数据驱动的科学工作流及其在生物医学中的应用实践*

□ 洪娜 钱庆 方安 吴思竹 杨林 / 中国医学科学院医学信息研究所 北京 100020

摘要：E-Science关注数字环境下的科研活动，然而随着生物医学大数据的爆发，数据密集型科学研究为e-Science带来了新的挑战。科学工作流通过形式化科学计算的流程，支持在一个专门的程序环境下自动协调多任务多步骤的处理，从而减少科研投入，提高科研效率。文章首先归纳并总结了科学工作流的相关定义，然后分析了当前主流的科学工作流系统，提出科学工作流处理如何应对密集的生物医学数据，并基于Taverna开展了生物医学领域的科学工作流实验。最后总结了科学工作流当前的研究现状和存在的一些问题。

关键字：科学工作流，e-Science，数据密集，共享工作流，Taverna
DOI：10.3772/j.issn.1673—2286.2014.03.004

1 引言

作为继理论和实验之后的第三种科研手段，科学计算已经在科研领域中发挥着非常重要且不可替代的作用[1]。科学计算通常需要组合涉及多个专业和领域的成千上万的仪器设备、应用程序、科研人员等对海量数据进行存储、查询、移动、处理、分析与可视化等操作，进而帮助科学家进行科学发现。如此复杂的计算流程和协同共享对目前的科研工作环境和科学计算支持工作平台提出了严峻的挑战。

科学工作流是指以数据驱动，面向科学实验过程的工作流。针对科学工作流，不同研究人员从不同的角度给出了不同的定义：在第四范式背景下[1]，C. Goble等人认为，“科学工作流是科研过程的精确描述，它包含一个能够协调多任务的多步骤处理过程，就像一个复杂的脚本”。M. P. Singh等人则认为[2]“科学工作流是描述求解科学问题中的一系列结构化活动和计算过程”；B. Ludascher等人将科学工作流定义为[3]“是完成一个科学目标的过程的形式化描述，过程表示了计算任务及任务间的依赖关系”。

2 科学工作流系统研究现状

科学求解过程的特殊环境和独特需求，使得科学工作流和一般事物型工作流在关键技术方面存在着明显的不同。科学工作流通过对复杂应用程序及各程序间的数据依赖关系进行组合，并控制各部分在时间、空间以及资源等约束条件下按序完成，为科学家进行科学数据管理、分析、仿真和可视化等提供流程组合和自动化运行的管理平台，已经成为复杂科学计算流程管理的必要手段，有效推动了科学研究的进展。在科学工作流的研究和发展历程中，多个大型的科研组织针对各自需求及研究背景建设了面向不同应用的科学工作流系统，其中比较著名的有Taverna、Pegasus、Triana、Kepler、KNIME、GridFlow、ICENI等。

2.1 Taverna

Taverna[4]是在英国e-Science研究框架下启动的项目myGrid中的一个子项目。myGrid主要为生物学和生物信息学领域的in silico实验进行开发，该实验的目

* 本文系国家“十二五”科技支撑计划项目课题“科技知识组织体系共享服务平台建设”（编号：2011BAHl0B03）、国家社会科学基金项目“关联数据中潜在知识关联的发现方法研究”（编号：11CTQ016）和中央级公益性科研院所基本科研业务费课题“面向大数据的医学科研支撑环境建设初步研究”（编号：13R0102）的研究成果之一。

标在于使用计算机的信息存储能力和分析能力来验证科学假设、论证理论推理、探索新模式或验证已知事实；而myGrid则旨在为生物学研究人员提供一个透明的基于网格的实验环境进行知识密集型任务的开发，从而减少科学家在与具体计算相关的工作上的投入。Taverna平台允许用户在远程与本地机器上构建复杂的分析工作流，并使用他们自己的数据来运行工作流并对计算结果进行可视化。

