchanged”. This definition distinguishes two aspects of interoperability: the ability to exchange information (data interoperability), and the ability to make use of the information (functional interoperability).

It is common to analyze interoperability as consisting of three dimensions or levels.

- 1) **Physical interoperability** is the ability to connect systems at the level of electrical signals.
- 2) **Syntactic interoperability** is the ability of systems to decoded message formats into a common set of symbols and to distinguish well-formed from malformed messages. In Section 3 of this paper, common identifiers were discussed, a core prerequisite for syntactic interoperability.
- 3) **Semantic interoperability** is the ability to share unambiguous meaning. In Section 4, common vocabularies were discussed, a core requirement for *semantic interoperability*.

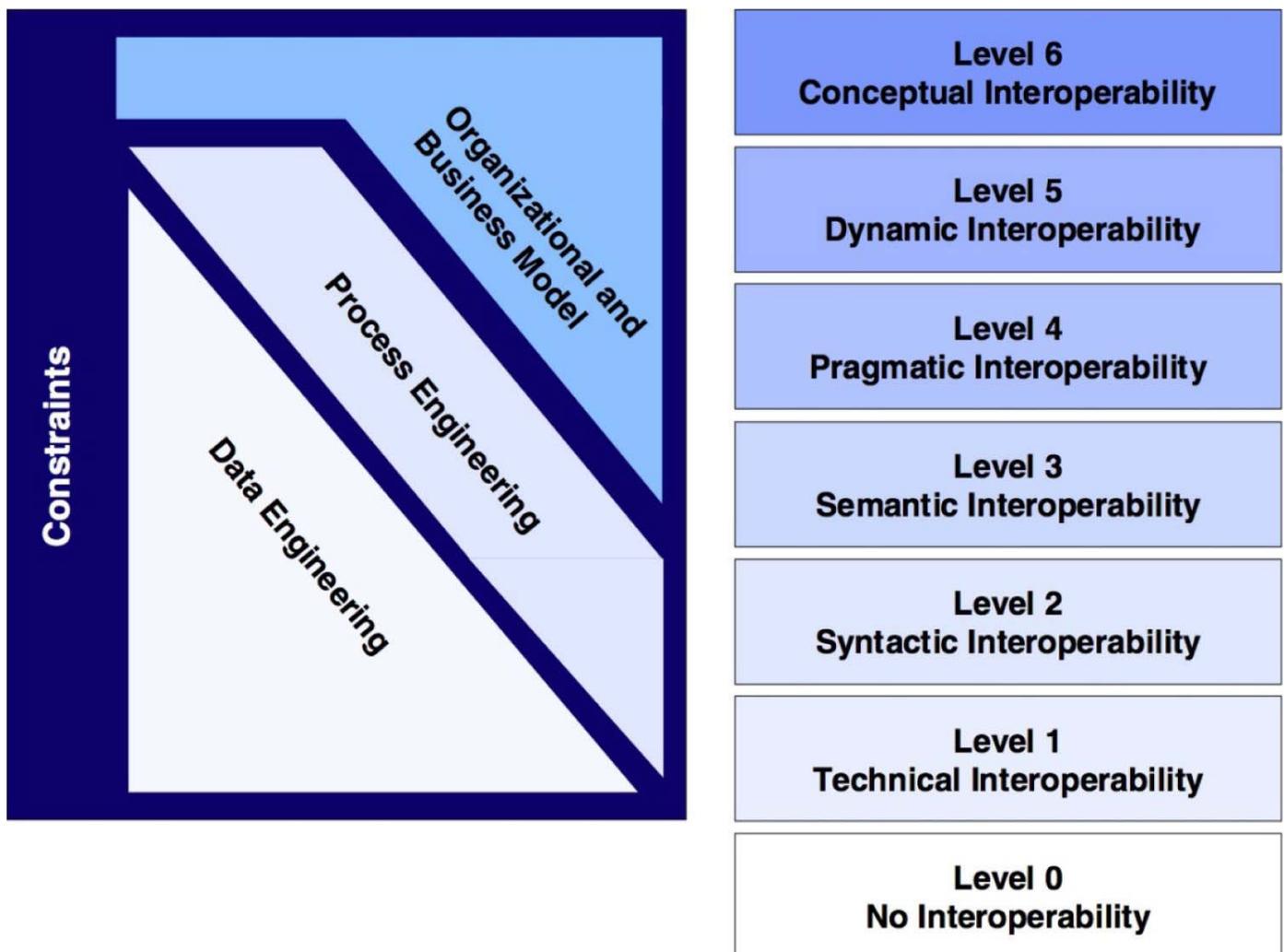


Figure 2: Layers of Conceptual Interoperability Model (LCIM) according to Tolk, developed in conjunction for DoD/NATO in the course studies of modeling & simulation of complex battlefield scenarios [Tolk 2009].

Yet in the past decade, as experience in complex systems in knowledge-based environments has accumulated, it has become apparent that these three dimensions of interoperability delineate only the lowest levels, mostly relevant to data interoperability. They fail to address higher dimensions of functional interoperability. In particular they fail to address questions of how knowledge can be rendered functionally

interoperable across business contexts, across temporal contexts, and across strategic contexts. In the case of the pathology surgical report use case, we showed in Section 1 that the report document has a “life-cycle” and that it is reused for a wide range of purposes: billing and financial, clinical research, public health, quality assurance, health planning, legal, and many others.

Investigations of Tolk and others for the USDoD and NATO regarding battle management modeling and simulation, and standards proposed by the USDoE for interoperable energy grid management [GridWise Architecture Council 2008] exemplify the recognition that in complex systems-of-systems—in particular systems that integrate multiple agents with different goals—functional (“substantive”) interoperability requires integration that extends well beyond the level of common syntax (“Level 2” in Figure 2), and even beyond the level of common vocabulary (“Level 3” in Figure 2).

This is because the relevant constraints on the functioning of the complex system at these levels arise not primarily from incompatibilities in the data, but from the way the data is used in a purposeful process, and from the roles of the executing agents in an organization (Figure 2, left side). [Tolk 2009] describes interoperability levels 4 through 6 of the Layers of Conceptual Interoperability Model (LCIM) illustrated in Figure 2 as follows:

Level 4: Pragmatic Interoperability is reached when the interoperating systems are aware of each other’s methods and procedures. In other words, the use of the data—or the context of its application—is understood by the participating systems ...[which] implies the awareness and sharing of a common reference logical model.

Level 5: At the Dynamic Interoperability level, interoperating systems are able to comprehend and take advantage of the state changes that...each other are making over *time [italics ours]*...[which] implies that systems understand how the symbols they exchange are used during run-time.

Level 6: Conceptual Interoperability...requires that systems share a common reference conceptual model that captures the *assumptions and constraints [italics ours]* of the corresponding real or imaginary object. [Tolk 2009]

Pragmatic interoperability extends semantic interoperability by awareness of the procedural context in which the data will be used, and therefore implies not merely a shared vocabulary but also a shared logic. Dynamic interoperability extends pragmatic interoperability by awareness of temporal changes in the meaning of the data, which (we noted in Section 3) is a computationally difficult problem [Zhou 2007]. Finally at the highest level of Conceptual Interoperability, assumptions and constraints must be captured, which implies awareness of policies and commitments among actors.¹¹

Within the U.S. national health domain—unlike the military and energy domains—systematization of these upper levels of the interoperability stack seem to have received little systematic attention. The Australian e-health effort has published a framework paper on a comprehensive interoperability framework [Anonymous 2005], but we are unaware of any attention devoted to this important problem area by in the United States.

