Research Data Governance, Roles, and Infrastructure

14

Anthony Solomonides

Abstract

This chapter explores the concepts, requirements, structures, and processes of data or information governance. Data governance comprises the principles, policies, and strategies that are commonly adopted, the functions and roles that are needed to implement these policies and strategies, and the consequent architectural designs that provide both a home for the data and, less obviously, an operational expression of policies in the form of controls and audits. This speaks to the "What?" and "How?" of data governance, but the "Why?" is what justifies the extraordinary efforts and lengths organizations must go to in the pursuit of effective data governance. This receives a fuller answer in this chapter; in brief, information is a valuable asset whose value is threatened both by loss of integrity, the principal internal threat, and by its potential for theft or leakage, compromising privacy, business advantage, and failure to meet regulatory requirements—the external threats. Internal and external threats are not quite so neatly distinguished in real life, as we shall see later in the chapter.

Keywords

Data governance · Research data governance · Information governance · Data integrity · Internal and external threats · Security · Privacy · Confidentiality Regulatory frameworks · HIPAA · Common rule

The American Medical Informatics Association (AMIA) Clinical Research Informatics Working Group (CRI-WG). Acknowledgements: Judy Logan, WG Chair 2014–2016; Abu Mosa, Monika Ahuja, Kris Benson, Shira Fischer, Lyn Hardy, Kate Fultz Hollis, Bernie LaSalle, Nelson Sanchez Pinto, Lincoln Sheets, Ana Szarfman, Chunhua Weng, Chair Elect 2018–2020.

A. Solomonides, PhD, MSc (Math), MSc (AI), FAMIA (✉)
Department of Family Medicine, NorthShore University HealthSystem, Research Institute, 1001 University Place, Evanston, IL, USA

R. L. Richesson, J. E. Andrews (eds.), *Clinical Research Informatics*, Health Informatics, https://doi.org/10.1007/978-3-319-98779-8_14

Introduction: A Conceptual Model

This chapter was originally conceived around a framework discussed by the members of American Medical Informatics Association's (AMIA) Clinical Research Informatics Working Group (CRI-WG). It finally crystallized in this form as a contribution to the present book. The framework is depicted in Fig. 14.1.

The schema in Fig. 14.1 places data and information at the center: the nature and context of data and information impacts the way it is governed, the functions that implement governance, and the underlying technology that houses, communicates, and defends it. The idea is that not only does each of these domains of activity demand attention in its own right, but the relationships and interactions between them also must be addressed. All relations are bidirectional: data governance adds to the data even as it "governs" it.

In the course of this chapter, we shall examine the qualities that give data its value, the life cycle of data, the vulnerabilities of data, and the implications of all these for the organization of "data governance."

What is data governance? As suggested in the model, it comprises the principles, policies, and strategies adopted, the functions and roles that—in the favored phrase of the domain—are "stood up" to implement these policies and strategies, and the

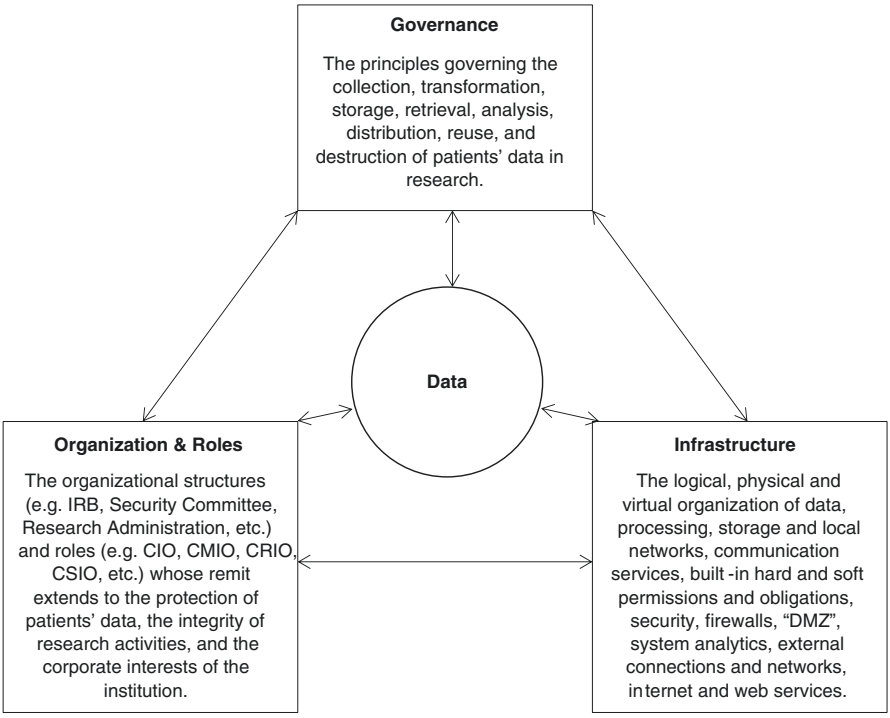

Fig. 14.1 The conceptual model. The three domains of data governance and their interactions

consequent architectural designs that provide both a home for the data and, less obviously, an operational expression of policies in the form of controls and audits.

This speaks to the "What?" and "How?" of data governance, but the "Why?" is what justifies the extraordinary efforts and lengths organizations must go to in the pursuit of effective data governance. This receives a fuller answer below, but in brief, information is a valuable asset whose value is threatened both by loss of integrity, the principal internal threat, and by its potential for theft or leakage, compromising privacy, business advantage, and failure to meet regulatory requirements—the external threats. Internal and external threats are not quite so neatly distinguished in real life, but we reserve this distinction for later in the chapter.

Research Data Governance

The principles governing the collection, transformation, storage, retrieval, analysis, distribution, reuse, and destruction of patients' data in research.