2.2 Triana

Triana[5]是由Cardiff大学在EU的资助下为GridLab实验开发的一个开源问题解决环境，在强大的数据分析工具中组合了一个可视化接口，已经被科学家广泛应用于信号、文本与图像处理等多个应用领域中。Triana最初是在1990年为GEO600设计的，此后在多个领域中进行了扩展，目前系统中已经开发了500多个应用程序。Triana的一个重要的特征是系统中不存在任何控制结构，任务间所有的依赖关系都通过数据流进行表达，而循环选择等复杂的控制结构则通过专门的组件来实现。Triana也支持图形化的流程设计方式，并将流程自动保存成基于XML的作业描述语言GJD（GridLabJobDefinition）。Triana系统中支持多种类型服务的组合，其中包括Web服务、网格服务、Gridlab服务以及第三方服务等。

2.3 Pegasus

Pegasus[6]是美国的威斯康辛大学为GriPhyN项目开发的一个子项目，它是一个典型的使用规划技术来支持流程动态生成的工作流系统，目前已经应用于生物信息学、生物学、宇航学、高能物理、地震波检测与地震科学等研究领域中。与Taverna类似，在Pegasus中也实现了一个基于语义进行流程组合和表达的工具Wings，科学家可以在功能层设计工作流而不需要考虑实际的执行环境是网格还是一系列的Condor池或是本地机器，并通过XDTM语言对数据集与工作流流程进行抽象描述形成抽象工作流。Pegasus架构在底层作业调度器Condor和DAGMan之上，采取基于任务聚类的作业调度算法，工作流引擎对工作流任务进行聚合后提交到合适的资源上，并由该资源上的DAGMan或Condor代理对作业进行提交。

2.4 Kepler

Kepler[7]是由美国国家科学基金（NSF）资助，由UC Berkeley和San Diego超级计算中心联合开发的基于Java的科学工作流管理系统，其目标在于提供给科学家一个开源的科学工作流管理系统以帮助科学家进行流程设计，并在网格资源上进行调度和执行，达到提高工作效率的目标。Kepler在Ptolemy II的基础上开发而来，继承了Ptolemy II面向角色建模的特点，能在单个科学工作流中组合不同的计算模型，计算模型通过相应的Director进行控制。Kepler中的任务由Actor进行表示，通过输入输出端口对多个Actor进行连接，形成科学工作流。Kepler利用内嵌的并行控制和工作流调度机制，将科学工作流的设计、执行、运行时交互、本地和远程数据访问、本地和远程服务调度无缝地组合起来。Kepler主要应用于生物学、生态学、天文学以及社会生态学等领域，也有多个应用系统在其基础上进行二次开发。

2.5 KNIME

KNIME（Konstanz Information Miner）[19]最初设计目标是建立一个具有友好操作界面、智能的、集数据处理、数据转换、数据分析和数据调查于一体的数据挖掘平台，目前也被用于一些场合的科学工作流建设，如OpenPHACTS项目。KNIME使用户以视觉化的方式创建数据流或数据通道以及工作流，可选择性地运行一些或全部的分析步骤，并可以对分析的结果进行图形处理以及交互式处理。

KNIME由Java写成，其基于Eclipse并通过插件的方式来提供更多的功能，用户也可以根据自己的需要，编写具有独特功能的节点。KNIME支持的处理操作广泛，从最基本的数据操作（例如为统计函数进行数据筛选、整合，如计算均值、标准差或是进行线性回归系数），到需要大量计算的数据处理任务（如聚类、决策树、神经网络）。此外，大多数的拥有建模功能的节点会为用户提供一个交互式的环境，帮助用户透过多种不同视图来探索产生的数据。KNIME的数据流程包含若干节点，节点之间通过流水线进行连接，数据或模型在这些流水线上传输。每个节点会处理到来的数据或模型，当需要数据输出时，节点会产生结果来满足要求。