How do these frameworks pertain to the pathology reporting use case? In the pathology report life cycle, a given data item has different functional meanings in different business contexts. For example, a pathologic diagnosis may appropriately determine the patient’s subsequent clinical care during an admission, but may represent misleading information in the context of determining the patient’s DRG for purposes of inpatient reimbursement. A diagnostic report which is complete in terms of required elements may be incomplete in terms of information content in determining eligibility for a clinical trial. Any framework for supporting reporting must allow for the pathologist to attach comments to any item in the report that may color, qualify or modify the templated diagnostic item; in determining treatment in areas where best practice is not clear-cut, or where multiple alternatives exist, these items may be the most important information in the report. All of these scenarios represent the effect upon semantics of information-in context (LCIM Level 4 “pragmatic interoperability”; GridWise Level 5 “business context”), which are not adequately addressed by syntactic standardization (common identifiers) or semantic standardization (common terminologies).

Temporality (LCIM Level 5 “dynamic interoperability”) modifies the semantics of pathology report items. A tumor classified in one way on biopsy may undergo reclassification upon resection, pathologic staging may change if new material is resected, and an initial diagnosis may subsequently be altered following internal or external review. In all these cases and in many more, data exists in the system which is true at one time point, but false at another and a particular assertion may “flip” from being true to false many times. Semantic interoperability is not the final stage of useful interoperability, but only a preliminary stage.

The implementation model currently most suitable to support higher-level interoperability is the Service Oriented Architecture (SOA), but current clinical systems and standards have not yet achieved the level of maturity necessary for such an architecture.

If true interoperability cannot be achieved by semantic standardization alone, what are the architectural prerequisites? Architecture is defined by the IEEE as the fundamental organization of a system embodied in its components, their relationships to each other and to the environment and the principles guiding its design and evolution. The best current model of the evolution of the architecture “stack” is provided by the Open Group Service Integration Maturity Model. The IMM divides large-scale software architecture evolution into two major stages, each consisting of several levels. The Service Foundation stage has three levels and mature Service stage has four levels. Figure 3 shows these maturity levels, arrayed in order from left to right. Each level consists of eight dimensions, shown on the diagram as the row headers.

In this model of architectural maturity, the dimension of Information standards does play a role, but not the sole role. Based on the discussion in Section 4, the state of maturity of the “Information” dimension in the U.S. healthcare environment today lies near Level 4. In some cases (like specialized clinical vocabularies) Information dimension maturity is still at Level 3, Level 2 or even Level 1. It clearly falls short of Level 5 (“Enterprise wide standard data vocabulary”), although some terminologies (SNOMED CT for example) express ambitions in this direction.

In the “Applications” dimension U.S. clinical information technology today is largely module-based, i.e. is at OSIMM Level 1. Clinical software systems today may not be object-based (Level 2) designs, or (if they are) typically do not expose their objects outside of their own implementation silo. Cross-enterprise clinical software today makes extremely scant use of reusable, cross-implementation components (Level 3), and in HL7 discussion of support for services (Level 4) is only lately underway.

Similarly, on the “Methods” dimension, much existing standards work is based on structured analysis and design (Level 1) methodology. Upper levels of HL7 version 3 (e.g. the RIM) were developed using an early object-oriented modeling methodology (Level 2), but focused on software interactions as XML message exchanges rather than as object calls. The elaboration of HL7v3 involves a hybrid structured analysis and object-oriented modeling process. Component-based development (Level 3) and services (Level 4) remain novel concepts with respect to many existing clinical information systems.

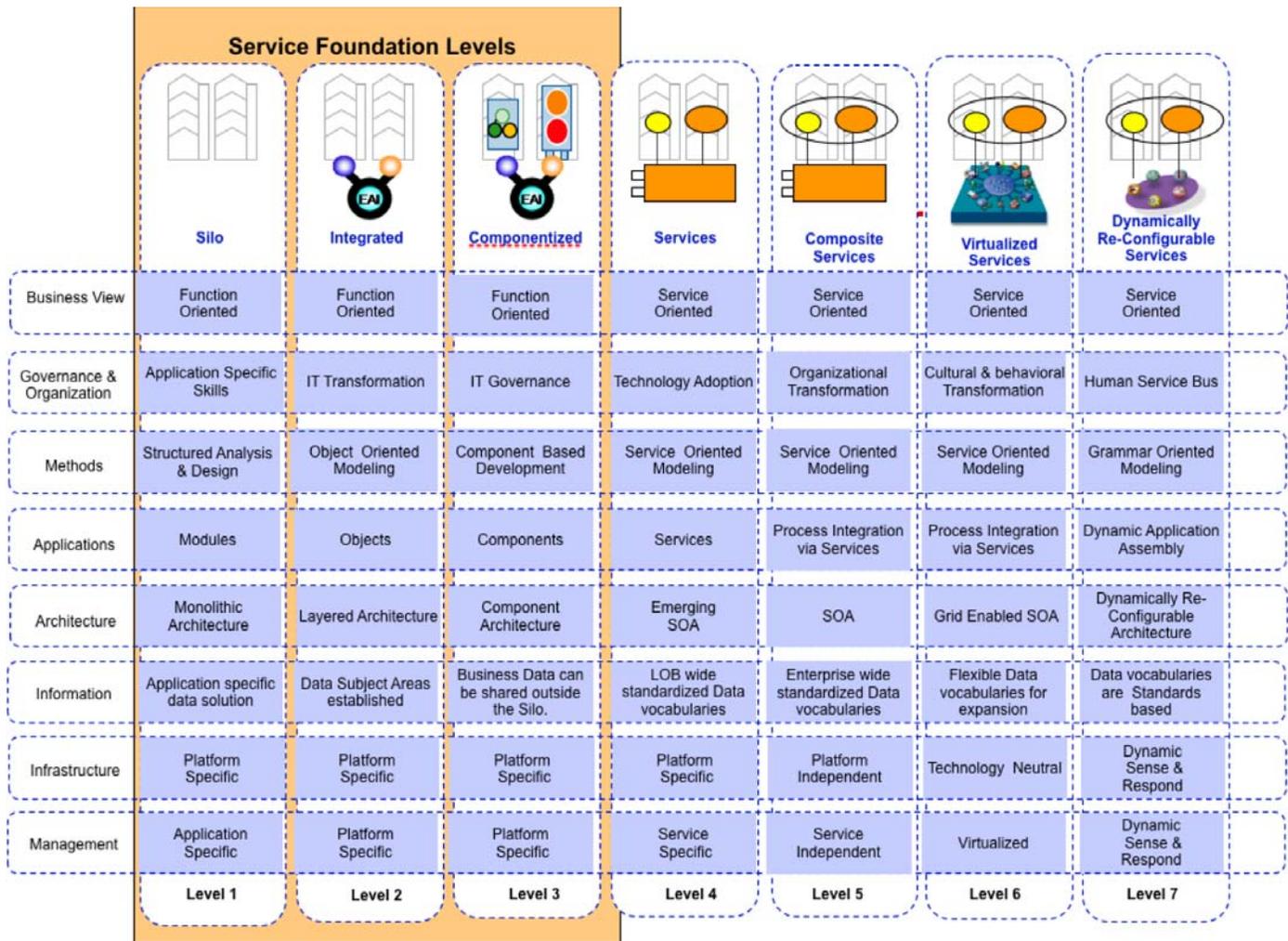


Figure 3: Open Group Service Integration Maturity Model (OSIMM) [Anonymous 2007a].

In summary, the architectural maturity level of the clinical information systems on the U.S. healthcare scene today spans OSIMM Level 1 -4 across various maturity dimensions. Functionally, most clinical systems today operate overall at Level 1 (“Silo”) or Level 2 (“Integrated”), but rarely achieve maturity Level 3 (“Componentized”) and have not yet moved out of the Service Foundation stage into the true Service stage of maturity.

Beyond traditional SOA, clinical information systems must evolve toward a form of “Commitment-Based SOA” in order to support the highest levels of interoperability.

