开放数据联盟链
Open Data Chain

开放数据联盟链（ODC）白皮书

中国科学院计算机网络信息中心

2021年1月

引　言

基于区块链技术的科学数据开放共享新模式通过分布式账本、数据精准确权、数据隐私安全、智能合约等机制有效解决了对等机构间数据共享的诸多问题。新模式可以应用在科学数据中心之间、科研机构之间、科研人员之间，实现跨学科、跨领域的数据开放共享应用。

为推动区块链技术与科学数据开放共享深度融合，形成发展共识，中国科学院计算机网络信息中心和北京大学等单位共同组织编写了《Open Data Chain（ODC）白皮书》。本白皮书深入解读了区块链内涵概念，广泛调研分析了区块链在开放数据领域的应用现状，提出了 ODC 联盟链体系，并介绍了 ODC 在数据追溯、确权凭证、可信共享、版本管理等领域展开的应用探索，最后总结了开放数据区块链发展建议与展望。

编写说明

指导单位：中国科学院网信办、国家科技基础条件平台中心

牵头编制单位：中国科学院计算机网络信息中心

参与编制单位：北京大学，中国科学院国家空间科学中心、微生物研究所、地理科学与资源研究所、计算技术研究所、信息工程研究所、微电子研究所，中国农业科学院农业信息研究所

主要编写人员：周园春、陈明奇、石蕾、刘佳、王姝、郭志斌、王丽娟、柳熠、佟继周、胡晓彦、吴林寰、郭学兵、孙毅、周启惠、陈曙东、郭雷风

目录

一、 区块链技术概述

（一）基本概念

区块链概念自 2008 年在比特币白皮书中被提出以来，引起全世界广泛关注，区块链技术是由多个互不信任的节点组成的分布式网络共同维护的、用于记录交易的难以篡改的分布式账本技术。

区块链普遍被认为是一种在数字世界进行可信交换的技术。因此，交易、区块、链就是区块链最基本的三个概念。交易（Transaction），即对账本的一次操作，导致账本状态的一次改变，如增加一条记录；区块（Block），一段时间内的所有交易和状态结果，对当前账本状态的一次共识，由共识算法生成新的区块；链（Chain），区块按顺序串链，是整个账本状态变化的日志记录。

区块链主要有两种分类方式。第一种分类方式根据发展阶段划分，分为区块链 1.0、区块链 2.0、区块链 3.0。第二种分类方式是根据对参与方访问权限和范围的不同进行划分，分为公有链、联盟链和私有链。公有链一般没有身份认证，任何实体都可以参与。公有链通常都有用于经济激励的原生数字货币，一般使用工作量证明（Proof of Work，简称 PoW）共识。联盟链是在一群有身份认证的参与者之间组建链。联盟链提供了一种实体间安全交互的方式，这些实体具有共同的目标但又无法完全信任彼此，比如基金交易、商品交易或者信息交易。私有链，仅使用区块链的技术进行记账，只有授权的部分人可以使用，信息不公开，一般为公司内部使用。

区块链技术最早来源于比特币，以比特币为代表的数字货币是区块链技术的典型应用。区块链技术后续被扩展应用于金融领域，例如支付清算、证券、数字票据等，同时也被延伸扩展至供应链、选举、司法存证、税务、物流、医疗健康、能源等多个垂直行业应用。

（二）研究范畴

区块链作为点对点网络、密码学、共识机制、智能合约等多种技术的集成创新，提供了一种在不可信网络中进行信息与价值传递交换的可信通道。当前，基于区块链的应用探索一直在加速推进，跨链、隐私保护、安全监管等区块链关键技术也正在成为研究热点。

1. P2P 网络

对等网络技术（Peer to Peer，简称 P2P）是区块链系统连接各对等节点的组网技术，学术界将其翻译为对等网络，在多数媒体上则被称为“点对点”或“端对端”网络，是一种建构于传输层的覆盖网络（overlay network）。不同于中心化网络模式，P2P 网络中各节点的计算机地位平等，每个节点有相同的网络权力，不存在中心化的服务器。

2. 密码学

哈希运算能够实现数据从一个维度向另一个维度的映射，通常使用哈希函数实现信息摘要，hash 函数碰撞概率极低，并且能够隐藏原始信息。区块链中哈希函数特性包括：函数参数为 string 类型，固定大小输出以及计算高效。常用的 hash 算法包括 MD5 和 SHA 系列

算法。

签名算法通过用私钥对信息进行加密变换以保证信息的不可否认性。当前区块链主要使用椭圆曲线数字签名算法（Elliptic Curve Digital Signature Algorithm，简称 ECDSA），该签名算法首先需要生成个人的公私钥对：(sk, pk) := generateKeys(keysize)，sk 私钥用户自己保留，pk 公钥可以分发给其他人；其次，可以通过 sk 对一个具体的 message 进行签名：sig := sign(sk, message) 这样就得到了具体的签名 sig；最后，拥有该签名公钥的一方能够进行签名的验证：isValid := verify(pk, message, sig)。

3. 共识机制

目前主要有几大类共识机制：PoW、PoS、DPoS、PBFT。

工作量证明（Proof-of-Work，简称 PoW），就是人们熟悉的比特币挖矿，通过计算出一个满足规则的随机数，即获得本次记账权，发出本轮需要记录的数据，全网其它节点验证后一起存储。可实现完全去中心化，节点自由进出，但挖矿造成大量的资源浪费，共识达成的周期较长，不适合商业应用。

权益证明（Proof of Stake，简称 PoS），是 PoW 的一种升级共识机制，根据每个节点所占代币的数量和时间，等比例的降低挖矿难度，从而加快找随机数的速度。PoS 还是需要挖矿，本质上没有解决商业应用的痛点。

