数据空间基础设施的技术挑战及数联网解决方案

罗超然[1,2]，马郓[1,2,3]，景翔[1,2,4]，黄罡[1,2,5]
1. 数据空间技术与系统全国重点实验室，北京 100091；
2. 北京大数据先进技术研究院，北京 100091；
3. 北京大学人工智能研究院，北京 100871；
4. 北京大学软件与微电子学院，北京 102627；
5. 北京大学计算机学院，北京 100871

摘要

数据空间是网络空间从“以计算为中心”向“以数据为中心”转型的一种新形态，蕴含着变革性重大科技问题和换道超车创新机遇。类似互联网是网络空间的主要基础设施，数据空间也需要“以数据为中心”的新型基础设施，其核心功能是实现数据的一阶实体化。从数据空间的视角出发，分析总结互联网、万维网和数字对象架构等主流技术体系对数据一阶实体化的支持和不足，给出数据空间基础设施的基本内涵与技术挑战。提出基于数据语用原理的数据一阶实体化方法，通过融合数字对象架构、分布式账本和智能合约等技术形成数联网解决方案，支撑互联网规模的数据空间基础设施构造和运行。

关键词

数据空间基础设施；数联网；数字对象架构；数据语用

中图分类号：TP311 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023024

Internet of data: a solution for dataspace infrastructure and its technical challenges

LUO Chaoran[1,2], MA Yun[1,2,3], JING Xiang[1,2,4], HUANG Gang[1,2,5]
1. National Key Laboratory of Dataspace Technology and System, Beijing 100091, China
2. Advanced Institute of Big Data Technology, Beijing 100091, China
3. Institute of Artificial Intelligence, Peking University, Beijing 100871, China
4. School of Software and Microelectronics, Peking University, Beijing 102627, China
5. School of Computer Science, Peking University, Beijing 100871, China

Abstract

Dataspace is the transformation of cyberspace from "computing centric" to "data centric", which contains great technological issues and innovative opportunities. Similar to the internet, which is the main infrastructure of cyberspace, dataspace also needs a new "data-centric" infrastructure, whose core function is to realize the first-class entity of data. From the perspective of dataspace, the supports and shortcomings of mainstream technologies such as the internet, the World Wide Web, and the

digital object architecture for the first-class entity of data were analyzed and summarized, and then the basic connotations and technical challenges of dataspace infrastructure were given. Finally, a first-class data substantialization method based on data pragmatics was proposed. Based on this method, a solution called the internet of data by integrating digital object architecture, distributed ledger, smart contract, and other technologies was proposed to support the construction and operation of internet-scale dataspace infrastructure.

Key words
infrastructure of dataspace, internet of data, DOA, data pragmatic

0 引言

互联网以开放式体系结构和标准化协议在物理网络之上构建了一个由主机（host）组成的虚拟数据传输网络，从而屏蔽了下层物理网络的通信细节，在异构的物理网络之上形成了全球一体化的网络空间。随着互联网的发展和大数据时代的到来，数据成了互联网中最具价值的资源，高效地发现、访问、使用互联网中的数据也成了用户对互联网的主要诉求。然而，互联网数据的分散性及互联网环境本身开放、动态、异构的特点，使得访问和使用互联网数据面临协调成本高、责权利难保障，以及低效、易错、难复盘等挑战[1]。当前互联网“以计算为中心”的技术体系将计算资源视为一阶实体，围绕某个数据计算任务，以计算资源为中心调用所需数据，并临时性地赋予其身份标识，使其成为一个局部实体。在这个过程中，数据作为计算资源的附属，对外是不可见的，且其生命周期也会随着计算任务的结束而消亡，数据的价值通常被限定在一个个既定的计算任务之中[2]。随着大数据时代的到来，数据成了与计算资源同等甚至更重要的战略资源，大数据技术的发展将会在互联网和其他网络之上催生出一个虚拟的数据空间。

数据空间是网络空间从“以计算为中心”向“以数据为中心”转型的一种新形态。数据空间以数据为一阶实体，基于数据的自然属性构建数据的逻辑模型，并将其抽象为直接可见、可用的独立实体。一方面不依赖下层软硬件，软硬件环境的改变不会导致数据实体的变化；另一方面独立于上层应用，应用场景的变化不会导致数据自然属性的改变。数据空间“以数据为中心”的特征势必需要一套“以数据为中心”的新型基础设施将网络空间中资源的表征和组织从计算架构转变为数据架构，支撑数据的一阶实体化表达。