In any enterprise, and in a healthcare organization more than most, data is literally an asset and, metaphorically, also a significant liability. The value of data can be realized in better business and care delivery decisions, in fulfilling a public health mission alongside provision of best care, in discovery of new knowledge through research, in improving quality and safety of patients, and in informing the healthy on how to maintain and enhance their health. The trouble with data is its vulnerability. If stolen by a competitor, it can damage a business irreparably, whether by identifying weaknesses in services offered or potential clients to be enticed away. In healthcare, if patients' data is disclosed without authorization, there are consequences beyond loss of business and patients' loss of confidence in the system: regulatory breaches bring fines and large settlements in their wake.

As a discipline, data governance delineates the (kinds of) principles, policies, strategies, functions, and actions that can guide and support the establishment of a coherent data governance program. As a practice, data governance aims to defend the value of the data in an organization, facing both inwards and outwards. The task for the institution is to assure the integrity of the data so that it does not lose its informational value. The task external to the institution is to protect the data from deliberate theft, accidental leakage, and inappropriate disclosure.

This chapter reviews more specifically the question of data governance for electronic patient data that is to be used for research. It would be more accurate to say, of course, "the questions" in plural form. To begin, there is no universal agreement on what constitutes data for research rather than data for the effective delivery of care, data for quality assessment or improvement, or even data for administrative transformation, e.g., through analytics. Thinking particularly of patients' medical records, it is not even clear who "owns" it, notwithstanding ownership rights asserted both by patients and by providers. There is considerable variability on what is interpreted as "human subjects" research in different places, with consequences for informed consent requirements. (Indeed, as of this writing, there is some

uncertainty as to the exact requirements for consent following changes to the Common Rule[1] by the last and current administrations.[2]) Thinking of data, we must qualify our scope to "mainly" electronic patient data; some of the data may not be readily recognized as "electronic" (e.g., scanned paper documents whose content is not machine-readable). Further, powerful technologies and massive semipublic data repositories, including those of the social media giants, mean that secure de-identification of protected health information (PHI) remains an elusive goal.

What Does Data Governance Govern?

A succinct description of data governance may be framed in three dimensions: structures, processes, and results. The similarity to Donabedian's dimensions of "quality" is not accidental [1]. Structures and processes are amenable to identical definition; "results" is broader than "outcomes." Outcomes matter, but in the governance of information so do other results, such as aberrant behaviors and workarounds. A search through the literature has not surfaced many publications that elide the preeminent framework for quality improvement with data governance, but it is not hard to see the parallels. Data governance is often paired with (and then barely distinguished from) master data management, and it is again the case that what these two have in common is the concern with the quality of data and data processes. Some, notably American Health Information Management Association (AHIMA), address these issues under the title of "Information Governance"[2]; [AHIMA] we briefly turn our attention to the ambiguity between data and information.

We shall draw—and blur—the distinction between data and information. Whether we speak of data governance or information governance, there are times when it is necessary to draw a distinction: data streaming from a device, for example, in the absence of a framework, is meaningless and may be thought of as simply data: *make of it what you will!* The moment that data stream is accommodated in a data structure that confers meaning to it—e.g., a column headed "Hourly Temperature" or more obscurely, "^{O}C"—it becomes information. What has complicated this naive picture is the advent of data science in all its forms, from simple analytics to data mining and machine learning: with a little information about context, the possible meanings of a naked data stream may be guessed, so even where a useful distinction may be drawn in theory, it may be blurred in practice. Just as no

[1] Code of Federal Regulations 45 CFR part 46, subpart A, is known as the **Federal Policy for the Protection of Human Subjects** or the **Common Rule**. It is shared verbatim by a number of departments, hence "common." See https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html.

[2] As of this writing, the status is described in the announcement "HHS and 16 Other Federal Departments and Agencies Issue a Final Rule to Delay for an Additional 6 Months the General Compliance Date of Revisions to the Common Rule While Allowing the Use of Three Burden-Reducing Provisions during the Delay Period" (https://www.hhs.gov/ohrp/final-rule-delaying-general-compliance-revised-common-rule.html).

physician would have difficulty guessing what the sequence 39.4, 39.8, 39.4, 38.9, 38.6, 38.2, … likely means, a sophisticated machine learning algorithm would probably get there too.

Information is, in our definition, data organized in a way that imparts or reflects meaning. This gives information an abstract spatial quality. In this light, information means not only the (raw) data, but the meaning that renders it into information. This forces us to consider metadata on a more or less equal footing as data itself. This is reflected in the data manifold (see Fig. 14.2 above). A note of 144/102 in a patient's chart may give the appearance of a vulgar fraction, but to the knowing eye it has as very specific, indeed, highly significant meaning. How that meaning will be translated into machine-readable form—a form in which a software application can take it as its input and generate some valid output—is the result of a cascade of design decisions which also ultimately impact the governance process. Likewise, social scientists, especially social constructivists, may assert with some justification that all data is theory-laden. Grounded theory [3] notwithstanding, most data is collected with a theory of some sort in mind. We shall evade this dilemma by our convention that data becomes information in the light of a theory, however lightly that theory may be asserted—perhaps only implicitly through the headings at the top of columns of data.

In the temporal dimension, information governance spans the life cycle of the artifacts called *information*, including their creation (or capture), organization, maintenance, transformation, presentation, dissemination, curation, and destruction. The information governance process therefore treats data not only in its spatial aspect but also through its temporal dimension.

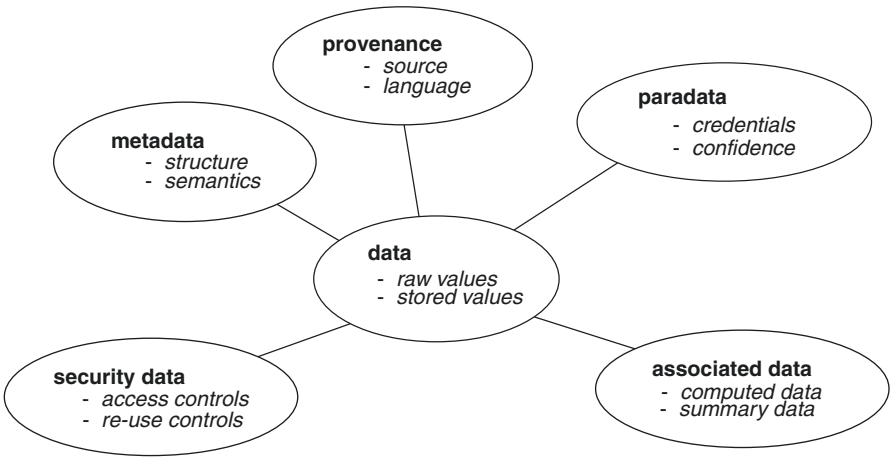

Fig. 14.2 The data manifold. Data is characterized not only by its values but also by what is loosely termed its "metadata," which can be analyzed into metadata proper, provenance data, paradata (e.g., concerning the credibility of the data), security data, and various computed summaries, and so on

