第 2 章
数据、人员和系统

“只要联盟仍可通过剑和刺刀维持，并且冲突和内战取代了兄弟之爱和仁慈，它对我就没有吸引力。”

——罗伯特·李将军[⊖]（1861）

“有些人喜欢工作中不断的纷争，
喜欢整天钩心斗角，
但这并不是我的人生梦想，
我觉得这样的人是疯了。”

——詹姆斯·韦尔登·约翰逊
《Lazy》（1917）

2.1 目的

本章定义了一组与数据和数据管理相关的角色。很多数据质量的词汇都起源于制造业的质量控制流程。把数据作为产品的中心比喻用类似于制造业中使用的术语描述有关数据的角色：**数据生产者、数据消费者、数据代理、数据管理员、数据质量项目和利益相关者**（Strong、Lee 和 Wang，1997 年）。在许多组织中，个体扮演多种角色。同一个人可以在一种情况下充当数据生产者，而在另一种情况下充当数据消费者。同样重要的是，充当这些角色的并不总是人类。由于系统也生产和消费数据，我们也将很简单地探讨一些关于它们与质量之间的关系的一般考虑。本章也将解决与数据所有权有关，并有时在承担信息技术（IT）角色的人与那些承担业务（非技术）角色的人之间产生的具有挑战性的关系问题。

2.2 企业或组织

在此处，术语**企业**（enterprise）和**组织**（organization）可互换使用，指的都是生产和使用数据，以实现他们的目标的企业或其他组织。在当今世界，这包括几乎所有或大或小的公司、非营利组织、教育机构，以及政府部门及其下属部门。我倾向于把**组织**这个术语作为更加通用的词。当确实要表示一个特定组织的全面视图时，用**企业**应该比用组织更合适。但是，人们使用**企业**这个术语时，他们往往是指一个组织的子部门（通常是自己所在的子

⊖ 罗伯特·李（Robert E. Lee）将军是美国内战中南方军主帅。——译者注

部门）。

由于数据跨越组织边界，提高数据质量就需要企业的全盘考虑。这项工作很有挑战性，因为当今的许多组织都是庞大而复杂的。

2.3 IT 与业务

大多数组织都对技术和非技术职能和部门加以区分。信息技术（IT）是指任何充当一种技术角色的人（包括数据库管理员、程序员，基于 IT 的业务分析师和信息系统经理）。术语**业务人员**（the business）应该是指那些做的工作直接关系到组织业务目标的实现的人，而不是那些配角，如在一个专门负责开发和维护技术系统的部门中工作的人。在实践中，**业务人员**经常被用来作为任何不直接在一个组织的 IT 部门工作的人的 IT- 称呼。[⊖]

IT 和业务之间的关系是许多组织的紧张关系的根源，特别是关系到数据管理时。这种紧张关系往往表现在数据质量的定义，以及谁对数据质量负责的问题上（Klein 和 Callahan，2007 年；Pipino、Pipino、Lee 和 Wang，2002 年）。

当数据被认为对大多数组织的成功至关重要，并且 IT 负责管理数据和容纳它的系统时，IT 往往被认为（并经常自认为）与业务相分离。这种说法好比是把循环系统看成与身体相分离。是的，虽然循环系统本身是一个东西，它的零部件可以理解为它本身的零部件。但它不能在身体之外存在，而没有它身体也不能生存。更重要的是，其存在的根本目的（保持血液在身体中流动）对于身体是不可分割的。

数据管理协会（DAMA）称，数据管理整体应该是一个共同的责任（DAMA-BOK，2009 年）。但是，许多企业疲于应对如何使它成为一个整体。对于数据质量的问题，大多数专家都认为，需要业务人员（而不是 IT 人员）来定义什么是高质量的数据。这种想法在措辞方面往往采用**所有权**（ownership）一词，如“企业应该拥有数据”。不过，业务人员需要有人帮助他们来改善存储在系统内的数据的质量。IT 人员负责这些系统。信息系统和它们所包含的数据对于当今的组织运营是不可或缺的。IT 之所以存在，是因为企业需要技术来运营。IT 需要把自己看成与所要支持的业务流程有更紧密的关系。该关系包括对数据内容具有更好的理解，以保证更高水平的数据质量。