股份授权证明机制（Delegated Proof of Share，简称 DPoS），类似于董事会投票，持币者投出一定数量的节点，代理他们进行验证和

记账，其整个共识机制还是依赖于代币，很多商业应用是不需要代币存在的。

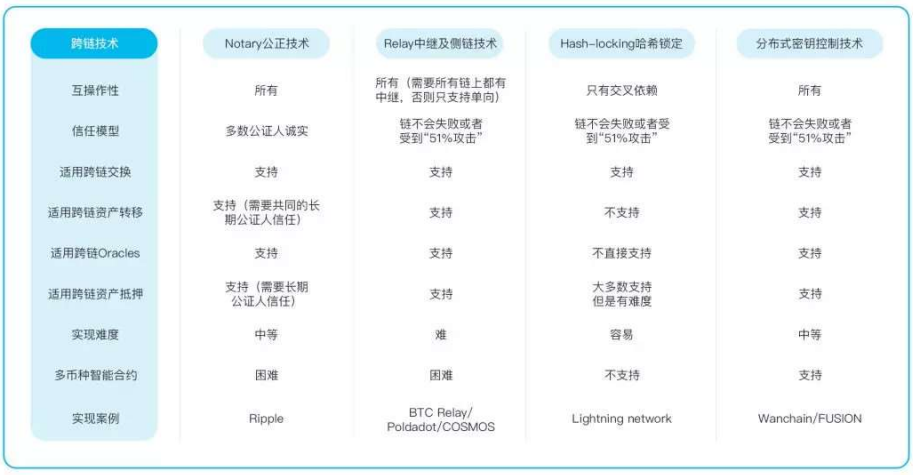
实用拜占庭容错算法（Practical Byzantine Fault Tolerance ，简称PBFT），是一种状态机副本复制算法，即服务作为状态机进行建模，状态机在分布式系统的不同节点进行副本复制，每个状态机的副本都保存了服务的状态，同时也实现了服务的操作，尽管可以存在多于 $3f + 1$ 个副本，但是额外的副本除了降低性能之外不能提高可靠性。

4. 智能合约

智能合约负责将区块链系统的业务逻辑以代码的形式实现、编译、部署，完成既定规则的条件触发和自动执行，最大限度的减少人工干预。智能合约包括事务处理和保存的机制，以及一个完备的状态机，用于接受和处理各种智能合约。比特币只支持简单的脚本语言，以太坊拥有图灵完备的智能合约语言，但是智能合约的拟定和部署十分繁琐，且容易受到攻击。

5. 跨链技术

跨链技术使区块链适合应用于场景复杂的行业，以实现多个区块链之间的数字资产转移，如金融质押、资产证券化等。目前主流的跨链技术包括：Notary 公证技术、Relay 中继/侧链技术、哈希锁定（Hash-locking）和分布式密钥控制技术。



跨链技术	Notary公正技术	Relay中继及侧链技术	Hash-locking哈希锁定	分布式密钥控制技术
互操作性	所有	所有（需要所有链上都有中继，否则只支持单向）	只有交叉依赖	所有
信任模型	多数公证人诚实	链不会失败或者受到“51%攻击”	链不会失败或者受到“51%攻击”	链不会失败或者受到“51%攻击”
适用跨链交换	支持	支持	支持	支持
适用跨链资产转移	支持（需要共同的长期公证人信任）	支持	不支持	支持
适用跨链Oracles	支持	支持	不直接支持	支持
适用跨链资产抵押	支持（需要长期公证人信任）	支持	大多数支持但是有难度	支持
实现难度	中等	难	容易	中等
多币种智能合约	困难	困难	不支持	支持
实现案例	Ripple	BTC Relay/ Poldadot/COSMOS	Lightning network	Wanchain/FUSION

图 1 区块链跨链技术对比

6. 隐私保护

通过加密交易来保证内容的机密性，只有利益相关方能够对齐进行解密并执行。业务逻辑（通常使用智能合约实现）也作加密处理确保数据安全，并且只有在运行的时候才能加载、破解。

零知识证明（Zero Knowledge Proof，简称 ZKP），指的是证明者能够在不向验证者提供任何有用的信息的情况下，使验证者相信某个论断是正确的。区块链的隐私将通过使用“零知识证明”得到进一步提升，除了声明的有效性，这个验证方法并不会透露出其他的信息。

身份混淆是将在区块链上交易用户的身份隐匿起来。Fabric 使用交易证书（TCerts）即每个交易的短期证书，满足一次一密、不可伪造、无关联性和可跟踪性。使得用户不仅以匿名方式参与到系统中，而且阻止了交易之间的关联性。

账本隔离是将具有不同隐私需求的账本，分别存放到不同的分布式账本上。使用多通道（子链模式），隔离账本。Fabric 利用多通道（Channel）的机制，实现账本隔离保护隐私性。Channel 代表了一个

私有的广播通道，保证了消息的隔离性和私密性，不同的链码（Chaincode）关联主体只知道自己 Chaincode 相关交易和执行交易验证，共识服务只接收相关主体的广播请求和执行对相关主体的消息送达，节点只记录与其相关的 Chaincode 的状态。

7. 安全监管

交易的隐私性固然重要，但是区块链商业用途也需要遵守一定的规章制度，确保监管方能够访问调查交易记录。安全监管包括数据上链内容监管以及基于区块链网络的监管两个部分。

通过区块链系统的特定数据内容快速检测发现和预警技术，有害信息的受控回滚技术，保障数据内容安全上链；通过区块链系统的多中心监管机制和分级治理机制，进行分布式、协作联动的监管中心有效治理；通过区块链行为的关联分析，结合网络流量特征分析，实现区块链匿名节点的身份追踪。

通过区块链交易分析和数据流转分析，实现交易的全流程监测和数据的全生命周期溯源；通过区块链的智能化分析和监控，实现面向特定目标的异常行为检测和预警；通过高效的联盟链监管技术，设计区块链系统的第三方中心监管机制，支持多层次的、联动协作的监管监测，实现对可疑交易和有害数据的快速监测和有效监管；针对重要行业，研究基于区块链的应用服务，实现透明化的行业监管监测。

二、 开放数据区块链现状

科学研究和学术创新的需求下，政府机构、科研单位、学会协会

和出版者共同推动科学数据共享，试图把开放存取信息资源的范围扩大到科学数据，构建一个科学数据开放共享体系。基于区块链的数据共享技术为数据共享的创新开拓了新的思路，我们调研分析了国际和国内在科学数据领域区块链应用的研究现状。

（一） 国际现状与分析

1. 美国

美国作为最早对开放数据进行科学管理与共享运动的领导者，2009 年，美国奥巴马政府推出了最重要的数据开放平台 Data.gov，这也是美国“开放政府”承诺的关键部分，涵盖了农业、气象、金融、就业、人口统计、教育、医疗、交通、能源等大约 50 个门类。2015 年，美国国家科学基金会（National Science Foundation，简称 NSF）发布《国家科学基金会公共获取计划》和《开放政府计划 3.5》以促进其资助科学数据等研究成果的开放获取。

