第1章

绪论

1.1 引言

信息技术的发展使个人数据管理问题日益突出，并日益引起数据库、信息检索、人机交互等多个领域的专家学者和产业界的广泛关注。据IDC统计，2006年全球新产生的数据量达到1610亿GB，2007年达到2810亿GB[1]，2010年达到12 000亿GB(1.2ZB)[2]，信息的爆炸性增长使人们日常需要处理的信息量迅速增长。微软公司的研究员Gordon Bell从2000年开始收集个人数据信息，包括阅读的文章、听过的音乐、建立的文档、访问的网页、个人医疗信息以及拍摄的照片等，到2007年其收集的个人信息量已经达到150GB[2]。个人信息量的迅猛增长使人们管理个人信息的负担日益加重。据IDC调查[3]，美国从事信息工作的人员平均每个月在个人信息管理方面浪费的时间约20小时。近年来随着Web 2.0、物联网、移动互联网等技术以及移动通信设备的发展，人们产生信息的方式更加多样化，人们的各种信息也都可以更为容易地记录下来，个人数据量会进一步增长，未来个人数据管理问题将更加突出。

1.2 个人数据管理的发展

实际上，当有个人信息出现的时候，就出现了如何有效管理个人信息的问题。在古代，个人信息主要以文字符号的形式存在，记录或存储的介质最初为树皮、贝壳等，后来发展到用纸张记录信息。在这种情况下个人的信息量还不大，个人信息管理的问题也不突出。随着电子技术的发展，信息的产生方式和存储方式都有了很大变化，信息量不断增加，信息管理问题也日益突出。据文献记载，最早提出个人信息管理(Personal Information Management，PIM)这一概念的是美国科学家万尼瓦尔·布什(Vannevar Bush)，他在 1945 年发表的《As We May Think》[4]一文中构想了一种能够帮助人们管理个人信息的工具 Memex，如图 1-1 所示，并将其描述为一种能够记录书籍、唱片等信息并能帮助人们快速查找所需信息的工具。

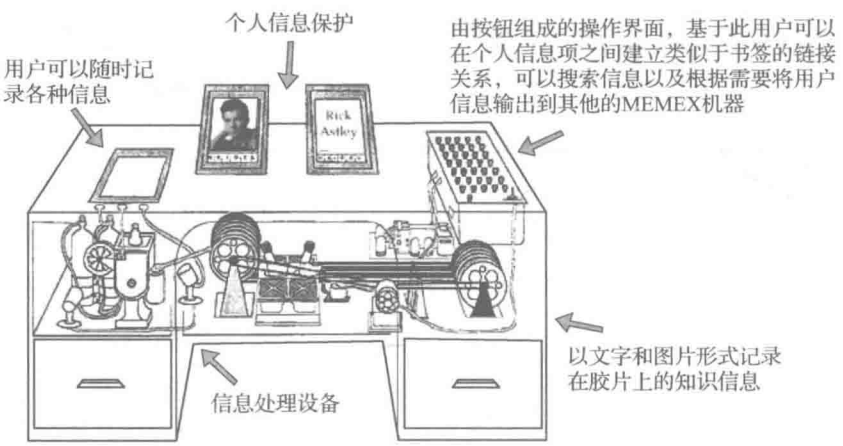

图 1-1 万尼瓦尔·布什提出的个人信息管理工具模型

万尼瓦尔·布什想象了一种如图 1-1 所示的能够帮助人们管理信息的工具 Memex，并对其进行了这样的描述：Memex 是一种能够记录所有书籍、唱片、交流信息的工具，它能够快速、自动、灵活地帮助人们

查找所需要的信息。布什只是为个人信息管理进行了一个形象化描述，随着信息科学技术的发展，一些学者从不同视角对 PIM 给出了定义：①PIM 是人们对于日常信息的处理、分类和访问(Lansdale，1988)；②PIM 是为用户创建的供其在工作环境中使用的系统，其包含获取信息的规则与方法、对信息进行组织与存储的机制、维持系统运行的一些规则与过程，以及对信息进行访问、处理、产生输出的方法和机制(Barreau，1995)；③PIM 的目的是存储信息以使其能够在以后被访问(Boardman，2004)。