Service-oriented computing is centered on provision of business services, and developed to facilitate processes based largely on the concept of discrete transactions. For example, SOA is by definition “stateless”, meaning that no service may be required to condition its results based on a “memory” the results of its previous transactions; each transaction must be handled entirely based upon the input that is supplied with the request. Participants in our expert panel discussions recognized that adequate support for activities of clinical medicine involves abstractions more complex than discrete, stateless transactions. Decision-making regarding appropriate treatment in medicine, for example, typically involves long-term consultative engagements and negotiations among multiple autonomous agents (physician, patient, consultants, specialists, insurer). These engagements have an intrinsic wholeness that is difficult to capture as a series of discrete, independent transactions.

The software community recognizes this limitation of the service model. Emerging architectural alternatives to SOA for such capturing complex interaction structures in software include multiagent systems, but these are still at an early stage of development. Some extensions of the SOA model replace the centrality of the transaction with higher-level abstractions that, while adhering to SOA design principles, can nevertheless capture some aspects of diachronic engagements.

Commitment-based SOA (CSOA) [Singh 2007] is one such set of design patterns that layers well on top of

existing SOA infrastructure. In CSOA, architecture components should be business services whose connectors are patterns that support key elements of service engagements. CSOA gives primacy to the business meanings of service engagements, which are captured through the participants' commitments to one another. Each participant is modeled as an agent with interacting agents carrying out service engagements by creating and manipulating commitments to one another.

Conclusion & Future Directions

The cancer pathology report is a paradigmatic example of a complex medical report with a life-cycle that involves many data consumers with a variety of interests. Because of the availability of content standards maintained and published by the College of American Pathologists, it appears particularly suitable for electronic rendering as a structured document. The concept of “structured document” encompasses a broad range of formats whose suitability for automated processing varies widely. It is commonly argued that a structured document format could improve data quality and accessibility for many communities of data consumers by enhancing the accuracy of coded renderings in a variety of vocabularies.

The literature does contain some evidence of such value. Structured reporting enhances the quality of data input by pathologists. While introduction of structured reporting produces an initial workflow disruption for pathologists, it is likely that it ultimately improves pathologist productivity. Downstream benefits are less certain. Existing medical software systems are not designed to take full advantage of structured reporting formats. While clinicians find structured pathology reports easier to understand than traditional narrative reports, documenting statistically significant improvements in patient outcomes is a challenge. It is likely that structured reporting makes it easier to reuse cancer diagnostic data for research. Public health experts favor the use of structured reporting for public health surveillance, although specific examples of benefits are elusive. There is currently no substantial evidence that structured pathology reporting results in overall cost savings in medical care.

One prerequisite for optimum use of structured reports is a system of common identifiers for individual entities: persons, events and things—a problem that is different from, and prior to, the issue of common terminology. Technical requirements for maximally useful cross-enterprise identifiers are well-established. Unfortunately, common identifier frameworks used today in healthcare represent a patchwork of alternative syntaxes, and coverage is limited. Identifiers based on the URI standard stand out among available options because of strong existing infrastructure, easy migration path from existing systems, and proven scalability.

A second prerequisite is availability of sharable concept schemes (vocabularies, terminologies, ontologies) that support granular and interoperable cross-enterprise data storage, query and retrieval. One metric of concept scheme adequacy is its ability to yield reproducible manual encodings of medical documents. Available studies show that today’s concept schemes tend to have less-than-satisfactory coding reproducibility. If even manual encoding is poorly reproducible, then automated encoding is apt to be even more challenging, and will probably require a combination of techniques including text indexing, natural language processing and constrained entry. Another metric of adequacy is the ability of a concept scheme to cover multiple workflow contexts. The literature review did not demonstrate a single terminology that appears to provide adequate coverage for all contexts. Indeed, it is likely that the best solution is to allow the existence of multiple, independently curated schemes that can be inter-mapped in a common framework. Existing mapping frameworks are technically cumbersome and yield poor quality results. Reformulation of existing concept schemes into more expressive, logically-based modeling and mapping frameworks may facilitate interoperability and improve mapping quality.

Finally, while semantic interoperability in pathology reporting may be achievable by a combination of structured documents, entity identifiers and shared vocabularies, it seems clear that full support for the complex life-cycle of these artifacts requires higher levels of interoperability. Pragmatic, dynamic and conceptual interoperability levels have been characterized in analyses performed by DOD and DOE as “supra-semantic” levels that are relevant in complex systems. Support for these supra-semantic forms of interoperability can only be guaranteed within modern software architecture such as SOA. Even conventional SOA may be an insufficient framework to support some kinds of highly contextual information processing tasks that occur within the medical document lifecycle, and “commitment-based SOA” or multi-agent systems architectures may be required.

What questions remain to be answered?

- 1) Studies should be conducted to determine the value of capturing data as a text block or discrete data elements. If data is captured as discrete data elements, studies should determine the integrity of the data as it is used by others besides clinicians.
- 2) How long does a cancer report persist downstream with the original intent maintained. Is there a difference in interpretation by a clinician reading through a narrative text versus a synoptic one? Does

one form or the other enhance correct interpretation? Or, are both needed at least some of the time?

- 3) Future studies might want to address the change in work flow process. What is the degree of change needed and how does it affect the day to day operations of the pathology lab? Does this change eliminate some steps in case sign-out process or does it add more steps? If steps are eliminated, such as the need for transcription, does that increase the onus on pathologists by increasing their time spent per case?
- 4) It would be of interest to have further studies continue to establish the reproducibility of metadata annotation. Are the issues confined to the ontologies themselves or to the terminology browsers used in the studies? Did coders use different browsers? Does this make a difference? Or, are the forms used to code data in need of improvement or better structure? Studies are needed to assess the nature of case report forms data collection and question modeling as well as the coding of these elements. Are there issues in the structure of the report form that influence coding?
- 5) It may be of interest to look at coding practices for a patient record versus coding for epidemiologic or surveillance activities. Is there a difference in coding thought processes and/or techniques for different use cases when coding using SNOMED CT? Studies to examine the coding consistency across trained individuals applying standardized terminologies to represent clinical research data in pure research settings are needed.
- 6) It may be of interest to look at coding practices for a patient record versus coding for research or surveillance activities. These studies point to the importance of education and certification in implementing large scale coding programs on a national scale. To be maximally useful ontologies must also be logically structured in a consistent manner thus allowing navigation between concepts in a predictably and meaningful manner, and in a fashion that allows modelers to validate logical consistency. Studies should be conducted to ascertain these features.

Bibliography

Agarwal S, Lamparter S, Studer R. Making Web services tradable A policy-based approach for specifying preferences on Web service properties. *Web Semantics: Science, Services and Agents on the World Wide Web* 2009;7(1):11-20.

Andrews JE, Richesson RL, Krischer J. Variation of SNOMED CT coding of clinical research concepts among coding experts. *J Am Med Inform Assoc* 2007;14(4):497-506.

Anonymous. National E-Health Transition Authority. Towards an Interoperability Framework. Version 1.8. Sydney (AU): National E-Health Transition Authority; 2005. Available from: http://www.nehta.gov.au/Component/Docman/doc_download/26-Towards-an-Interoperability-Framework-v18.

Anonymous. The Open Group. Business scenario: identifiers in the enterprise. San Francisco: The Open Group; 2006. Report No. K061.

Anonymous. The DOI Handbook. Oxford (UK): International DOI Foundation; 5 October 2006 [cited 27 April 2009]. p. 49-57. Available from: http://www.doi.org/handbook_2000/resolution.html.

Anonymous. Welcome to Apache Hadoop! [Internet]. 2007 [updated 2007; cited 21 April 2009]. Available from: <http://hadoop.apache.org/index.html>.

Anonymous. The Open Group. San Diego: The Open Group; 1 February 2007 [updated 1 February 2007; cited 29 April 2009]. Available from: www.opengroup.org/projects/osimm/uploads/40/12647/OSIMM_-_WG_Update_2-01-07.ppt.

