Introduction

Introduction

Introduction

Overwhelmed with Information

Today more than ever before in history, we live in an age of information-driven science. In all areas of the life sciences, in particular, well-organized and well-funded research groups are carrying out sustained and systematic research into areas of fundamental biological concern, yielding ever-larger quantities of information that is accessible only with the aid of computers. Vast amounts of information are being produced daily as a result of new types of high-throughput technology in areas such as next generation sequencing, molecular screening, and 2-, 3-, and 4-D imaging at multiple scales from molecules through cells and cell populations up to whole brains. At the same time the contents of scientific journals are increasingly being made available in forms that make them accessible to automated search and processing.

Already, the sheer quantity of available scientific information is becoming overwhelming, and this new information can be used effectively only if there is some strategy for ensuring its progressive integration with the information already existing, and for making it readily available in formats understandable to both computers and to human beings. The progress of science requires that the results being achieved in Pittsburgh or Berkeley should be able to build on the results already achieved in Peking or Bangalore. For these reasons scientific information needs to be stored, standardized, processed, and made available in a way that overcomes the idiosyncrasies of particular research groups and technologies.

Computers are able to store massive amounts of information, more than any single human memory could possibly retain, and they are able to retrieve specific information in focused ways, and to perform logical operations—to reason, in a sense—across this information in ways that go beyond what would be achievable by any human mind. It is thus by now self-evident that the problems of storing, organizing, retrieving, integrating, and making available the ever-growing quantity of scientific information must be met by exploiting the powers of computers. Indeed, many are now pursuing a vision of the future in which the knowledge possessed by experts working in the same or related fields would be organized and stored in interconnected computer repositories in such a way that it would become accessible to any human being or to any suitably equipped computer anywhere in the world, in real time, and be continuously updated in light of new results.

For this to occur, however, databases must be created in such a way that their contents can be shared and used in equally effective ways not only by those who have created them, and by those who populate and maintain them, but also by as yet unidentified populations of external users. For reasons that we will discuss, this goal has not yet been attained. A recent survey of healthcare data scientists finds that it is the diversity of data available, not the quantity, that poses the primary challenge in making use of electronic medical data and information.1 Similarly, a recent report from the Office of the National Coordinator for Health Information Technology points out that “despite significant progress in establishing standards and services to support health information exchange and interoperability, it is not the norm that electronic health information is shared beyond groups of health care providers who subscribe to specific services or organizations.”2

This same point was reiterated in expert witness testimony before the U.S. Congress in July 2014.3 Electronic information needs to be interoperable, shareable, and reusable. Think of a doctor specializing in a certain rare disease with immediate access to the most current information about all the patients suffering from this disease and about novel treatment outcomes. Imagine, still more ambitiously, a single, integrated biomedical knowledge base, a kind of Great Biomedical Encyclopedia, comprehending all biomedical knowledge within one constantly evolving system. Such possibilities are not beyond our reach. Experiments are currently being made under headings such as “semantically enhanced publishing” or “Big Data to Knowledge” (BD2K), and the potential benefits of success of such ventures are easy to appreciate. These benefits can be achieved, however, only through radical improvements in the degree to which the information systems involved are capable of interoperation.

What is needed in order to advance interoperability in a way that would address the shortcomings just described? First, consistency in the way standard sorts of data are described—consistency in units of measure, in terms for chemical, biological, and clinical entities of different sorts, in tagging of sequence and image data, and many more. Second, however, it requires consistency at the level of scientific assertions: when one biologist identifies a certain enzyme as exercising a function of a certain type when in a certain sort of cellular location, then his assertions have to be comparable to—and logically combinable with—the assertions of other biologists who have identified, for instance, that enzymes of that sort interact with certain other proteins whenever such a function is realized. The task of collecting and organizing scientific data in a way that supports such comparisons and combinations in the era of information-driven science is, it turns out, much more difficult than was anticipated in the era when scientific reasoning was carried out by human beings without the aid of computers. In the pages that follow we address some of the obstacles to interoperability and attempt to show how ontology can help overcome them.