由以上定义可以看出，PIM 的定义与信息技术的发展有密切关系，Lansdale 只是对 PIM 给出了一个宏观的描述；Barreau 指出 PIM 中应包含获取信息的规则、方法，以及存储信息的策略、机制；到 2004 年，Web 技术的成熟和存储技术的发展，使海量信息数据的存储成为可能，Boardman 认为 PIM 的核心是数据的存储和再访问。这些关于 PIM 的描述成为进一步研究、定义 PIM 的基础。

2005 年，在美国西雅图举办了第一届国际个人信息管理技术研讨会(PIM Workshop 2005)，来自世界各地的专家学者对 PIM 研究中的一些基本概念、基本的科学问题及其挑战性等进行了研讨，提交了一份研究报告[5]。在这份报告中，对个人信息空间、个人信息管理等基本概念，PIM 研究内容、面临的机遇与挑战等进行了以下阐述。

PIM 研究聚焦于信息世界的一个信息子集，其中每个信息元素对于主体都有一定的影响能力。即 PIM 所研究的信息对于主体是有用的，这种有用性可以是现实的，也可以是潜在的。例如，一个人到某地旅游时需要选择旅馆，关于旅馆的信息会有很多，如位置、价格、经理、员工数目、营业状况等，如果对该位旅客做出选择产生影响的因素只有位置和价格，那么在其 PIM 系统中关于旅馆的信息可以只包含旅馆的位置、价格信息。因为主体的需求是动态变化的，因此 PIM 的信息集合也是变化的，但具有相对稳定性。在 PIM 研究中，个人信息(PI)包括以下三层含义：①个人保存并为自己所用的信息；②与个人有关但被其他实体控制的信息，如被医疗保险机构所掌握着的健康信息；③一个人经历过但

不为自己所控制的信息，如访问过的网页。

个人信息项(Personal Information Item)：信息项是与主体相关的信息包。在传统的以纸为介质的个人信息管理系统中，一篇文章、一封信都可以看作信息项。现在的信息中包含大量的数字信息，因此一个信息项可以是一封电子邮件、一个电子文档、一张图片等。每个信息项有一个信息框(Information Form)，信息框与具体的应用和工具有关，这些应用和工具用来命名、移动、修改、复制、组织信息项，也可以为信息项赋予一些属性，如 Outlook 可以看作一个信息框，通过该信息框可以实现对邮件的访问。

个人信息空间(Personal Space of Information，PSI)：个人信息空间是指主体能够控制或名义上能够控制的所有信息项组成的集合。这里所提到的控制并不是指排他性的专属，可以与其他用户共享，如一个实验室服务器上供所有成员访问的文档信息。一个 PSI 往往包括一个人的书籍、文档、邮件、访问过的网页或其他存储在不同计算机上的与主体有关的文件。

PSI 是可供用户通过多种方法利用的潜在的数据源。对 PSI 中信息的有效访问与重用，可以大大提高个人的工作效率。个人信息管理的目的就是实现对个人信息空间的有效重用。

个人信息管理(PIM)：PIM 本质上是一系列操作行为的集合，其行为目的是建立、使用和保持个人信息及用户需求之间的映射。对个人信息管理有关的行为可以归为三种：输入行为、存储行为和输出行为。在此基础上提出了一个如图 1-2 所示的个人信息管理系统概念框架[5]。

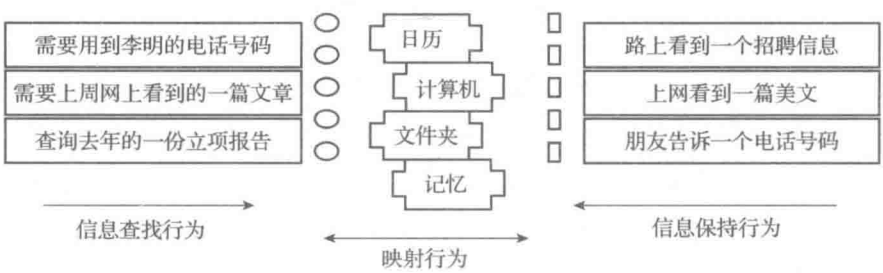

图 1-2　个人信息管理系统概念框架

由图 1-2 所示的个人信息管理系统概念框架可以看出，其涉及的行为可以分为三类：信息保持行为、信息查找行为和映射行为。

1. 信息保持行为

即影响个人信息空间中数据输入的一系列行为。具体来说，是指完成从信息到需求所进行的行为。例如，当用户遇到某个信息的时候，如访问了某个网页、获得了某人的联系方式等，往往要将这些信息保存下来以备将来使用。这类行为包括信息的分析、分类、记忆、存储等。由于信息的隐蔽性、数据源的多样性、遇到信息的偶然性以及主体自身因素，信息保持技术涉及诸多新的研究问题。