"Knowledge" is beyond our consideration, but it is often confused with information or placed in the putative hierarchy of "data-information-knowledge" (to which "-wisdom" is also added to make matters even more obscure). Knowledge is a human attribute: to quote Laurence Prusak, one of the founders of the knowledge management school, "there is no knowledge but that which a knower knows" [4]. What concerns us in this review is information in the sense of data whose meaning is derivable from its form, whether that form was deliberately constructed or imputed through some algorithmic process. Information is capable of being processed by machine. Except where the distinction makes a difference, we shall use data and information as synonymous, and, when necessary, the distinction will be made explicit.

Why Data Governance?: The Value of Data

Information is a resource. Decisions can be made at least in part on the basis of information in our possession. Information can be mined for patterns that lead to hypotheses about how something works, or fails to work, or how a pattern of behavior may contribute to the development of a condition. The value of this resource, therefore, depends on certain characteristics it may possess to an absolute degree (e.g., correctness) or in some measure (e.g., relevance). These characteristics have been painted slightly different by different authors, but in broad brush they agree that data must be accurate, valid, reliable, timely, relevant, and complete.[3] Data must also be available, if it is to be useful, but along with many others, we treat that as a system rather than a data characteristic.

Accuracy Inaccurate data can scarcely be described as "information," so accuracy or correctness is essential, but this criterion encompasses certain subsidiary characteristics. Any units that are used must be explicitly defined. The data must be sufficiently granular and precise to the degree necessary for its use; ideally, it must also be unique in the sense that a system must afford a single source of truth.

Validity Data must conform with any restrictions on the values it may take, and any relationships that are prescribed between such values: in database parlance, the data must conform with certain integrity constraints. **Legitimacy** is sometimes added to this category; it is all the more important here in the context of governance. To that end, it is often desirable to be able to reconstruct a trail back to the source of the data, a form of metadata known as provenance.

Reliability Data must be both self-consistent—integrity constraints and single source of truth contribute to this—and consistent with its environment, such as the

[3] This simple list was promoted to public bodies in the United Kingdom by the now dissolved Audit Commission. The elaboration in this chapter is the author's, based on contributions from numerous authors.

applications that must use it. Where a transformation is necessary to address the requirements of an application, the validity of that translation must be assured and the transformation itself be logged in provenance.

Timeliness Data is often time-stamped, meaning that the time of its collection or entry into the system is itself recorded. Any significant time lags or delays, or any gaps, affect the usefulness of the data, especially if any data-driven decision is to be made. There should thus be minimal delay between any event and its record and minimal latency in providing the record for use.

Relevance Data is normally collected for a purpose. It is both good practice and a common regulatory requirement that a **principle of parsimony** be adopted in data collection: all the data that is required—all salient data—and only those. The **accessibility** of data, including the navigability of the architecture holding the data, is considered by some to be an aspect of relevance.

Completeness Complementing the principle of parsimony is the principle of completeness which asserts the need for the data—more precisely, for the data model—to be comprehensive, i.e., to provide as complete a picture of the entity it relates to as is necessary for its purpose. As in all modeling activity, salient features must be selected for inclusion; this is a matter of purpose and subject to scientific disagreement.

In our conceptual model of data, the data manifold (Fig. 14.2), we have distinguished between what may be termed raw values and a collection of what are often loosely called metadata—data about data—but classified into categories reflecting a purpose: *provenance*, to show how the data came from or was created; *metadata* proper, which portrays the semantic relationship between content and structure, for example, the relationship between attribute names and values; *paradata* which may be associated with confidence in the data; *security and privacy* data, reflecting access and use privileges; and *associated* data, mainly summaries of raw data.

The Life Cycle of Data

We have asserted that data governance principles, policies, structures, and functions address all phases of the data life cycle. Typically, we consider these to be collection (or creation), transformation, storage, retrieval, analysis, dissemination (or distribution), transmission, reuse, and destruction of data. In our case, we may think of these specifically as patients' or subjects' data in research.

Each of these phases in the life of data entails some threat to the integrity of the data. Poor collection practices threaten both the legitimacy and the accuracy of data; data from an inappropriately credentialed laboratory may be worthless; poorly maintained instruments may compromise precision; a copy of data collected on a portable device may remain insecurely in that device even after it has apparently

been uploaded to a secure system—the very word "uploaded" gives a false sense of security. Data is not like boxes on a dock being uploaded onto a van.

Considering creation and transformation, we know that software does not always function as intended or as designed. Even at the creation stage, habitual users of software are aware of invisible transformations that may occur when entering data (think, e.g., of presentation vs. storage formats for dates in Excel; consider the metadata needed to ensure that a date entered in US format reads correctly in a European-installed copy of the program). Data transformations undertaken in the service of analysis or dissemination likewise can cause problems. Notoriously, mix-ups between unit systems can cause catastrophic failures.

Storage in the relatively short term is highly reliable, but long-term storage is technology dependent and may provide another source of error or effective loss. If an organization considers that data still to have value, then appropriate curation is necessary to ensure its retention. When the data no longer has value or there is no legitimate reason to keep it, the data must be securely destroyed: description of the method of destruction and oversight that the necessary steps are taken often falls to a data governance function. Encryption of stored data is often required as a minimal defense against theft or leakage outside a secure perimeter.