上述的5个系统有各自的特色，应用的领域也有

所不同。Taverna大量整合了分子生物学领域的工具和数据库，支持针对特定领域的数据文档的处理，并在myExperiment[8]中共享了大量已开发的工作流，具有较好的领域适用性。Kepler能操作很多格式的数据，既可以本地运行，也可以联网运行。强大的网络能力使Kepler软件能帮助用户分享、复用那些由科学社区开发的数据、工作流和构件，从而满足一般的公共需求。所以需要远程交流的用户可以考虑选择Kepler。Pegesus系统都是基于分布式的，面向大数据量的计算，如果有许多数据需要计算，比如DNA序列的生物研究，还有天文学的相关研究，都需要进行大量的数据计算，可以考虑使用上述系统。

3 数据驱动的生物医学研究

3.1 密集数据带来的科研困扰

目前，生物医学数据呈现大规模、快速增长的态势。 随着基因测序、高通量筛选等技术的快速发展，大量的候选基因被识别；基于QTL分析，每个染色体区域可以产生超过200个基因；微阵列基因表达研究可以将整个基因组嵌入到一个芯片上；而且这些基因信息在不断地变化，这些现状都为生物医学领域的科学研究带来巨大挑战，当前的科学研究方法也呈现出了一定的局限性，主要体现在：

（1）数据规模过大，导致研究人员难以分析；

（2）为了开展研究，通常进行数据的筛选，而这些筛选往往带有科研人员的主观色彩，甚至是采用不成熟的筛选策略；

（3）大多数情况下，科研人员仍然遵从假设驱动的数据分析；

（4）科学数据的更新加快，常常需要对变化的数据进行重新分析；

（5）有时候为了数据分析，采用并不恰当的方法；

（6）错误被逐层放大，这很可能是由于在某一个环节出现了人为错误，或者是上述任何一种原因导致。

面对如此多的数据问题，科学研究过程迫切需要借助计算机技术来延伸科研人员的数据处理能力，这就需要用自动化的方式来分析数据。科学工作流成为了解决这种数据困扰的途径之一。

3.2 科学工作流助力生物医学研究

生物医学研究的一个重要环节是科学实验，而科学实验的一个重要特点则是实验过程的可重复性和实验结果的可验证性。在数据密集型数据实验过程中，科学工作流将会逐步成为数据驱动的科学研究的重要核心，它提供一系列的技术手段用于支持科学实验。科学工作流将会成为一种将常规科学研究过程进行系统化、精确化、可重复执行的理想模式。

从抽象层次来看，科学工作流是一种在计算机中实现的具备明确、准确、模型化表示的科学实验操作协议，在多层次上来支持数据驱动的的科研，包括数据、服务、方法和工作流模板等，科学工作流将在以下方面发挥重要的作用[9,10]：

（1）科学工作流的动态生成和自由组合功能能够很好地支持科研人员将即刻的需求转化为可操作的实践；

（2）通过将现有工作流重新配置或改造成新的组件，加速科学实验设计；

（3）科学工作流提供了一种系统化和自动化的途径，用来对各种不同的数据集进行分析并支持多类型的应用；

（4）科学工作流捕捉到了科研的形式化过程，从而使科研结果可以重现，科研方法可以被检验和重复利用，以及二次改造；

（5）科学工作流往往都提供一个可视化的界面帮助用户操作，科研人员可以在不具备底层编程知识的背景下创建这些流水线，同时科研人员也不需要掌握所有的专业知识，只需要对各种功能节点进行组装；

（6）超越数据集成，科学工作流固化了数据挖掘、知识发现、算法的参数控制等高级过程；

（7）在一个通用软件平台和共享框架下，科学工作流将各种科学应用以明确和可重用的规范集成进来；

（8）科学工作流平台是一个不断增长的资源池，在开放共享的科研趋势下，大量的独立资源会被不断加入到这个资源池中，便于广泛范围下科研人员的访问和使用。

3.3 科学工作流系统的领域适用性分析

由于科学求解过程的不同和数据类型的差异，一般情况下，科学工作流系统都有各自倾向的适用领域。由于Taverna大量整合了分子生物学领域的工具和数据库，且能够支持大部分的生物数据处理，被认为是生物