Given the continuing and accelerating developments in science and information-based technology, a perspective geared toward enhanced accessibility of databases is necessary, especially in highly data-intensive and rapidly evolving fields such as biomedicine. The attempt to achieve these ends is being carried out in the work of many current institutions and initiatives, including the Gene, Cell and Protein Ontology Consortia, the ISA Commons, the Neuroscience Information Framework (NIF) Standards, and the cROP (Common Reference Ontologies for Plants) and Open Biomedical Ontologies (OBO) Foundry initiatives. In addition, the National Institutes of Health and its counterparts in other countries are providing significant grant support to researchers in the hope of making information resulting from biomedical research more easily accessible through new publishing policies and through the use of new types of unified information systems.

Obstacles to Accessibility: Human and Technical Idiosyncrasy

As we have already indicated, there are a number of obstacles to achieving effective accessibility, interoperability, and reusability of data and information. The first is that the members of the community of scientific (including clinical) researchers use different, and sometimes incommensurable, specialist terminologies and formats in describing the results of their research. The second is that they also use a variety of different computer technologies to encode and store their results, in part because these technologies are themselves rapidly evolving; in part because software and database developers have incentives to create new artifacts rather than to reuse what already exists. These two issues can be labeled the *human idiosyncrasy* and the *technological idiosyncrasy* problems, respectively. Different researchers encode their research using different terminologies and coding systems, and using different computing formats and software.

These problems are compounded by the fact that clinicians, informaticians, and researchers do not always use terms systematically or consistently. To give just one example, the chemokine identified in the Protein Ontology as

PR:000001987 C-C motif chemokine 15

is referred to, in the literature of different research communities, using all of the following labels in table 0.1.

[Insert Table 0.1]

Such incongruities cause problems not only for human experts—who are typically familiar with the usage only in their own disciplinary communities—but also for computers. Even small imprecisions of this sort, if multiplied throughout a database or information repository, can lead to problems in successfully finding, organizing, and using relevant information by means of computers. They can lead not only to faulty classifications of biomedical phenomena and to failures in communication of research results, but also to faulty diagnoses.

The Computer Limitations Problem

In addition, the very virtues of computer implementations—unlimited memory, efficient retrieval and reasoning, widespread availability—create a further issue, which we shall refer to as the *computer limitations problem*.

This problem arises from the fact that computers are, in familiar ways, unintelligent; they do not understand themselves, their programming, or the intended interpretation of the representations that they contain and manipulate. For this reason the use of computers in addressing the issues of scientific information management creates problems of its own. Above all, the very success of ontology-based approaches to the integration of data has led to a *multiplication of ontologies* in ways that threaten to recreate in a new form the very problems of interoperability that ontologies were themselves designed to solve.

Some Implications of Computer Limitations for Information Representation and Management

Computers are limited, first, in the sense that they cannot tell good from bad data. If data entered into a computer are described in vague or ambiguous or incoherent ways, then the computer programmed to reason with such data will likely produce results that are equally vague or ambiguous or incoherent. For example, one early version of the International Classification of Nursing Practice (ICNP) defined “water” as “a type of Nursing Phenomenon of Physical Environment with the specific characteristics: clear liquid compound of hydrogen and oxygen that is essential for most plant and animal life influencing life and development of human beings.”4

A definition such as this, while clearly incorrect—water is not “a type of nursing phenomenon”—may seem relatively benign (after all, the definition is not completely incorrect). However, from the definition of “water” just provided, it would follow by relatively standard reasoning that “there is a nursing phenomenon that is essential for most plant life.” To a human mind this is a puzzling and clearly false claim, but to an automated reasoning program it is a simple consequence of the definition. Such mistakes can severely compromise the goals of information aggregation, but they will not normally be detectable even by a perfectly functioning computer.