Anonymous: Scalable knowledge composition [Internet]. Stanford, CA: Stanford University; 2008 [cited 23 April 2009]. Available from: <http://infolab.stanford.edu/SKC/index.html>.

Anonymous. College of American pathologists -cancer protocols and checklists 2009: Available from: http://www.cap.org/apps/cap.portal?_nfpb=true&cntvwrPtlit_actionOverride=%2Fportlets%2FcontentViewer%2Fshow&_windowLabel=cntvwrPtlit&cntvwrPtlit%7BactionForm.contentReference%7D=committees%2Fcancer%2Fcancer_protocols%2Fprotocols_index.html&state=maximized&pageLabel=cntvwr.Accessed 6 April 2009.

Anonymous. IHTSDO: International health terminology standards development organization [Internet]. Copenhagen: IHTSDO; 2009 [cited 23 April 2009]. Available from: <http://www.ihtsdo.org/>.

Anonymous. ASN.1 & OID project. 2 April 2009: Available from: <http://www.itu.int/ITU-T/asn1/>. Accessed 8 April 2009.

Anonymous. Domain counts & internet statistics [Internet]. Reston, VA: Name Intelligence, Inc.; 2009 [cited 12 April 2009]. Available from: <http://www.domaintools.com/internet-statistics/>.

Anonymous. UMLS Documentation. Washington (DC): National Library of Medicine; 2009. Available from: <http://www.nlm.nih.gov/research/umls/umlsdoc.html>.

Appavu SI. Analysis of Unique Patient Identifier Options: Final Report. Washington, DC: U.S. Department of Health & Human Services; 1997. Available from: <http://www.ncvhs.hhs.gov/app0.htm>.

Assistant Secretary of Planning and Evaluation (ASPE). A unique health identifier for individuals: a white paper. Washington (DC): U.S. Department of Health & Human Services; 1998. Available from: <http://aspe.hhs.gov/ADMNSIMP/nprm/noiwp1.htm>.

Barnhill W. XRI as relative URI [Internet]. Billerica (MA): OASIS; 31 October 2008 [cited 12 April 2009]. Available from: <http://wiki.oasis-open.org/xri/XriAsRelativeUri>.

Bizer C, Cyganiak R. Quality-driven information filtering using the WIQA policy framework. Web Semantics: Science, Services and Agents on the World Wide Web 2009;7(1):1-10.

Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness DL, Patel-Schneider PF, Stein LA. OWL Web Ontology Language Reference. Dean M, Schreiber G, editors. Boston: World Wide Web Consortium (W3C); 2004.

Berners-Lee T, Fielding R, Masinter L. Uniform resource identifier (URI): Generic syntax. Reston (VA): The Internet Society; 2005. (Request for Comments (RFC); no. 3986). Available from: <http://labs.apache.org/webarch/uri/rfc/rfc3986.html>.

Berners-Lee T, editor. Notation 3: a readable language for data on the web. Boston: World Wide Web Consortium (W3C); 2006. Available from: <http://www.w3.org/DesignIssues/Notation3.html>.

Bodenreider O. Issues in mapping LOINC laboratory tests to SNOMED CT. AMIA Annu Symp Proc 2008:51-5.

Brown SH, Elkin PL, Bauer BA, Wahner-Roedler D, Husser CS, Temesgen Z, Hardenbrook SP, Fielstein EM, Rosenbloom ST. SNOMED CT: utility for a general medical evaluation template. AMIA Annu Symp Proc 2006:101-5.

Centers for Medicare & Medicaid Services; U.S. Department of Health & Human Services. HIPAA administrative simplification: standard unique health identifier for health care providers. Final rule. Fed Regist 2004;69(15):3433-68.

Ceusters W, Smith B. Tracking referents in electronic health records. Stud Health Technol Inform 2005;116:71-6.

Ceusters W, Smith B. Strategies for referent tracking in electronic health records. J Biomed Inform 2006;39(3):362-78.

Ceusters W, Elkin P, Smith B. Referent tracking: the problem of negative findings. Stud Health Technol Inform 2006;124:741-6.

Ceusters W, Elkin P, Smith B. Negative findings in electronic health records and biomedical ontologies: a realist approach. Int J Med Inform 2007;76 Suppl 3:S326-33.

Chan NG, Duggal A, Weir MM, Driman DK. Pathological reporting of colorectal cancer specimens: a retrospective survey in an academic Canadian pathology department. Can J Surg 2008;51(4):284-8.

Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE. Bigtable: A distributed storage system for structured data. In: Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI '06); 6-8 November 2006; Seattle (WA). Berkeley (CA): USENIX: The Advanced Computing Systems Association; 2006.

Chaudhuri S, Ramakrishnan R, Weikum G. Integrating DB and IR technologies: What is the sound of one hand clapping. In: Stonebraker M, Weikum G, DeWitt D, editors. Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR); 4-7 Jan 2005; Asilomar (CA). CIDR Organizing Committee; 2005. p. 1-12. Available from: <http://www.cidrdb.org/cidr2005/>.

Chen ES, Hripcsak G, Friedman C. Disseminating natural language processed clinical narratives. AMIA Annu Symp Proc 2006;Annu Symp Proc:126-30.

Chiang MF, Hwang JC, Yu AC, Casper DS, Cimino JJ, Starren JB. Reliability of SNOMED-CT coding by three physicians using two terminology browsers. AMIA Annu Symp Proc 2006:131-5.

Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, Cooper J, Guan W, Groen PCD. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009:1-13.

Cooper BF, Ramakrishnan R, Srivastava U, Silberstein A, Bohannon P, Jacobsen HA, Puz N, Weaver D, Yerneni R. Pnuts: Yahoo!'s hosted data serving platform. *Proceedings of the VLDB Endowment Archive* 2008;1(2):1277-88.

Cross SS, Feeley KM, Angel CA. The effect of four interventions on the informational content of histopathology reports of resected colorectal carcinomas. *J Clin Pathol* 1998;51(6):481-2.

DeCandia G, Hastorun D, Jampani M, Kakulapati G, Lakshman A, Pilchin A, Sivasubramanian S, Vosshall P, Vogels W. Dynamo: Amazon's highly available key-value store. In: *Proceedings of the 21st ACM SIGOPS Symposium on Operating Systems Principles*; 14-17 Oct 2007; Stevenson (WA). New York: Association for Computing Machinery; 2007. p. 205-20.

Ding J, Erdal S, Dhaval R, Kamal J. Augmenting Oracle Text with the UMLS for enhanced searching of free-text medical reports. *AMIA Annu Symp Proc* 2007:940.

Dividino R, Sizov S, Staab S, Schueler B. Querying for provenance, trust, uncertainty and other meta knowledge in RDF. "Web Semantics: Science, Services and Agents on the World Wide Web" 2009:1-16.

Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, Speroff T. Evaluation of the content coverage of SNOMED CT: Ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clin Proc* 2006;81(6):741-8.

Erdal S, Kamal J. An indexing scheme for medical free text searches: a prototype. *AMIA Annu Symp Proc* 2006:918.

Erinjeri JP, Picus D, Prior FW, Rubin DA, Koppel P. Development of a Google-Based Search Engine for Data Mining Radiology Reports. *J Digit Imaging* 2008

Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. *Brain Res* 2004;11(5):392-402.

Genesereth MR, Fikes RE; Stanford University Logic Group. Knowledge Interchange Format Reference Manual. Stanford, CA: Stanford University; 1992. (Logic Group Technical Report; no. 92-1). Available from: <http://logic.stanford.edu/kif/Hypertext/kif-manual.html>.

GridWise Architecture Council. GridWise Interoperability Context-Setting Framework. Version 1.1. Richland (WA): GridWise Architecture Council; 2008.

Hanauer DA. EMERSE: The Electronic Medical Record Search Engine. *AMIA Annu Symp Proc* 2006:941.