万维网（World Wide Web，WWW）和数字对象架构（digital object architecture，DOA）是互联网环境下“以数据为中心”的两大主流技术体系。二者从各自的数据应用场景出发，以超文本和数字对象的模型表示数据，赋予数据唯一的标识使其可直接访问，基于开放式软件体系结构组织数据，并通过标准协议规范数据的交互行为。经过数十年的发展，WWW和DOA已经在网络空间中形成了互联网规模的网页空间和数字出版物空间，支撑了大量的数据融合应用。

本文从WWW和DOA切入，分析其需求动机和技术发展脉络，归纳总结数据空间基础设施的技术特征和关键挑战，进而提出一种数据空间基础设施解决方案——数联网。

1 数据空间视角下的互联网、万维网和数字对象架构

互联网面向计算机之间的数据传输场景，将计算机抽象为主机，以IP地址作为主机的识别符和地址，以传输控制协议（transmission control protocol，TCP）建立起主机之间的虚拟通信链路，进而在链路中传输数据。在互联网的数据传输场景下，数据被抽象为计算机之间传输的数据包（packet），通过计算机之间的虚拟通道顺序传输。这种对数据的抽象并未将数据视为一阶实体，数据包依附于机器之间建立的传输通道，其生命周期也会随着TCP连接的释放而结束。

数据的一阶实体化是数据空间的核心理念：一阶代表数据的直接性，即数据是直接可见、可用的，应用直接通过数据的标识访问数据而非通过接口调用等间接方式获取数据；实体代表数据的独立性，即数据是自然存在的，不依附于其他任何实体。“以数据为中心”的数据空间势必需要一套“以数据为中心”的新型基础设施，构建一阶数据实体的资源模型和访问架构，将分散在网络空间中的数据资源有效组织起来，从而高效地发现、访问和使用数据。

从数据空间的角度来看，WWW和DOA是目前互联网上两个主流的“以数据为中心”的技术体系，二者从各自最初的数据使用场景出发，逐渐泛化对目标数据资源的定义、扩展数据模型、演进系统架构，最终实现了互联网规模的数据发现、访问和使用。

1.1 万维网：从网页到资源表征

WWW诞生于1989年，当时就职于欧洲粒子物理研究所的 Berners-Lee T设计WWW的初衷是使实验组里各国的高能物理学家能通过计算机网络方便地传递、共享科研信息[3]。WWW将数据抽象为HTML文档（网页），不仅包含了数据本身的内容，还包含视觉展示内容的语义标签。将浏览器/服务器架构作为系统实现的模型，并制定了用于二者间传输网页的HTTP 1.0协议，以可见、可读的文本作为HTML文档的序列化方式。在这个场景下，WWW解决的是人与人之间信息共享的问题，因此WWW的设计也主要侧重于如何便捷地公开数据及如何方便地浏览信息。对于数据提供方而言，仅需一台连接互联网的机器，将数据封装为HTML文档，并提供一个HTTP访问端口，即可公开数据。而对于数据使用方而言，仅需知道数据的统一资源定位器（uniform resource locator，URL），即可通过浏览器访问、使用目标数据。

访问WWW数据的前提是知道目标数据的URL。尽管可以通过网页之间的超链接跳转来发现新的网页，但随着WWW上信息规模的爆发式增长，如何高效地发现所需信息成为WWW面临的关键挑战。人与人之间信息的开放与共享是WWW的核心理念，因此WWW上绝大部分数据是公开可访问的文本数据。搜索引擎基于爬虫技术，通过网页之间的超链接爬取大量网页、建立文本索引并提供搜索服务，基本解决了WWW公开数据搜索、发现的难题。在WWW最初的基础设施系统架构中，并没有搜索引擎这一角色，然而随着WWW规模的增长，受高效发现数据的需求驱动，目前搜索引擎已经成了WWW不可或缺的一部分。

随着WWW的进一步发展，人与人之间的信息共享已不是WWW的唯一用途。如何使机器也可以利用WWW基础设施、

使用WWW数据，构建跨互联网的分布式应用，成为WWW的另一大数据使用场景。面向这一使用场景，WWW出现了两个技术发展方向：Berners-Lee T提出的语义网[4]（semantic web）及Fielding R T提出的表述性状态转移[5]（representational state transfer，REST）。语义网的思路是通过为网页添加机器可解释的语义标签，使网页的数据可以被机器解释、使用。REST则不再使用网页来表示数据，而是将WWW上的数据都抽象为资源，通过交换资源的表征（representation）访问资源的内容、修改资源的状态。从实际发展的角度来看，REST无疑是更成功的技术方案，并且其设计理念也在HTTP1.1中得到了体现和标准化。REST是一种软件体系结构风格，其架构的核心约束包括以下几点：

