

文章编号: 1000-5641(2019)02-0077-20

农业大数据基础设施开发的参考模型方法

赵志明¹, 廖晓锋¹, 王晓玲², 阮春晟³, 朱轶峰³, 丰东升⁴

(1. 阿姆斯特丹大学 信息技术研究所, 阿姆斯特丹 1098 XH;

2. 华东师范大学 上海市高可信计算重点实验室, 上海 200062;

3. 上海农业物联网工程技术研究中心, 上海 200050;

4. 上海市农产品质量安全中心, 上海 200003)

摘要: 连续的环境观测数据是复杂系统, 如天气建模、智慧型科技创新和宏观系统层科学研究的重要驱动力; 数据源、应用领域和应用需求的不同, 使得管理千差万别的实时数据, 并提供有效查找、识别、融合和重用服务变得极富挑战性. 数据基础设施通过为数据提供全生命周期的管理服务, 为上层数据驱动的科学研究和应用创新提供标准化的数据查询、读取和处理服务. 然而数据基础设施的建设常局限于特定领域和历史遗留的前期系统, 缺少统一参考, 以及基础设施之间数据和服务的互通困难, 极大地制约了上层应用, 特别是跨学科研究发展的需求. 针对大数据基础建设中的上述挑战, 借鉴欧洲环境大数据参考模型在环境大数据基础设施建设方面的经验, 提出了农业大数据参考模型, 为我国农业大数据基础设施建设和所涉及的数据互通挑战提供参考. 两个案例展示了所提出的农业大数据参考模型在大数据基础设施的需求分析改进、历史遗留系统的数据互通接入等方面的作用.

关键词: 农业大数据; 参考模型; 基础设施

中图分类号: TP391 **文献标志码:** A **DOI:** 10.3969/j.issn.1000-5641.2019.02.009

A Reference Model approach for developing agricultural big data infrastructures

ZHAO Zhi-ming¹, LIAO Xiao-feng¹, WANG Xiao-ling², RUAN Chun-sheng³,
ZHU Yi-feng³, FENG Dong-sheng⁴

(1. *Informatics Institute, University of Amsterdam,*
1098 XH Amsterdam, Netherlands;

2. *Shanghai Key Laboratory of Trustworthy Computing, East China*
Normal University, Shanghai 200062, China;

3. *Shanghai Agriculture Internet of Thing (IoT) Development and*
Research Center, Shanghai 200050, China;

4. *Shanghai Agriculture Products Quality and Safety Center, Shanghai 200003, China)*

Abstract: Big data infrastructures provide services for the management of data over the course of their lifecycle, and offer users the ability to effectively discover and access

收稿日期: 2018-03-22

基金项目: 上海市科技兴农重点攻关项目(沪农科攻字(2016)第2-1号)

第一作者: 赵志明, 男, 博士, 研究员, 研究方向为云计算、大数据管理、软件工程.

E-mail: zmzhao@ieee.org.

data for different application purposes. These emerging infrastructures essentially enable system-level data-centric research; third-party innovation, however, often requires data from different sources. The construction of big data infrastructures faces important interoperability challenges arising from the diverse nature of data acquisition, annotation, and identification performed in different research domains. Moreover, the evolution of different infrastructures is often driven by the specific interests of researchers, in their respective domains, and the constraints of legacy technology. The ENVRI Reference Model (ENVRI RM) is an output of the EU H2020 ENVRI and ENVRI PLUS project, targeting the aforementioned challenges in the context of environmental sciences by modeling environmental research infrastructures with a multi-viewpoints framework; these viewpoints include science, information, computation, engineering, and technology. Each viewpoint describes concrete aspects of a system definition and forms a mechanism to improve the interoperability across the whole system as well as alignment with existing legacy systems. The challenges encountered in the Shanghai Agricultural Big Data Infrastructures construction work are similar to those detected in the ENVRI RM, which provides an ideal place to test the generalizability of the ENVRI RM to other domains. Using the ENVRI RM as a reference, this paper presents an Agricultural Reference Model, which includes the five aforementioned viewpoints, but with consideration of the specifics of the agricultural domain, to address the problems encountered in revising and upgrading the Shanghai Agricultural Big Data Infrastructures. Two use cases are introduced to demonstrate its effectiveness. One is to improve the requirement engineering procedure with the community and role context captured using the Agricultural Reference Model. The other is to upgrade the large volume of existing systems to increase interconnections via the interoperability mechanisms provided by the Agricultural Reference Model.

Keywords: agriculture big data; reference model; infrastructure

0 引 言

基于实时数据的建模和科学研究方法在和人工智能、物联网、云计算等技术相结合后成为农业、智慧城市、医疗等很多应用领域创新的重要手段,但这也要求数据源能提供有效采集、管理和读取的机制^[1].近年来传感器、物联网和物理信息系统在应用领域中的大量部署,虽然极大地降低了实时数据采集和生产过程的数字化成本,然而因为采集标准、使用目的、管理规范的不同,这些大量增长的数据常常只能服务于特定的应用,并处于孤岛中而难以融合.这种数据发现、获取、融合的难度极大地阻碍了物联网和大数据第三方创新应用的开发,很多数据无法充分发挥其应有的利用价值.

以农业为例,精细化生产管理、精准和自动化流程控制、供应链跟踪是优化农业资源使用、提高农产品质量和品质的重要途径^[2].然而农业精准控制的建模极为复杂,有时只能通过大量的历史观察数据来寻找规律并建立数值模型;而基于数值计算的模型不仅需要大量的计算和观测数据来调节、完善模型参数,还需要大量不同领域的知识来完成不同模型的融合,比如研究气候对农业的影响,需要土壤、水、气象、动植物、生态系统等多个领域的协同.

如何有效采集、管理和使用快速增长的各行业数据,加强现有和在建信息化系统的互

通互联,并使其成为有效支撑数据驱动的应用创新和系统层科学研究的基础设施,正成为人们不得不面对的挑战.在过去的10多年里,这类数据基础设施引起了各国政府的高度重视,并将其视为提升各国自身科研创新能力的重要手段.比如欧洲的研究基础设施(Research Infrastructure, RI)^[3]温室气体监控(ICOS)^[4]、农业生态系统分析和实验(ANAE)^[5]、欧洲大陆板块(EPOS)^[6],以及未来的欧洲开放数据云(European Open Science Cloud)^[7]、美国的地球观测(DataOne)^[8].

通过对需求和挑战的分析以及对目前研究现状的调研,本文以欧洲一些在研的参考模型方法为基础,以农业大数据为实际的应用场景,提出了我国农业大数据基础设施参考模型;以农业信息化为实际案例给出了农业大数据基础设施的参考模型描述,并通过案例加以验证.参考模型的方法重点针对:①从不同的视角对大数据基础设施的功能和体系结构进行建模描述;②对符合农业大数据管理特征的生命周期进行分析建模;③为系统体系架构和标准的选取提供指导.

本文的主要贡献如下.

(1) 分析调研了大数据基础设施的建设挑战和研究现状.

(2) 重点介绍了欧洲环境大数据基础设施以及其参考模型的建设方法,并以此为基础提出了农业大数据基础设施的参考模型.

(3) 以基础设施的需求工程和历史遗留数据系统的互通接入,演示了参考模型的指导意义,对大数据基础设施的互通提出了建议.

本文后续内容安排如下:第1节分析列举大数据基础设施实施过程中的若干互通挑战;第2节从数据和信息化基础设施、概念模型、体系架构等3个方面进行相关研究综述;第3节详细介绍农业大数据基础设施参考模型的多视角分析方法及其5个视角;第4节通过两个案例展示农业大数据基础设施参考模型在需求工程改进和遗留系统互通升级改造上的运用;第5节进行总结并展望未来工作.