2015年3月18日，美国政府发布了名为“今日的数据，明日的发现”公共获取计划，该计划概述了一系列活动，以增强源自NSF资助研究的科学出版物和数字化科研数据的公共获取。该计划对版本记录、同行评议学术期刊中已经接受的最终版手稿、评审过的会议论文集或汇报中的文章、数据管理计划等作出了要求。

在科学数据区块链研究方面，2015 年 9 月美国麻省理工学院数字货币计划（Digital Currency Initiative，简称 DCI）成立，该计划汇集了高科技行业的资深人士，加密程序员，教职员工，学生和研究科

学家，重点研究出版物和区块链技术的开源应用并进行试点测试。2018 年 12 月推出了一个加密货币与区块链跨领域研究同侪评论——Cryptocurrency Research Review（(加密货币研究评论），以填补加密货币行业学术领域仍缺乏共识机制的空白。

2018 年 9 月，美国圣地亚哥超级计算机中心（San Diego Supercomputer Center，简称 SDSC）联合 IBM、Intel 等公司成立了区块实验室（BlockLab），专注区块链在科研和工业领域的应用。目前的主要成果是建立了开放科学链（Open Science Chain，简称 OSC），该项目得到美国国家科学基金会 81.8 万美元的资助。OSC 采用了 Hyperledger Fabric 底层搭建区块链平台，同时提供中间件和 Web 门户。科研人员可以通过 Web 门户进行数据的注册和验证，并更新、检索数据，提供历史记录进行追踪。

2. 欧盟

欧洲各国同样也针对政府与企业出台了不少科学数据管理与共享的相关政策，旨在保障数据的质量与精度。

2008年，欧盟开始第七研发框架计划（7th Framework Programme，简称FP7），启动了研究数据开放先导性计划（Open Research Data Pilot，简称ORD Pilot)，要求FP7资助的科研项目成果实施开放存取，包括与欧洲研究委员会（ERC）的合作项目及2010年开展的欧洲开放获取基础设施研究项目（Open Access Infrastructure Research for Europe，简称OpenAIRE）项目等。此外，欧盟还成立了科学数据高级专家组，从事科学数据相关工作的环境研究。

2014年，欧盟启动了为期两年的促进欧洲研究开放科学培训项目，旨在帮助科研人员、研究生、图书馆员和其他利益相关者在其现行的科研方法中融入科研数据开放存取的方法和途径。

2015年6月，欧盟召开“开放创新时代”会议，提出制定科学数据的管理、交互和质量相关标准，以促使2020计划中对科研数据开放的充分重视。

2016年，欧盟启动了“欧洲开放科学云”计划，致力于为欧盟170万科研人员和7000万从事科技创新活动的在职人员创造一个共同的虚拟在线环境，将首先在欧洲乃至全球合作伙伴的科技界实施，然后逐步向其它公共部门和各行各业用户拓展。

在科学数据区块链研究方面，欧洲机构共同认为区块链对其他领域具有巨大潜在影响，开展了“欧盟区块链：区块链促进产业转型”（Blockchain4EU）研究项目。该项目是对基于区块链和其他分布式账本技术在各行业领域中现有、新兴和潜在应用的前瞻性社会技术探索。

2019 年 9 月欧盟科学中心在区块链论坛上发布报告《Blockchain Now And Tomorrow》（区块链：当前和未来）。该报告汇集了欧盟委员会的科研机构在不同学科的研究成果，以探索区块链和分布式账本技术的多种潜力。

（二） 国内现状与分析

我国对科学数据共享与开放获取重视较早，建立了一些科学数据共享平台，制定了一系列数据共享的政策法规。早在 2001 年，我国

就启动了“科学数据共享工程”，在资源环境、农业、人口与健康、基础与前沿等领域共 24 个部门开展科学数据共享工作，并起草和制定了《科学数据共享条例》《国家科技计划项目科学数据汇交办法》等。

随后科技部发布了《2004-2010 年国家科技基础条件平台建设纲要》《科学数据共享工程“十一五”建设规划》《中华人民共和国政府信息公开条例》《国家科技资源共享服务平台管理办法》等一系文件加强对科学数据开放共享的政策支持。

2018 年 3 月，国务院办公厅正式印发《科学数据管理办法》，旨在加强和规范科学数据管理，保障科学数据安全，提高开放共享水平，更好支撑国家科技创新、经济社会发展和国家安全。2019 年 2 月，为贯彻落实国务院办公厅颁发的《科学数据管理办法》，中国科学院印发了《中国科学院科学数据管理与开放共享办法(试行)》，并启动中国科学院科学数据中心体系建设，加快推动科学数据汇交、管理、共享和服务的工作。

中国科学院在 1982 年开始启动“科学数据库及其信息工程”，开始了中国科学院科学数据库的建设工作。1986 年 6 月，国家计委批准该项目纳入国家重点建设计划。2001 年，中国科学院启动信息化专项，科学数据库作为科研信息化的重要工作之一列入了专项支持。2011 年和 2016 年，科技数据资源整合与共享工程和科学大数据工程分别纳入到中国科学院“十二五”和“十三五”信息化专项重点支持的项目。

在科学数据区块链研究方面，虽然近几年区块链技术在我国得到

了快速发展，但在科学数据研究领域，对于构建基于区块链的科学数据安全体系还处于探索阶段。随着我国科技创新能力和投入不断增强，科技创新能力不断提升，科学数据呈现出快速增长趋势，探索科学数据+区块链应用的相关问题是为科学数据开放共享寻找新的方法与途径，这对推动实施国家大数据战略，推进数据资源整合和开放共享，保障数据安全具有极其重要的意义。

三、 ODC 联盟链体系

（一）ODC 体系架构

开放数据联盟链（OpenDataChain，简称 ODC）是面向科学数据中心构建的数据联盟链，致力于为科技成果发布、确权和开放共享提供技术保障。ODC 通过为每个数据中心创建联盟链节点，支持数据所有者、数据发布者、数据唯一标识、数据版本、实体数据摘要等数据上链。用户通过 ODC 可以进行数据离线或在线完整性校验，在数据仍存储在科学数据中心本地的情况下保证科学数据链上安全共享、一致性验证、数据追溯、版本管理、数据确权、使用记录和贡献量统计。