Second, two databases that store the same or related information, but use different terminologies or different organizational principles, cannot by themselves come to some kind of agreement about what they are referring to or carry out by themselves the task of aligning the information they contain. In such cases the data in the two repositories are siloed. Human beings, unlike computers, can comprehend the meaning as well as the vocabulary and grammar of a language. Thus, the information in two repositories such as those just described will be separated by a gap that can be crossed only with human intervention, for example through the provision of an explicit set of instructions for the translation (or “mapping”) between the two terminologies. The National Institutes of Health refer in this connection to the phenomenon of “data wrangling,” encompassing activities that make data more usable by changing their form but not their meaning: “Data wrangling may involve reformatting data, mapping data from one data model to another, and/or converting data into more consumable forms. Such data wrangling activities make it easier to submit data to a database or repository, load data into analysis software, publish to the Internet, compare datasets, or otherwise make data more accessible, usable, and shareable in different settings.”5

All of these activities are designed to address what we might call the *Tower of Babel problem* of computers and information systems talking past each another in a mutual misinterpretation of information, which can arise even where computers use the same terms, if sameness of meaning has not been previously guaranteed.

What such problems show is that, while computers and information technology will indisputably form a central part of the solution to the problem of finding, organizing, integrating, and sharing the Big Scientific Data of the future, they are by no means the entire solution.

The Problem of Imprecise Thinking

The underlying goal, in all of the preceding, is the interoperability of data and information systems, or in other words a situation in which data collected in one system would be usable within the framework of a second system without further intervention by human beings. One specific obstacle to interoperability will be of primary importance in what follows—and is arguably the root cause of the other problems already discussed. It can be summarized under the heading of *imprecise thinking*, by which we mean a family of interrelated errors of logic and of language use—that have repeatedly affected attempts to design information technology systems thus far. Avoiding these types of errors is important because, as discussed, it is ultimately human beings who are responsible for designing the ways in which data and information are imported into information systems.

Existing repositories of scientific data often contain not only ambiguities and inconsistencies in the use of technical terms, but also basic errors of logic. One reason for this is a matter of human resources; creating precise definitions of terms to be used in describing data, and ensuring consistent use of terms in accordance with these definitions, is expensive. It is also slow, and constrains the freedom to explore new kinds of data, a factor of considerable importance in areas of science that are advancing at a fast pace. In biomedicine and similar domains, therefore, we are caught in the horns of a dilemma. Each research community seeks to resolve issues of data representation in the most economical way, often by resorting to ad hocshortcuts or workarounds, and the result is that basic errors of definition, of fact and of logic, creep into the systems that are developed in ways which degrade the quality and accessibility of the information they contain. This book is an attempt to help people break out of this dilemma.

An Example: The BRIDG Model

The following are examples of imprecise thinking taken from early versions of a database of biomedical information known as the Biomedical Research Integrated Domain Group (BRIDG) model. The BRIDG is a product of work by researchers from the National Cancer Institute (NCI), the U.S. Food and Drug Administration (FDA), the Clinical Data Interchange Standards Consortium (CDISC), and the Health Level 7 (HL7) Regulated Clinical Research Information Management Technical Committee (RCRIM TC) whose goal was to produce a “shared representation of the dynamic and static semantics” associated with protocol-driven research in the biomedical sciences and its associated regulatory artifacts. Part of BRIDG’s work has included the creation of definitions for biomedical terms, many of which are taken over from its HL7 partner organization, including the definition of *living subject* as “a subtype of Entity representing an organism or complex animal, alive or not.”6

The first problem with this definition is that it will cause logical problems when used in the coding of data about (for example) corpses, since it must leave room for coding data about an entity that is at one and the same time both living and not living.

The second problem is that it conflates an object (a living subject) with its representation (an “Entity representing”). This is like being told that a mammal is a *representation* of an animal that feeds its young with milk. It is an example of what philosophers refer to as the use/mention confusion, illustrated in its most blatant form by a sentence such as “Sleeping is healthy and contains two vowels.” The word “mammal” is a noun that stands for or represents things, but mammals themselves are not nouns. Mammals are things, specifically living things that feed their young with milk.7

Another example is taken from an early version of BRIDG’s definition of the class *performed activity* as “the description of applying, dispensing or giving agents or medications to subjects.”8 Here an activity, which is a kind of event that happens in the world, is defined as a “description” of certain other kinds of events. Here again there is an obvious confusion of an entity with the description of an entity.