1 大数据基础设施的互通挑战

作为一个分布式的信息系统,大数据基础设施的研究和开发面临理解需求、优化重用现有技术、系统融合和互通、可持续等挑战.本节分析列出了上海农业大数据基础设施建设中所面临的3类挑战.

1.1 需求的采集和分析挑战

应用领域的信息化系统和大数据应用呈现多样化的趋势,除传统的政务管理型平台,越来越多的以物联网、移动终端、智慧终端为节点的数据采集、用户互动、智能服务的增值应用不断涌现.而这些应用对数据和服务的需求常存在巨大的差异,如建模预测需要数据实时融合、产品的追踪溯源需要数据完整历史和演化链、单个用户的数据订阅需要定制化的数据信息内容生成处理.

需求采集和分析挑战主要体现在:①数据基础设施的涉众众多,关键涉众的甄别不完全,例如,农户在农业数据基础设施中常被当成是一个用户而忽略其自身也是数据采集者和提供者的角色;②相关涉众对数据基础设施的部分前瞻功能理解的局限性,导致无法精准描述需求,如数据采集者、数据编目员、大数据应用开发者在农业应用中对采集什么范围的数据和数据的质量精度因各自视角不同而难以共识;③项目中需求常常只是在开始采集一次,而后缺少回访,难以捕捉需求的变化.

1.2 现有标准、技术选择和使用挑战

近年来针对大数据的管理和处理涌现了大量的标准、算法, 以及软件工具. 以农业为例, 我国已经定义的和农业有关的各类标准就有近 300 个, 其中包括产品追溯^[9]、信息安全^[10]、传感网络^[11]、农村电商^[12]、字符编码^[13]、系统架构^[14]等; 而国外的数量则更多, 常用的包括传感器配置和数据采集 (Sensor Model Language)^[15]、数据审编流程如英国 DCC 标准^[16]、数据格式标准 NetCDF^[17]、元数据标准的 ISO 19115^[18]和 CERIF^[19]、地理信息系统服务标准 OGC OWS^[20]、系统服务架构 ODP^[21]、联合国粮农组织的农业词汇标准^[22], 以及 Research Data Alliance (RDA) 的研究数据管理标准^[23]等.

尽管有这些标准和现有技术, 但在实际农业信息系统开发和集成的过程中, 选取和指定适合的标准仍然面临很多困难和挑战. 第一, 由于缺少有效的标准全景图和查询手段, 系统设计者和开发者常无法有效地根据需求和体系架构发现和挑选最佳的现有标准; 第二, 由于现有的信息化和数据管理系统对体系架构缺少标准化建模, 农业中的不同行业如蔬菜业、种植业、畜牧业、水产业, 以及农业机械的信息系统在建设中对标准的选取和使用常因系统架构和模型的差异而难以简单重用相互实践经验, 系统间集成和互通困难; 第三, 由于缺少对数据管理、服务流程的深入理解, 很多需要标准化的环节如业务分类规格、目录颗粒度, 地理空间表达、区域范围表达、涉及企业编码、涉及人员编码、计量单位、数据采集周期、数据分类统计、数据归集流程、数据应用权限等没被足够地重视和认识; 第四, 由于缺少有效的使用标准和升级指南, 历史遗留的农业信息系统和新系统之间的兼容难题使得信息系统在标准升级换代中成本高昂, 导致现有的数据管理平台缺少开放性, 系统和平台的接入缺少有效的管理和审核机制, 第三方大数据创新困难.

1.3 系统互通和第三方创新挑战

很多行业在过去的几十年里都采集了大量的观测和监测数据; 然而由于缺少有效的信息技术手段, 这些数据很多没有以基础设施的形式为公众提供服务, 很多在研和在建的基础设施也大多从特定领域出发; 同类系统之间的互通因为查询元数据标准、控制词汇、服务接口等的差异而难以互通. 由于缺少有效的安全和服务标准, 应用数据的采集和拥有者对数据分享和第三方创新无法有效参与. 一方面大量采集的数据, 如历史遗留的信息管理系统采集的数据, 因缺少标准化的管理流程而无法有效地通过审编、编目集成到应用中产生其价值而被闲置; 另一方面应用创新, 如精准控制和决策却又因为大量的信息鸿沟无法有效获取所需的特定时间、空间的数据, 并对其进行融合、处理和发掘. 以上这些对该行业信息化、智能化的深层创新和经济利用极其不利.

文献 [24] 从数据、应用 workflow、服务、基础设施、管理模式等不同的角度对基础设施的互通做了分析. 由于缺少有效的标准框架系统, 系统集成中产生的依赖复杂度随系统数量的增加而快速加大; 而这些系统对底层计算和存储技术的依赖却不尽相同. 随着底层技术和部分系统标准的更新, 整个信息系统的上层服务的升级和可持续变得非常具有挑战性. 建设中的上海农业信息公共信息服务平台就是这样的例子: 以不同分支行业的政务管理信息系统为基础, 融集成农机、绿叶蔬菜、水产等为一体的农业实时观测数据平台. 然而这个服务平台中很多遗留系统是科研项目的成果, 在建设过程中都往往未能严格采用软件工程方法规范其系统的设计和质量控制; 此外系统中的元数据结构、系统架构、服务界面等缺少通用性、延展性, 在扩充这类系统为基础设施的时候, 常受特定应用案例愿景或应用领域的引导而对其他同类基础设施之间可能的互通需求缺少充分的分析. 同时, 不同基础设施在开发

中,因各自进度的差异,且标准的选取决策受不同领域社群的影响,相互之间又缺少必要的沟通,导致在系统后期进行互通时,尽管元数据标准一致,但控制词汇的差异和对标准不同的扩展,互通起来十分困难。

2 相关研究综述

2.1 数据和信息化基础设施

实时数据基础设施通常是一个需求复杂的计算机信息系统,其建设需求因所服务人群以及应用差异而多样。数据基础设施通常由特定科研用户群通过自发组织、数据共享,并根据特定的研究兴趣热点而建设。2007年的一项欧盟调查显示,当时欧洲各类研究类型基础设施就已超过700个^[25]。一种常见的数据基础设施建设模式是从若干已有的特定信息化系统和政务平台开始,通过升级、集成和再开发,完成与实时数据管理系统的对接,从而实现基础设施的建设。

数据基础设施作为信息化基础设施的一部分,其功能的边界在不同系统中有着不同的定义:很多是以支撑大型和前沿科学研究为目的,如高能物理、环境全球挑战等。在过去的20多年里,可以看到有很多大型科学研究支撑系统开发出来。根据这些支撑系统所负责的具体功能,可将其大致分成3个主要大类^[26]。

(1) 以特定应用开发管理为中心的虚拟研究环境^①:管理科学研究问题的求解过程及其生命周期管理,典型系统如虚拟研究环境(Virtual Research Environment)^[27]、科学门户(Science Gateway)^[28]、虚拟实验室(Virtual Laboratory)等。

(2) 以数据管理为中心的大数据基础设施^②:管理大数据的生命周期,提供数据管理和数据处理的基本服务,典型系统如欧洲哥白尼^③、美国的OSDC^④。在这些基础设施中,也常包括和数据采集有关的设备。

(3) 以通用计算、存储和网络资源管理为中心的信息化基础设施^⑤:管理计算、存储和网络资源并提供通用服务管理底层资源服务的生命周期,典型系统如EUDAT^⑥、EGI^⑦、中国国家网格^⑧、ESNET^⑨等。

在实际中,这3类支撑系统之间并不一定存在一一对应的上下层次关系,数据基础设施可能运用自己内部的数据中心(计算和存储资源),虚拟研究环境也可能和大数据基础设施集成在一个独立服务中。