2. 信息查找行为

即影响个人信息空间中信息输出的一系列行为。具体来说，是指完成从需求到信息所进行的行为。例如，当用户需要用到个人信息空间中的某项信息(如某电话号码、邮件、图片等)的时候，将个人需求提交，并从个人信息空间中获得该信息。这类行为涉及查询接口、人机界面、搜索技术、信息分析、自动提醒等技术。需要指出的是：这里所说的信息查找和通常所说的 Web 搜索不同。其指的是从个人信息空间中重新查找曾经见到过的信息项，而 Web 搜索指的是在 Web 数据空间中搜索所希望的数据项，用户并不知道该数据项是否真的存在。

3. 映射行为

即影响和实现个人信息空间中信息映射的一系列行为。要高效地完成上面两种行为的映射，需要解决信息的存储、索引、安全性、一致性等一系列问题，这类行为就主要针对解决这些问题。

PIM 研究聚焦于个人信息管理中与信息保持、信息存储、信息查找有关的一系列技术，以提高个人信息管理的水平。在 PIM Workshop 2005 的研究报告中阐述了未来个人信息管理面临的主要研究问题，包括：个人信息识别与保存；个人信息的组织模型；个人信息查找与自动提醒；个人信息管理技术评价方法；个人信息的安全性与隐私保护；主

体记忆模式对信息映射方法的影响等。

目前，计算机和互联网技术的发展使数据日益成为重要的信息承载形式，大量的信息以数据的形式存储在各种各样的系统和设备中，在很多场景下，个人信息管理往往表现为对个人数据的管理，因此本书主要介绍个人数据空间管理的相关知识。

近年来国际上召开了多次个人信息管理研讨会，部分研讨会与 SIGIR 2006、SIGCHI 2008 等不同领域国际学术会议一同举办，在 SIGMOD、VLDB 等数据库领域重要学术会议上也陆续有一些关于个人数据管理的相关研究工作发表。这说明个人数据管理已经引起不同领域学者的广泛关注，且成为一个跨信息检索、人机交互、数据库等多个学科的研究领域。具体的研究题目涉及个人数据空间模型、数据索引、数据查询、桌面信息检索、人机交互界面设计、系统实现等多个方面。表 1-1 对不同领域学者对于个人信息管理这一问题所持有的观点进行了归纳[6]。

表 1-1　不同领域关于个人信息管理的基本思想

领域	基本观点	方　法
信息检索	将个人信息看作非结构化文本信息，基于文本信息管理方法管理个人数据	基于关键字索引和检索技术组织、存储和搜索个人信息
数据库	利用已有的数据库技术组织和管理个人数据信息	应当适应大规模、异构数据管理需求，采用与之相适应的数据存储、索引和查询方法
人机交互	个人信息管理应当以为用户提供快捷、舒适的操作体验为目标	从人机交互角度，通过分析用户行为，分析个人信息管理工具需要满足的约束条件

尽管不同领域的学者从不同的角度来看待个人信息管理，但他们所持有的观点并不矛盾。总的来说，未来的个人信息管理系统需要综合信息检索技术在非结构化数据管理方面、数据库技术在结构化数据管理方面、人机交互技术在界面设计方面的技术优势和成果，设计能够满足各种人群需要的个人信息管理系统。

目前大数据管理日益成为一个重要的研究领域。随着移动互联网、物联网、车联网、智能家居等技术的发展及各种可穿戴设备的普及，各

种与人相关的数据信息会不断地产生并被集成起来，这些个人数据将会成为名副其实的大数据。未来大数据的核心将是围绕人的数据，很多大数据应用也将围绕着人的各种需求。个人数据管理也将成为未来大数据管理的重要研究课题。

1.3 个人数据特征

个人数据具有以下特征：大规模、多样性、分散性、分布性、聚集性、不确定性、数据对象粒度的不均衡性、数据对主体的依赖性[6]，这些属性决定了在个人数据集成与管理中需要采用不同的方法与策略。