从哲学上讲，似乎有理由问，“IT 不是总是一个大型企业的一部分吗？在企业的所有部分不应该为更大的利益共同工作吗？”答案当然会是“是”。但是，IT 和业务之间的关系，经常由涉及预算和其他资源的决定的组织路线形成。这样的决定可能是高度政治化的，并且成为提高数据质量的障碍。在《Data Driven》一书中，Thomas Redman 指出了数据和信息资产的有效管理的 12 个障碍。在这些障碍中，包括数据共享和数据所有权的政治，管理和数据流的错位，以及把数据和信息管理混入技术管理（Redman，2009 年，第 161 页）。[⊜]

测量的实用性也需要被考虑在内。虽然有关数据的期望应该来自数据消费者，但有时很难让消费者从自己制定测量的角度明确表达这些期望。这种困难有时会导致 IT 不应该或不能测量数据质量的说法。Matthew West 博士指出：“信息与其他产品的区别是，存在确定其

⊖ IT 与业务之间的界限是模糊的。我自己的数据质量架构师角色就是一个例子。我通过一个 IT 组织来报告工作。被我清楚地看作业务人员的人认为我是“技术人员”。但是，在我自己的报告组织结构中，更直接的技术人员把我看作业务人员（the business）。甚至在技术性和非技术性的角色内部，也有更多的差别出现，例如，数据人员（data people）和其他人员（other people）之间的差别。

⊜ Redman（2009），第 7 章。

质量的特定属性，这些属性独立于数据所表示的东西”（West，2003 年，第 6 页）。那些他确定的属性有准确性、及时性、完整性和一致性。他进一步指出，“我们在 IT 和信息管理中做的所有事情都与提供信息质量有关”（West，2003 年，第 7 页），这句话强调了一个事实，即数据生产者与数据消费者共同承担理解他们的产品质量的责任。在依赖数据处理的数据质量维度的情况下，IT 远比数据的消费者更能够测量数据质量。在这种情况下，关键是数据消费者要明白为什么测量到位使他们对数据有基本的信心。但同样重要的是，IT 要在生产过程中承担数据管理的责任。

在这本书中引用的 IT 和**业务人员**将对担任技术角色的人员和那些非技术角色的人员加以区别。为了提高数据质量，技术和业务专长二者缺一不可。最相关的问题不是“谁该负责数据质量，IT 还是业务人员？”，而应该是“业务人员在哪些方面有助于提高数据质量，而 IT 要在哪些方面做出自己的贡献？”

下面定义的术语——**数据生产者、数据消费者、数据代理**被用来描述与数据不同的关系。从这个意义上说，它们指向跨部门的职能，并且最好在一个数据链或信息的生命周期中理解它们。相对于一个数据集的数据消费者可被视为相对于另一个数据集的数据生产者。

2.4　数据生产者

数据生产者是创建数据的人和系统。数据生产者可能为了使数据可供使用而明确地创建数据，或者它们也可能作为另一个过程的副产品而产生数据。尽管生产者能控制他们创建的东西，但他们不能控制其数据的可能用途。他们可能为一种目的创建数据，但该数据随后被用于其他用途。

数据生产者的一个重要子集是创建数据的流程的业务所有者。无论数据是否被立即消费或输送到下游流程，业务流程所有者都是信息链中的关键环节。他们对自己的流程的目的和功能有认识。他们可以修改流程，以确保它们产生的数据更好地满足数据消费者的需求。

2.5　数据消费者

数据消费者是在信息生命周期的任何时刻使用数据的人员与系统。我不喜欢**数据消费者**这个术语，因为它似乎忽略数据的一个重要特点，即数据实际上不是采取使用其他资产的方式“消费”的。多个人或系统可以使用同一数据集而不妨碍其他人这样做。然而，该术语适用于数据是在信息生命周期中由数据生产者创建并由数据的消费者使用（应用）的理念。我宁愿用**数据消费者**（data consumer）来替代术语**用户**（user）、**最终用户**（end user）和**客户**（customer）。因为在引用数据客户和企业的其他客户时会产生混乱。我还没有把术语**知识工作者**（knowledge worker）作为一个通用术语来采用，因为知识工作者只是数据消费者的一个子集（并非所有消费数据的人都是知识工作者），并且因为**数据消费者**的概念既包括人员又包括系统。