Data analysis is often carried out using specialized software packages, including statistical tools, data analytics, de-identifiers, natural language processors (from simple concordances to highly sophisticated NLP tools), visualization, and more. The integrity of these processes is, of course, a concern and a matter for the researcher, but they also pose a challenge to a data governance function to ensure that there is no inadvertent leakage or disclosure through the use of these tools. Since these are often proprietary and function as a "black box," it is necessary to trial such software under controlled conditions in a suitable "test harness" that captures all traffic in and out of the application.

One of the principles of grid computing, and subsequently cloud computing, is the notion that when the data cannot be sent to the algorithm for whatever reason—in the case of healthcare, because it may be protected health information—there is provision for the algorithm to be sent to the data. There are some issues with this, both in terms of licensing—do all the sites need a license for any proprietary software involved?—and in technical terms, can the distributed results be legitimately aggregated? Some remarkable work has been emerging in this area [5].

Data sharing and publication are a particular challenge to a data governance function. Poor programming practices can lead to information leakage and to vulnerabilities in, for example, publication through a website, including the possibility of intrusion, malware injection, and other forms of attack. Other means of sharing, such as direct transmission of data, pose well-known security problems, including interception and corruption. Just as secure storage is typically encrypted, encryption of data for transmission provides a degree of security. However, technological advances threaten even this defense. Data may also be compressed prior to transmission to reduce its volume; depending on the nature of the data, a decision has to be made about the degree of "loss" of definition that can be tolerated in compression.

Why Data Governance?: From Data Protection to Research Ethics

While data in all its life cycle stages must be protected from error and unintended loss of integrity, it must also be defended against deliberate attack and against careless mishandling resulting in disclosure. Data needed to support business functions is not only valuable to the owner organization but is also of considerable interest to its competitors. This includes very basic data, such as details of patients and the conditions they suffer from or the specialist physicians they see. The pervasiveness of security requirements is a consequence of the digital transformation of business and of healthcare in particular. When records took the form of paper files, inappropriate disclosure meant misplacing a file and information theft meant stealing it. When we spoke of security, we meant physical security—locks and keys. The digital economy has brought with it a need for a security function of a very different kind, but the jargon of physical security has been extended to the digital variety.

By far one of the largest concerns in a healthcare organization is the protection of personal health information. The complexities of research (such as the need to "blind" studies) makes biomedical and healthcare research data management all the more fraught. This is the case in virtually all developed healthcare systems, although the jargon may differ from place to place. We shall adopt US usage, where such information is described as *protected health information* (commonly, *PHI*). In the American context, two regulatory frameworks weigh heavily on the policies and practices of healthcare organizations that engage in research: the **HIPAA** rules and the **Common Rule**. Although at the time of writing there is some uncertainty concerning the final shape of the Common Rule, the general principles, which would apply, suitably translated, in most jurisdictions with a research culture, can be outlined with some certainty.

The Health Insurance Portability and Accountability Act [6] formalized privacy requirements for any "*covered entity*" that handles patient information in electronic form. Covered entities include all providers who transmit patient data in electronic form, health plans, and healthcare information clearinghouses. When a third party is employed by a covered entity to process any PHI on its behalf, it must enter into a binding business associate agreement (BAA) with that third party, so that its handling of PHI is also ruled by HIPAA. For example, some academic medical centers that are not an integral part of their associated university have a BAA to enable academics to work with—and in particular to do research using—PHI. Pharmacy benefit managers and health information exchanges also normally operate subject to a BAA with their associated covered entities.

The **HIPAA Privacy Rule** is designed to protect individuals from harm that may be sustained through the inappropriate disclosure or illegitimate use of personal information. The scope of this protection is considerable: the individual may suffer harm from causes ranging from identity theft, through medical insurance fraud, to denial of health insurance coverage because of "known" (i.e., disclosed) existing conditions—including now genetic information which has complicated matters further still. The Privacy Rule allows for the possibility of de-identification of patient information: this may be accomplished by one of two methods—one is the so-called

Safe Harbor method which requires the removal of 18 specified types of identifiers as well as any other data that may lead to reidentification. The second method is through Expert Determination: a statistical expert must testify that by application of scientific principles, it has been determined that there is negligibly small risk that the anticipated recipient of the data would be able to identify an individual.

Supporting the goals and implementation of the Privacy Rule, HIPAA adds a Security Rule. This requires the operational, logical, and physical structure of the information function to be secured against known and foreseeable challenges. We term the function that defends against deliberate attack, inappropriate disclosure, and leakage of information the security function. By the very nature of the asset we are seeking to protect—information—security has to take many forms and be implemented at many levels, from low-level protection systems in the sense of close to the physical infrastructure, through authentication protocols for authorized users, to authorization processes and allocation of access rights, finally to an individual or, more likely, a committee charged specifically with high-level decision-making on the release of data. Since, as implied here, security also encompasses infrastructure systems and networks, the entire information architecture, physical, logical, and operational, is subject to the requirements and dictates of security. We shall see that the various demands of privacy and security (and confidentiality, as we shall add) have led to the creation of a number of distinct roles in healthcare organizations, all of whom bear the words "information officer" in their title, sometimes leading to confusion as to their exact purpose and responsibilities. We shall argue below that provided role descriptors are clear and any overlap in duties is managed, none of these roles is superfluous.

We now turn to the second framework with direct relevance for research, that of the Common Rule, as codified in Federal Regulation 45 CFR part 46. The Common Rule is so-called because it is adopted "in common" by 18 agencies, although its development is normally led by the Department of Health and Human Services (HHS).[4] The primary purpose of the Common Rule is to protect human research subjects in studies funded by any of these 18 agencies, but in practice most institutions apply the Common Rule to all research, irrespective of funding source. The Common Rule offers protection against physical and informational harms: in particular, it encompasses all the stages in the life cycle of data—collection, use, maintenance, and retention—and how these may impact a research subject's physical, emotional, or financial well-being or reputation.