- 所有资源需要被资源描述符标识；
- 所有访问资源的操作语义需要统一；
- 对资源的操作通过交换其表征来实现；
- 交换资源表征的消息需要是自描述的，包含所有本次交互所需的信息。

Fielding R T在参与HTTP1.1协议制订和Apache服务器软件开发的过程中将REST风格应用其中：采用统一资源标识符（uniform resource identifier，URI）标识资源；基于HTTP1.1中定义的8种方法统一表示对资源的操作语义；以XML或JSON这类机器可解释的格式交换资源表征。REST简洁、可扩展的设计理念及Apache软件的成功使REST成为目前Web上最主流的应用程序接口（application programming interface，API）标准。

REST的成功得益于其与WWW的深度融合，但也受到了WWW技术体系的限制，具体如下。

- REST采用URI来唯一标识数据资源，然而URI扩展自DNS，本质上是对互联网机器的标识，当数据位置发生移动或机器域名失效时，URI对应的数据便无法访问。
- REST缺乏有效的数据发现机制，这一问题源自WWW的最初设计。搜索引擎为WWW解决了此问题。然而，WWW中的数据是公开的、可索引的网页文本，REST中的数据则是抽象的资源，且REST缺乏网页超链接这种可以互相发现的机制，虽然有类似Programmable Web这类REST API注册平台，但该平台上注册的API数量也仅有24 000个左右，远远达不到应有的规模，并且该平台已于2022年10月31日停止运营。
- REST缺乏对数据提供方的权益保障机制，这一问题同样源自WWW的最初设计。WWW设计之初的目标是人与人之间的数据共享，其网页数据大部分是公开可访问的，因此并未有相应的机制保障数据提供方的权益。

1.2 数字对象架构：从数字出版物到数字对象

DOA起源于互联网发明人Kahn R E在1988年主持的由美国国防高级研究计划局（Defense Advanced Research Projects Agency，DARPA）资助的数字图书馆项目[6]。该项目的主要目标是为美国大学的计算机院系搭建数字图书馆系统，以将其现有的技术报告电子化、接入互联网，并保证数字图书馆系统之间的互联互通。在该项目中，Kahn R E提出了一种信息基础设施的开放式体系架构，基于此架构实现的系统称为数字图书馆系统（digital library system，DLS）。DLS将数字化的文献视为数字出版物，如何保障出版物的知识产权是DLS设计主要考虑的问题之一。DLS中的数据管理核心子系统包括：为数字出

版物分配标识的注册系统、存储出版物实体的数据库系统、全局的出版物索引目录系统等。

在DLS的设计中，数字出版物实体存储于所有者本地的数据库服务器，通过全网唯一的标识和全局的编目、索引服务，使用者可以发现、寻址并访问所需的数字出版物。此外，为了保证对数据的使用都在所有者的控制范围内，DLS提出了智能代理“Knowbot”的概念。智能代理是一个可以在不同DLS之间移动、执行的活动程序，DLS通过智能代理发现、使用数据。智能代理以数字出版物的标识为数据输入，在执行过程中根据所需出版物的标识移动至目标所在的数据库服务器，从本地访问目标数据，从而保证数据的使用在其所有者的控制范围内。

扩展数字出版物这一使用场景，将数字图书馆泛化为信息系统，将技术报告泛化为任意数据资源，Kahn R E在2006年提出了DOA的概念[7]。彼时DOA仅完成了数字对象全局标识解析系统的设计和实现，搭建了数字对象标识基础设施Handle系统，却遗留了大部分开放性问题有待解决。经过十余年的发展，DOA最终形成了包括1个模型、3个构件、2个标准协议在内的架构模型，具体如下。

- 数据模型方面。DOA以数字对象为其体系结构中的基本元素抽象数据资源。一个完整的数字对象可以分为3个部分：标识、元数据、实体。
- 系统构件方面。针对数字对象模型的3个组成部分，DOA提出了3个构件：数字对象标识系统、数字对象注册表及数字对象仓库，分别用于管理数据的标识、元数据及数字对象实体。
- 标准协议方面。DOA制订了两个标准协议，分别是：用于数字对象搜索和访问的协议DOIP[8]（digital object interface protocol）和分配、解析标识的数字对象标识解析协议DO-IRP[9]（digital object identifier resolution protocol）。