Harvey JM, Sterrett GF, McEvoy S, Fritschi L, Jamrozik K, Ingram D, Joseph D, Dewar J, Byrne MJ, Group KCPOCS. Pathology reporting of breast cancer: trends in 1989-1999, following the introduction of mammographic screening in Western Australia. *Pathology* 2005;37(5):341-6.

Hayes P, McBride B, editors. RDF Semantics. Boston: World Wide Web Consortium (W3C); 2004. (W3C Recommendation). Available from: <http://www.w3.org/TR/rdf-mt/>.

Horner MJ, Ries LAG, Krapcho M, Neyman N, Aminou R, Howlader N, Altekruse SF, Feuer EJ, Huang L, Mariotto A, Miller BA, Lewis DR, Eisner MP, Stinchcomb DG, Edwards BK, editors. SEER cancer statistics review, 1975-2006 [Internet]. Bethesda (MD): National Cancer Institute; 2009 [cited 23 June 2009]. Available from: http://seer.cancer.gov/csr/1975_2006/.

International DOI Foundation. Key facts on the digital object identifier system [Internet]. Oxford (UK): International DOI Foundation; 30 June 2009 [cited 1 July 2009]. Available from: <http://www.doi.org/factsheets/DOIKeyFacts.html>.

International Organization for Standardization. Information technology — Common Logic (CL): a framework for a family of logic-based languages. Geneva: ISO/IEC; 2007. 59 p. (Approved Standard; no. 24707:2007).

ITU (International Telecommunication Union) . Generation and registration of Universally Unique Identifiers (UUIDs) and their use as ASN.1 object identifier components. Geneva: International Telecommunication Union (ITU); 2004. 34 p. (ITU-T Recommendation; no. X.667).

ITU (International Telecommunication Union). ASN.1 & OID project [Internet]. International Telecommunication Union (ITU); 2 April 2009 [cited 8 April 2009]. Available from: <http://www.itu.int/ITU-T/asn1/>.

Jacobs I, Walsh N, editors. Architecture of the World Wide Web, Volume One. World Wide Web Consortium (W3C); 2004. Available from: <http://www.w3.org/TR/webarch/>.

Jagannathan V, Mullett CJ, Arbogast JG, Halbritter KA, Yellapragada D, Regulapati S, Bandaru P. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *Int J Med Inform* 2009;78(4):284-91.

James BC, Hammond ME. The challenge of variation in medical practice. *Arch Pathol Lab Med* 2000;124(7):1001-3.

Karim RZ, van den Berg KS, Colman MH, McCarthy SW, Thompson JF, Scolyer RA. The advantage of using a synoptic pathology report format for cutaneous melanoma. *Histopathology* 2008;52(2):130-8.

Klyne G, Carroll JJ, McBride B, editors. Resource Description Framework (RDF): Concepts & abstract syntax. Boston: World Wide Web Consortium (W3C); 2004. (W3C Recommendation). Available from: <http://www.w3.org/TR/rdf-concepts/>.

Liu Z, Ranganathan A, Riabov A. Specifying and enforcing high-level semantic obligation policies. *Web Semantics: Science, Services and Agents on the World Wide Web* 2009;7(1):28-39.

Madden JF, Albarracin N, Kennedy MF, deBaca M, editors. Synoptic-Report: XML framework for synoptic medical diagnostic reports [Internet]. 2009 [cited 6 April 2009]. Available from: <http://code.google.com/p/synoptic-report/>.

Manola F, Miller E, editors. RDF Primer. Boston: World Wide Web Consortium (W3C); 2004. Available from: <http://www.w3.org/TR/REC-rdf-syntax/>.

Mealling M. A URN namespace of object identifiers. Reston VA: The Internet Society; 2001. (Request for Comments (RFC); no. 3001).

Mohanty SK, Piccoli AL, Devine LJ, Patel AA, William GC, Winters SB, Becich MJ, Parwani AV. Synoptic tool for reporting of hematological and lymphoid neoplasms based on World Health Organization classification and College of American Pathologists checklist. *BMC Cancer* 2007;7:144.

Moskovitch R, Martins SB, Behiri E, Weiss A, Shahar Y. A comparative evaluation of full-text, concept-based, and context-sensitive search. *Brain Res* 2007;14(2):164-74.

Murari M, Pandey R. A synoptic reporting system for bone marrow aspiration and core biopsy specimens. *Arch Pathol Lab Med* 2006;130(12):1825-9.

Network Applications Consortium Distributed Management Task Force. Core Identifier Framework Matrix: Technical Guide. Reading (UK): The Open Group; 2007.

Paskin N. Digital object identifiers for scientific data. *Data Science Journal* 2005;4:12-20.

Patel C, Cimino J, Dolby J, Fokoue A, Kalyanpur A. Matching Patient Records to Clinical Trials Using Ontologies. In: Aberer K, Choi K-S, Noy N, Allemang D, Lee K-I, Nixon L, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudré-Maroux P, editors. *The Semantic Web. Sixth International Semantic Web Conference*; 11-15 November 2007; Busan (Korea). New York: Springer; 2007. p. 816-29. (Lecture Notes in Computer Science; no. 4825).

Patel CO, Cimino JJ. A scale-free network view of the UMLS to learn terminology translations. *Stud Health Technol Inform* 2007;129(Pt 1):689-93.

Pathak J, Solbrig HR, Buntrock JD, Johnson TM, Chute CG. LexGrid: A Framework for Representing, Storing, and Querying Biomedical Terminologies from Simple to Sublime. *Brain Res* 2009

Pickens S, Solak J. National Provider Identifier (NPI) planning and implementation fundamentals for providers and payers. *J Healthc Inf Manag* 2005;19(2):41-7.

Prud'hommeaux E. Matching graph patterns against stem graphs [Internet]. Version 1.43. Boston: World Wide Web Consortium (W3C); 24 July 2008 [updated 9 December 2008; cited 2 July 2009]. Available from: <http://www.w3.org/2008/07/MappingRules/>.

Qu Z, Ninan S, Almosa A, Chang KG, Kuruvilla S, Nguyen N. Synoptic reporting in tumor pathology: advantages of a web-based system. *Am J Clin Pathol* 2007;127(6):898-903.

Rao J, Sardinha A, Sadeh N. A meta-control architecture for orchestrating policy enforcement across heterogeneous information sources. *Web Semantics: Science, Services and Agents on the World Wide Web* 2009;7(1):40-56.

Rector A, Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. *J Am Med Inform Assoc* 2008;15(6):744-51.

Rector AL. Clinical terminology: why is it so hard? *Methods Inf Med* 1999;38(4-5):239-52.

Richesson RL, Andrews JE, Krischer JP. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. *J Am Med Inform Assoc* 2006;13(5):536-46.

Roessler T. W3C technology and society domain: security home [Internet]. 4 June 2009 [cited 8 September 2009]. Available from: <http://www.w3.org/Security/Activity>.

Rothschild AS, Lehmann HP, Hripcsak G. Inter-rater agreement in physician-coded problem lists. *AMIA Annu Symp Proc* 2005:644-8.

Sauermann L, Cygniak R, editors. Cool URIs for the semantic web. Boston: World Wide Web Consortium (W3C); 2008. (W3C Working Group Note). Available from: <http://www.w3.org/TR/cooluris/>.

Singh MP, Chopra AK, Desai N; Department of Computer Science. Commitment-Based SOA. Raleigh (NC): North Carolina State University; 2007.

Thompson HS, Orchard D, editors. URNs, Namespaces and Registries. Boston: World Wide Web Consortium (W3C); 2006. Available from: <http://www.w3.org/2001/tag/doc/URNsAndRegistries-50>.

Thompson JF, Scolyer RA. Cooperation between surgical oncologists and pathologists: a key element of multidisciplinary care for patients with cancer. *Pathology* 2004;36(5):496-503.