An institution may obtain a *Federal-Wide Assurance (FWA)* asserting that any research funded by the 18 agencies (or all research, for that matter) will be conducted in full compliance with the provisions of the Common Rule. The Office of Health Research Protections (OHRP), an office of the DHHS, describes the FWA as "the only type of assurance currently accepted and approved by OHRP," through

[4] At the time of writing, the Common Rule is subject to revision. A revised rule had been approved on the very last day of the Obama administration, but this was suspended for review by the incoming Trump administration. Recent (April 2018) indications are that the Obama rule may be amended before it is implemented.

which "an institution commits to HHS that it will comply with the requirements in the HHS Protection of Human Subjects regulations at 45 CFR part 46." A critical step in obtaining a FWA is the registration of an *Institutional Review Board (IRB*) who must approve all research involving human subjects, whether it involves a clinical trial or processing of subjects' identified personal health information. As an alternative, it is also possible for an institution to nominate an established IRB as the one on which the institution will rely for approval of its research. Either way, the IRB must approve all research using identifiable data of living individuals with the aim to establish new knowledge. Approval by an IRB ensures that subjects will be informed of the nature, process, and risks of the research and that on the basis of this information, subjects freely consent to participate and know that they have a right to withdraw at any time. Consent may include an indication of future work that may be undertaken using the same data. However, "broad consent," in the sense that it allows researchers freedom to use the data for other studies without returning to the subjects for a fresh consent, has not hitherto been allowed.[5] Some studies undertaken with a view to quality assessment or improvement and not primarily intended to generate new knowledge may be exempt from IRB approval. Likewise, studies regarded by the IRB as posing minimal risk, or using fully de-identified data and so deemed not to be human subjects research, may be exempt from, or subject to a lighter "expedited," IRB review. The IRB is charged with continuing to monitor research studies both for noncompliance and for any unanticipated risks that arise in the course of a study. Through the mechanism of FWA and IRB review, the OHRP retains considerable powers to discipline any noncompliant entity. IRBs are subject to periodic review and are accountable for their record.

As well as PHI, privacy frameworks recognize a further category of data, *personal identifying information (PII)*. The distinction from PHI is implied in the descriptor: many of the data elements that Safe Harbor requires to be removed are PII. Personal demographics, dates of birth, telephone numbers, and so on do not impart health information but can readily identify an individual. De-identification in some cases has to be done in a way that can be reversed under very strict conditions. For example, a patient whose record appears suitably redacted with a randomly generated identifier may need to be contacted, either because something very serious has been observed (a so-called incidental finding) or because he or she meets certain criteria and is therefore a candidate to be consented for a deeper study. The linking information is sometimes entrusted to a neutral role in the institution, often approved through the IRB: the *honest broker*. The honest broker is entrusted with the link between the institutional identifier of a patient (e.g., the medical record number) and that patient's randomly generated pseudo-identifier. It is possible to arrange for the honest broker to know nothing more than that link, i.e., no PHI at all. This also provides a means to protect confidentiality.

[5] The Obama rule and the revision still under current consideration do allow for broad consent in some cases. As embodied in this rule, broad consent is thought to place a considerable burden on the institution to maintain awareness and monitor its application.

Confidentiality of personal health information extends the concept of privacy to a principle of parsimony concerning the sharing or dissemination of data. Simply stated, confidentiality requires data to be disclosed on a strict need-to-know basis. Initially this arose from considerations concerning certain stigmatizing conditions: does a medical assistant rooming a patient for a visit need not know that he has suffered from severe depression in the past? Indeed, certain kinds of data are often treated as privileged—HIV status, mental health—but this is not uniform. However, when subjects are involved in a clinical trial involving an intervention and the trial is itself "double blind," the emergency room physician faces a real problem when that subject reports to the ER with an acute neurological complaint of no known cause. Within integrated systems, the patient's electronic record may include a flag indicating that a patient is indeed involved in a trial, so that in a worst case scenario the record may be unblended to provide the treating physician with knowledge of what treatments, especially medications, the patient had received prior to his being taken ill.

Theories of Information Governance

> *In its most abstract sense, governance is a theoretical concept referring to the actions and processes by which stable practices and organizations arise and persist.*
>
> Wikipedia—entry on Governance

To govern is to manage, to control, to direct, and to steer. We tend to think of "government" as made up of the authoritative structures of regulation and control, while "governance" reflects the *process* of regulation and management. In this chapter, we have taken this broad view of the term as our scope, so as to provide a wide perspective that captures all the activities that may fall under the term, at least in as far as it relates to research. In this section, we venture a little further into the realm of legal and socioeconomic analyses of privacy so as to locate information governance in its broader context.

We can argue that information governance is driven by two forces: what may be loosely called data management or data stewardship—maintaining the integrity and safety of the data—and *privacy and business protection*, defending sensitive data from disclosure, leakage, and theft. Security, as an active program to defend the business from attack, touches on both. While the operational structures to maintain the integrity of the data are readily seen as necessary, the concept of privacy as a driver for information governance is often misconstrued. Is the entire governance "enterprise" really necessary? How does the need for privacy arise? Concerns about identity theft and medical fraud on one hand and a patient's "ownership" of her medical record each contributes, but they have their roots in alternative conceptions of privacy.

James Whitman [7] describes a dichotomy between two privacy cultures which he codifies as dignity vs. liberty and locates, respectively, in Europe and the United States. This thesis begins with the observation that many authors have difficulty

defining privacy in exact terms, often relying on allusion to make the case for privacy: "It is the rare privacy advocate who resists citing Orwell when describing these dangers"—threats to "fundamental rights [7]." The slipperiness of the concept can also be made "by citing a large historical literature, which shows how remarkably ideas of privacy have shifted and mutated over time [7]." And the contrast between European and American sensibilities is pressed home: "Why is it that French people won't talk about their salaries, but will take off their bikini tops? Why is it that Americans comply with court discovery orders that open essentially all of their documents for inspection, but refuse to carry identity cards?" Whitman traces these differences to "intuitions that reflect our knowledge of, and commitment to, the basic legal values of our culture."[7].