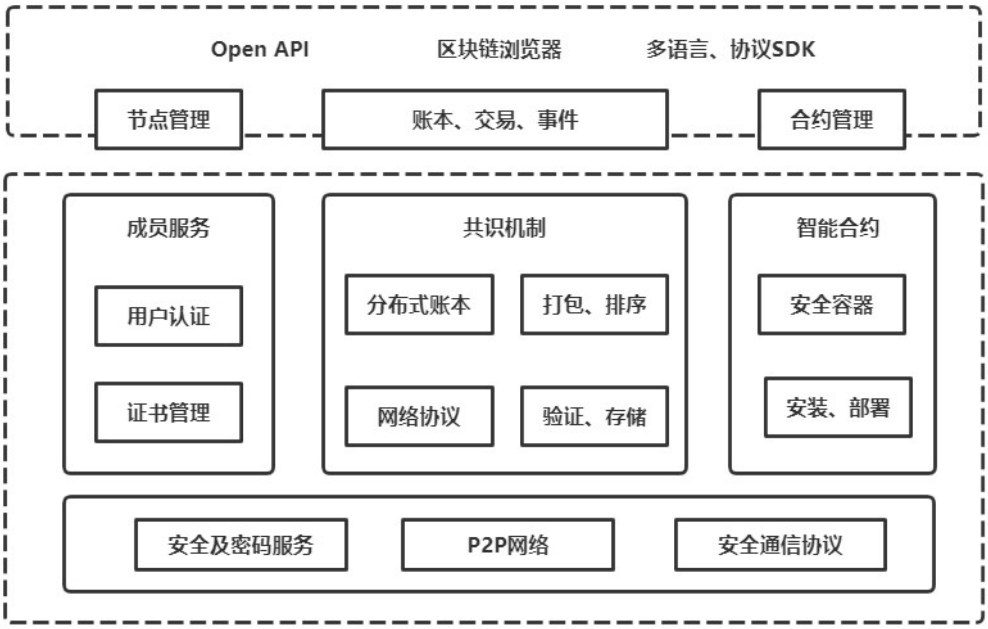


图 2 ODC 技术架构图

1. 节点与账本模型

区块链的网络由节点构成，每个对等节点都拥有一份账本（Ledger）和智能合约（Smart Contract）的拷贝。网络 N 由对等节点（Peer）P1、P2、P3 构成，每个节点都保存着属于它们自己的分布式分类账本 L1 的实例，P1、P2、P3 都使用码链（ChainCode）S1 来访问自己保存的账本。

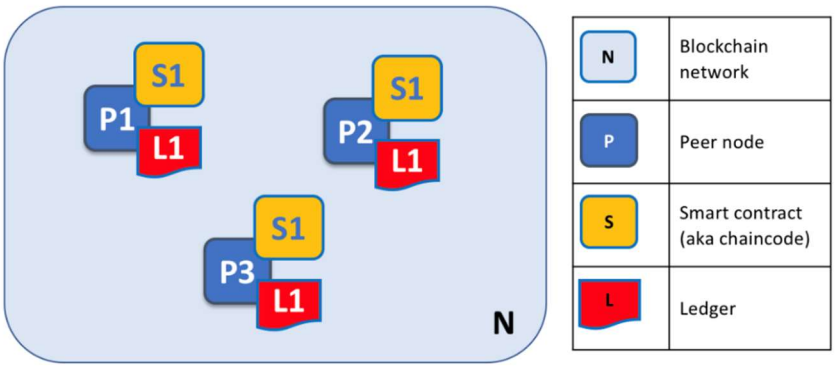


图 3 区块链网络和节点示意图

开放数据联盟链 ODC 把数据账本模型映射为“KEY-VALUE”结

构，为数据的存储提供更好的伸缩性。还定义了标准的持久化服务接口（Persistent Service Interface），能够适配不同的数据库存储引擎以满足不同数据中心的业务需求。

2. 密码算法

密码算法的选择需要满足安全和合规的要求，同时面临源自实际业务场景的多样性要求。开放数据联盟链 ODC 在密码方面的关键任务是设计可插拔的密码框架，定义标准的安全服务接口（Security Service Interface）。新版本默认支持国密算法以满足国家商用区块链安全性要求。

3. 共识机制

共识协议的核心任务是保障区块链网络中有效节点的状态一致性。另外在选择共识协议时，还需要考虑业务场景中的安全性要求、时效性要求和节点规模等诸多因素。开放数据联盟链 ODC 在共识协议方面的关键任务是设计可插拔的共识框架，解耦共识协议与数据账本模型，以满足不同数据中心业务场景的多样化需求。目前采用的是 PBFT 协议实现，后续会加入更多共识算法模型实现。

4. 接入机制

终端接入是开放数据联盟链 ODC 的基本功能，在确认终端身份的同时提供连接节点、转发消息和隔离共识节点与客户端等服务。ODC 服务网关接入为保障安全隐私，一方面通过具有隐私保护功能的密码算法和协议，来进行隐藏端到端身份信息，脱敏处理数据信息，防止无权限客户端访问数据信息等操作；另一方面，通过 ODC 服务

网关的隔离作用使外部实体无法干预内部共识过程，保证共识和业务之间的独立性。ODC 服务网关支持 HTTP、TCP、UDP、MQTT 等通讯协议，保证不同设备信息上链能够适配区块链节点的 API，实现各节点在不同协议之间的互操作。开放数据链 ODC 还提供跨平台的客户端降低区块链的使用成本。

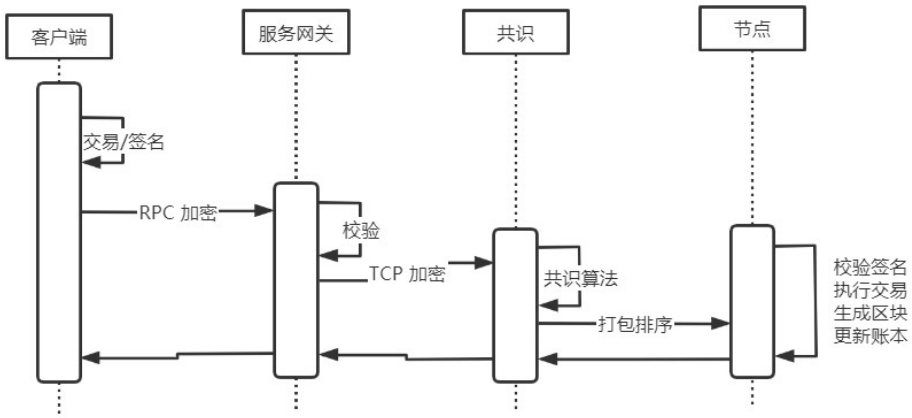


图 4 ODC 接入机制

5. CA 认证

开放数据联盟链 ODC，采用联盟链技术，所有数据中心需要注册会员服务来获取身份，从而进行网络访问或数据上链等服务。注册服务可以控制并管理开放数据联盟链 ODC 参与者的权限，身份管理服务能够提供管理保障，根据 ODC 参与者的身份角色来进行授权，审核服务能够帮助授权用户之间的数据建立关联关系并监控其运行情况。