Tobias J, Chilukuri R, Komatsoulis GA, Mohanty S, Sioutos N, Warzel DB, Wright LW, Crowley RS. The CAP cancer protocols--a case study of caCORE based data standards implementation to integrate with the Cancer Biomedical Informatics Grid. *BMC Med Inform Decis Mak* 2006;6:25.

Tolk A, Muguira JA. The levels of conceptual interoperability model. In: Proceedings of the 2003 Fall Simulation Interoperability Workshop; 14-19 September 2003; Orlando (FL). Orlando (FL): Simulation Interoperability Standards Organization (SISO); 2003. p. 7. Available from: <http://www.sisostds.org/index.php?tg=fileman&idx=get&id=2&gr=Y&path=Simulation+Interoperability+Workshops%2F2003+Fall+SIW%2F2003+Fall+SIW+Papers+and+Presentations&file=03F-SIW-007.pdf>.

Tolk A. What comes after the semantic web? PADS implications for the dynamic web. In: Proceedings of the 20th Workshop on Principles of Advanced and Distributed Simulation; 24-26 May 2006; Singapore. New York: IEEE; 2006. Available from: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1630709&isnumber=34195>.

Tolk A, Blais CL. Taxonomies, ontologies, and battle management languages--recommendations for the coalition BML study group (05S-SIW-007). In: Proceedings of the Spring Simulation Interoperability Workshop; 3-8 April 2005; San Diego (CA). Orlando (FL): Simulation Interoperability Standards Organization (SISO); 2005. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.8165&rep=rep1&type=pdf>.

Tolk A, Diallo SY, King RD, Turnitsa CD. A Layered Approach to Composition and Interoperation in Complex Systems. In: Tolk A, Jain LC, editors. *Complex Systems in Knowledge-Based Environments: Theory, Models and Applications*. Springer; 22 January 2009. p. 41. (Studies in Computational Intelligence; vol. 168).

Tonkin E. Persistent identifiers: considering the options. *Ariadne* 2008(56):unpaginated. Available from : <http://www.ariadne.ac.uk/issue56/tonkin>.

Tsan CJ, Serpell JW, Poh YY. The impact of synoptic cytology reporting on fine-needle aspiration cytology of thyroid nodules. *ANZ J Surg* 2007;77(11):991-5.

URI Planning Interest Group W3C/IETF. URIs, URLs, and URNs: Clarifications and Recommendations. Boston: World Wide Web Consortium (W3C); 2001. Available from: <http://www.w3.org/TR/uri-clarification/>.

Vikström A, Skånér Y, Strender LE, Nilsson GH. Mapping the categories of the Swedish primary health care version of ICD-10 to SNOMED CT concepts: rule development and intercoder reliability in a mapping trial. *BMC Med Inform Decis Mak* 2007;7:9.

Wade G, Rosenbloom ST. The impact of SNOMED CT revisions on a mapped interface terminology: Terminology development and implementation issues. *J Biomed Inform* 2009:1-4.

Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Brain Res* 2009

Wang Y, Wei D, Xu J, Elhanan G, Perl Y, Halper M, Chen Y, Spackman KA, Hripcsak G. Auditing complex concepts in overlapping subsets of SNOMED. *AMIA Annu Symp Proc* 2008:273-7.

Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *AMIA Annu Symp Proc* 2003:699-703.

Weikum G, Kasneci G, Ramanath M, Suchanek F. Database and information-retrieval methods for knowledge discovery. *Comm ACM* 2009;52(4):56-64.

Wilcox AB, Hripcsak G. The role of domain knowledge in automating medical text report classification. *Brain Res* 2003;10(4):330-8.

Windley PJ. *Digital Identity*. Cambridge (MA): O'Reilly Media; 2005.

Wingo PA, Howe HL, Thun MJ, Ballard-Barbash R, Ward E, Brown ML, Sylvester J, Friedell GH, Alley L, Rowland JH, Edwards BK. A national framework for cancer surveillance in the United States. *Cancer Causes Control* 2005;16(2):151-70.

Zhou L, Hripcsak G. Temporal reasoning with medical data—A review with emphasis on medical natural language processing. *J Biomed Inform* 2007;40(2):183-202.

Appendix 1: Unstructured (Traditional) Pathology Report

Patient: JONES, JOHN Admit MD: SMITH, SANDRA MD
ID #: H9876543 Order MD: SMITH, SANDRA MD
DOB: 01/01/1930
Sex: MALE

Clinical History

Abnormal colonoscopy. Biopsy 12/1 showed adenocarcinoma.

Gross Examination

Received a container labeled "left colon" containing a 13 cm long left ascending colectomy specimen, with appendix and a 1 cm segment of terminal ileum. The retroperitoneal margin is inked blue. The opened specimen exhibits a 3.5 cm papillary tumor in the cecum, 0.5 cm from the ileocecal valve and 9 cm from the distal specimen margin. Sectioning through the demonstrates induration of the muscularis propria underlying the tumor, but no gross extracolonic extension. Pericolonic fat is dissected for lymph nodes.

Blocks Submitted

A1 distal margin
A2 ileal margin
A3 retroperitoneal fat margin underlying tumor
A4-9 representative tumor
A10-15 apparent lymph nodes, closest to tumor in A10

Microscopic Examination

Microscopic examination shows a moderately differentiated adenocarcinoma arising in a villous adenoma. No medullary or mucinous component is identified. There is a marked intratumoral lymphocytic response, but only a mild-to-moderate peritumoral response. Carcinoma invades to the outer one-half of the muscularis propria of the cecum, but does not perforate into the surrounding fat. All margins are negative, with the closest margin being the deep margin, which is 0.8 cm from the nearest tumor. Seven lymph nodes from the posterior cecum are negative for carcinoma. Also noted in a small separate tubulovillous adenoma in the ascending colon. Based on these findings, the tumor is staged as pT2N0MX.

Diagnosis

MODERATELY-DIFFERENTIATED ADENOCARCINOMA OF COLON (CECUM).

SEE MICROSCOPIC DESCRIPTION.

Signed: John Johnson, MD, Pathologist

Date: 1/4/2009

Appendix 2: Structured Pathology Report

Tumor Synopsis

Clinical findings: MASS
BIOPSY POSITIVE FOR CARCINOMA

Specimen:

Procedure: SEGMENTAL COLECTOMY
Sites included: CECUM, LEFT COLON
Tumor site: CECUM

Tumor findings:

Histologic type: ADENOCARCINOMA
Histologic grade: HIGH GRADE

S/O microsatellite instability:

High grade histology: POSITIVE
Medullary component: NEGATIVE
Mucinous component: NEGATIVE

Immune response:

Intratumoral: MARKED
Peritumoral: MILD TO MODERATE
Perforation of colon wall: NEGATIVE

Accessory histologic findings:

Discontinuous tumor extension: NEGATIVE
Pre-existing polyp: POSITIVE (VILLOUS ADENOMA)
Lymphovascular invasion: INCONCLUSIVE
Venous invasion: NEGATIVE

Tumor extent:

Size: 3.5 CM
Deepest invasion: MUSCULARIS PROPRIA

Surgical margins:

Distal: NEGATIVE
Proximal: NEGATIVE
Circumferential: NEGATIVE
Closest margin: CIRCUMFERENTIAL (0.8 CM)

Regional lymph nodes:

Posterior cecal nodes: SEVEN LYMPH NODES, NEGATIVE FOR CARCINOMA (0/7)

AJCC Pathologic Stage:

T: 2

N: 0

M X

Additional findings: SEPARATE ADENOMATOUS POLYP

Signed: John Johnson, MD, Pathologist

Date: 1/4/2009

Appendix 3: XML Pathology Report

```
<?xml version="1.0" encoding="UTF-8"?>
<synopsis pert:schemaLocation="http://purl.oclc.org/medicalreporting/report/cap-
cancer/resection/resection@colon.rng" pert:version="O.I" xmlns="
http://www.cap.org/pert/2009/01/" xmlns:colon="http://www.cap.org/pert/2009/01/colon/"
xmlns:pert=" http://www.cap.org/pert/2009/01/">

  <clinical>
    <clinicalFindings> <c1inicalFinding value="mass"/>
    <c1inicalFinding value="biopsy positive for adenocarcinoma"/>

  </clinicalFindings>
</c1inical>
<specimen>
  <procedures>
    <procedure value="segmental colectomy"/>
  </procedures> <sites>

    <site value="cecum"/>
    <site value="left colon"/>
  </sites> <tumorLocations>

    <tumorLocation value="cecum"/>
  </tumorLocations>
</specimen> <tumor>

  <histologicTypes>
    <histologicType value="adenocarcinoma"/>
  </histologicTypes> <grade value="high"/>
  <colon:suggestMicrosatelliteInstability>

    <colon:highGrade value="positive"/>
    <colon:medullaryComponent value="negative"/>
    <colon:mucinousComponent value="negative"/>
  </colon:suggestMicrosatelliteInstability>
  <colon:immuneResponse> <colon:intratumoralResponse
value="marked"/> <colon:peritumoralResponse value="mild to
moderate"/>
</colon:immuneResponse>
```

```

    <colon:tumorPerforation value="negative"/>
  </tumor>
  <accessory
  > <colon:discontinuousExtramuralExtension value="negative"/>
    <colon:preexistingPolyp value="villous adenoma"/>
    <lymphovascularInvasion value="inconclusive"/>
    <venousInvasion value="negative"/>
  </accessory>
  <extent>
  <invasion>
    <colon:deepestInvasion value="muscularis propria"/>
  </invasion>
  <tumorSize dimension-l="3.5" unit="cm"/>
  </extent> <margins>

    <margin location="distal" status="negative"/> <margin location="proximal"
    status="negative"/> <margin c1osest="true" location="circumferential"
    status="negative">
      <distance unit="cm" value="0.8"/>
  </margin>
  </margins>
  <nodes>
    <nodeGroup location="posterior cecal" >
    <nodeStatus count="7" value="total"/>
    <nodeStatus count="0" value="positive"/>
    </nodeGroup>
  </nodes>
  <stage>
    <T value="2"/>
    <N value="0"/>
    <M value="X"/>
  </stage>
  <additional>
  <findings>
    <finding value="other polyp"/>
  </findings>
  </additional>
  </synopsis>

```

Appendix 4: Semantic Web (RDF/N3) Pathology Report

```
@prefix cap: <http://www.cap.org/pert/2009/01/> . @prefix colon:
<http://www.cap.org/pert/2009/01/colon/> . @prefix rdfs: <http://www.w3.org/200/01/rdf-
schema#> . @prefix cmh: <http://www.centerville-memorial-hospital.org/pid#> . @prefix
cmhsp: <http://www.centerville-memorial-hospital.org/pathology/sid#> . @prefix cmhmd:
<http://www.centerville-memorial-hospital.org/md-id#> .
```

```
@prefix dc: <http://purl.org/dc/elements/1.1/ > .
```

```
<#patient>=<cmh:X875693> .
<#specimen> = <cmhsp:SL-09-12345> .
<#> dc:author <#pathologist>.
<#pathologist>=<cmhmd:9876-1> .
```

```
<#patient>
  <ca p:c1inicalFinding> [a <cap:Mass>J,
    [<cap:diagnosedBy> <cap:Biopsy>; a <cap:Cancer>]
```

```
<#patient> <cap:sourceOfSpecimen> <_:ThisSpecimen>.
```

```
<#specimen> =
  [
    a <cap:Specimen>; <cap:derivedFromProcedure> [a
      <cap:segmentalColectomy>]; <cap:includesAnatomicPart> [a
        <anat:Cecum>, <anat:leftColon>]; <cap:involvedByDisease>
        <_:ThisTumor>
```

```
< :ThisTumor>
  a <cap:Carcinoma>; <cap:site [a <anat:Cecum>];
  <cap:histologicType> <cap:Adenocarcinoma>;
  <cap:histologicGrade> <cap:highGrade>; <cap:hasTumorFinding>
  <_:microsatelliteInstabilityFindings>;
  <colon:intratumoralResponse> <cap:marked>;
  <colon:peritumoralResponse> <cap:mildToModerate>;
  <colon:tumorPerforation> <cap:negative>;
  <colon:discontinuousExtramuralExtension> <cap:negative>;
```

```
<colon:preexistingPolyp> [a cap:Concept; rdfs:label "villous adenoma"];
<cap:lymphovascularInvasion> <cap:inconclusive>; <cap:venousInvasion>
<cap:negative>; <colon:deepestInvasion> [a cap:Concept; rdfs:label
"muscularis propria"]; <cap:size> [<cap:dimension> 3.5; <cap:unit>
[<rdfs:label> "em"]]
```

```
<_:microsatelliteInstabilityFindings> =
  ([ <colon:highGrade> <cap:positive>],
  [<colon:medullaryComponent> <cap:negative>],
  [<colon:mucinousComponent> <cap:negative>]
```

```
<#specimen> <cap:margin>
  (
```

```
    <cap:location> <cap:distalMargin>;
    <cap:marginStatus> <cap:negative>
```

```
  [ <cap:location> <cap:proximalMargin>;
    <cap:marginStatus> <cap:negative>
```

```
  [
    <cap:location> <cap:circumferentialMargin>;
    <cap:marginStatus> <cap:negative>; a
    <cap:ClosestMargin>
```

```
< :ThisTumor>
  <cap:regionalLymphNodes>
    [ <cap:nodeGroup> [<rdfs:label> "posterior cecal"];
    <cap:totalNodes> 7; <cap:positiveNodes> 0
```

```
< :ThisTumor>
  <cap:T> "2";
```

<cap:N > "0";
<cap:M > "X"

<_:ThisSpecimen>
 <cap:additionalFinding> <cap:polyp> .

Appendix 5: Relevant Tumor Registry organizations & projects

1. National Program of Cancer Registries (NPCR)

The National Program of Cancer Registries (NPCR) was established by Congress through the Cancer Registries Amendment Act in 1992, and administered by the Centers for Disease Control and Prevention (CDC). NPCR collects data on the occurrence of cancer; the type, extent, and location of the cancer; and the type of initial treatment from central cancer registries. These data represent 98 percent of the United States population. Participating central cancer registries and affiliated hospitals are required to report and use a standard nationally defined set of specific data items and codes.

2. Surveillance, Epidemiology and End Results (SEER) Program

The Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute collects information on cancer incidence and survival in the United States. SEER collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 26 percent of the United States population.

3. National Cancer Data Base (NCDB)

The National Cancer Data Base (NCDB), a joint program of the Commission on Cancer (CoC) and the American Cancer Society (ACS), is a nationwide oncology outcomes database for more than 1,400 Commission-approved cancer programs in the United States and Puerto Rico. Some 75% of all newly diagnosed cases of cancer in the United States are captured at the institutional level and reported to the NCDB. These data are used to explore trends in cancer care, create regional and state benchmarks for participating hospitals, and to serve as the basis for quality improvement.

4. Commission on Cancer (CoC)

The American College of Surgeons (ACoS) Commission on Cancer (CoC) is a consortium of professional organizations dedicated to improving survival and quality of life for cancer patients through standard setting, prevention, research, education, and monitoring quality of care.

North American Association of Central Cancer Registries (NAACCR)

The North American Association of Central Cancer Registries (NAACCR) is a collaborative umbrella organization for cancer registries, governmental agencies, professional organizations, and private groups in North America that are interested in enhancing the quality and use of cancer registry data.