But what is it that must be kept private? The foundational paper on privacy by Warren and Brandeis [8] was conceived on the advent of photography and the danger that one's image may be captured unawares. From here, it is a fairly straightforward leap to the loss of privacy through the inappropriate disclosure of personal health information. Curiously, there is a quasi-symmetrical concern with the person being forced to witness something inappropriate about others, as in the occasional system message that images have been removed from an email to protect privacy. Loss of privacy in these senses appears to mean, primarily, a loss of dignity, from an image of the subject with company he may wish not to acknowledge, to a revelation of an embarrassing condition in the medical record.

lives, and the personal health record is not so different from one's home.

The instinctive response to this is to claim ownership of the personal health record, a tenet apparently bolstered by the law, although the complexity of who owns and who is the custodian of the record muddies things considerably. Positions on this are easy to polarize. How can the culture of the "learning health system" be promoted if citizens claim ownership of their health data and wish to hoard them? How can an individual claim that her data has been "stolen" if she does not own her medical record? But if the patient owns her medical record, what was the physician's intellectual contribution to that record? After all, the patient did not diagnose herself—it was a physician with 7 years' solid training and more years' experience who did that.

This observation gives us a handle on the second contrast we must reckon with. This is presented here in terms of Viktor Mayer-Schönberger's opposition of a systems-based theory of information governance to the prevailing rights-based view [9]. Mayer-Schönberger turns his attention to the protection of intellectual property (IP) as a means to break the deadlock over privacy rights. Like Whitman, he begins by observing differences between continental European conceptions of privacy rights and American ones, and in the interests of an international information economy, he seeks commonalities between them. In Europe he recognizes complementary moral and economic dimensions to information rights, while in the United States, he notes a trend toward "propertization." European modes of control over information relating to an individual, such as the legal "right to be forgotten," are expressions of a moral commitment. American legislation is a diffuse mix of federal, state, and case law which makes control over personal information all but

impossible in practice. Notwithstanding these differences, he finds little empirical evidence that these rights are much acted on in the courts. He comments wryly, "Perhaps hoping for individuals to enforce their rights through costly court action is too ambitious a vision, and thus the problem lies in the governance mechanism used to afford information privacy" [9].

Looking at the United States, he finds a more interesting contrast between the ways in which information rights and privacy rights are codified. Intellectual property rights serve twin purposes: economic and moral protection of the author, on one hand, and a stimulus to trade, on the other. While privacy rights are essentially inalienable, IP rights can be transferred—sold or licensed—even as the author retains the *moral* right to be identified as such. What would be the conditions under which personal information could be treated as property, as something title to which can be meaningfully transferred? As far back as 1998, writing in a computing journal, Kenneth Laudon proposed a market for private information: "Who owns and controls personal information in national data networks? Why not let individuals own the information about themselves and decide how the information is used? A regulated National Information Market could allow personal information to be bought and sold, conferring on the seller the right to determine how much information is divulged [10]." Chronologically, this coincides with proposals for personal health records, in the sense of records that may be "banked" by individuals, much as they bank their money and protect their financial interests, which were put forward both in the United States and the United Kingdom. Mayer-Schönberger's argument is complex and nuanced, but in essence he advocates for information rights that are governed by a "systems-based" approach. This envisions a "thick network" of professionals and formal and quasi-formal bodies that would mediate informational transactions, much as various bodies handle copyrights and patents. Exemplars of such individuals and bodies are drawn from European practice, where there are "information commissioners," "data guardians" (cf. Caldicott[6] Guardians in the UK NHS), and others that play an active role in the maintenance of privacy through audits and public reporting. This approach has the potential, both to secure privacy rights by enforcing protections and to allow the economic value of the data to be realized in a fair marketplace. Indeed, Mayer-Schönberger notes that intellectual property rights are not typically defended by individuals, but by organizations established for that purpose.

Data Governance Organization and Roles

The organizational structures (e.g., IRB, Security Committee, Research Administration, etc.) and roles (e.g., CIO, CMIO, CRIO, CSIO, etc.) whose remit extends to the protection of patients' data, the integrity of research activities, and the corporate interests of the institution.

[6] Instituted following The Caldicott Committee. Report on the Review of Patient-Identifiable Information. December 1997. UK Department of Health.

Management and regulatory oversight duties and functions in an institution are likely to be distributed among senior post-holders and committees, the former where direction is perceived to be a senior management responsibility, the latter where expert consensus as well as executive fiat may be necessary.

The commonest roles in most institutions are those of the Chief Information Officer (CIO) and the Chief Medical Information Officer (CMIO). The CIO is typically a career technical administrator who has risen to a "C-suite" executive role, while the CMIO is typically a medically qualified and still active physician who acts as a bridge between the technological functions of the organization and the body of physicians (often a medical group) in whose service the technology has been introduced but who are often highly critical of it. The CMIO likely reports to the President or Chair of the medical group as well as to the Chief Executive or Chief Operating Officer. In recent times, the role—even the very concept of the CMIO—has been challenged as too narrow. Some have advocated for the broader concept of Chief Clinical Information Officer (CCIO) which would encompass also the Chief Nursing Information Officer and rarer roles, such as the Chief Pharmacy Information Officer. Elsewhere, the role of CMIO has been redefined as that of the Chief Health Information Officer (CHIO), implying a higher level role, not so much in terms of the organizational hierarch as in terms of the types and breadth of information that the post-holder should be concerned with. Few institutions have all these roles, but almost all have at least one—which one reflecting history and organizational preferences. In the commonest setting, where there are both a CIO and a CMIO, they are likely to divide their attention, respectively, between systems, networks, and technical employees for the CIO and the conceptual design of the information, its integrity, and the way it is entered by and presented to physicians.

Among committees, the Internal Review Board (IRB) is necessary wherever research is done; since quality assurance and quality improvement work often includes elements of research, the IRB is essentially universal. In many organizations, a data governance committee may be established, whose role is significantly narrower than the "data governance" discussed in this chapter: it is restricted to determinations of whether data has been sufficiently de-identified, or the intended recipient of the data is appropriately credentialed, and other granular decisions of this nature. A Security Committee may oversee data requests and releases from the viewpoint of technical security or, more likely, from the viewpoint of business sensitivity.