1)**大规模。**一般情况下人们很难把个人数据和大数据关联起来，认为个人数据不过就是个人计算机、手机中的数据，这样的数据量也就在几百兆的数据量之内，怎么可以说是海量数据呢？实际上，随着信息技术和可穿戴设备的发展，人们的一言一行都有可能被记录下来，包括人们开车的信息、位置的信息、从网上购物的各种信息等。试想一下，如果一个人每时每刻的言行举动都会被记录下来，那么会有怎样规模的数据量。

2)**多样性。**个人数据的多样性是指个人数据类型的多样性，包括传统数据库、文本、邮件、图片、音频、视频等。产生多样性的主要原因包括以下方面：一方面信息技术的发展使得不断地产生出新的类型的数据，比如关系数据库的发展产生了关系型数据，互联网技术的发展产生了 HTML 类型网页数据，XML 技术的发展产生了 XML 类型数据。另一方面是指主体个性化引起的多样性。数据空间的主体是人，不同的人由于职业、年龄、文化背景、民族等的不同，所需要管理的数据也不尽相同，比如一个从事化学研究的学者和一位音乐专业的学生所管理的数据、一位作家和一位摄影师所管理的数据的类型会有一些差异。

3)**分散性。**数据的分散性是指数据存放在不同的数据源中。数据分散的原因是个人应用的多样性和随之而来的数据源的多样性。随着信息技术的发展和各种个人信息管理设备的普及，大量的个人应用软件或工

具开始出现并获得推广，这些软件或工具由不同的人员或部门开发出来，使用的数据存储方式和数据模式不尽相同，从而导致个人信息分散存储在多个不同的数据源中，形成一个个“信息孤岛”。比如每个人都有邮箱、个人通讯录、个人图片、个人文档、个人收藏夹等，这些信息分散存储在不同的系统中，无法进行跨越不同数据源的信息检索。

4)**分布性。**个人数据分布存储在不同的物理设备上。例如个人邮件会存储在互联网中的邮件服务器上，个人文档会存放在个人计算机上，通讯录等信息会存储在手机等设备上，有的通讯录会和邮件一起存放在网络邮箱服务器上；个人访问网页的收藏夹也会存放在个人计算机的浏览器的目录下；个人照片有的存放在个人计算机上，有的存放在个人手机中。这种物理存储的分布性对个人数据空间管理和数据安全提出了挑战。

5)**聚集性。**虽然个人数据信息分布存储在不同的设备和位置，这些位置看起来是零乱、无序的。但是观察发现，这些信息的分布也表现出一定的聚集性。人们为了记忆、查询的方便，往往会按照个人的习惯进行分类存放，而且相关的数据往往会聚集在一起。比如，用户的个人照片往往会集中存放在个人计算机的某个目录下；个人关于某个任务的文档信息也往往会存放在特定的文件夹中。这个规律可以用来提高数据集成和查询的效率。

6)**不确定性。**个人数据的不确定性包括两种，一方面是由于客观原因造成的不确定性。例如，有些数据信息是从网页、邮件、文档中采用自动的方式抽取出来的，由于数据抽取、模式匹配技术等技术原因的局限性，使得抽取的数据具有不确定性。另一方面是由于主观原因造成的不确定性。当用户遇到一个数据项的时候，往往很难准确判定其与主体的关系和价值，有时也不容易对其进行准确的分类。用户经常遇到这样的情况，当需要保存一个文档的时候，往往会为应当保存到什么文件夹下而犹豫不决。比如一篇论文研究的问题可能跨越数据库、物理学等不同学科方向，如何将它保存在个人计算机中适合的位置并不容易确定，主观随意地分类往往为数据查询带来麻烦。

7)**数据对象粒度的不均衡性。**数据对象的粒度是指一个数据管理对

象的大小。个人数据管理所针对的对象，既包括几字节的数据对象，如电话号码、个人密码等，又包括视频文件等大小超过 100MB 的数据对象；从数据对象属性多少的角度，既包括属性较少的简单的数据对象，如某个人的联系方式，也包括一些逻辑结构复杂的数据对象，比如一篇章节结构复杂的论文。这种粒度的不均匀特性也为个人数据存储模式和逻辑模式的确定带来了困难。因此如何用一种统一的数据模式来描述这些不同格式、不同粒度的数据信息成为极具挑战性的问题。