2.2 概念模型

对大数据生命周期、功能模块和系统架构进行概念建模是大数据基础设施建设的重要基础。文献[29-30]对数据管理生命周期做了调研,甄别了数据生命周期中的关键步骤,但不同模型对这些步骤的侧重和组织不同。下面是具体的5个模型。

(1) DataONE(Data Observation Network for Earth,地球数据观测网络)主要为科学研究者提供开放、持久和安全的地球观测数据的访问服务,数据采集(如计划、采集、保证、

① Virtual Research Environment

② 研究基础设施 Research Infrastructure (RI) 互用 <https://ec.europa.eu/research/infrastructures>

③ <http://www.copernicus.eu/>

④ <https://www.opensciencedatacloud.org/>

⑤ 常称为 e-Infrastructure 或者 cyber infrastructure

⑥ <https://eudat.eu/>

⑦ <https://www.egi.eu/>

⑧ <http://www.cngrid.org/>

⑨ <https://www.es.net/>

描述)和长期保存、数据应用(如数据发现、集成、分析等)融合在了一个管理周期中。

(2) 英国 Digital Curation Center (DCC) 模型针对数据长期保存, 强调数据审编和保存所牵涉的数字对象的描述、元数据、长期保存计划和社群参与的必要性; 围绕审编和长期保存这两个基本过程, 将数字对象的生命周期分为创立、评价、选择、长期保存、存储、读取、使用和转换等关键步骤。

(3) 美国联邦地理数据委员会提出的空间数据共享生命周期模型^[31]将地理和空间数据的生命周期分为定义、库存/评价、获取、读取、维护、使用/评价、归档等阶段, 通过评估来修正这些阶段的活动。

(4) BLM 模型^[32]是针对土地信息和资源等相关数据管理提出的以“质量控制 (QA/QC)”为核心的模型, 包括计划、采集、维护、读取、评价、归档这些核心步骤。

(5) 欧洲环境大数据参考模型 (ENVRI RM)^[33]是欧盟 ENVRIPLUS^[34]及其前期项目 ENVRI 中的科研成果, 其对欧洲环境和地球科学有关的 21 个大数据基础设施的共性做抽象, 抓住了数据中最核心的 5 个环节: 数据采集、审编、发布、处理和应用。

类似的数据生命周期模型还有很多, 但通过上面的 5 个不同模型, 还是很容易看到, 数据生命周期模型是和领域特征、基础设施的服务密不可分。DataONE DCC、BLM 模型强调数据的长期保存, ENVRI RM 则侧重数据采集、管理和使用的周期。

2.3 体系架构

大数据基础设施在建设时, 常有两种不同的架构模式: ①集成垂直服务的一体式数据中心模式, 如 D4Science^[27]; ②以开放服务平台的形式, 将基础设施部署于不同的底层 IT 基础设施或者环境中, 从而建立特有的大数据基础设施, 如欧盟的 EOSC^[7]。但在实际开发中, 很难有一个通用的大数据基础设施架构可以通用于不同的应用领域。

参考模型 (Reference Model) 通常用于对某个具体领域内一类系统或者一定组织的体系结构、共性特征、描述术语等的抽象, 如 ISO 标准化组织的网络开放参考模型 (Open System Interconnection (OSI))^[35]和开放分布式处理参考模型 (Open Distributed Processing (ODP))^[21]、OASIS 标准化组织的面向服务体系架构参考模型 (Service Oriented Architecture (SOA))^[36]、工作流管理联盟 (WfMC) 的工作流参考模型 (Workflow Management Coalition)^[37]等。

欧洲环境大数据基础设施参考模型 (ENVRI RM) 是欧盟 FP7 ENVRI 项目和 H2020 ENVRI PLUS 两个项目的研究成果^[33]。该参考模型研制的核心目的是为了提升欧洲不同大数据研究基础设施建设的效率和促进互通。FP7 ENVRI 项目是欧盟针对环境和地球科学的 7 个泛欧洲研究基础设施互通且为期 3 年的集群项目, 参与该项目的 7 个基础设施来自环境和地球科学中 4 个不同子领域: 大气、海洋、生态和地质。

在这个背景下 ENVRI RM 基于开放分布式处理参考 (ODP) 模型, 描述了环境大数据基础设施的共性特征、大数据基础设施的抽象结构模型、环境大数据管理生命周期建模等所需的词汇和结构。该模型在 ENVRI 后续项目 ENVRI PLUS 中得到了进一步充实, 并运用到了 22 个环境大数据基础设施中, 这 22 个环境大数据基础设施中绝大多数都是欧盟泛欧洲环境大数据建设蓝图 (ESFRI) 的重点项目。

ENVRI RM 以本体的形式存在, 在其设计过程中, 使用了 UML 模型对来自不同基础设施的需求和原始设计文档进行理解和提取, 形成了相关的 UML 图例。本文主要介绍上海农业大数据基础设施建设中对 ENVRI RM 的借鉴。限于篇幅, 本文仅基于 UML 模型在项目的

相关实现文档中进行介绍。

3 农业大数据基础设施参考模型

农业大数据基础设施参考模型根据农业物联网云平台需求分析, 运用欧洲环境大数据基础设施参考模型 (ENVRI RM) 的最新成果开发而成。该参考模型力争对农业大数据基础设施的数据管理、功能模块、系统架构等方面从不同的视角给出描述词汇和本体。

3.1 多视角分析方法

开放分布式处理 (ODP) 是计算机科学领域的一个参考模型, 是国际标准化组织 (ISO)、国际电工委员会 (IEC) 和电信标准化组织 (ITU-T) 联合提出的一个标准, 主要用于分布式、网络化信息系统的建模和设计。

ODP 的核心思想是从企业、信息、计算、工程和技术这5个不同的视角来对一个复杂计算机系统进行建模。ODP 提供不同视角的一系列建模和描述的机制, 可让不同专业背景的人以适合的视角对系统进行描述。图1体现了这不同视角之间的关系。



图1 ODP的基本视角

Fig. 1 Viewpoints of open distributed processing

- (1) 企业视角描述系统的需求、功能、应用领域和约束条件等。
- (2) 信息视角描述系统信息演化过程中的数据、信息、类型和语义等。
- (3) 计算视角描述系统的功能模块、模块集成和接口界面等。
- (4) 工程视角描述系统的系统架构、模块互动和功能分布。
- (5) 技术视角描述系统开发中针对不同功能和过程所需的技术选项。

这些视角相互关联但又可以独立用来描述一个系统。在实际使用过程中, 视角之间可以通过关联、灵活组合对复杂系统进行建模。ENVRI RM的多视角分析继承了ODP的多视角分析理论。

3.2 农业大数据基础设施的关键过程分析

根据项目需求和应用分析, 农业大数据基础设施的业务定位可涵盖数据管理、创新应用、服务部署运营。在参考模型上, 数据管理的生命周期采用3部分的融合: 数据、创新应用和服务运营。在内涵上农业大数据基础设施涵盖农业物联网、云平台、农业大数据管理、农业大数据应用等方面。

围绕应用的生命周期, 大数据基础设施应支持用户描述应用需求 (Application Requirement)、发现服务和数据发现 (Data and Service Discovery)、开发和集成应用 (Application Development)、运行应用 (Execution)、分析执行结果 (Result Analysis)。数据管理聚焦数据的数据采集 (Data Acquisition)、数据审编 (Data Curation)、数据发布 (Data Publishing)、数据处理 (Data Processing)、数据应用 (Data Use)。在应用的执行过程中, 计算和存储资

源的请求 (Resource Request)、选择 (Resource Selection), 建立服务和运营等级协定 (Service Level Agreement, Operational Level Agreement), 分配和配置 (Resource Provisioning and Deployment) 资源, 对计算调度 (Computing Scheduling)、计算作业进行监控管理 (Monitoring and Management).