Consider, now, what happens when the two definitions introduced so far are considered together for purposes of reasoning. Here are the definitions again:

• *Living subject:* A subtype of Entity representing an organism or complex animal, alive or not.

• *Performed activity:* the description of applying, dispensing, or giving agents or medications to subjects.

Suppose (as is plausible) that *administering insulin* is a *performed activity*, and that “subject” in the preceding refers to a *living subject*. Then, combining these two definitions, we get: *administering insulin is the description of applying, dispensing or giving agents or medications to a subtype of Entity representing an organism or complex animal, alive or not*.9

BRIDG, or at least its early version, is just one example of a terminology and modeling resource that has encoded such mistakes. In a more recent version *performed activity* is redefined as “an activity that is successfully or unsuccessfully completed.”10 Here the use-mention error has been corrected, but unfortunately it seems that a new error has entered in. To help the user, BRIDG provides two examples of usage of *performed activity*, as follows (where “CBC” stands for complete blood count):

CBC that is performed on a specific StudySubject on a given day.

A scheduled blood draw that is missed by a specific ExperimentalUnit on a given day.

We assume these to be examples of “successful” and “unsuccessful” completion, respectively. From the second example, however, it seems that we can infer that a blood draw that is *not* *performed* (because it is “missed”) is an example of a *performed* activity.

Another common mistake in artifacts such as BRIDG is the inclusion of *circular definitions*. A circular definition is a definition that uses the term defined, or a close synonym, in the definition itself, thus rendering the definition (at best) uninformative. For example, the First Healthcare Interoperability Resources Specification (FHIR), again drawing on HL7, defined *container* as “a container of other entities” and *food* as “naturally occurring, processed, or manufactured entities that are primarily used as food for humans and animals.”11 Such circular definitions are not false (they amount to saying simply that a thing is what it is); but they are uninformative and therefore contrary to the intent of designing information resources that will be helpful repositories of information and will serve to constrain incorrect coding by providing information helpful to the coder. In some cases the definitions are worse than circular, for example when a term such as “health chart” is defined as “the role of a material (player) that is the physical health chart belonging to an organization (scoper).”12 The best that can be said about definitions of this sort is that they certainly will not help ensure that the terms in question are used correctly by the human beings who need them in order to code healthcare data.

These are just a few examples of the kinds of imprecise thinking that can and do occur in biomedical informatics and other fields. Many of these errors in definition are familiar from subjects dealt with in logic and philosophy: problems of ambiguity, circular definitions, use-mention confusions, and confusions of reality with our thoughts or perceptions of reality. It is thus important to stress that there are principles of reasoning and methods—long familiar to students of logical philosophy—for constructing definitions and classifications of information in ways that are designed to avoid such errors. What is needed is for these principles to be articulated and brought to bear in a systematic way in the context of information systems.

Ontology as Part of the Solution

In philosophical contexts, “ontology” has traditionally been defined as the theory of what exists (or of “being *qua* being”): the study of the kinds of entities in reality and of the relationships that these entities bear to one another. This study includes in principle the entities dealt with by the special sciences (physics, chemistry, biology, and so on); but philosophical ontology has a universal focus. It is directed at the most basic or most general features of reality, features common to all domains, including, but not restricted to, the domains covered by science.

Examples of such general or common features of reality might include: unity and number, identity and difference, part and whole. The goal of philosophical ontology is to provide clear, coherent, and rigorously worked out accounts of such basic features, and to argue for these accounts, for example in light of their relative simplicity and logical coherence, and also in light of their coherence with the special sciences (or, in earlier times, for their coherence with theology).

In recent times use of the term “ontology” has become prominent in computer and information science, and has been especially successful in areas of bioinformatics. The term “ontology”—as in “Gene Ontology,” “Infectious Disease Ontology,” “Plant Ontology,” and so forth—refers to a standardized representational framework providing a set of terms for the consistent description (or “annotation” or “tagging”) of data and information across disciplinary and research community boundaries.