2.6　数据代理

数据代理（data broker）[⊖]的概念有时在生产者 / 消费者模式中被忽略了。**数据代理**是数

⊖　感谢 Jim Locke 对数据中间人的评论。Gartner 还使用术语数据处理程序（data handlers）来探讨一种角色，它与确保当数据在系统与应用之间移动时仍然是安全的相关联。

据管理的中间人。虽然他们不产生数据，但他们使得他人能够使用它。他们类似于制成品的分销商。关键是要认识到代理会影响数据的质量，因为他们是信息链的一部分，并且可以影响数据内容和数据的结构形式，及其可用性和及时性。

2.7 数据管家和数据管家工作

术语**管家工作**（stewardship）被定义为“对他人财产的管理（management）或照看（care）”（NOAD）。数据管理员是负责数据的照看和管理的个人。虽然**数据管家**（data steward）和**数据托管人**（data custodian）基本上是同义词，但最好采用**数据管家**这个术语，因为对许多人而言，术语**托管人**（custodian）是**看门人**（janitor）的代名词，而不是简单地意味着照看别人的东西的某人。理想情况下，管家工作意味着承担的一种角色——这个术语的“管理和照看”部分。良好的管理工作涉及对数据内容，以及对数据在组织中的作用、用途和价值的理解。管理员应该了解数据，帮助他人了解和使用它，并支持改善数据的状况。

DAMA 把管家工作定义为纯粹的业务角色。虽然数据管理专业人士是“数据资产的保管人和技术保管人……数据管家是数据管理中分配的业务职责的问责机制”（DAMA，2009 年，第 5 页），但在业务**管家**和 IT **托管人**或**保管人**之间划清界限基本没有好处。IT 具有一个管家的工作角色。此角色包括纯粹的数据管理（management）职能，如数据库管理（administration），但它也承载着对数据内容和用途至少有一个工作认识的义务。DQAF 被研制出来的一个原因就是帮助说明 IT 的管家工作角色的有关细节。

综上所述，人们希望有另一个词来指代这个照看和管理的角色，这是因为，管家工作的概念还未被很好地理解。人们不知道究竟如何成为好的管家，而大多数组织都没有有效地授权的管家。根据企业的性质，数据管家几乎总是还承担着其他角色。他们可能是主要的生产者、消费者，或数据质量项目团队的一部分。他们可能是一个正式的数据治理结构的一部分，或者可能仅仅是试图使其组织的数据更可靠且更可理解的善良而勤奋的人。理想的情况下，与给定的企业的成功有利害关系的所有人都应该是一个企业的数据的好管家，就像他们应该是属于一个组织的所有其他资产，从曲别针到部门预算及时间。

2.8 数据所有者

管理员照看其他人的财产，因此询问谁是数据所有者是合乎逻辑的。**数据所有者**（data owner）的概念对于大多数组织都是具有挑战性的。数据不是有形的，它并不总是被理解为一种资产。当它被确认为一项资产时，它通常被认为是一个组织的资产，但是这种看法可能最终意味着没有人照看它。大多数其他组织的资产本身并没有所有者。设备管理部可能负责一个物理设备，但不拥有它。同样，会计部负责总账，但也不拥有它。资产是由组织本身所拥有的。数据和其他资产之间的差别在于，其他资产存在明确的问责（而不是所有权），它通常驻留在组织内的一个部门或功能中。与此相反，数据是在部门和职能（Redman，2008）之间移动的。虽然 IT 显然要负责驻留数据的系统，但 IT 历来否认承担在这些系统中的数据的责任。