The organization of the infrastructure may also entail the creation of certain particular roles. Some of the functions of the narrow "data governance committee" just described may be delegated to an individual role, often designated the "honest broker." Sometimes honest brokers are appointed through a formal IRB process, but in many places the designation is ad hoc. Honest brokers are most frequently associated with project- or program-specific repositories, as in the case of PCORnet data marts or where a so-called "pluripotent" database, i.e., one capable of supporting many projects, is established.

Most institutions with a developed health IT infrastructure are able to differentiate a number of different components: a transactional system, used by providers and

administrators to record patient-related data, often based on an encounter or on a report from a lab, from pathology, or radiology. This will generally tend to be reorganized directly or overnight into a well-structured database. This may be good enough to serve as the "single source of truth" but is often so exquisitely normalized that its navigation is extremely laborious, and so queries run very inefficiently. Thus there is a need for a data warehouse, i.e., a collection of denormalized, flat, materialized views on the data, the so-called data marts. These can be searched very efficiently either through a standard query language or through a specialized tool, thus making it accessible to non-expert programmers. Finally, research needs may be addressed in a variety of ways. At some institutions, the same enterprise data warehouse also serves for research but is therefore highly restrictively controlled. Elsewhere, a de-identified copy of the data warehouse is created especially for research. Here access is less restricted, but access to patients for consent requires additional steps. Finally, there are several national projects which have promulgated specific data models which must be adopted in order to participate. These include the PCORnet Common Data Model, OHDSI/OMOP mandated by the All_of_Us "precision medicine" project, and i2b2 adopted by several CTSAs and other collaborations. In addition to these are the numerous ad hoc collaborations which succeed in data sharing though data sharing and data use agreements. In all these cases, some governance mechanism is deployed to ensure ethical, data release, and security approvals are obtained.

Implementation: An Effective Data Governance Structure

The design, deployment, and maintenance of an effective data governance program is a major undertaking. Addressing all the relevant issues in an enterprise-wide set of structures and processes requires an in-depth understanding of concepts and requirements from so many domains that it is almost always best left to a team with diverse backgrounds, each expert in his or her domain. There are rare professionals who have specialized in this area and whose expertise is highly valued. An informatician charged with implementing a data governance program, especially one in healthcare, would be well served by a comprehensive guide book and a team of knowledgeable fellow professionals who can cover the specialist topics: knowledge of legal aspects of data protection, knowledge of security frameworks, and knowledge of the enterprise and its culture, not least the often competing constituencies within a single enterprise. There is a choice of guidebooks on data governance.[7]

[7] John Ladley. *Data Governance: How to Design, Deploy, and Sustain an Effective Data Governance Program*. Morgan Kaufmann, 2012. A readable, comprehensive guide to the broad spectrum of data governance—recommended.

David Plotkin. *Data Stewardship: An Actionable Guide to Effective Data Management and Data Governance*. Morgan Kaufmann, 2013. Puts the onus for data governance on data stewards; this may be somewhat narrow for healthcare institutions.

Helmut Schindlwick. *IT Governance: How to Reduce Costs and Improve Data Quality through the*

The building blocks of an effective strategy: Case Study

In a report to the AAMC conference on Information Technology in Academic Medicine in 2016 and again in an AMIA CRI-WG Webinar [11], a university medical center team reported that when they began work on the creation of a data warehouse without a parallel data governance effort, they were hampered by a number of problems. These were in essence the common problems that have led to the establishment of data governance structures and processes in many organizations, reflecting both the need to protect the data from error, redundancy, and inconsistency, as well as to defend the data from accidental or malicious disclosure. Crucial headline issues they identified included ill-defined responsibility and ownership of data, along with a lack of standards and consequent mistrust of the data; they also found data replicated across multiple silos, with inconsistent integration, and noted that there was no enterprise-wide data quality audit, so that errors were not systematically traced back to their origin; and there was no information life cycle management. They were also troubled to find little understanding of the data across business lines—the clinical enterprise, the research enterprise, the academic/student enterprise, and even the finance enterprise. Thus they were persuaded of the need for a data governance process. More accurately, they understood that their need for consistent, reliable data across all business units led inexorably to the initiation of a broad data and information governance program that would address consistency in the management and use of institutional data, as well as transparency on its provenance and semantics. Moreover, it would also lead to better performance in responding to users and improved business analytics. This section of the chapter relies heavily on the experience of this team and its very well-laid out history of the development of its data governance program.

The team adopted Gartner's definition [12] of information governance as "the specification of *decision rights* and an *accountability framework* to ensure appropriate behavior in the valuation, creation, storage, use, archiving and deletion of information. It includes the *processes*, roles and policies, *standards* and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals" (emphasis added) [@REF6]. Basing their program on this definition, they addressed the first component and determined questions of accountability and specified ownership, roles, and responsibilities. They then engaged key stakeholders across the institution to ensure that decisions would be adhered to. Their second focus was on standards: they specified expected data quality standards and a pragmatic margin of tolerance. Their standards addressed information consistency, the models of the data and their contexts, protection, and the life cycle of the data to ensure liveness while retaining a manageable volume. Finally, in what they saw as the most important component, they turned to processes: decision-making guidelines and protocols, agreement on escalation process for decision resolution,

Implementation of IT Governance. CreateSpace, 2017. Highly recommended by some business leaders, it seems to restrict its purview to IT-related matters.

Robert S. Seiner. *Non-Invasive Data Governance*. Technics Publications, 2014. Appears rather more authoritarian than its title may suggest.

communication and workflows, and change management. The team asked three questions: *What* decisions need to be made? *Who* makes them? *How* are they made? These questions focused the team's attention on data as an enterprise asset and that it is worthwhile investing in its stewardship.

Focusing these questions on particular domains, four basic domains were identified: data, metrics, tools, and funding. In the case of data, a number of decisions had to be made: which is the system of record for source data? What is the tolerance threshold for different types of data—patient counts may need to be accurate plus or minus N, perhaps, but financial data must be as accurate as possible. What data transformations are allowed, and what relationships must be preserved? What access approvals are required, and who is authorized to grant such approvals? If, as is the case in many academic medical centers, there are multiple coexisting enterprises—clinical, educational, research, business—how is consistency maintained between them? In this particular case, the local decision grants the data steward at the source continuing stewardship of those particular data as it migrates, e.g., to the data warehouse.