Ontologies are designed to promote greater consistency in description of data. To this end the terms in an ontology are provided with both textual definitions (to ensure consistency on the side of human beings maintaining and using the ontology) and logical definitions (to aid computer access and quality control). The result is organized in a graph-theoretic format (see chapter 4), in which terms serve as nodes in the graph and ontological relations (as between a type and its subtypes) serve as edges. The terms are then used for purposes of annotating or tagging heterogeneous data contained in multiple electronic information repositories.

In this way an ontology brings multiple advantages. It provides a common mode of access to the data. It promotes intertranslatability of the content of different data repositories, thereby promoting the cumulation of science. It helps to identify incompatibilities between different bodies of data, thus leading to new scientific questions, which may need to be addressed by new experiments. It enables the development of more powerful software tools for the mining of valuable scientific information from aggregated data stores and thereby also enables the formulation of more powerful queries and analytical methods.

For all of this to work, however, the ontologies themselves have to be developed in a mutually consistent fashion, they have to be used aggressively in annotations to multiple different sorts of data and information, and at the same time they have to be adjusted over time in such a way as to keep pace with scientific advance. The solution to the ontological problem is thus not simply a matter of agreeing on a common vocabulary and using it to annotate data. Rather, it requires a coordination of research groups simultaneously developing ontology resources in distinct but interrelated areas. Such coordination involves multiple different sorts of activity, including for example programming, project management, and user-interface development. Some of these activities, however, are of a philosophical nature. For experience has shown that the coordination of ontology developers working in different fields can be advanced if the ontologies they create and maintain through time can employ a common set of basic categories, which can be used as common starting point by developers of ontologies for different domains. This common set of basic categories helps in determining where entities of different kinds should be positioned within an ontology, to determine what relationships hold between them, and to determine how the corresponding terms in the ontology should be defined. The development and application of such a common set of basic categories is, however, itself a nontrivial exercise, and it requires the addressing of problems that are at least closely analogous to many of the traditional problems of philosophical ontology.

There is thus an affinity between the two senses of the term “ontology” we distinguished, and the recognition of this affinity allows us to define an approach to the building of computer-based ontologies on the basis of tried and tested logical and philosophical principles. This book is intended as an introduction to the foundations and methods of this approach.

A New Organon for the Information Age

In the following pages, we will present and explain the basic elements of the ontological approach to solving the problems of information management. In addition, we will put forward specific recommendations and principles that are intended for use by individuals interested in constructing ontologies in new domains.

While we believe that our proposals in this book are applicable in many other areas where terminologies and ontologies are being used in the organization of data and information, including industry, finance, government administration, manufacturing, and defense, we have directed our remarks primarily to those working in areas of information-driven scientific (including clinical) research. We view science as the systematic attempt to describe and explain what exists on the basis of experiments. Our primary focus will be on the construction of ontologies for the purpose of representing the sorts of entities that are of concern to scientific research. What is said can be applied, however, without restriction, to any domain where information is being assembled and used concerning what exists.

This document is thus conceived as an *organon*, by which is meant an *instrument*, for the conduct of scientific research, especially as concerns the representation of the results of such research in digital form. The first organon was written by Aristotle in the fourth century BC, and included his works on deductive reasoning (logic and the syllogism). Francis Bacon’s *Novum Organum* (1620) was conceived as an extension and correction of Aristotle’s *Organon* in light of the success of the experimental method introduced into science almost two thousand years later. Bacon focused on what is involved when inductive reasoning is used as part of a gradual approach to the understanding of nature, which in his eyes involves moving by degrees from particular cases and attempting to discover general axioms from those observations.

The increasing use of computers in scientific research, both for representing information and for acquiring novel results, has raised in its turn a host of new questions about the nature and methods of science. Our proposals in the chapters that follow are intended to constitute portions of a new instrument of scientific method for the information age, one that will clarify basic principles and methods of computer-based and computer-assisted scientific representation and research, with the goal of creating a more successful collaboration between scientists and informatics researchers.