根据前期的需求分析, 特别是针对农业物联网云平台的业务需求, 提出了这类基础设施应该支持的核心操作, 如图 2 所示.

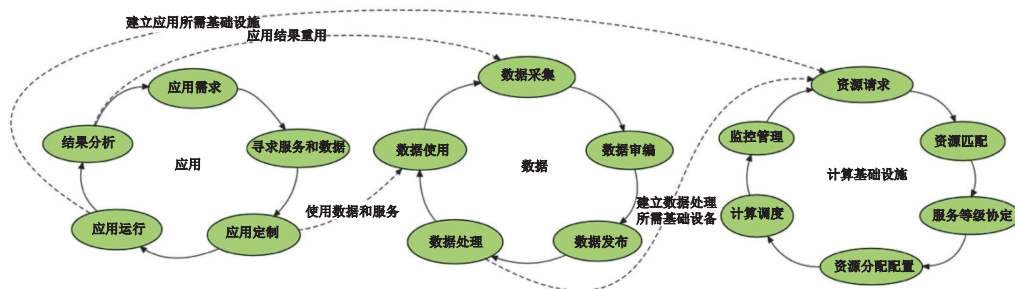


图 2 农业大数据基础设施的相关活动

Fig. 2 Activities in agricultural big data infrastructures

根据不同基础设施的调研结果, 将农业大数据基础设施对贯穿应用、管理数据和资源生命周期的关键功能进行建模, 如图 3 所示.

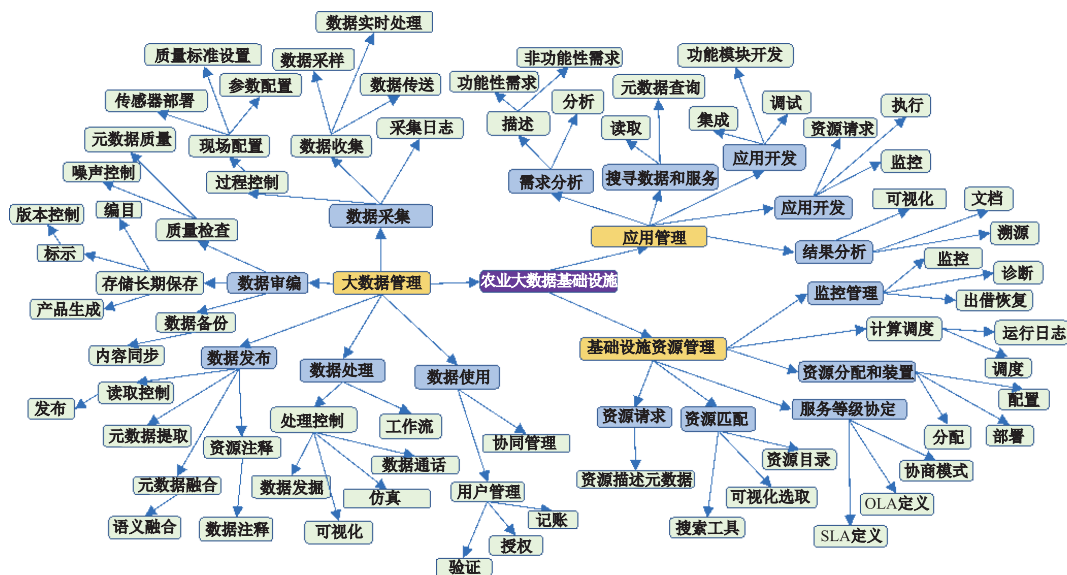


图 3 农业大数据基础设施的基本概念

Fig. 3 Basic concepts of agricultural big data infrastructures

3.3 分解视角的建模

3.3.1 科学视角 (Science Viewpoint, SV)

科学视角的建模是分析农业大数据基础设施的组织结构和功能, 通过这些组织内部关系, 以及组织和相关个体互动的行为推导出农业大数据基础设施所包含对象的原理和特性. 该视角定义描述了农业大数据基础设施涉及的社群、角色、特征行为, 以及农业大数据基

基础设施的高级行为. 该视角的目的如下.

(1) 为描述农业物联网大数据云的相关涉众 (Stakeholder)、社群 (Community)、社群行为 (Behavior), 以及各种参与角色、运行的管理策略等方面提供通用词汇.

(2) 根据通用词汇, 结合农业物联网大数据云平台的需求、前期工作以及相关的应用场景, 抽象出特征的社群、社群行为、涉众、角色等具体描述词汇.

涉及的基本的概念如下.

- 相关涉众: 农业大数据基础设施的利益相关方, 如图4所示.

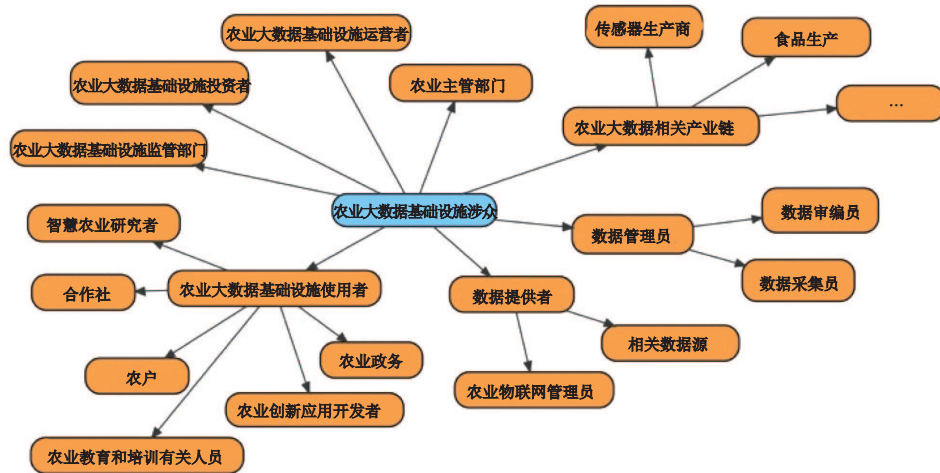


图4 农业大数据基础设施相关涉众

Fig. 4 Stakeholders in agricultural big data infrastructures

- 社群: 与大数据基础设施的数据、服务、运营相关的人员根据不同业务或需要而组成的群体, 如图5所示.

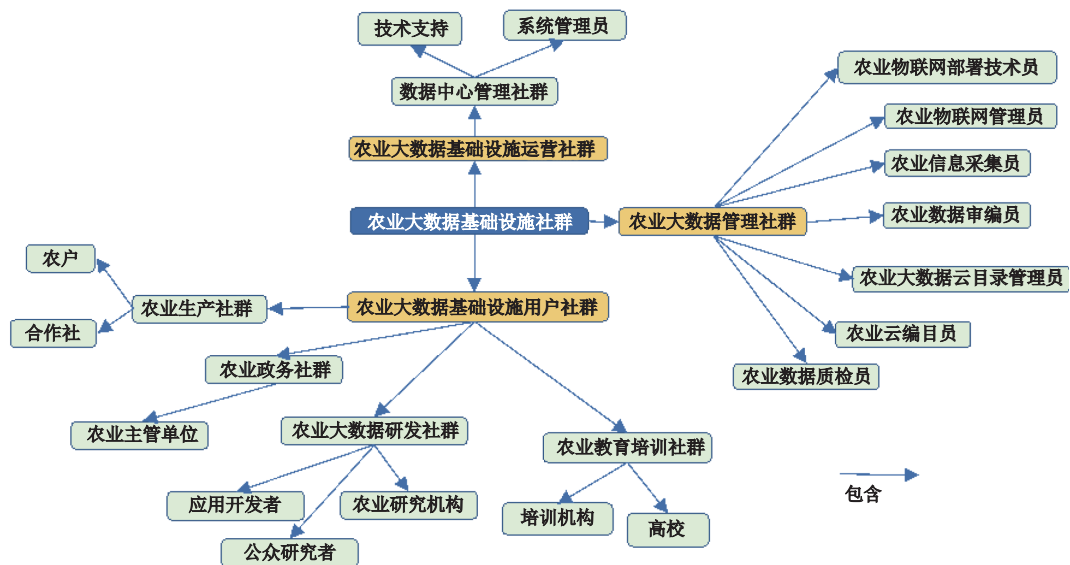


图5 农业大数据基础设施用户社群

Fig. 5 User community of agricultural big data infrastructures

- 社群行为: 大数据基础设施特定社群的共性行为.