起源于DLS的DOA，继承了对数据安全、权益相关方面的考量。首先，数字对象实体存储于由数据提供方控制的数字对象仓库中，对数字对象实体的访问均发生在数据提供方控制的环境下；其次，数字对象的标识可视作对数字对象实体的引用，应用程序可以使用数字对象标识声明其所使用的数字对象，在执行时通过标识寻址到所需数字对象实体；最后，元数据和数字对象实体分离的结构使得即使在数字对象实体不可访问的前提下，也可通过元数据发现所需目标数字对象。目前，数字对象标识解析系统和DO-IRP已有较大规模应用，支撑了覆盖全球的Handle系统的建设与运行，催生了如数字出版DOI®系统[10]、中国工业互联网标识解析系统等多个大规模标识解析应用[11]。但数字对象仓库系统、注册表系统及DOIP由于推出时间较晚，暂时还未得到非常广泛的应用。

作为目前互联网环境下主流的数据空间基础设施，WWW和DOA由于各自的起源不同，二者技术和协议的侧重点也有所不同，信息的开放、共享是WWW的主要目标，而如何保障数据所有者的权益则是DOA主要关注的问题。

2 数据空间基础设施的基本内涵与技术挑战

数据空间是网络空间从“以计算为中心”向“以数据为中心”转变的新形态，是构建在互联网及其他网络之上的一体化虚拟空间，其基础设施的首要目标是能在互联网规模上以一阶实体的形式有效组织并

高效使用分散在网络空间中的数据资源。然而，互联网开放、复杂、动态、难控的特征给数据空间基础设施带来了巨大的挑战。作为目前能够在互联网规模上实现数据发现、访问的主流技术体系，本文通过对WWW和DOA的分析，归纳总结了两条技术路线的共同技术特征，为数据空间基础设施的设计提供参考。

总体而言，WWW和DOA的技术要点可以总结如下：异质同构的数据元素、开放式架构的基础软件、标准化的操作协议。

- 异质同构的数据元素。数据元素是数据基础设施中的基本资源单元，每个数据元素都是独立的个体，能够与其他数据元素区别开来。与此同时，数据元素又是异质同构的，即遵循同样的数据模型、数据结构，但其内容来自各自的实际数据，可以互不相同。
- 开放式架构的基础软件。基础软件是WWW和DOA的运行实体，也是数据元素操作的执行者。WWW和DOA实质上由大量基于开放式体系结构的软件系统共同组成，系统可以任意加入、退出，且不会对其他系统的运行造成过多影响。
- 标准化的操作协议。标准化的协议是访问、使用数据元素的统一指令。用户通过操作协议向基础软件系统发送指令，基础软件系统根据指令中的操作语义执行操作，访问数据元素并将结果返还给用户。

依据此模型审视WWW和DOA可以发现，二者都对数据元素进行了统一的抽象，提供了数据标识、数据访问的基础功能，但各自面向不同的数据使用场景又有一些差异。Web1.0面向人与人之间的信息共享，将数据元素抽象为人类可读的网页；REST面向分布式Web应用的需求，将数据元素抽象为自描述、无状态的资源表征；语义网面向机器可处理的需求，以富含机器可解释标签的RDF文档[12]抽象数据元素，并且提出了本体定义语言（web ontology language，OWL）[13]和RDFS[14]来统一标签语义；DOA面向数据所有者权益保护的需求，将数据元素抽象为元数据、实体分离的数字对象。与此同时，这些技术的基础设施和标准协议也有所不同：由于元数据分离机制，DOA在基础设施中增加了数字对象注册表构件以管理数字对象的元数据，并且在标准协议中也明确了数字对象搜索的操作语义，使得数据搜索成了DOA的基础功能之一。而WWW中的搜索引擎是平台化的而非开放式的，访问方式也是基于平台接口而非标准化的操作语义，因此搜索引擎不能算严格意义上的WWW基础软件。互联网数据基础设施技术体系对比见**表1**。

异质同构的数据元素、开放式架构的基础软件、标准化的操作协议是WWW和DOA的共同技术特征，也是二者能够在互联网规模实现数据发现、访问和使用的关键。作为构建在互联网和其他网络之上的数据空间，其基础设施也势必需要参考和借鉴互联网相关技术体系的特点，但又需要根据其面向场景的不同做出相应的取舍和扩展。

3 数联网：数据空间基础设施的解决方案

针对数据空间的战略需求和技术挑战，结合互联网相关技术的要点，提出了面向数据空间的基础设施解决方案——数联网。

3.1 基于数据语用原理的数据一阶实体化方法

数据一阶实体化是数据空间的核心。数据空间将数据作为一个个直接可见、可

表1 互联网数据基础设施技术体系对比

类型	对比项	Web1.0	RestFul Web2.0	语义网	DOA
数据元素	元素抽象	网页	资源表征	RDF文档	数字对象
	元素标识	URL	URI	URI	Handle
	表现形式	HTML	JSON/XML	RDF/XML	JSON + Bytes
基础软件	标识系统	DNS服务器	DNS服务器	DNS服务器	标识解析系统
	访问系统	Web Server	Web Server	Web Server	数字对象仓库
	搜索系统	无	无	无	数字对象注册表
标准协议	标识协议	DNS	DNS	DNS	DOIRP
	访问协议	HTTP 1.0	HTTP 1.1	HTTP 1.0	DOIP 2.0
	搜索协议	无	无	无	DOIP 2.0