识别数据**所有者**（owner）的愿望可以被看作一个解决问题的愿望，特别是由于组织的数据不符合其需要的事实或看法。当人们认为其组织的数据不充分、不一致或混乱时，他们就会寻找一个所有者以将数据置于更好的控制之下。正如组织的其他资产一样，数据也需要进行管理。管理包括知道组织拥有什么数据，使用数据来推进组织目标的实

现，并尽量减少与数据使用相关的任何风险。数据可能会很难管理，因为它不仅是无形的，而且还很容易复制，还因为许多组织没有在管理数据和管理驻留数据的系统之间划清界限。

信息技术人员与业务人员之间的冲突关系往往使得数据和系统之间的边界甚至更加模糊。因为人们可以清楚地看到 IT 系统的成本，但并不总是能看到自己组织的数据所带来的好处，做出有关数据管理决策时，平衡成本和收益会是相当困难的。一个自然的结论是：只要我们有一个数据的拥有者，那么所有的问题都将得到解答！而管理数据将会很容易。不幸的是，数据管理的复杂挑战没有一个简单的解决方案。不过，也有有效的办法。其中之一是在整个信息生命周期中对组织内的数据有明确的问责制。如果组织把这种问责命名为**数据所有权**（data ownership）是有益的，那么他们就应该这么做。

2.9　数据所有权和数据治理

数据治理研究所的 Gwen Thomas 指出，数据治理的目标包括实现更好的决策、降低运营摩擦和保护数据的利益相关者的需求，所有这些都是为了提高企业的整体效能。数据治理通过实现可重复的流程、通用的方法和协调工作——换句话说，通过简化组织必须完成的工作（Thomas，2011），也应有助于提高运营效率。数据治理的工作方式应该像很多其他形式的治理一样。但是，对于数据管理和数据治理，大多数机构仍然处在政治理论阶段。没有就如何落实治理理念达成一致，有的人担心其影响。与把数据治理理解为降低风险的途径相反，人们往往将它反倒作为另一种形式的风险。当治理从顶层规定而没有对现行做法进行估值时，它可能就像一组局外人在可能工作得很好地满足局部目标的流程上面强加一个不切实际的模型，而原来的流程如果进行一些调整甚至有可能促进整个企业的战略。当数据治理是强加的时，就无法获得员工的全力支持，并会滋生不信任，因为目前尚不清楚谁将会从中受益，谁会付出代价，或者新管理体制会是什么样子。⊖

换句话说，数据治理可能会煽动权力斗争。正如 Thomas Redman 指出的，“数据和信息引起了热烈的激情和残酷的政治。有经验的人知道，阻碍组织更好地管理和利用其数据和信息资产的努力不是硬性的技术问题，而是组织、政治和社会方面的软问题”（Redman，2008 年，第 159~160 页）。

2.10　IT，业务和数据所有者，终极版

Redman 的观察使我们又回到了 IT 和业务之间关系的问题：在一个基本水平上，数据治理的努力尝试调解这种关系。调解不是一件简单的事情，因为无论是业务还是 IT 都是复杂的实体，它们都应该是向着共同目标执行和协调多种职能的组织，但情况往往不是这样，它们总是在争夺有限的资源（这意味着数据治理也需要在不同的业务部门之间，以及业务和 IT 部门之间斡旋）。数据所有权成为争议的焦点，因为不同的利益相关者对数据具有不同的目标，并因此有不同的方式来对数据进行控制和决策。David Loshin 在他讨论数据估值的不同概念、数据和隐私、争夺地盘、恐惧，以及官僚主义的关系时，总结了使数据所有权的讨论复杂化的因素（Loshin，2001 年，第 28 ～ 30 页）。

Loshin 还通过把数据的所有权的问题分解成可管理的部分，并让我们看到各部分如何相

⊖　数据治理的风险是我喜欢 Robert Seiner 的“非侵入式”方法的原因之一。

互关联来提供了一个符合逻辑和常识的方法。他指出，“所有权的问题实质上是一个控制的问题，控制信息流、信息的成本、信息的价值”（Loshin，2001 年，第 28 页）。他概述了在创建信息时扮演不同角色的人。“在信息工厂中的参与者”包括数据供应者、收购者、创作者、加工者、包装者、配送代理、消费者，以及保证人们都在做自己的工作、系统在正常运转，并对该组织的总体目标和战略进行决策支持的不同层次的管理者（第 26 页）。随后，他详细说明了数据所有权的责任。这些责任包括定义数据、保护数据安全、使数据能够访问、支持终端用户社区、包装数据、维护数据、测量数据的质量，以及管理业务规则、元数据和标准，并对数据的供应者进行管理（第 31 ～ 33 页）。