- 备份: 将数据复制到其他数据存储区, 以便在发生数据丢失后可恢复原始数据. 长期保存是一种特殊类型的备份.
- 质量检查: 验证数据质量的操作, 可包括消除噪音、删除显然是错误的数据、计算校准. 质量检查可以在数据生命周期的不同阶段进行.
- 数据挖掘: 执行一系列的元数据/数据请求、数据处理、解释结果, 做一个新的请求. 通常这个操作有助于加深对数据的了解, 可以通过特殊的软件来支持这个操作.
- 最终审查: 检查要发布的数据, 这些数据不会再被更改. 该操作触发数据状态的变化——“最终复查”.
- 自由文本注释: 给信息对象添加一个简短的解释或意见.
- 执行概念模型映射: 执行值的转换规则 (从一个单元到另一个单元的映射) 或概念的转换规则 (将含义从一个概念模型转换为另一个概念模型, 例如转换代码列表).
- 进行观察和数据采集: 测量参数或观察事件. 测量或观察的性能产生测量结果.
- 数据处理: ① 转换和生成数据产品; ② 计算, 例如统计过程、模拟模型; ③ 可视化.
- 发布数据: 使数据公开.
- 发布元数据: 将注册的元数据提供给公众.
- 查询数据: 向数据存储发送请求以检索所需的数据.
- 查询元数据: 向元数据资源发送请求以检索感兴趣的元数据.
- 注册元数据: 将元数据输入到元数据目录中.
- 解析注释: 检索解析数据对象所包含的一组注释.
- 设置概念映射规则: 指定数据和/或概念的映射规则.
- 描述数据采集规范: 如测量或观测 (地点) 的地理位置.
- 描述测量或观察: 描述观测/测量方法的细节.
- 存储数据: 以持续的方式归档或保存数据, 确保持续的可访问性和可用性.
- 追溯: 自动生成和存储有关操作和数据状态更改的元数据并作为起源实例.
- 数据和元数据状态: 用来描述数据或元数据在生命周期中的不同阶段. 数据可以有原始 (数据被最初采集)、已注释 (对数据记录进行相关信息的补充、注释)、已质检 (对数据内容和质量进行检查)、已赋元数据 (根据预定的元数据标准对数据单元建立相应的元数据)、已复核 (对数据的质量、结构等方面进行检查审核)、已概念模型映射 (根据概念模型映射规则已完成数据和相关概念模型做映射)、已公布 (基于大数据云的管理规范将数据单元对外公开) 等操作状态. 元数据可以有原始 (元数据被创立时候的最初描述, 此时的内容可能是自由文本)、已登记 (在系统中 (如元数据目录) 创立相应的元数据记录)、已编目 (在元数据目录中编录相关元数据记录)、已公布 (数据内容对外公开) 等状态. 数据和元数据状态可以通过状态转换图和数据操作加以关联, 如图7所示.

3.3.3 计算视角 (Computational Viewpoint, CV)

计算视角对农业大数据基础设施中的计算对象以及它们可以调用的接口进行建模, 模型中计算对象通过接口调用基础结构中的其他对象. CV 的描述分为3部分: 对象、CV 对象和子系统、CV 集成点, 其中对象包括以下几类.

- CV 界面对象: 便于用户访问农业大数据基础设施的计算对象.
- CV 代理对象: 充当用户或者客户端访问农业大数据基础设施中的数据或者服务的

媒介, 这类计算对象执行语义解释或者复杂查询管理.

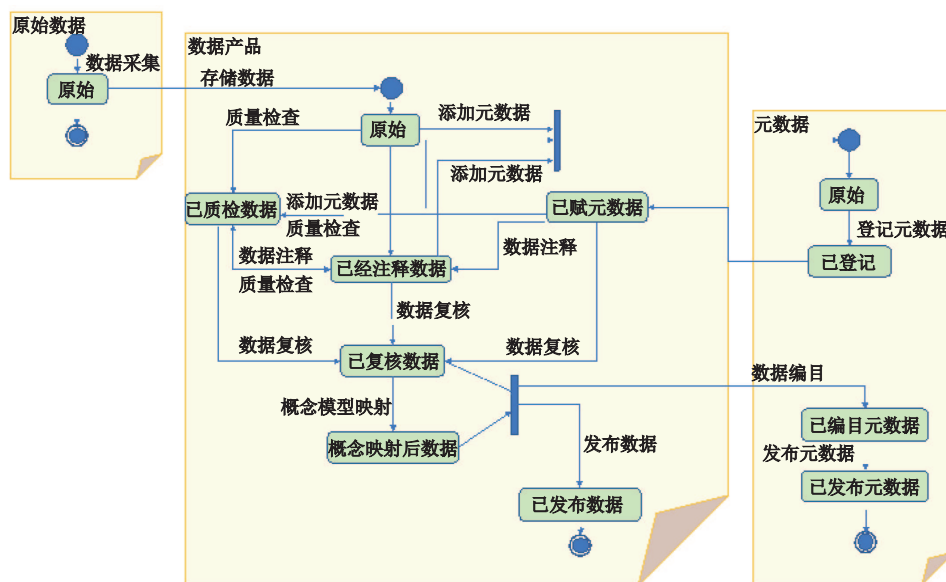


图7 农业大数据基础设施数据状态和数据操作

Fig. 7 Data states and operations in agricultural big data infrastructures

- CV服务对象: 提供对农业大数据基础设施系统和各类资源(内部和外部)的访问、集成、编程的计算对象.
- CV组件对象: 提供对后端对象访问的计算对象.
- CV后台服务器端对象: 包含农业大数据基础设施系统和用于访问研究或派生数据产品的资源的计算对象.

3.3.4 工程视角 (Engineering Viewpoint, EV)

在参考模型中, 工程视角从系统设计的角度根据信息视角、计算视角和科学视角中的数据对象、数据流、计算对象, 以及社群行为等的实现, 提出合理的模块分解和体系架构. 在参考模型中, 工程对象 (Engineering Object) 和对象配置 (Object Configuration) 是基本概念. 工程对象中的一个重要元素就是容器对象 (Container Object). 容器对象本身也是一种工程对象, 但可以是组织和包含其他工程对象并形成结构. 以微服务为主的工程对象是目前软件系统开发的重要趋势.

在工程视角里, 根据信息视角和计算视角定义了一系列基本工程对象 (Basic Engineering Object, BEO), 这些对象分为标识、目录、处理、追溯、传送和存储几大类. 每个BEO一对一映射一个CV对象.

3.3.5 技术视角 (Technology Viewpoint, TV)

技术视角涵盖了应用于计算过程必须执行的现有计算平台的现实约束 (例如对用于实现该系统的设施和技术限制). ENVRI RM中工程视角 (EV) 的规范旨在保证由其他视角 (科学视角、信息视角、计算视角) 提供的模型与要实施的技术和标准之间的灵活对应.

4 应用案例

本节结合上海农业大数据基础设施参考模型项目中的实际工作, 通过两个实际案例来

讨论参考模型是如何帮助实现需求分析改进和遗留系统升级改造的.