6. 智能合约

区块链技术为智能合约提供了安全可信的执行环境，促成了智能合约概念的实现。智能合约是由事件驱动的、具有状态且运行在一个可复制、可分享的账本之上并能够保管账本上资产的程序，其目的是

让一组复杂的、带有触发条件的数字化承诺能够按照参与者的意志，正确执行。智能合约不仅可以接收和储存价值，也可以向外发送信 息和价值，整个过程可以在无中心，无信任的前提下，自动化、智能化的执行。使用安全容器来存储、运行智能合约，智能合约的开发是区块链应用的主要功能，开放数据联盟链 ODC 提供完备的智能合约集成开发调试环境，并提供数据标识、元数据版本管理、数据上链、历史数据追溯等公共合约，可以减轻数据中心开发压力，以更便捷快速的方式使用智能合约。

（二）ODC 建设原则

开放数据联盟链 ODC 在建设过程中，遵从简单易用、灵活扩展、安全可靠、高效运维和合作开放的设计原则。

1. 简单易用

为科学数据中心用户实现自动化配置、部署区块链应用，并提供区块链全生命周期管理，让数据中心用户能够更加容易的使用区块链技术，专注于科学数据应用的开发和创新，极大的降低区块链的使用门槛。

2. 灵活扩展

开放数据联盟链 ODC 在设计过程中，通过分层的架构设计、可插拔的模块设计和面向接口的软件设计，将网络构建、加密、共识、资源管理、用户管理、运维管理等功能模块分开设计实现，并通过云和容器技术，实现计算资源、存储资源和网络资源的动态扩展。

3. 安全可靠

开放数据联盟链 ODC 具有有效的防篡改机制、清晰的崩溃容错

安全边界、安全的数据管理和隔离机制，具备高速的网络连接能力和海量的数据存储能力，提供完善的用户管理、权限管理和密钥管理，提供可靠的网络安全能力、分类分级的故障恢复能力和运营安全能力。

4. 高效运维

开放数据联盟链 ODC 提供故障分类分级报警体系和运维方法，提供必要的运维接口和运维授权能力，为智能合约和链上应用提供全方位的资源监控能力。

5. 合作开放

开放数据联盟链 ODC 专注于底层技术和平台的服务能力建设，和各数据中心携手合作，共同打造科学数据中心的区块链解决方案和区块链生态，共同推进区块链场景落地，帮助数据中心解决数据共享、验证、版本追溯等共性问题。

（三）ODC 技术特点

ODC 基于区块链和数据标识融合创新技术，以数据生产要素为核心，以数据确权、数据追溯等应用需求为牵引，推动数据可信共享和确权归属，为营造积极开放的数据生产流通使用环境提供技术支撑。

1. 部署方式灵活

ODC 联盟链支持科技云平台部署，同时也支持异地部署。隔离了底层云平台的差异性，为用户提供了一致的使用体验。

2. 安全隐私保障充分

ODC 联盟链支持 Fabric 的多通道数据隔离、国密算法等全部的安全和隐私保护功能，同时也提供在操作系统、云平台等层级的全方位安全加固保障。另外 ODC 提供异地部署，可以进一步帮助用户解

决数据共享和安全保护的矛盾。

3. 创新的 ODC 客户端

ODC 客户端为中心提供了本地数据验证、元数据上链等功能，用户需要使用 ODC 客户端自行创建、导入和管理区块链的各种证书和私钥。 通过使用 ODC 客户端，用户管理证书和私钥不仅更加方便而且更加安全。

4. 在线管理功能丰富

ODC 客户端提供了完全可视化的区块链在线管理功能。结合管理用户私钥证书的 ODC 客户端，用户可以轻松快速地在平台上创建节点，进行各种区块链和安全参数的设置。

5. 安装部署时间短

为了让安装部署时间不随节点数目线性增长，设计了一个自动化的引擎，仔细梳理了区块链的安装配置流程，将节点的安装部署高度自动化。

四、ODC 最佳实践

ODC 联盟链通过分布式记账方式、多节点共识机制、非对称加密和智能合约等多种技术手段，在数据中心之间建立强大的信任关系和价值传输网络，使其具备分布式、不可篡改、价值可传递和可编程等特性。在应用方面，ODC 应用场景不断铺开，在数据溯源、确权凭证、可信共享、版本管理等领域持续探索。

（一）ODC +数据溯源

1. 科学数据亟需新的溯源模式

在大数据背景下，科学数据成为重要的数字资产，这些科学数据可以分为两类：一类是原始数据，另一类是对原始数据进行加工后生成的数据。在科学研究过程中，需要对数据进行一系列处理，由于中间过程缺少必要的透明度，很难判断其来源和可靠性。科学数据溯源涉及生成一段数据或事物的实体、活动以及人的信息，可用于评估数据或事物的质量、可靠性或可信度。传统的数据溯源系统一般采用中心化方式存储数据，在遭受内部、外部攻击时，存在单点故障等风险问题，威胁数据安全性。

2. 区块链技术保障数据可信溯源

科学数据在开放共享的大趋势下，在学术、科学及教育等整个生命周期内对科学数据进行数据溯源，通过重现及组织数据以供当前科研活动获取，进而用于未来再发现及再利用。科学数据溯源目的是为了实现数据的追溯，确保单个溯源记录的真实性及记录顺序不可被修改。而区块链具备数据的防篡改、可追溯特性，两者的结合对于科学数据溯源具有重要作用。科学数据标识技术将为科学数据提供互联网环境下访问的便利途径，确保数据溯源信息互联互通，实现对数据全生命周期的跟踪与溯源。

将合约执行的关键点通过账本进行存证。包括合约的调用者、调用时间、运行节点情况、合约输入、合约关键点、合约返回值等。通过监控数据流转的全生命周期，保障科学数据可信流通可溯源，支持

数据共享开放过程可信、可管、可控。

3. ODC 科学数据溯源

科学数据溯源是通过记录数据流转的信息来实现数据溯源，但是记录信息本身也是数据，同样存在安全问题。为了防止有人恶意篡改数据溯源记录的相关信息，实现对溯源记录信息的分布式可信存证，利用区块链技术有效地防止恶意篡改联盟链中的溯源记录，对科学数据在生命周期内修改行为的记录，按时间先后组成溯源链，通过溯源链记录数据上链、更新等数据全生命周期信息。