鉴于这些角色和职责，Loshin 描述了所有权本身具有的不同范式。每个范式都是基于价值与信息相关联或从中得出的方式。数据可以由它的创建者、消费者、编制者、企业、出资的组织、解码器、打包者、阅读者、主体、购买者 / 执证者所拥有，或每个人都是所有者。Loshin 的模型识别的是一组不同的与数据之间的关系，每个关系都包含一定程度的控制、决策责任，或针对数据的其他权力的现实。如果不使用**治理**（governance）这个词，他建议创建一个数据所有权的策略来阐明不同的关系，明确职责，并提供解决利益相关者之间的争端的标准（第 38 ～ 40 页）。他的做法如此透彻简单，以至于人们会对组织中仍在进行争夺所有权的战斗感到惊讶。但如前所述，许多组织仍然处在数据治理的政治理论阶段。地盘、恐惧和官僚主义继续盖过了常识。

之所以业务人员渴望拥有数据“所有权”，是因为他们能从控制数据中获益，虽然这个说法简单，但基本正确。但是，如果他们不“拥有”处理和驻留数据的系统，那么他们就没有对数据的控制感。IT 人员否认数据的所有权，因为他们不控制数据内容，如果数据内容不能满足业务需求，他们就不希望被追究责任。与此同时，他们负责处理和驻留数据的系统，并且，由于这一事实，IT 在很大程度上控制了任何给定系统中的数据，以及系统之间的数据移动。而数据的这种水平移动，使数据所有权更加靠不住。当数据沿着数据链移动时，它可以被复制并转化成另一种数据集，潜在的“拥有”，即由并非其原始系统的其他系统，或原来不负责它的其他团队控制（Redman，1996 年，第 235 页）。所有权和治理的问题被分解成两个相互关联的问题：控制与决策。数据质量测量的优点之一是，它提供了一种有关特定类型数据的知识，这使得组织能够了解他们的数据是否已受控制。受控制的数据更容易被信任，并且不太可能成为政治斗争的对象。

2.11 数据质量项目组

术语**数据质量项目组**（data quality program team，简称 **DQ 组**）[⊖]是指正式承担数据质量活动的人员。数据质量活动包括数据评估和测量、数据质量问题管理，以及一系列推动改进数据，并有助于创造一种用于生产高质量数据的组织文化的战略和战术活动。在这个意义上说，DQ 组成员始终发挥着管理作用，但并非所有的数据管理员都是 DQ 组的一部分。在最好的情况下，DQ 项目组成员都在注重实施数据质量战略，无论是直接作为动手的分析师，还是间接作为数据质量方法的传播者和主题专家。并非所有的组织都有一个数据质量项目，也并非所有实现这些目标的工作组都将在其名称中包含**数据质量**这个词。然而，为了使数据

⊖ Redman（1994）使用术语数据保管人（data keeper）来指“负责实施给定的数据库的多个共同的数据质量功能的人员。例如，数据保管人应管理编辑活动”（第 292 页）。

质量管理有显著的进展，最好是指定一些个人来领导这些工作。[⊖]

2.12　利益相关者

术语**利益相关者**（stakeholder）是指受组织的数据质量影响的个人、团队和角色的集合。利益相关者包括这里提到的所有人：来自业务和技术团队的直接和间接的数据生产者、消费者、中间人和管家，数据质量项目组成员，以及公司管理层和所有者。从前面的讨论中，我们了解到，利益相关者与数据具有广泛的关系。当利益相关者相互作用时，尤其是当他们做出有关数据的决策时，他们应该知道对方的角色并了解对方的观点。