4.1 参考模型指导下的需求采集

农业信息化一直存在需求采集难的难题: 一方面系统开发前期, 开发人员对基础设施发展趋势的认识不足, 需求采集问题的设置常不得要领; 另一方面基层农户、数据采集站对系统缺少全局认识, 采集来的需求针对性不强. 针对这些问题, 本文在上海农业数据基础设施建设中, 运用参考模型对传统需求工程做了指导和增强, 对农机、畜牧等现有系统的原需求采集做了改进, 如图 8 所示.

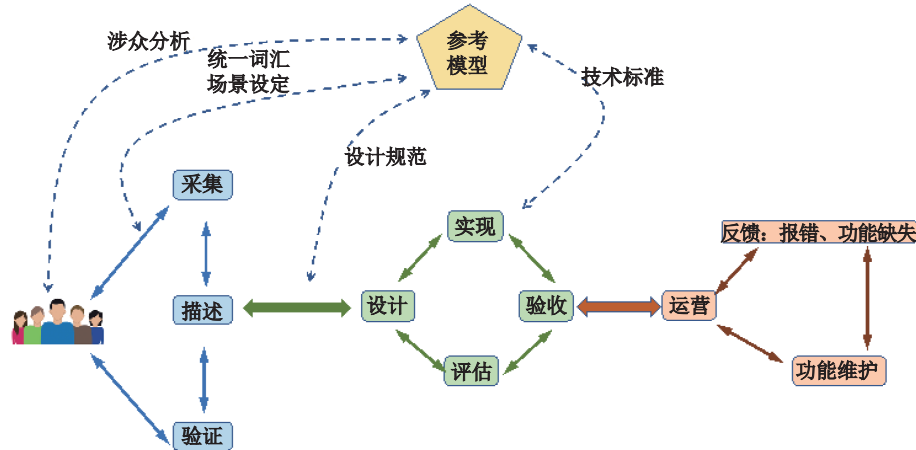


图 8 参考模型在需求工程和设计中的作用

Fig. 8 The Reference Model's involvement in requirement engineering and system design

4.1.1 如何确定需求的采集面

从确定相关涉众入手, 根据参考模型对数据采集、审编、处理的相关涉众进行角色划分, 得出农业数据采集、审编、处理和使用过程中的角色, 如图 9 所示. 根据图 9, 从数据采集社群、管理社群和使用社群入手, 对案例中主动角色如农户、物联网数据采集员, 以及被动角色如农机传感器、农机 GPS、数据采集站等进行细致梳理.

4.1.2 如何有效地设置需求问题

针对不同的涉众, 根据参考模型及不同视角来设置需求问题, 让问题本身更接近相关涉众的实际背景, 也让问题的范围和描述更具代表性和针对性; 根据参考模型的数据管理生命周期, 让问题的设置覆盖更完整的系统功能范围. 在实际中, 针对农机、畜牧数据采集, 审编不同角色的问题, 并依据参考模型加以设置.

4.1.3 如何设计有效的需求采集方式

在需求采集过程中, 除了单纯的采访, 根据参考模型的功能模型做基于场景的问题描述, 可极大地减低双方在沟通中的难度. 从参考模型的不同视角将问题的设置顺序与一些现有软件原型相结合, 可避免采集过程过于抽象. 在问题设置上, 应充分考虑相关角色的背景, 比如农机政务管理的问题可从科学和企业视角进行设置; 相关信息和计算的问题则重点针

- 4) 联系不到合适的农机服务.
- 5) 操作技术不够熟练.

随着农户对需求问题回答的逐步推进, 假设农户在回答 Q3 时如果选择 4) “联系不到合适的农机服务”, 此时又可以在参考模型的指导下进一步对农户与合作社的隶属关系进行需求考察, 进而给出另一个更贴近农户需求的问题 Q4.

Q4: 您家庭参加了农民专业合作社吗?

- 1) 参加了.
- 2) 没参加.
- 3) 不清楚.

这个动态确定需求问题的过程如图 10 所示.

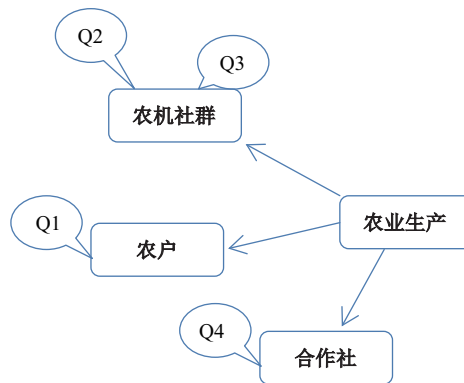


图 10 参考模型指导下的动态确定用户社群的需求分析

Fig. 10 Dynamic community requirement analysis supervised by the Reference Model

限于篇幅, 本文仅从科学视角进行描述, 通过分析确定出相关涉众, 如农业科研人员、涉农企业、各级涉农政府管理部门、农业生产人员 (农产品运销大户、种养加工大户、合作社社员、县乡农技推广人员)、系统使用人员等. 对上述涉众的各类需求进行归纳整理, 在参考模型的指导下进行上下文感知的需求问题提取, 同时考虑用户对需求了解的逐步深入, 最后设置相应的需求问题. 图 11 是基于参考模型指导实现的动态需求分析系统的运行界面截屏.

4.2 参考模型指导的系统升级

大数据基础设施的建设一般不会从零开始, 很多正在使用的遗留系统将成为基础设施的重要起点. 如何升级这些现有系统使其满足基础设施的需求是需要解决的重要问题. 以上海农业信息化为例, 过去 20 多年里开发的和农业有关的特定行业信息化政务和数据管理系统就有 20 多个, 然而大多数系统只考虑了该行业的特定功能 (主要是政务管理、自动化).

该案例的核心思想是将上海农机管理系统政务信息、农机实时信息、农机租用以及补贴等多个不同系统的信息进行有机融合, 在此基础上对农机资源分配、补贴审核等事务做优化.

上海农机综合管理平台实现了农机业务的全面信息化管理, 涉及农机的基本信息、补贴、厂家、供应商、业务流程、综合调度等各方面, 并已上线运行多年, 积累了较全面的农机管理信息数据. 在这些前期系统中, 普遍都针对各自的目标进行过相应的需求分析; 然而这些系统平台普遍不同程度地存在覆盖面较低、兼容性差、数据共享弱、信息服务能力受

限等问题. 另外由于缺乏标准化开放数据端口, 各系统之间很难互通数据, 无法实现涵盖范围更广的应用开发.



图 11 基于参考模型指导实现的用户社群和角色分析系统的运行界面截屏

Fig. 11 Screenshot of the Revised Community and Actor Analysis System

该应用案例的一个重要目标就是, 将政务的资源 and 人员等相对“静态”的信息管理和农机实时监控数据, 通过构建一个开放数据平台实现联网. 该平台涵盖政务管理、农户服务、合作社应用这 3 大系统, 可以提高农机化生产管理与服务的信息化水平, 农机化生产主体和管理部门能够及时、便捷地获取和发布有效的农机作业信息, 解决农机作业供需市场的信息不对称, 更好地为农机管理系统、机手和农民服务, 促进农业增效和农民增收.

在现有系统的基础上, 制定更全面的智能农机大数据应用规划, 利用云计算、大数据等技术, 将农机作业信息、农业地理信息、气候土壤信息、智能农机决策支持信息等应用数据集成, 建立农机行业统一的信息管理平台, 实现农业和农机大数据的协同管理和综合利用. 数据融合, 碎片化的孤岛数据实现互通, 具体有如下的功能. 为实现更高级的智能应用提供数据支持. 具体有如下的功能.