Turning to values and metrics, it is necessary to pay attention to different ways of defining units in different business areas: a faculty "FTE" (full-time equivalent) in academics may not be the same as a faculty FTE in clinical; dates and times of events is another well-known area of divergent definitions. There are data benchmarks, both internal and external; again, a choice has to be made on who will be responsible for maintaining these. In the present case study, the relevant source data steward retains this responsibility and so ensures continuity. This responsibility stays with the steward for that element of data right up to when it contributes to a dashboard report to management. For the last two domains, tools and finance, in this case study, the recommendation is, first, to make sure that technical professionals are involved in all tool choice decisions and that business management is on board when there is likely to be a need for funding.

Drilling down into greater detail, the team created a "decision matrix" with a horizontal axis of the four domains (data, metrics, infrastructure and tools, infrastructure funding), each broken down further by the enterprise area (system-wide, education, research, clinical, faculty) so that there are 20 columns in all. The vertical axis represents the data stewards and possible decision-makers in the organization: some c-suite executives with informatics or operational responsibilities, deans, associate vice-presidents with relevant portfolios, etc. In each box in the matrix, an entry identifies members, decision-makers, veto-holders, and information providers, and those must be informed of any relevant decision. This tool provides the medium of negotiation of roles and determination of who should be the data steward for each element. In reality, each data element requires attention in this way, so the process has to break down responsibilities at least one more time to get to a clear determination of who has ownership of what. Indeed, in conclusion, the team has observed that there are three rings of data, the inner ring of master data which is shared across all business areas and has to be governed collectively; the middle ring of shared application data which may belong to one functional area and governed locally; and finally, the outer ring of single application data, managed by the small number of concerned individuals. A sophisticated approach quantifies

responsibilities for data elements and so assigns the role appropriately. Master data is determined by exclusion as well as by inclusion: certain data elements may be useful or important, but they may not be "master data" because they change frequently or relate to specific attributes.

Acknowledgments In addition to the members of the AMIA CRI-WG, I must acknowledge a number of sources. The section on "Defense of Data" has benefited greatly from the American Statistical Association's Committee on Privacy and Confidentiality and its comparison of the HIPAA Privacy Rule and the Common Rule [13]. The section on roles owes a great deal to the paper by Sanchez Pinto et al. [14] and in particular to the three CRIOs who spoke at the workshop from which the paper was developed, Bill Barnett, Peter Embi, and Umberto Tachinardi. Also fellow panelists at AMIA Summit 2018, Harold Lehmann, Kate Fultz Hollis, Bill Hersh, Jihad Obeid, Megan Singleton, and Umberto Tachinardi. The work of John Holmes [15–17] was also influential. The implementation section benefited from Adam Tobias and colleagues' work at USF [11]. Of course, none of these authors bears any responsibility for errors or misunderstandings that may have crept into this chapter.

References

1. Donabedian A. Evaluating the quality of medical care. Milbank Q. 2005;83(4):691–729. Reprinted from The Milbank Memorial Fund Quarterly 44:3.2;166-203 (1966)
2. AHIMA. Information Governance Principles for Healthcare (IGPHC). Available at: www.ahima.org/~/media/AHIMA/Files/HIM-Trends/IG_Principles.ashx.
3. Martin PY, Turner BA. Grounded theory and organizational research. J Appl Behav Sci. 1986;22(2):141.
4. Fahey L, Prusak L. The eleven deadliest sins of knowledge management. Calif Manag Rev. 1998;40(3):265–76. ("Error 3"). This precise formulation was given—repeated twice for emphasis—at a HICSS2000 keynote.
5. Her QL, Malenfant JM, Malek S, Vilk Y, Young J, Li L, Brown J, Toh S. A query workflow design to perform automatable distributed regression analysis in large distributed data networks. eGEMs. 2018;6(1):1–11.
6. Health Insurance Portability and Accountability Act of 1996. Public Law 104–191. US Government Publishing Office. 1996. Available at: https://www.gpo.gov/fdsys/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf
7. Whitman JQ. The two western cultures of privacy: dignity versus liberty. Yale Law J. 2004;113:1151–221. Available as Faculty Scholarship Series, Paper 649 at http://digitalcommons.law.yale.edu/fss_papers/649
8. Warren SD, Brandeis LD. The right to privacy. Harv Law Rev. 1890;4(5):193–220.
9. Viktor Mayer-Schonberger. Beyond privacy beyond rights – toward a systems theory of information governance. Calif Law Rev, 98:1853–1885 (2010). Available at http://scholarship.law.berkeley.edu/californialawreview/vol98/iss6/4.
10. Laudon KC. Markets and privacy. Commun ACM. 39, 9:92–104
11. Tobias A, Chackravarthy S, Fernandes S, Strobbe J AAMC Conference on Information Technology in Academic Medicine, Toronto, June 2016; also presented as an AMIA CRI-WG Webinar, October 2016.
12. https://www.gartner.com/it-glossary/information-governance.
13. American Statistical Association. Committee on privacy and confidentiality. Comparison of HIPAA Privacy Rule and The Common Rule for the Protection of Human Subjects in Research. 2011.

14. Sanchez-Pinto LN, Mosa ASM, Fultz-Hollis K, Tachinardi U, Barnett WK, Embi PJ. The emerging role of the chief research informatics officer in academic health centers. Appl Clin Informat. 2017;8(3):845–53.
15. Brown JS, Holmes JH, Shah K, et al. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010;48(6., Supplement 1: Comparative Effectiveness Research: Emerging Methods and Policy Applications):S45–51.
16. Holmes JH, Elliott TE, Brown JS, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. JAMIA. 2014;21:730–6.
17. Maro JC, Platt R, Holmes JH, et al. Design of a national distributed health data network. Ann Intern Med. 2009;151:341–4.