用且独立的逻辑实体。向下不依赖软件和硬件环境，通过和下层软硬件数据载体的解耦，在多样、异构、动态的软硬件环境上维持一个共性、同构、稳定的数据使用环境；向上不依附于应用和业务逻辑，将数据本身的自然属性显式化，并以与业务无关的数据语义理解和操作目标数据，在复杂、变化的业务逻辑中保持数据使用的简化、统一。目前互联网上大多采用通道式的数据使用模式，数据应用基于数据源提供的数据接口调用数据，并在应用内部进行数据的处理。通道式的数据使用模式并未将数据作为一阶实体，数据仅作为应用的附属，对外不可见，并且其生命周期也会随着应用的结束而消亡。

针对此问题，基于数据语用原理的数据一阶实体化方法，为数联网的构建提供理论基础被提出了。数据语用即数据在不同应用中的含义和使用方式，代表数据应用和数据源可分离的关键理念[15]。基于数据语用原理，数据空间的运行行为可以建模为不同场景下数据应用对数据资源的访问和使用。通过将数据资源从数据源中解构出来作为一阶实体化的数据元素，同时将对数据的使用从数据应用中解构出来作为灵活可调度的运算单元，并根据具体的应用需求、使用场景按需进行运算单元和数据元素的重构，从而实现数据的一阶实体化并支撑数据空间的运行，如**图1**所示。

数据语用原理为数联网提供了理论支持，在此基础上本文融合了数字对象架构、智能合约、分布式账本技术，提出了数据空间基础设施的解决方案——数联网。其主要包括3个关键技术：基于数字对象架构的一阶数据实体模型及交互技术、基于语用合约的一阶数据实体使用技术、基于关系链的一阶数据实体可信保障技术。

3.2 基于数字对象架构的一阶数据实体模型及交互技术

数据空间构建在互联网和其他网络之上，其基础设施也将至少达到互联网规模。作为互联网上典型的“以数据为中心”的技术体系，WWW和DOA对于数据空间基础设施而言具有较高的参考价值。然而，数据空间中的数据是一阶实体，要求数据既要直接可访问又要独立于软硬件环境。WWW以URL为数据资源的标识，而URL本质上是对机器的标识，数据并未完

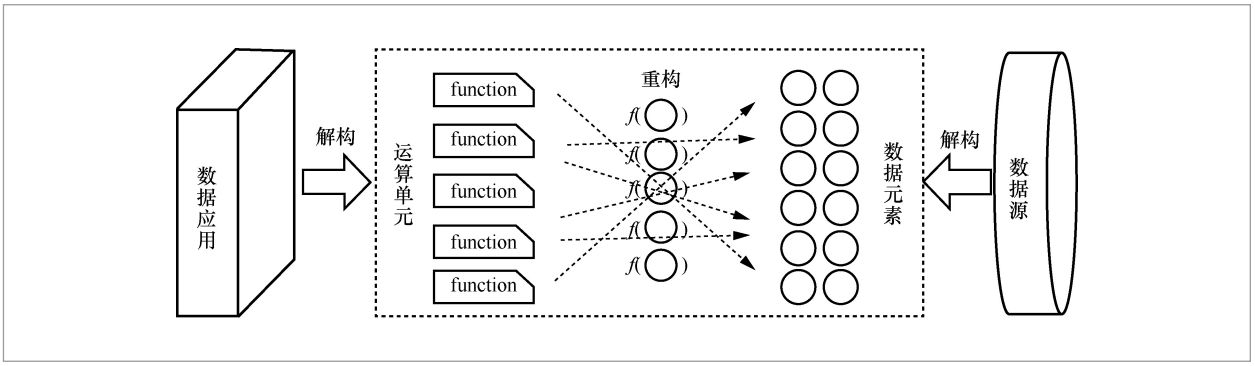

图1 基于数据语用原理的数据一阶实体化方法

全独立于其所属的机器，数据物理环境的迁移或者机器URL的失效均会导致数据不可用。此外，虽然WWW开放、共享的特点极大促进了数据的产生和使用，加速了大数据时代的到来，但同样也带来了数据安全、隐私、信任等潜在风险。在大数据时代特别是人机物融合的新环境下[16]，数据不再仅仅是现实世界的电子化记录，更是控制和影响现实世界的工具，其所面临的安全、隐私、信任等挑战更加严峻。DOA中标识、元数据、实体三要素分离的架构特点可以在保障数据所有者相关权益的前提下高效地发现、访问分散的数据资源，有助于数据空间中数据价值的充分释放。然而，作为数据空间的数据基础设施，其需要解决的问题和面临的挑战均与互联网环境下的DOA有较大不同，因此本文对其进行了相应的扩展，使其能够成为数据空间中组织、管理一阶数据实体的基础设施。数联网架构示例如**图2**所示。