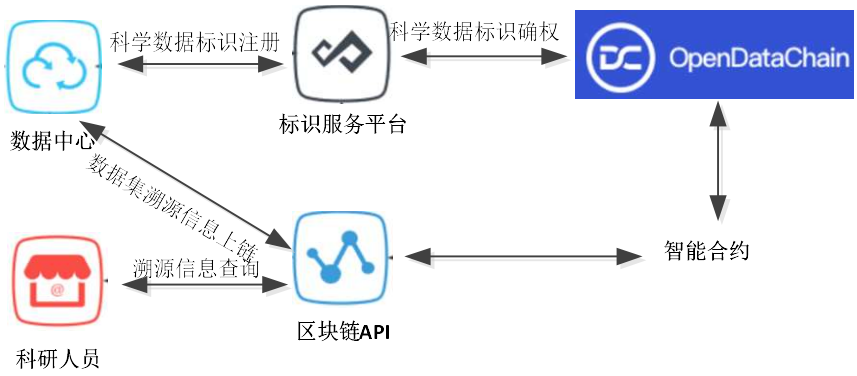


图 5 ODC+数据溯源应用场景

科学数据溯源上链流程：

(1) 科学数据上链数据准备。上链数据包括：数据集标识、版本信息、元数据、数据发布者信息、关联标识、关键字、数据集签名等。

(2) 数据集标识注册。数据集标识是科学数据的永久唯一标识，可实现科学数据的定位、追溯、引用、统计与评价。数据中心首先向标识平台申请标识前缀，提交数据中心名称、统一社会信用代码、联系人等信息，经标识平台审核同意后分配

标识前缀。然后数据中心向标识平台注册数据集标识，通过标识解析可以获取数据集的 URL 和标识信息等。

(3) 数据集签名。对数据集中每个文件内容进行签名，哈希算法包括:MD5、SHA-256、SM3 等。更新数据集或数据集合时，将跟踪所有上链数据的更改，包括该数据集合中每个文件的签名，从而使用户可以查看该数据集随时间的详细演变历史。

(4) 调用智能合约写入数据。数据中心调用智能合约将科学数据上链信息写入到区块链账本。在正式写入账本前，智能合约将验证标识平台的身份及权限信息，验证通过后在区块链账本记录科学数据上链信息。

科学数据溯源链可以支持研究人员有效地验证数据集的真实性，查看历史数据并验证所有权信息，跟踪来源并安全存储有关科学数据的元数据和验证信息，以可验证的方式跟踪数据更改,以可独立验证的方式促进数据重用，从而推动科学数据开放共享的可持续发展。

（二）ODC +数据确权凭证

1. 科学数据确权需求分析

2020 年 4 月 9 日，中共中央、国务院下发《关于构建更加完善的要素市场化配置体制机制的意见》，首次正式降数据纳入生产要素范围，同时提出要加快培育数据要素市场，推进政府数据开放共享。在促进数据流动的过程中首先要进行数据确权，将采集的数据进行分类整理，确认并分离所有权和使用权。

数据确权是区块链的最基本应用，区块链作为解决数据确权、定价交易的必要手段，所有溯源、交易等其他应用都要建立在数据确权的基础上。科学数据确权一般是确定科学数据的权利人，即谁拥有对数据的所有权、占有权、使用权、受益权，以及对个人隐私权的保护责任等。在研究科学数据确权时，主要聚焦于科学数据的所有权，即科学数据归属问题。

科学数据确权的发展可以加速推进交易生态的形成，交易生态的扩大倒逼更完善的确权技术和服务能力。因此区块链可以从存证、确权、交易、交换、信用服务和溯源等方面加速数字化转型。

2. 科学数据标识区块链应用

区块链可对科学数据进行唯一标识，确保存储在账本上的数据信息不可篡改，可追溯，并且以一种去中心化的方式在全网获得共识。区块链将科学数据封装为可上链的数据对象，通过唯一的赋码机制确保数据唯一性，为每个科学数据做可信认证。基于数据标识和区块链技术相结合的科学数据确权方案，具有确权的公平性、完整性和不可欺骗性。科学数据标识上链，通过数据标识解析技术，唯一确定科学数据集。

3. ODC 科学数据确权

基于 ODC 对科学数据进行确权是通过在区块链上存储数据的唯一标识和所有权信息实现的，通过数据集标识可以唯一确定一项数据集。

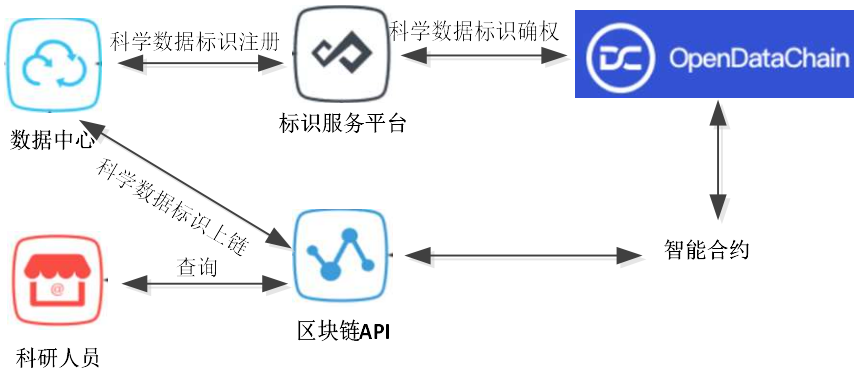


图 6 ODC+数据确权应用场景

数据中心在标识服务平台进行科学数据标识的注册，注册成功后，每个数据集将具有唯一标识，通过标识服务平台可以解析数据集唯一标识，并获得数据集对应的地址信息。为了确保数据集的归属和所有权，数据中心可以将数据集标识信息通过区块链 API 上链到 ODC 联盟链中。ODC 联盟链可以通过标识服务平台进行数据确权，并提供区块链 API 为广大科研人员提供数据查询服务。