5. Cancer Care Ontario (CCO)

Cancer Care Ontario, established in 1997, is a provincial agency responsible for cancer services. As the government's cancer advisor, Cancer Care Ontario directs and oversees close to \$700 million public health care dollars to hospitals and other cancer care providers to deliver cancer services. Their mission also includes working with cancer care professionals and organizations to develop and implement quality improvements and standards and to ensure the use of electronic information and technology to support health professionals.

CCO has endorsed the College of American Pathologists (CAP) cancer reporting standards. The CCO initiative has created an automated, secure, electronic pathology reporting system which collects cancer-related pathology information from hospitals and private labs across Ontario for reporting to CCO. To facilitate the timely collection of pathology data, CCO provides an electronic data collection system called the Pathology Information Management System (PIMS). PIMS is an information management initiative supported and funded by the Ministry of Health and Long-Term Care.

CCO's Pathology Project is working towards the goal of receiving 90% of pathology reports in discrete synoptic format using the CAP cancer checklist reporting standard. In the initial implementations, for 2009/10, hospitals will implement synoptic reporting tools for the top 5 cancer resection reports (breast,

lung, prostate, CRC and endometrial resections). CCO and the hospital will update PIMS for reporting in discrete data field format. In 2010/12, hospitals will expand synoptic pathology reporting for all cancers using the CAP checklists. CCO and the hospital will augment PIMS to send data to CCO with SNOMED CT/LOINC codes.

CCO also administers a report audit that includes pathology reporting and is used as a foundation for several Surgical Oncology indicators that assess quality of care and appropriateness of interventions. For example, for the colorectal cancer disease site, pathology report data is used to assess the percent of colorectal cancer resection specimens with at least 12 nodes examined (where the examination of at least 12 nodes, where possible, has been shown to be necessary for accurate detection of cancer spread into lymph nodes). As such the Pathology report audit is used to provide additional educational support to both surgeons and pathologists.

The results of a 2005/06 pathology audit unequivocally demonstrated that pathology report completeness rates dramatically increased by a factor of two- to three-fold for all disease sites when synoptic or synoptic-like templates were used for pathology reporting. These results are shown below:

Total Volume (% Complete)			
Disease Site	Total Reports/Cases	Synoptic(-Like)	Narrative
Prostate	828	674 (97%)	154 (50%)
Lung	535	442 (86%)	93 (34%)
Breast	1746	1517 (80%)	229 (43%)

From: Cco Pathology Checklist Reporting Project Overview, 2007

6. Canadian Partnership Against Cancer (CPAC)

CPAC is an independent organization, federally funded, to accelerate action on cancer control for all Canadians. Part of the initiative is to standardize pathology reporting of cancer cases. The Partnership is actively promoting the adoption and implementation of synoptic reporting as a standard, not only in pathology, but also for surgical data, imaging data and clinical notes. Studies have shown that completeness improves with synoptic reporting thereby facilitating clinical decision making and appropriate treatment. The mandatory (required) elements of the CAP cancer checklists are a minimum reporting standard. This national population-based collaborative stage data collection is for cancer cases diagnosed on or after January 1, 2010 for four cancer sites: colorectal, breast, lung and prostate. This part of the project comprises a \$17M (Canadian) investment that links together pathologists and cancer registries.

It is expected that in July 2009 the Canadian Association of Pathologists will publish a formal endorsement of the CAP cancer checklists. Currently, all provinces are in various stages of synoptic cancer reporting implementation. For examples, in British Columbia/Yukon, pathologists conform to the CAP checklists and work is underway to leverage provincial infrastructure to implement electronic synoptic reporting. In 2009, Alberta/NWT, a Cancer Synoptic Reporting Working Group was formed. Their mandate includes a requirement to use synoptic reporting based on CAP checklists. In Ontario, standardized synoptic reporting is being rolled out. Currently, 90% of cancer cases are reported electronically to Cancer Care Ontario, with 30% of those reports received in a standardized synoptic format. In Quebec, several of the CAP cancer checklists have already been translated into French.

Endnotes

¹ <http://www.cdc.gov/nchs/FASTATS/lcod.htm>

² Currently under contract #1U58DP001596

³ Most of the information cited in this chapter refers to surgical pathology reporting. Relatively few articles were found regarding the use of synoptic reporting on cytology fine needle aspiration specimens. [Tsan 2007] compared two periods of performance of reporting thyroid FNA. The synoptic cytology reporting system divided the FNAC into five diagnostic groups accompanied by implications and recommendation for patient management. The results enabled clear communication of complex and varied findings. The authors found that sensitivity, accuracy and overall false-negative rate improved with the introduction of synoptic cytology reporting. Synoptic reporting encouraged the cytopathologists to reach more conclusive and consistent diagnosis with more consistent terms. The synoptic reports were simple to understand without substantial changes in a surgeon's practice.

⁴ The CRC handbook discusses English as the only language for aircraft maintenance technicians (in any part of the world) and the costs of maintenance in industry as at about 80% of the total burden --- this includes everything from resolving miscommunications, resolving impedance mismatches between the protocols of communications between different facilities and suchlike --- all of which could be seen as analogs for the healthcare industry. Engineering Maintenance: A Modern Approach, B.S Dhillon, CRC Press, 2002. (ISBN# 1-58716-142-7)

⁵ Classically, URIs were conceived as including at least two "classes", Uniform Resource Locators (URLs) and Uniform Resource Names (URNs). The former were conceived as designating the location of a resource, the latter as representing a location-independent name. Within each class, any number of "schemes" could be defined, distinguished by In the In the contemporary view of the URI system (reviewed in [URI Planning Interest Group W3C/IETF 2001]), the classical distinction between URLs and URNs is viewed as of decreasing significance, and all schemes are seen now to exist in a flat URI space.

⁶ These terminologies are typically hierarchically structured lists of terms with associated alphanumeric codes, each in a terminology-specific format. The preferred term associated with each code may be accompanied by one or more alternate terms or synonyms. In some but not all terminologies, cross-hierarchy links are included that reflect association types other than parent-child among terms to provide additional information that may be useful to humans or computers. None of the terminologies is modeled in a standard modeling language such as UML, nor are any published in a format that is immediately compatible with standard machine inferencing interface languages such as KIF [Genesereth 1992], Prolog, LISP, Common Logic, the N-triples family or the RDF family. In some but not all cases, English-language descriptions expounding the intended meaning of some or all terms may be part of the official terminology; in other cases, questions regarding appropriate usage of the terms is not an official part of the terminology, but comes into existence as part of the "lore" of the user community.

While it is a common requirement to apply terms from multiple vocabularies to a single document—for example, it is often the case that a medical document may carry both CPT codes and ICD codes—it is uncommon to mix vocabularies in a single code string; for example it would be unusual to use a CPT code as a modifier of an ICD code.

⁷ <http://www.jfsowa.com/talks/cnl4ss.pdf>

⁸ That the words "bite", "dog" and "man" occur near each other in a sentence does not capture the entire meaning unless we know whether the relationship intended is that a dog bit a man, a man bit a dog, a bite dogged a man, a dog manned a bite, etc

⁹ Taxonomic data can certainly be stored in a patient database as one more item of interest; but it does not have sufficiently well-defined computability relation to serve as the schema for the database.

¹⁰ The IEEE name was originally an acronym for the Institute of Electrical and Electronics Engineers, Inc. Today, the organization's scope of interest has expanded into so many related fields, that it is simply referred to by the letters I-E-E-E (pronounced Eye-triple-E).

¹¹ A similar but alternate model of these upper layers of interoperability due to USDoE effort is the GridWise framework which, unlike the LCIM, does not call out dynamic (temporal) interoperability as separate layer, but GridWise Layers 7 and 8—interoperability concerning objectives and policies—bear similarity to LCIM's assignment of assumptions and constraints to its highest Conceptual Interoperability level