- (1) 农机在线监管、在位情况分析、作业状态控制、智能管理等功能.
- (2) 提供农机管理指挥调度中心监控展示系统、农机监控管理系统、农机智能调度管理系统、农机计划作业管理、农机机具管理系统、农田地理信息系统.
- (3) 对粮食作业的收割机、插秧机、拖拉机等农机信息统一化管理, 包括农机在线监管、在位情况分析、作业状态控制、智能管理等功能.
- (4) 智能测亩功能以及自动计产功能. 在开始作业时进行作业面积测量, 收割结束时自动计算作业面积, 同时自动计算粮食产量.

升级过后的系统需求分析如图 12 所示. 新系统对分散的独立系统的需求进行了合并调整及补充, 构建了一个开放的全数据平台, 覆盖原有的政务管理、农户服务、市场应用等 3 大系统, 对从涉众、服务和数据这 3 个层面进行了扩展. 图中绿色代表新增部分, 红色代表在原有模块上重新包装转换, 蓝色代表对原有模块进行了扩充.

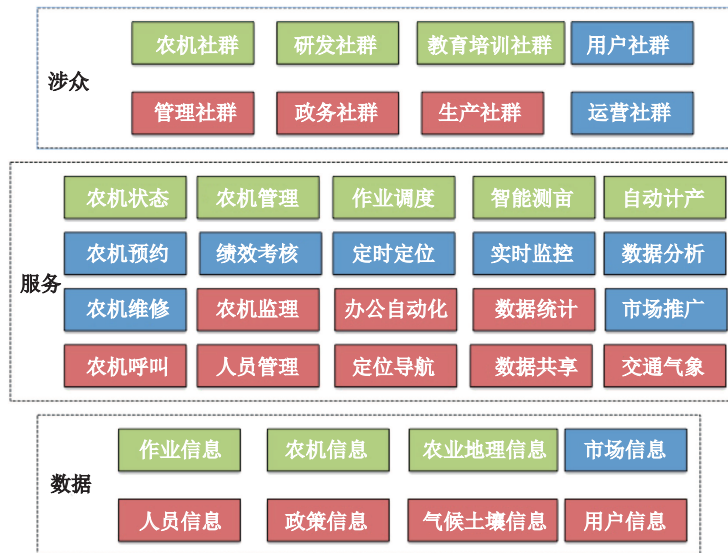


图 12 参考模型指导下的系统升级模块示意图

Fig. 12 Module upgrade supervised by the Reference Model

参考模型在升级过程中的作用主要体现在如下几个方面.

- (1) 整合约束条件, 对原系统需求进行了分析调整
- (2) 整合数据, 通过制定统一抽象模型、统一数据标准, 保证了数据的准确性、完整性、一致性.
- (3) 调整现有的数据周期, 结合参考模型的 5 个视角分析了实现多源数据的汇总融合以及共同利用.
- (4) 对现有系统模块结构和功能行为元素进行了层次化多精度描述, 并参照统一标准对原系统功能组件进行了调整、转换或组合.

针对各系统数据字段格式不一致、兼容性差的问题, 通过制定标准化开放数据接口, 实现了各系统之间数据互通和共享, 提高了信息服务能力. 采用与平台无关、语言无关的 Web Services 技术作为传输方式, 并制定统一接口. 通过归类整理原有的多个不同系统中的政务信息、农机实时信息、农机租用以及补贴等相关业务数据类别及数据资源, 制定了需要进行数据交换共享的指标项, 并规范了指标项目的属性、数据类型以及约束条件. 信息数据交换接口基于数据库中的表结构, 按照 Web Service 规范进行定义接口. 图 13 是升级后系统农机调度部分的运行示意图, 其中上半部分为系统日常管理界面图, 左下部分为作业面积测量功能截图, 右下部分为农机轨迹追踪, 可以用于实现电子围栏.

5 总 结

本文讨论了大数据基础设施的建设挑战和现阶段的紧迫性. 参考国外一些现有工作, 结合我国的实际情况, 以农业大数据基础设施为例提出了基础设施建设的参考模型方法. 通过多视角的建模方法, 对大数据基础设施的核心过程、基本描述词汇、体系结构的基本特征加以建模. 通过上海农业大数据基础设施建设中的需求工程改进和遗留系统升级两个案例演示了参考模型的使用价值.

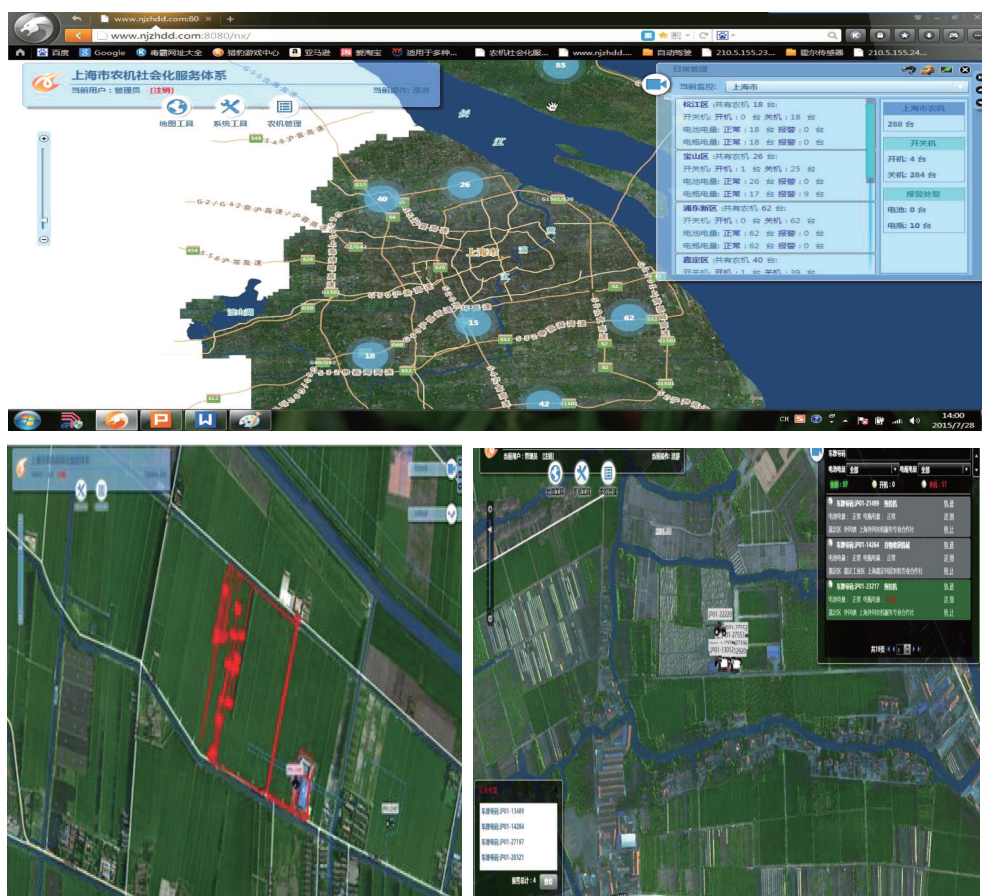


图 13 升级后的农机调度功能运行界面; 上半部分为日常管理界面图, 左下部分为作业面积测量, 右下部分为农机轨迹追踪(电子围栏)

Fig. 13 Screenshots of the Upgraded Agricultural Machinery Management System; the upper figure is the general user interface; the bottom left figure shows the automatic acres measurement function; and the bottom right is the trajectory analysis, which can be used for the electronic fencing function

我国大数据创新的高速发展, 与之相应的大数据基础设施将成为很多应用领域的创新支撑关键. 在欧洲研究基础设施的建设中可以看到各国对此的重视. 参考模型的方法不是着眼于一个特定的基础设施开发, 而是针对不同数据基础设施在建设中如何运用共同参考实现数据、服务和底层资源的互通, 从而最终实现对系统科学的支持.