（三）ODC +数据可信共享

1. 科学数据可信共享需求分析

科学数据是科技创新、经济发展和国家安全的重要战略资源。科学数据的有序管理、安全保障、开放共享与有效应用具有重要的科学研究和社会经济意义。目前，我国科学数据共享管理工作已初见成效，但仍有很多工作亟待推进。尤其是在跨部门、跨平台、跨学科的数据交换共享与综合治理过程中，针对科学数据及其元数据信息的规范性、一致性、可信度和可溯源性等问题尚未形成成熟的技术方案与系统平台。同时，科学数据的应用情况追踪与最终成果效益评价等方面的技

术基础也较为薄弱，目前科学领域重大项目和其他来源科学数据应用情况的追踪仍主要基于科研团队的主动上报反馈以及数据中心团队的人工整理与统计分析，未能有效触及所有使用了相关科学数据的科研成果与行业应用，也耗费较多的人力成本。

2. 区块链技术保障数据可信共享

区块链技术已在其他领域与行业中形成了若干有效应用示范，其应用场景与科学数据的可信共享中的需求具有相似性，因此基于区块链技术有望在科学数据的综合治理、发布共享、可信性验证及应用效益跟踪中得到有效应用，解决参与交换共享的科学数据信息的安全性、规范性、一致性、可溯源性等问题，解决对科学数据应用情况的挖掘分析与统计评价。

区块链技术为科学数据提供信息共享能力，包括支持建立若干子链，实现子链中各节点间科学数据元数据的上链、共享与自动同步；支持子链加入开放共享联盟，使该子链数据与信息进入更广泛的共享范围，并基于区块链的技术特征，保证数据信息的长期安全可用。同时，支持不少于 DOI（Digital Object Identifier）、PID（Persistent Identifier）和 CSTR（China Science & Technology Resource）等三类科学数据唯一标识符的解析能力和基于科学数据唯一标识符对所共享科学数据提供引用情况统计分析功能。

针对一些涉及数据安全和隐私保护问题的科学数据，科研机构不能直接将原始数据开放共享给科研人员，科研人员可以通过提交数据分析算法的方式只获取数据分析结果，而不需要直接下载原始数据，

同时也需要保证提交的算法可以如约执行。科研人员也可以通过直接提交数据分析代码到科研机构的数据中心以合约的方式运行，保证原始数据不脱离科研机构物理控制，只返回数据分析结果，同时通过智能合约和账本保证算法如约执行。通过上述两种方式，打消科学数据开放共享的顾虑，全面提升科学数据研究价值与社会价值。

3. ODC 可信数据共享

基于 ODC 对科学数据进行可信共享是通过在区块链上存储科学数据元数据实现的，通过在 ODC 联盟链中开放共享元数据，原始数据存储在数据中心本地，实现科学数据的可信共享。

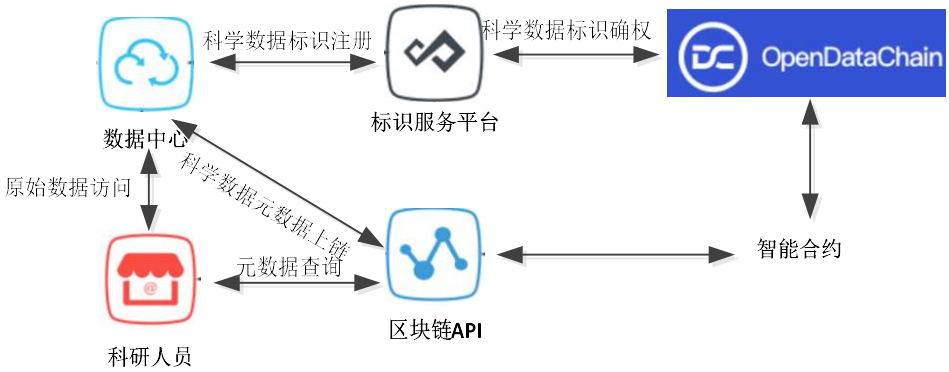


图 7 ODC+数据可信共享应用场景

科学数据中心在成功注册科学数据标识后，数据集相关元数据信息同步注册到标识服务平台，通过数据标识可查询到具体元数据信息。数据中心将数据集标识和元数据等信息发布到区块链上，在正式写入账本前，智能合约将验证数据中心的身份及权限信息，验证通过后在区块链账本正式记录该信息。元数据信息正式写入账本后，数据中心可以通过输入数据集标识作为参数调用智能合约查询元数据信息，如果其他数据中心也注册了引用该数据集的论文标识，通过智能合约也

可以查询到数据引用信息。

（四）ODC +数据版本管理

1. 数据版本管理新方式

版本管理已广泛应用于软件开发、文档管理、数据管理等领域。目前多数的版本管理集中在项目内部或小范围应用中，在科学数据开放共享的范畴内亟需解决版本管理问题，其中数据版本管理的核心是设计用来记录和跟踪数据集版本变化的系统。利用区块链的可追踪性和不可篡改性帮助数据中心掌握数据集版本管理的记录。

2. 区块链记录数据版本信息

数据版本管理就是记录存储数据变化的历史，以实现回溯，再现，主要强调管理的数据对象上发生了变化，记录和跟踪数据集的变化。

在使用基于区块链技术的科学数据集版本管理系统时，为了方便用户查询数据以及操作记录，结合智能合约设计了用户合约，存储和用户相关的数据操作记录的详细信息。

区块链在数据写入和变更时，都会记录操作的时间戳，使得各种操作都有具体时间可查。例如，在版权保护方面，作品的发布时间是处理版权纠纷时的关键参考依据，区块链的时间戳技术可以为纠纷处理提供强有力的佐证。区块链可追溯、不可篡改的特性，适用于证据链的保存，使得最终提供的不再是单点证据，而是完整的证据链，有利于增强证据的可信性。

3. ODC 科学数据版本管理

基于 ODC 对科学数据进行版本管理是通过在区块链上存储科学数据版本记录实现的，通过在 ODC 联盟链中记录数据集版本信息、签名、时间戳等内容实现科学数据版本的管理。