（1）扩展数字对象模型，提出数字对象第四要素——关系

数据空间中的一阶数据实体并不是彼此独立、毫无关联的。相反，数据实体之间的使用关系也是数据空间的重要组成部分，例如WWW网页中的超链接、DOA数

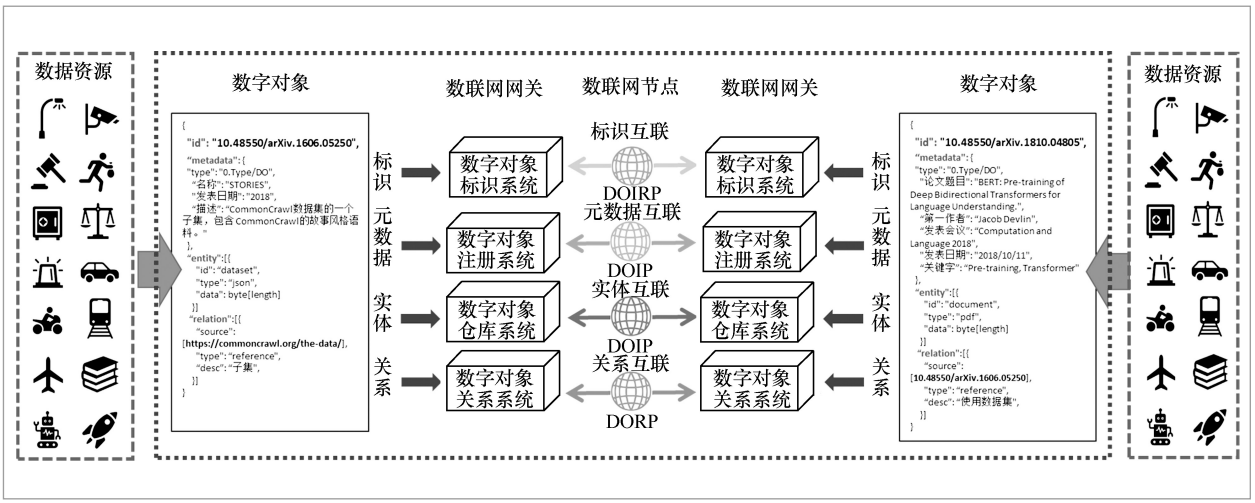

图2 数联网架构示例

字出版物中的引文。WWW基于超链接发现网页，催生了Google、百度等搜索引擎巨头；DOA基于出版物之间的引用保障数据所有者的数据权益，支撑了论文标识被引量成为科研成果评价的核心指标之一。在数据空间中，数据实体之间的使用关系是明确数据责权利、评估数据价值的重要依据。但是，无论是网页超链接还是论文的引文，其关系内置于数据实体中而并未显式化，需要获取数据才能知道该数据与其他数据之间的使用关系。对于网页、论文这类公开数据，可以通过大量爬取数据、提取数据中的链接以恢复数据之间的使用关系，这种方式代价大且需要所有数据的访问权限。针对此问题，对DOA进行了扩展，提出了数字对象关系显式化机理，将关系作为与标识、元数据同等重要的数字对象第四要素从数据实体中解耦出来，从而在没有全局数据视图的情况下仍然能够较为完整地获取数据使用关系的全局视图。基于关系显示化机理，笔者研制了数字对象关系系统，将每次对数字对象的创建、访问、删除操作都视作一次数字对象使用关系的创建或强化，并基于标准协议提供关系的访问和追溯，为厘清数据空间中数据价值、责任、权益提供根本性支撑。

（2）泛化DOIP，解耦数据访问的网络环境

一阶实体化的数据是数据空间的基础元素，一阶数据实体既不依赖于上层应用，也独立于下层软硬件环境。因此，面向一阶数据实体的访问协议也需要具备相应的独立性。然而，互联网环境下的DOA将数字对象访问与下层通信协议进行了绑定，依赖于传输层安全（transport layer security，TLS）协议提供的通信链路访问目标数字对象。而数据空间基于的网络环境是互联网和其他网络（包括电信网、移动网、物联网等）的交汇融合，其中存在大量异构的通信协议，与TLS紧耦合的DOIP也将数字对象与下层通信网络进行了深度绑定，不符合数据空间数据一阶实体化的基本特征。