参考模型的开发只是第一步, 其能被社群认可, 并成为相关数据基础设施建设的指导, 还需要一段很长的路.

[参 考 文 献]

- [1] DEMCHENKO Y, ZHAO Z M, GROSSO P, et al. Addressing big data challenges for scientific data infrastructure [C]//Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on. IEEE, 2012: 614-617.
- [2] 葛晶. 我国智慧农业的管理模式、问题及战略对策 [J]. 生态经济, 2017, 33(11): 117-121, 133.

- [3] MARTIN P, CHEN Y, HARDISTY A, et al. Computational challenges in global environmental research infrastructures [C]// Terrestrial Ecosystem Research Infrastructures: Challenges and Opportunities. Boca Raton, FL USA: CRC Press, Taylor & Francis Group, 2017: 305-340.
- [4] ICOS. Integrated carbon observation system [EB/OL]. [2018-03-05]. <https://www.icons-ri.eu/>.
- [5] ANAEE. Analysis and experimentation on ecosystems [EB/OL]. [2018-03-05]. <https://www.anaee.com/>.
- [6] EPOS. European plate observation system, 2017 [EB/OL]. [2018-12-26]. <https://www.epos-ip.org/>.
- [7] European Commission. European open science cloud [EB/OL]. [2018-03-05]. <https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>.
- [8] DataOne. Data observation network for earth [EB/OL]. [2018-12-26]. <https://www.dataone.org/>.
- [9] 中华人民共和国农业部. 农产品追溯编码导则: NY/T 1431-2007 [S].
- [10] 公安部. 信息安全技术网络通讯安全审计数据留存功能要求: GA/T 695-2007 [S].
- [11] 北京市质量技术监督局. 物联网感知设备通用信息安全技术要求: DB11/T 1285-2015 [S].
- [12] 浙江省质量技术监督局. 农村电子商务服务站(点)管理与服务规范: DB33/T 982-2015 [S].
- [13] 国家标准局. 文献数目信息交换用数字字符编码字符集: GB/T 6513-1986 [S].
- [14] 江苏省质量技术监督局. 江苏省农村综合信息服务平台建设通则: DB32/T 2290-2013 [S].
- [15] OGC. Open Geo Spatial Standards, SensorML [EB/OL]. [2018-03-06]. <http://www.opengeospatial.org/standards/sensorml>.
- [16] DCC. Curation Lifecycle Model [EB/OL]. [2018-03-06]. <http://www.dcc.ac.uk/resources/curation-lifecycle-model>.
- [17] OGC. OGC network Common Data Form (netCDF) standards suite [EB/OL]. [2018-03-06]. <http://www.opengeospatial.org/standards/netcdf>.
- [18] Geographic information-Metadata standard: ISO19115 [S/OL]. [2018-03-06]. <https://www.iso.org/standard/53798.html>.
- [19] euroCRIS. The Common European Research Information Format, CERIF [EB/OL]. [2018-03-06]. <http://www.eurocris.org/cerif/main-features-cerif>.
- [20] OGC. OGC Web Services Context Document (OWS Context), OWS [EB/OL]. [2018-03-06]. <http://www.opengeospatial.org/standards/owc>.
- [21] LININGTON P F, MILOSEVIC Z, TANAKA A, et al. Building Enterprise Systems with ODP: An Introduction to Open Distributed Processing [M]. [S.l.]: Chapman & Hall/CRC Press, 2011.
- [22] AIMS. AGROVOC Multilingual agricultural thesaurus, AGROVOC [EB/OL]. [2018-03-06]. <http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>.
- [23] RDA. Research Data Alliance, RDA [EB/OL]. [2018-03-06]. <https://rd-alliance.org/>.
- [24] ZHAO Z M, MARTIN P, GROSSO P, et al. Reference model guided system design and implementation for interoperable environmental research infrastructures [C]//e-Science (e-Science), 2015 IEEE 11th International Conference on. IEEE, 2015: 551-556.
- [25] ESFRI. European Strategic Forum for Research Infrastructures [EB/OL]. [2018-03-06]. <http://www.esfri.eu/>.
- [26] ZHAO Z M, BELLOUM A, BUBAK M. Special section on workflow systems and applications in e-Science [J]. Future Generation Computer Systems, 2009, 25(5): 525-527.
- [27] CANDELA L, PAGANO P, CASTELLI D, et al. Realising virtual research environments by hybrid data infrastructures: the d4science experience [C]//International Symposium on Grids and Clouds (ISGC) 2014. SISSA Medialab, 2014, 210: 022.
- [28] MILLER M A, PFEIFFER W, SCHWARTZ T. The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources [C]//Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond. ACM, 2012: 39.
- [29] CEOS. Working group on information systems and services [EB/OL]. [2018-03-05]. http://ceos.org/document_management/Publications/Governing_Docs/WGISS_Terms-of-Reference_Oct2015.pdf.
- [30] BALL A. Review of data management lifecycle models [R/OL]. (2012-02-13)[2018-03-05]. <https://purehost.bath.ac.uk/ws/portalfiles/portal/206543/redm1rep120110ab10.pdf>.
- [31] FGDC. Stages of the geospatial data lifecycle pursuant to OMB circular A-16 [EB/OL]. [2018-03-06]. <https://www.fgdc.gov/policyandplanning/a-16/index.html>.
- [32] BLM. The bureau of land management handbooks [EB/OL]. [2018-03-06]. <https://www.blm.gov/policy/handbooks>.
- [33] ENVR. Environmental RI reference model[EB/OL]. [2018-03-05]. <http://envri.eu>.
- [34] ENVRIPLUS.EU H2020 ENVRIPLUS project [EB/OL]. [2018-03-06]. <http://www.envriplus.eu>.
- [35] Information technology-Open Systems Interconnection-Basic Reference Model: Naming and addressing: ISO/IEC 7498-3: 1997 [S/OL]. [2018-03-06]. <https://www.iso.org/standard/25022.html>

- [36] OASIS. OASIS SOA reference model TC [EB/OL]. [2018-03-06]. https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=soa-rm.
- [37] WORKFLOW MANAGEMENT COALITION. Workflow management Coalition, WfMC reference model [EB/OL]. [2018-03-06]. <http://www.wfmc.org/2-uncategorised/53-reference-model>.

(责任编辑: 李 艺)

(上接第 76 页)

- [7] CHEN Q, QIAN J H, ZHU L, et al. Optimal temporal path on spatial decaying networks [J]. Journal of Applied Analysis & Computation, 2016, 6(1): 30-37.
- [8] CHEN Q, QIAN J H, ZHU L, et al. Optimal transport in time-varying small-world networks [J]. Physical Review E, 2016, 93(3): 032321.
- [9] GUIMERÀ R, ARENAS A, DÍAZGUILERA A, et al. Dynamical properties of model communication networks [J]. Physical Review E, 2002, 66(2 Pt 2): 026704.
- [10] ZHAO L, LAI Y C, PARK K, et al. Onset of traffic congestion in complex networks [J]. Physical Review E, 2005, 71(2): 026125.
- [11] KRIOUKOV D, FALL K, YANG X. Compact routing on internet-like graphs [C]//IEEE Conference on Computer Communications. IEEE, 2004: 209-219.
- [12] WANG W X, WANG B H, YIN C Y, et al. Traffic dynamics based on local routing protocol on a scale-free network [J]. Physical Review E, 2006, 73(2): 026111.
- [13] WANG W X, YIN C Y, YAN G, et al. Integrating local static and dynamic information for routing traffic [J]. Physical Review E, 2006, 74(2): 016101.
- [14] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks [J]. Science, 1999, 5439(286): 509-512.
- [15] TAN F, WU J, XIA Y, et al. Traffic congestion in interconnected complex networks [J]. Physical Review E, 2014, 89(6): 062813.

(责任编辑: 李 艺)