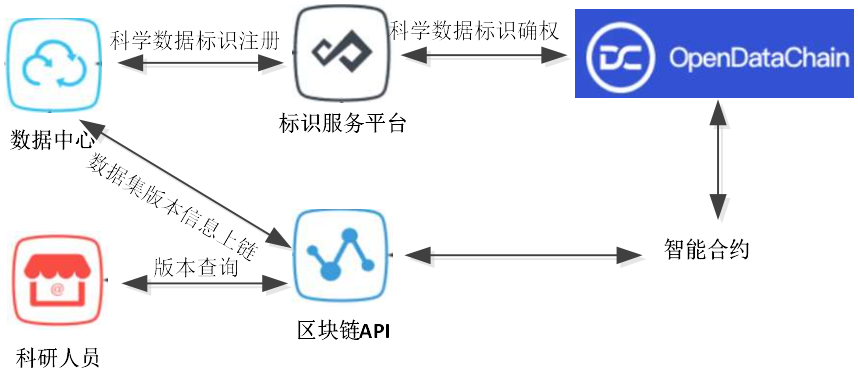


图 8 ODC+数据版本管理应用场景

科学数据中心在成功注册科学数据标识后，将数据集标识、版本信息、签名、时间戳等发布到区块链上，在正式写入账本前，智能合约将验证数据中心的身份及权限信息，验证通过后在区块链账本正式记录该信息。数据集版本信息正式写入账本后，数据中心可以通过输入数据集标识作为参数调用智能合约查询具体数据集版本记录信息。这些记录都存储的ODC联盟链上，通过查询对应数据集，可以获取数据集版本变化的记录过程。ODC联盟链一方面是确保单个数据集版本记录不是伪造的，另一方面确保记录顺序不会被修改。

五、发展建议与展望

（一）加强关键技术研究，推动开源自研平台建设

充分发挥区块链原始创新能力，密切关注国内外技术发展动态，加快推进包括共识机制、密码学、互操作、隐私保护等在内的核心关

键技术研发，开展产品开发和集成测试，适度推进标准制定；

强化区块链发展条件保障，充分利用国家科技计划（专项、基金等）和市场的支持，开展区块链重要基础理论研究、关键核心技术创新和重大技术试验验证；

重视区块链技术生态培育，整合政府、高校、企业、行业组织等资源优势，建设完善区块链创新生态；

鼓励区块链开源社区发展和开源自研平台建设，引导“区块链+”科研生态圈的发展，促进科研区块链开源社区构建。

（二）提升开放共享能力，创新科学数据区块链应用

建立健全科学数据整合共享机制，加强与科学数据开发利用以及共享政策的相互协调；

完善科学数据共享的配套服务体系，提高面向研发创新、科技孵化、成果转化的专业性科学资源机构服务能力；

积极拓展区块链技术在科学数据开放的应用场景，提高应用的深度和广度，使区块链技术与公共服务更加紧密、有机结合；

鼓励科研机构和企业共同开发区块链应用产品，通过项目实践不断加强对区块链技术的掌握，以助推区块链技术进步；

加快区块链和人工智能、大数据、物联网等前沿信息技术的深度融合，推进多领域协作融合发展，不断增强自主创新能力。

打造面向科学数据的共享开放平台和生态，完善平台数据全生命周期管理能力，丰富和完善平台数据来源、相关数据分析算法和工具，

吸引更多专业科研人员以及专业数据分析人员参与生态建设。

（三）强化科学数据监管理念，提升安全防护水平

加强和完善科学数据安全监管方面的法律制度建设，明确各方职责分工，加强监管部门间的协调性，提高监管机构内部建设和业务水平；

明晰数据权属，提升数据主体的保护意识，保障数据主体的参与性；加强知识产权保护，对科学数据使用者和生产者的行为进行规范；

积极鼓励通信、金融、医疗、教育、交通等重点领域制定行业自律规范，创造数据共享平稳健康发展的生态；

完善区块链技术应用立法，降低科学数据开放风险，明确区块链应用中的管理架构、职责分工、安全保障、推广应用，兼顾技术价值与数据安全与监管间的协调，明确权限与监管职能。

（四）扩大国际合作范围，提升国际影响力

在全球普遍重视大数据发展和数据开放共享的背景下，准确把握大数据时代科学数据发展趋势，尤其是当前新基建正不断加速，科学研究对于数据的存储、管理、传输都有了更高的要求。这些高要求，激活了更多科学数据区块链技术的应用。

我们要紧跟国际发展步伐，充分利用国内外已有条件，建设更多的数据驱动科学发现的领域科学大数据社区。积极参加国际合作项目，加强与相关科研机构、重大科学数据组织的交流与合作，充分借鉴国

内外先进经验和成熟做法，运用区块链、大数据、云计算等技术手段，全方位提高我国科学数据工作水平，提升国际影响力。

附录

术语

- **联盟链**

联盟链只允许特定某个群体的成员和有限的第三方参与，其参与者是被提前筛选出来或者直接指定的，数据库的读取权限可能是公开的，也可能像写入权限一样只限于系统的参与者。

- **共识机制**

共识机制是区块链网络用来达成交易确认共识的协议，确保交易顺序的一致性、账本一致性、节点状态的一致性等。

- **智能合约**

智能合约负责将区块链系统的业务逻辑以代码的形式实现、编译、部署，完成既定规则的条件触发和自动执行，最大限度的减少人工干预。

- **国密算法**

国家密码局认定的国产密码算法，即商用密码。主要有 SM1、SM2、SM3、SM4。

- **CA 认证**

CA 认证，即电子认证服务，是指为电子签名相关各方提供真实性、可靠性验证的活动。

- **跨链技术**

跨链技术指实现各区块链之间的原子交易、资产转换、区块链内部信息互通的技术，是连接各种类型区块链的桥梁。

参考文献

1. 黄如花, 陈闯. 美国政府数据开放共享的合作模式[J]. 图书情报工作, 2016, 60(19): 6-14.
2. Peter Wittenburg(GEDE) Blockchain Technology and FAIR Digital Objects - what is needed in science? V2.0, Status of November, 2019.
3. 王姝，晏敏，刘佳，周启惠，郭志斌，王雅哲，周园春. 基于区块链的科学数据标识技术创新应用模式[J]. 数据与计算发展前沿，2019，1(2): 62-74.
4. 田野, 王姝, 郑宁. 物联网标识关键技术及应用[M]. 中国质检出版社, 2019.
5. 国家科技基础条件平台中心. GB/T 32843-2016 科技资源标识[S]. 中国标准出版社，2016.
6. Hana Pergl Sustkova, Kristina Maria Hettne, Peter Wittenburg, Annika Jacobsen. FAIR Convergence Matrix: Optimizing the Reuse of Existing FAIR-Related Resources[J]. Data Intelligence. 2019,1(11):158-170.
7. Androulaki, Elli，Barger, Artem，Bortnikov. Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains[C]. EuroSys 2018 conference,2018.
8. Cachin C. Architecture of the Hyperledger blockchain fabric[C]. Workshop on Distributed Cryptocurrencies and Consensus

Ledgers, 2016.