具体而言，面向互联网设计的DOIP v2.0采用分隔符的方式传递数字对象访问消息，依赖于TLS协议建立可信、可靠的通信链路。当底层通信协议不可信、不可靠时，访问消息存在丢包、乱序、泄密、篡改等风险。为此，笔者对现有DOIP进行了扩展，针对现有基于分隔符的明文DOIP消息面临的可靠、安全和隐私问题，提出了一种基于包和字节的DOIP消息序列化方式，将可靠、安全和隐私保障机制内置于DOIP消息本身，从而实现DOIP与底层通信协议解耦，进一步支撑数字对象的一阶实体化。目前，新版DOIP v2.1已得到数字对象体系架构应用技术与标准促进组织（DOA Application Technology Standardization Development Organization，ATSD）的采纳，并正式发布[17]。

3.3 基于语用合约的一阶数据实体使用技术

数据的一阶实体化需要全新的数据使用机制。在WWW和DOA中，数据的使用方式通常由数据的提供方决定，数据的需求方需要按照提供方提供的数据接口获取数据内容，如对于网页就是通过HTTP请求获取网页的内容，对于数字出版物就是通过DOIP请求检索内容。然而在大数据时代，数据的使用价值是由需求方决定的，数据所能发挥的价值通常远远大于数据被提供时的预期价值。提供方定义的数据使用方式较大程度上制约了数据价值的充分释放。数据空间将数据一阶实体化，使数据作为直接可见、可用的自然实体暴露在外，使其不再受特定场景、特定应用需求的限制，数据需求方能够根据其应用需求，

直接在所需场景中访问、使用所需一阶数据实体，虽极大释放了数据的潜在价值，但也给数据的使用带来了较大的挑战。

在传统的数据使用模式中，数据需求方从数据提供方提供的数据接口中获取数据实体并交给数据应用进行处理，对于数据提供方而言，数据应用具体怎么使用数据是不可知的。这种数据使用方式的不可知，一方面导致提供方出于对数据安全、隐私等权益的担忧不敢提供数据，另一方面也导致需求方难以自证清白从而充分使用数据，限制了数据价值的释放。针对此问题，提出了基于语用合约的一阶数据实体使用机制，基于数据使用和数据应用可分离的数据语用机理，将数据需求方对数据的使用从数据应用中独立出来，并进行标准化、透明化，从而确保供需双方对数据的使用方式达成共识。进一步，自研了DOIP、DOIRP原生的智能合约语言和协同执行引擎，数据需求方将对数据的使用以语用合约的形式进行描述，合约引擎根据所需目标数字对象自适应地选择多个数据供方节点、公证节点协同执行，确保数据使用过程对需求方可见、对提供方可控，从而填补供需双方在数据使用上的认知鸿沟，解决数据空间中一阶数据实体使用的不可知挑战。

3.4 基于关系链的一阶数据实体可信保障技术

数据的一阶实体化需要更强的可信保障。数据空间以数据为一阶实体，不再预设数据的使用场景、使用范围，极大释放了数据的潜力，但也会导致更加复杂的数据使用关系，给数据的可信带来了巨大的挑战。具体而言，数据空间的参与方必然是多主体的，清晰的数据使用关系对于明确数据的责权利至关重要。数据提供者需要知道数据被谁用、怎么用、用完之后产出了什么；数据使用者需要知道数据是谁的、从哪来、是否真实可靠。而数据作为生产要素被使用时具有获得的非竞争性、使用的非排他性、价值的非耗竭性、源头的非稀缺性等特征，一数多用、多数合用、加工后复用将是数据使用的常态，这也直接导致了数据的使用关系是极为复杂的，且数据可复制、易复制的特征使得数据之间的关系更加难以追溯。

针对这一问题，设计了基于分层共识的关系链系统，将每次对数字对象的创建、访问、删除操作都视作一次数字对象使用关系的创建或强化，并采用区块链数据结构对其进行序列化存储，在多主体的数据空间中可信地表达全网范围的数字对象关系。针对区块链技术固有的代价高、效率低等问题，原创了分层随机共识技术，根据关系之间的串联、并联特征分别采用有序、无序的分层共识，最大限度提高区块链的吞吐量，在互联网环境下万级普通节点规模能够支撑每秒百万级的数字对象关系记录，实现了数据关系的可信、可靠存证，有效支撑了一阶数字对象实体的可信保障和责权利追溯。