We will focus on basic theoretical elements, on principles, and recommendations associated with best practices for the building of domain ontologies. Our subject, therefore, is the human contribution to the realization of the tasks involving ontology support in science, as opposed to issues associated with particular computational implementations. This is, first of all, because issues concerning idiosyncrasy and imprecision in human use of terms are logically prior to issues of computer encoding and implementation (the scientific information to be encoded must be properly understood, defined, and classified before it can be successfully encoded in a computer language). But it is also because the definitions, theory, and principles of best practice for ontology design that we will be discussing will in almost all cases be applicable regardless of the particular software framework in which the information is to be embedded. Our focus here is thus on giving researchers the theoretical principles they need to build useful ontologies that will be stable, we believe, even through successive generations of computer software and hardware.

Suggested Further Reading

For additional information on ontology and information ontology, see the website of Barry Smith, which includes ontology tutorials as well as links to both introductory and advanced essays on a range of ontological topics (<http://ontology.buffalo.edu/smith/>). We particularly recommend the following papers as introductions to basic issues of ontology and information ontology that reiterate and expand on the themes covered here in the introduction.

<jrn>Feigenbaum, Lee, Ivan Herman, Tonya Hongsermeier, Eric Neumann, and Susie Stephens. “The Semantic Web in Action.” *Scientific American* 297 (2007): 90–97.</jrn>

<edb>Grenon, Pierre. “A Primer on Knowledge Management and Ontological Engineering.” In Applied Ontology: An Introduction*,* ed. Katherine Munn and Barry Smith, 57–82. Frankfurt: Ontos Verlag, 2008.</edb>

<edb>Munn, Katherine. “Introduction: What Is Ontology For?” In Applied Ontology: An Introduction*,* ed. Katherine Munn and Barry Smith, 7–19. Frankfurt: Ontos Verlag, 2008.</edb>

<edb>Smith, Barry. “Ontology.” In Blackwell Guide to the Philosophy of Computing and Information, ed. Luciano Floridi, 155–166. Oxford: Blackwell, 2003.</edb>

<jrn>Smith, Barry, and Werner Ceusters. “Towards Industrial Strength Philosophy: How Analytical Ontology Can Help Medical Informatics.” *Interdisciplinary Science Reviews* 28 (2003): 106–111.</jrn>

<edb>Smith, Barry, and Bert Klagges. “Bioinformatics and Philosophy.” In Applied Ontology: An Introduction, ed. Katherine Munn and Barry Smith, 21–38. Frankfurt: Ontos Verlag, 2008.</edb>

<eref>Smith, Barry, Waclaw Kusnierczyk, Daniel Schober, and Werner Ceusters. “Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain.” In *Proceedings of the 2nd International Workshop on Formal Biomedical Knowledge Representation* (KR-MED 2006), vol. 222, ed. Olivier Bodenreider, 57–66. Baltimore, MD:KR-MED Publications, 2006. Accessed December 17, 2014. http://www.informatik.uni-trier.de/~ley/db/conf/krmed/krmed2006.html.</eref>

<edb> Smith, Barry, Lowell Vizenor, and Werner Ceusters. “Human Action in the Healthcare Domain: A Critical Analysis of HL7’s Reference Information Model.” In Johanssonian Investigations: Essays in Honour of Ingvar Johansson on His Seventieth Birthday, ed. Christer Svennerlind, Jan Almäng, and Rögnvaldur Ingthorsson, 554–573. Berlin/New York: de Gruyter, 2013.</edb>

Table 0.1

Labels for chemokine in different research communities

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| CCL15 ccl-15 HCC-2 hcc-2 hcc2 hmrp-2b HMRP-2B LKN1 LKN-1 | mrp-2b MIP5 lkn-1 MIP-5 mip5 mip-5 MIP-1D mip-1d MIP-1 delta | mip-1(delta) mip1(delta) mip1delta mip-1-delta mip-1delta mip1d MRP-2B Mrp-2b NCC-3 | ncc3 NCC3 ncc-3 SY15 SCYL3 scyl3 scya15 SCYA15 | CC motif chemokine 15 c-c motif chemokine 15 chemokine CC2 chemokine CC-2 chemokine cc-2 chemokine (c-c motif) ligand 15 inducible cytokine subfamily A (CysCys), member 15 leukotactin-1 leukotactin 1 | mip-related protein macrophage inflammatory protein 5 macrophage inflammatory protein-5 new CC chemokine 3 small ccl-15 small inducible cytokine A15 small-inducible cytokine a15 small-inducible cytokine A15 |