8)**数据对主体的依赖性。**个人数据是与特定主体有关的所有数据对象的集合，是否与主体相关是判定一个数据项是否应当属于某个主体的数据集合的唯一标准。但是这种相关性的定义和计算则是一个需要探究的问题。比如一个用户访问过的文件是否算作与其相关；一个用户没有访问过的网页但确是关于该用户的信息，其是否应该认定与用户相关等。这种对于主体的依赖性，使得个人数据模型、数据更新、存储、索引、查询等技术和方法，都要将主体作为一个需要考虑的重要因素。

以上是个人数据的一些静态特征。作为个人数据空间的所有者、管理者和最终用户，主体对数据的访问也呈现出一些特点。

1. 用户对数据的许多访问是“再访问”

人们对个人数据的访问大部分都是“基于确定或不确定线索的再访问”。人们保存数据文件的目的大都是为了将来对它们的重新使用，这与 Web 搜索不同。对于 Web 搜索，用户往往不知道所搜索的结果是否存在。基于这一结论，如果能够将用户访问过的数据信息集中起来，并基于用户访问模式区别对待，将会大大提高“再访问”操作的效率，从而提高总的访问效率。

2. 个人数据访问的局部性和连续性

分析发现，用户对于数据项的访问具有一定的连续性，即在一段时间内，人们往往会用到并访问某些特定的数据项。其原因是因为用户的行为或任务往往具有连续性，比如人们在从事一件工作的时候，往往要频繁访问与之相关的信息。又比如，用户在写毕业论文期间，可能需要

反复查阅相关的文献，反复修改相关的文档、图表等信息。这一结论可以用来帮助预测用户访问行为，从而缩小查询范围，提高数据访问效率。

3. 用户需要基于若干模糊的记忆线索查找数据对象

对于数据库查询，用户知道数据对象的存在，并且也往往知道其确切的查询线索。例如在学生管理系统中，当用户查询一个学生的时候，往往知道该学生的学号或姓名。而对于个人数据而言，当查找一个数据对象的时候，用户必须能够回忆起相关的信息。由于时间、地点、访问频率等多种因素的影响，用户针对不同的数据对象，往往能够回忆起若干不同的线索。例如，时间信息、相关任务或事件、关键字、数据产生者、存储位置等。这些线索有时是模糊的、不确定的，而且用户有时需要将多个模糊的线索组合起来进行查询。

4. 任务在个人数据管理中扮演重要角色

用户对个人信息的访问和处理往往是以活动或任务为中心进行的。研究表明，任务在个人数据管理中扮演着重要的角色，人们经常需要基于任务查询、访问、复制、分析个人数据信息。人们经常遇到以下场景：①在实际工作中，有时需要并行处理多项任务。这样就需要在不同任务之间进行切换，每次更换任务时，总是希望能够快速找到与其相关的数据信息。②当用户重新打开计算机的时候，总希望能够快速地浏览目前正在做的几项任务，选定一件任务后也希望快速地访问与其相关文件。③当用户更换工作地点的时候(如出差或回家)，有时需要复制与当前任务有关的文档以便继续进行目前的工作。④当面临一个新的任务的时候，用户往往需要查看以前是否完成过类似的任务，以便参考其文档信息，节省时间。⑤当一个用户被其他用户咨询曾经参加的某个任务的相关信息时，也需要查询该任务及相关数据信息。⑥当用户进行工作总结的时候，往往需要查询在某个阶段完成的任务情况。这种情况表明，任务应当作为一种用来组织、索引个人信息的线索或依据，从而使得个人信息管理工具支持基于任务的个人信息

查询。

那么目前的个人数据管理的情况是什么样呢？总的来说，人们在个人信息管理方面还面临很多问题，具体如下：

1)**总体效率不高。**调查发现，很多人都曾经遇到在查询自己的个人文档时由于记忆信息的模糊性导致查找时间成本过高的问题。尽管人们试图通过分类、加标签、利用数据库存储等各种方式对个人数据信息按照语义进行结构化，但是，由于在个人信息管理方面尚有一些基础问题未解决，因此总体效率不高。