4 结束语

数据空间是网络空间从“以计算为中心”向“以数据为中心”转型的一种新形态，是在互联网和其他网络之上形成的由海量一阶数据实体组成的一体化虚拟空间。以数据为一阶实体，有效组织分散在互联网和其他网络之中的数据资源，支撑高效的数据发现、解析、寻址和访问是数据空间基础设施的基本要求。通过对WWW和DOA的分析总结，提出了互联网规模的数据空间基础设施需具备的技术特征和面临的挑战。进一步，面向数据空间

基础设施的需求，提出基于数据语用原理的数据一阶实体化方法，以及数据空间基础设施解决方案——数联网：基于数字对象四要素构建一阶数据实体的数据模型和交互模型，基于语用合约达成供需双方在一阶数据实体上的使用共识，最后基于关系链保障数据空间运行时一阶数据实体的可信。

秉承着开放、开源、共建、共治的理念，数联网和数据空间的建设、发展、应用将是一个由所有用户共同参与、共同合作、共同推广的持续且漫长的过程，其所蕴含的巨大价值、产生的巨大利益也将由所有参与者共同分享。

参考文献:

[1] 黄罡. 数联网：数字空间基础设施[J]. 中国计算机学会通讯, 2021, 17(12): 60-62.
HUANG G. Internet of data: infrastructure of the digital space[J]. Communication of China Computer Federation, 2021, 17(12): 60-62.

[2] MEI H, HUANG G, XIE T. Internetware: a software paradigm for Internet computing[J]. Computer, 2012, 45(6): 26-31.

[3] BERNERS-LEE T, CAILLIAU R, LUOTONEN A, et al. The world-wide web[J]. Communications of the ACM, 1994, 37(8): 76-82.

[4] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic web[J]. Scientific American, 2001, 284(5): 34-43.

[5] FIELDING R T. Architectural styles and the design of network-based software architectures[D]. Irvine: University of California, Irvine, 2000.

[6] KAHN R E, CERF V G. The digital library project volume 1: the world of knowbots (DRAFT)[J]. Corporation for National Research Initiatives, 1988.

[7] KAHN R, WILENSKY R. A framework for distributed digital object services[J]. International Journal on Digital Libraries, 2006, 6(2): 115-123.

[8] DONA. Digital object interface protocol version 2.0[Z]. 2018.

[9] SUN S, LARRY L, BRIAN B. Handle system overview[Z]. 2003.

[10] PASKIN N. Digital object identifier (DOI®) system[M]//Encyclopedia of library and information sciences, 3rd edition. Boca Raton: CRC Press, 2009: 1586-1592.

[11] 工业互联网产业联盟. 工业互联网标识解析标准化白皮书[R]. 2021.
Alliance of Industrial Internet. White paper on standardization of industrial internet identification resolution[R]. 2021.

[12] W3C. RDF primer[S]. 2004.

[13] W3C. OWL web ontology language overview[S]. 2004.

[14] MCBRIDE B. The resource description framework (RDF) and its vocabulary description language RDFS[M]//Handbook on ontologies. Heidelberg: Springer, 2004: 51-65.

[15] 滕腾, 黄罡, 陈兴润, 等. 网构软件数据语用的一种动态支撑方法[J]. 软件学报, 2008, 19(5): 1160-1172.
TENG T, HUANG G, CHEN X R, et al. An approach to dynamically operating data pragmatics for internetware[J]. Journal of Software, 2008, 19(5): 1160-1172.

[16] 黄罡. 面向人机物融合的泛在系统软件技术专题[J]. 中国计算机学会通讯, 2020, 16(4).
HUANG G. Special topic of ubiquitous system software technology for human-machine integration[J]. Communication of China Computer Federation, 2020, 16(4).

[17] ATSD. Digital object interface protocol specification version 2.1[Z]. 2023.

作者简介

罗超然（1990– ），男，博士，北京大数据先进技术研究院助理研究员，主要研究方向为系统软件、数字对象架构、数联网、数据空间。

马郓（1989– ），男，博士，北京大学人工智能研究院助理教授、博士生导师，主要研究方向为Web系统、服务计算、移动计算等。

景翔（1979– ），男，博士，北京大学软件与微电子学院副研究员，中国计算机学会（CCF）会员，主要研究方向为操作系统、分布式计算。

黄罡（1975– ），男，博士，北京大学计算机学院教授、博士生导师，CCF杰出会员，主要研究方向为系统软件、软件自适应、数联网、数据空间。

收稿日期： 2023-02-21
通信作者： 黄罡，hg@pku.edu.cn
基金项目： 北京高等学校卓越青年科学家计划项目（No. BJJWZYJH01201910001004）
Foundation Item: Beijing Outstanding Young Scientist Program (No. BJJWZYJH01201910001004)