医学领域的科学工作流代表性系统。

Taverna提供了一个用于流程定义的可视化操作平台，以及数据演化过程中的来源数据自动捕获和记录功能，可以支持科学实验的重运行。尽管Taverna并不能支持所有类型的医学数据，但它支持多种Web Services的服务调用，还支持可扩展的组件集成，特殊需求的组件可以自行开发后嵌入到Taverna系统中，与它现有的组件共同生成工作流，具有足够的扩展功能。可见，在生物医学领域，Taverna是一个较为理想的科学工作流系统。

4 基于Taverna的生物医学工作流实践

本文开展了基于Taverna的生物医学工作流实践，面向生物医学具体应用场景，基于Taverna平台设计了一个从关联数据中获取Alzheimer病相关信息的工作流实例。在实践过程中，本文详细分析了Taverna的工作机制，对其进行了配置、调用和测试。

Taverna提供了多达3000多个服务，在进行工作流组合时，为了便于在大量服务中进行查找和选择，Taverna提供基于语义的服务组合方法，通过GRIMOIRES对服务进行语义注册，并将语义元数据存储在KAVA中；在流程设计过程中，由Feta组件对KAVA中的服务进行语义查找并组合成抽象工作流；抽象工作流由Scufl语言进行描述，并由工作流引擎FreeFluo来进行解析和调度。图1为Taverna 2.X系列版本的平台模块图。

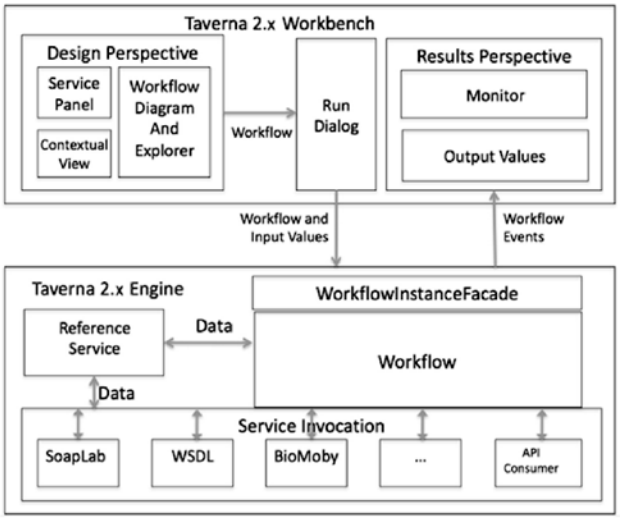

图1 Taverna 2.x平台结构图

当前采用的Taverna版本是2.2版，Taverna工作台分为三个大的区域，分别是服务调用区、工作流浏览器和工作流面板。

服务调用区提供了所有可以访问和使用的服务，以树形结构组织了一个服务目录，这些服务由本地服务和远程服务共同组成，能够支持大多数生物医学领域科学研究场景下的功能分解和流程组装，如图2所示。

工作流浏览器显示了用户编辑工作流的详细信息，包含输入、输出的默认值和描述，远程服务如何分配，也

图2 Taverna的服务目录

包括配置参数细节的显示，例如迭代和循环。Taverna还支持对工作流的验证，在执行一个工作流之前，Taverna会检查它的内部连通性以及服务可获得性等。

工作流面板是当前编辑工作流的可视化显示区域，它支持输入、输出、服务和数据流的显示，支持通过拖拽的方式连接服务或者编辑工作流，面板还支持对工作流的存储和共享。一个工作流面板中编辑的工作流实例如图3所示。