2)**数据一致性问题。**由于个人数据的分散性和分布性，有时会导致更新不同步问题，导致不同数据源中同一个数据对象的属性的描述不一致。比如在手机的通讯录和邮箱的通讯录中，同一个人的联系方式可能不一致。此外，在个人信息的版本管理方面也常常遇到问题，比如所找到的文件版本不一样而导致各种问题，有时造成比较严重的后果。

3)**数据安全和隐私数据泄露问题。**目前没有非常方便有效的措施对个人数据信息进行有效的保护，根本原因是没有有效的方法自动识别个人数据信息，并标记数据信息的价值，从而无法快速有效地对数据进行备份，因此，数据安全问题依然不容忽视。此外随着信息的分布存储，人们将越来越多的个人信息存放在云空间或个人移动设备上，设备的丢失、第三方服务商的非法操作等都会导致隐私信息的泄露，在新闻或网络上也会时常发现隐私泄露问题的相关报道。目前有许多学者在进行这方面的研究工作，但还有一些理论问题和基础性的技术问题尚未解决。

4)**个人数据查询问题。**因为目前个人数据分布在不同的数据源中，因此无法有效地进行跨数据源的查询，而这样的查询有时是必需的。此外，由于主体的个性化，比如年龄、性别、职业、民族等的不同所带来的个性化，使得个人数据查询接口、查询处理策略、查询优化方法等都需要采用不同于以往的技术。此外，用户记忆的局限性使得系统应当支持尽可能多种类的查询，以适应众多不同用户的需要。

针对个人数据及其主体数据操作的特征，人们提出了个人数据空间的概念。

1.4 个人数据空间的提出

广义上讲，个人信息管理的对象是存放在各种介质(包括纸张、胶片等)上的个人信息，第一届国际个人信息管理技术研讨会的研究报告[5]对个人信息管理的研究对象和个人信息特征进行了详细阐述。

随着计算机、互联网等技术的发展，个人信息更多以数据形式存放在各种电子设备中，个人信息管理主要表现为对个人数据的管理。2005年，Alon Halevy等学者针对海量、异构等新的数据特点提出了“数据空间”的概念[7]。与传统的数据库技术相比，其需要管理的是大规模、异构数据信息，因此在数据模型、数据操作方面都需要不同的方法和技术。

个人数据也具有大规模、异构的特点，其不仅包括结构化数据，也包括大量图片、网页、音频、视频等非结构化数据，因此一些学者针对个人数据特点提出了个人数据空间的概念[8]，相关研究工作日益得到大家的关注。除了上述特征之外，对主体的依赖性是个人数据管理系统区别于其他数据管理系统的重要特征之一，数据空间是与主体相关的数据及其关系的集合[9]，数据空间中的所有数据对于主体来说都是可以控制的。主体相关性和可控性是数据空间中数据项的基本属性，我们所说的数据空间实际是指主体数据空间，与之相对的是公共数据空间。图1-3显示了个人数据空间和公共数据空间的关系[9]，个人数据空间是公共数据空间的一个子集，随着主体需求的不断变化，数据项不断从公共数据空间纳入到主体数据空间中。

主体、数据集、服务是数据空间的三个要素[10]。主体是指数据空间的所有者，可以是一个人或一个群组，也可以是一个企业。对于个人数据空间来说，主体就是个人数据空间的所有者。数据集是与主体相关的所有可控数据的集合，其中不仅包括数据对象，也包括数据对象之间的