2.13　系统和系统设计

如前所述，系统也产生并消费数据。《新牛津美语辞典》把**系统**（system）定义为“形成一个复杂整体的一组相互关联的事物或部件”。更抽象地，系统是“一套据以做某些事的原则和程序，一个有组织的方案或方法”（NOAD）。质量先锋 W. Edwards Deming 把系统定义为“协同工作，试图实现系统的目标的相互依存的组件组成的网络。系统必须有一个目标。没有目标，就没有系统。系统的目标必须是在系统中的每一份子都很清楚的”（1994，第 50 页）。Deming 的表述强调，系统必须是有目标的。如果不是，它们就只是组件的一个集合。

技术系统是为了满足特定的目的而创建的，有时也被称为系统的应用程序。任何人类创建的系统都基于一组可以称为**模型**（model）的假设，它是规定系统应该做什么，应该如何做的一种范式（Weinberg，2001 年）。如我们将在第 3 章讨论的，模型总是简化的，因此依赖某个系统可能导致会丢失一些东西的风险。系统的成功，部分取决于我们对它有多么深思熟虑，如果系统不具有一个坚实的详细模型来负责它将被应用的方式，系统就可能无法满足其目标。

Kristo Ivanov，在他 1972 年对信息质量的讨论中，解决了系统模型不充分带来的挑战。他指出，需要将系统与它与之交互的世界之间（即一个系统必须考虑的真实世界中的对象和系统考虑它们的方式之间）保持对应和一致。当某个信息系统不符合这些条件时，改变系统常常比“迫使现实来适应模型”更有挑战性（Ivanov，1972 年）。因此，负责将信息输入该系统的人会变通地解决该系统的限制，以保持工作能进行下去。变通导致在系统中的信息与它假定来表示的对象、事件和概念是不符合的。

某个系统模型的其他不足也可能会导致产生质量低劣的信息。Ivanov 表示，通过这样的变通创建的信息是不准确的，这要么是将它输入系统的人造成的，要么是系统不允许他们以任何其他方式来考虑它的缘故。他指出，即使是看似简单的问题（真实的值是什么呢？），也不能在特定系统内定义的上下文之外得到回答（Ivanov，1972 年）。他的观点是，糟糕的系统设计（一个不充分的模型）造成的低质量的信息可能会与人为因素，如数据录入错误，同样多。他甚至认为，用错误率来表达的信息质量“可能是系统设计或模型是否充分的一个重要指标。到现在为止，这一直被视为编码 [数据输入] 和观察过程本身的主要指标”（Ivanov，1972 年）。本书的配套网站更充分地总结了 Ivanov 的思路。现在，我想强调他的基本看法：系统是否充分直接影响数据质量。系统的不充分性可能表现在其目标方面（系统的目标可能

⊖　参见 Pierce（2003 年），Redman（1994 年，2007 年），English（2008 年），以及 Yonke、Walenta 和 Talburt（2011 年）。

和组织的目标不一致），也可能表现在其功能方面（功能可能丢失或形式不正确）。

2.14 总结思考

生活在信息时代，我们都同意，电子数据对大多数组织的工作至关重要，但我们仍在努力，以满足生产高质量的数据的目标。业务 /IT 鸿沟使得能以满足不断变化的业务需求的方式处理数据的系统很难建立。但是，如果我们能够更好地表达和记录有关数据的期望，并且拥有一个生产出更高质量的数据的共同目标，那么就可以做出有助于生产更高质量的数据的组织决定。与其他产品的情况一样，我们应该能够提高产生、收集、处理、存储数据，并使数据可用的系统的质量。

业务和技术团队都对数据质量承担共同的责任，而数据质量测量则为成功地履行这一职责提供了一种手段。他们有共同的责任，只是因为两者都是同一个组织的一部分，而更重要的是，因为对于大多数企业来说，组织的成功依赖于高质量的数据。实现一定水平的合作是不容易的，在不同的组织中，它不会按照相同的模型实现。但正如在建立任何模型时那样，企业需要有意识地就他们如何制定其组成部分之间的关系作决定。