*Note:* Information from the ImmuneXpresso project (http://www.immport-labs.org); with thanks to Shai Shen-Orr and Nophar Giefman.

{Notes\_begin}

Introduction

1. Mik Miliard, “Data Variety Bigger Hurdle than Volume,” *HealthcareITNews*, July 3, 2014, accessed August 25, 2014, <http://www.healthcareitnews.com/news/data-variety-bigger-hurdle-volume?topic=02,06&mkt_tok=3RkMMJWWfF9wsRonuq3IZKXonjHpfsX87OQkWbHr08Yy0EZ5VunJEUWy2YIDT9Q%2FcOedCQkZHblFnVUKSK2vULcNqKwP>.

2. Karen B. DeSalvo and Erica Galvez, “Connecting Health and Care for the Nation: A 10-Year Vision to Achieve an Interoperable Health IT Infrastructure,” The Office of the National Coordinator for Health Information Technology, 2014, p. 4, accessed September 1, 2014, <http://www.healthit.gov/sites/default/files/ONC10yearInteroperabilityConceptPaper.pdf>.

3. Greg Slabodkin, “HER Interoperability Key to Modernizing Clinical Trials,” *HealthData Management*, July 10, 2014, accessed August 25, 2014, <http://www.healthdatamanagement.com/news/EHR-Interoperability-Needed-to-Improve-Clinical-Trials-48392-1.html>.

4. Susan J. Grobe, “ICNP Version 1: International Classification for Nursing Practice—A Unified Nursing Language System,” 2005, accessed August 30, 2014, [www.nicecomputing.ch/nieurope/S%20Grobe%20ICNP.pdf](http://www.nicecomputing.ch/nieurope/S%20Grobe%20ICNP.pdf).

5. U.S. Department of Health and Human Services, “Development of Software and Analysis Methods for Biomedical Big Data in Targeted Areas of High Need (U01),” 2014, accessed August 25, 2014, http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-14-020.html.

6. This definition is taken over from the HL7 Reference Information Model (RIM) Version V 02-07, as described in section 3.2.5, accessed July 25, 2014, http://www.vico.org/CDAR22005\_HL7SP/infrastructure/rim/rim.htm.

7. In a later version of its documentation HL7 corrects this second error. It now defines a *living subject* as “anything that essentially has the property of life, independent of current state (a dead human corpse is still essentially a living subject).” See Health informatics, HL7 version 3, Reference information model, Release 4, Document ISO/HL7 21731:2011(E), <http://www.hl7.org/index.cfm>. We note that the first error still remains.

8. BRIDG Version 1.0. Phase 1.0. Created on January 5, 2005; last modified December 14, 2006, accessed August 25, 2014. All releases are available at <http://bridgmodel.nci.nih.gov/>.

9. Compare also the case of Microsoft HealthVault, which defines “an allergy episode [as] a single unit of data that is recorded in Microsoft HealthVault,” accessed August 25, 2014, <http://msdn.microsoft.com/en-us/library/aa155110.aspx>.

10. BRIDG Version 3.2, created May 9, 2014, accessed August 25, 2014, http://bridgmodel.nci.nih.gov.

11. Both examples taken from the First Healthcare Interoperability Resources Specification (FHIR), accessed June 6, 2014. <http://www.hl7.org/implement/standards/fhir/v3/EntityClass/index.html>.

12. Health Level 7 FHIR Development Version, accessed September 29, 2014, <http://hl7.org/implement/standards/FHIR-Develop/v3/RoleClass/index.html>.

{Notes\_end}