关系。主体通过服务对数据空间进行管理，如数据分类、查询、更新、索引等，都需要通过数据空间提供的服务完成。

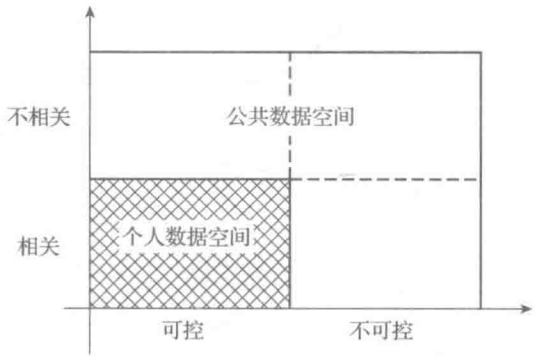

图 1-3 个人数据空间和公共数据空间的关系

以前研究界关注更多的是企业数据管理的问题，那么个人数据管理和传统的企业数据管理有何不同？个人数据管理与企业数据管理具有以下不同：①从面对的用户来看，个人数据管理面对的是使用计算机的用户，这些用户差异很大，其表现在年龄、职业、知识背景、使用计算机的习惯等很多方面，而企业数据管理系统主要面向具体的业务流程，比如成本管理，这种业务流程具有规范性，因此不必关注具体用户的差异性；②从数据依赖性来看，个人数据依赖于特定主体，而企业数据依赖于企业的业务流程；③从数据存储来说，个人数据分布在计算机、笔记本、手机等设备或邮箱、网盘、云存储空间等多种主体能够控制或不能完全控制的存储空间中，而企业数据则存储在企业控制的服务器上；④从数据输入方式来看，个人数据大部分通过实体识别等技术自动、透明地进行识别和保存，而企业数据往往是采用人工录入方式；⑤在数据查询方面，作为主体的人具有多样性，这种多样性表现在职业、年龄、教育背景、性别、记忆力等诸多方面，这决定了主体查询情景和方法的多样性，而企业数据查询则主要面向业务需求，具有相对稳定性。

由此可见，个人数据与企业数据并不完全相同，具有一些自己的特征，因此针对企业数据的管理技术也不能照搬到个人数据管理中，需要

基于个人数据特征研究与之相适应的数据管理技术。

1.5 个人数据管理系统框架

个人数据管理的最终目的是提高用户对个人数据的管理效率。基于此作者提出了个人数据管理系统框架，如图 1-4 所示。

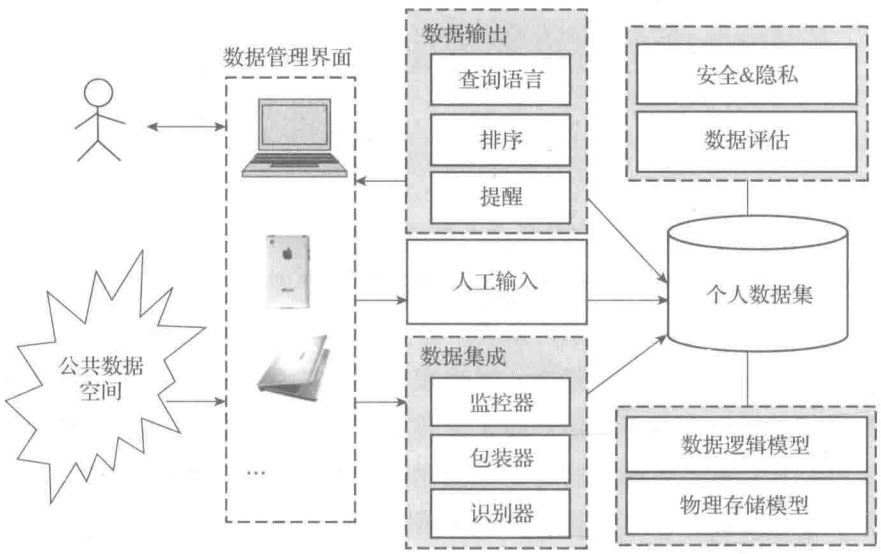

图 1-4 个人数据管理系统框架

个人数据管理系统主要包括数据集成、数据模型、数据输出、数据安全与质量保证四个模块。

1)**数据集成。**数据集成模块负责数据的输入，包括用户行为监控器、个人数据识别器和包装器。用户行为监控器自动监控用户行为，发现与用户相关的数据信息；个人数据识别器负责将个人数据实体及其属性识别出来并进行保存；包装器负责对特定数据对象的处理。因为个人数据来自不同的数据源，需要针对不同的数据类型设计包装器。由于信息的隐蔽性、数据源的多样性、遇到信息的偶然性、数据处理效率以及主体自身因素，个人数据集成需要用到自然语言处理、信息抽取等多方

面的知识。

2)**数据模型。**数据模型主要涉及数据逻辑模型、物理存储模型。采用什么样的逻辑模型来表示个人数据及其之间的关系、如何存储和索引个人数据等，都是需要研究的问题。

3)**数据输出。**数据输出指影响到个人数据输出的一系列行为，涉及查询、排序、提醒等方面的问题。

4)**数据安全与质量保证。**包括数据安全性保证策略和隐私信息保护策略，以及数据质量评价策略。

为了满足上述访问，个人数据系统需要提供如下功能。

1. 多种查询方式

在数据空间中，用户面临多种多样的查询场景，需要不同的查询方法。当用户需要查询一个经常访问的数据对象的时候，用户倾向于使用资源管理器以浏览的方式查询；查询一个很长时间没有访问过的数据文件时，用户需要基于关键字进行查询；当用户回忆不起文件的存储位置和关键字信息的时候，用户则需要基于一些模糊的信息查询所需要的数据文件。因此个人数据空间需要能够支持多种查询方式。

2. 简单的查询接口

个人数据管理系统的目的是帮助用户有效地管理个人数据资源。与传统的数据库管理系统相比，数据空间系统中不一定有专业的管理员，大部分用户没有很多的关于数据管理的专业知识。因此要求查询接口足够简单。

3. 基于任务的查询

个人数据管理系统应当能够提供基于任务的查询接口，用户可以查询所完成或正在执行的任务及其相关联的个人数据信息。

基于这一系统框架，本书将从数据模型、数据集成、数据存储、数据查询、数据安全与隐私保护、系统实现、新技术发展几个方面，对个人数据管理相关技术进行阐述。

参考文献

信息技术的发展使得个人数据信息急剧膨胀，个人信息具有数量大、多样、分散、分布、异构、依赖主体的特征，这些特征使得个人信息管理日益成为一个重要的极具挑战性的问题[1,2,3]。个人信息管理这一概念的提出可以追溯到1945年。美国科学家Vannevar Bush[4]构想了一种能够帮助人们管理个人信息的工具Memex。2005年第一届关于个人信息管理专题的研讨会在美国西雅图举办，并发表了大会报告[5]，其对个人信息、个人信息空间、个人信息管理等一些基本概念进行了阐述，引起信息检索、数据库等领域学者的广泛关注。目前电子数据成为信息的主要承载形式，因此个人信息管理主要表现为个人数据信息的管理[6]。针对目前数据呈现出海量、异构特性，并且传统的数据库技术已经不能很好地管理这些数据的情况，M. Franklin和A. Halevy[7]提出了"数据空间"概念，一些学者将这一概念与PIM相结合，进一步提出了"个人数据空间"的概念[7,8]，文献[9]对数据空间技术进行了综述分析，文献[10]对个人数据管理相关技术从数据模型、集成、存储、查询等方面进行了综述分析，提出了面向主体的数据集成框架。

[1] J Gantz，D Reinsel，Chute C，etc. The Expanding Digital Universe：A Forecast of Worldwide Information Growth Through 2010[EB/OL]. http：//www. emc. com/collateral/analyst-reports/expanding-digital-idc-white-paper. pdf.

[2] J Gantz，D Reinsel. The Digital Universe Decade - Are You Ready? IDC iView，May 2010[EB/OL]. http：//www. emc. com/leadership/digital-universe/expanding-digital-universe. htm.

[3] J Gantz，etc. Cutting the Clutter：Trackling Information Overload at the Source[EB/OL]. http：//www. xerox. com/assets/motion/corporate/pages/programs/information- overload/pdf/Xerox-white-

paper-3-25. pdf

[4] Bush V. As we may think[J]. The Atlantic Monthly, 1945.

[5] An NSF-Sponsored Invitational Workshop on Personal Information Management[EB/OL]. http: //pim. ischool. washington. edu/pim05home. htm.

[6] 李玉坤，任标，赵喜燕，等. 个人数据管理技术研究[J]. 计算机科学与探索，2014(11)：1281-1295.

[7] M Franklin, A Halevy, D Maier. From Databases to Dataspaces: A New Abstraction for Information Management[J]. ASM SIGMOD Record, 2005, 34(4): 27-33.

[8] J-P Dittrich, S MAV. iDM: A Unified and Versatile Data Model for Personal Dataspace Management[C]. In Proceedings. of the 32nd International Conference on Very Large Data Bases (VLDB 2006). 2006: 367-378.

[9] 李玉坤，孟小峰，张相於：数据空间技术研究[J]. 软件学报，2008，19(8)：2018-2031.

[10] Y Li, X Meng. Research on Personal Dataspace Management[C]. In Proceedings of the 2nd SIGMOD PhD Workshop on Innovative Database Research(IDAR 2008), 2008: 7-12.