以上工作流支持从Bio2RDF中获取有关Alzheimer疾病相关的基因、蛋白、遗传、PubMed文献等信息，该实例可以扩展到任何一种疾病相关信息的查询。该实例是基于myExperiment中的工作流进行的二次开发，myExperiment是一个优秀的工作流仓储，其中存储了大量的共享工作流，支持用户的二次利用。基于大量的共享工作流和Taverna的服务，我们可以认为任何科研流程的形式化表示，都可以通过Taverna实现，关键的问题是在复杂性和重用性之间进行取舍，从而形式化那些适合用工作流长期存储和执行的科研过程。

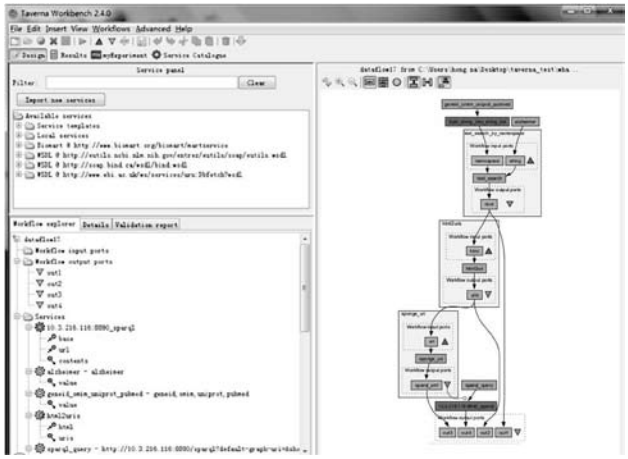

图3 工作流实例——查找Bio2RDF关联数据中有关Alzheimer病的相关信息

5 结语

尽管科学工作流已经在众多的科研领域开展了实验，但是其发展仍然受限于多种因素，诸如工作流系统之间缺少支持互操作的标准、用户不同层面的需求难以同时满足、来源数据的及时获取不易实现等问题，但是目前最迫切需要解决的问题是工作流的共享和利用，建立一个保障工作流长期保存和演化的机制，才能真正推动科学工作流的应用；同时利用用户共享和用户标注的机制，支持工作流的组织和获取；工作流开发者在共享工作流时应当尽量多地提供元数据和描述文档，从而有效支持工作流的利用和二次开发；为了保障工作流的灵活机制，应当尽可能地创建独立功能的小型工作流，多个小型工作流可以灵活组装成面向不同功能的大型复杂工作流。由此，科学工作流才能真正发挥其价值，将科学家从繁琐的常规数据处理工作中解脱出来，使他们集中精力关注研究内容，促进研究探索和科学发现。

参考文献

[1] GOBLE C, DE ROURE D. The impact of workflow tools on data-centric research [J]. 2009.

[2] SINGH M P, VOUK M A. Scientific workflows: scientific computing meets transactional workflows [C]// Proceedings of the NSF Workshop on Workflow and Process Automation in Information Systems: State-of-the-Art and Future Directions. 1996: 28-34.

[3] LUDÄSCHER B, ALTINTAS I, BOWERS S, et al. Scientific process automation and workflow management [J]. Scientific Data Management: Challenges, Existing Technology, and Deployment, Computational Science Series, 2009: 476-508.

[4] OINN T, GREENWOOD M, ADDIS M, et al. Taverna: lessons in creating a workflow environment for the life sciences [J]. Concurrency and Computation: Practice and Experience, 2006, 18(10): 1067-1100.

[5] MAJITHIA S, SHIELDS M, TAYLOR I, et al. Triana: A graphical web service composition and execution toolkit [C]// Proceedings, IEEE International Conference on Web Services, IEEE, 2004: 514-521.

[6] DEELMAN E, SINGH G, SU M H, et al. Pegasus: A framework for mapping complex scientific workflows onto distributed systems [J]. Scientific Programming, 2005, 13(3): 219-237.

[7] ALTINTAS I, BERKLEY C, JAEGER E, et al. Kepler: an extensible system for design and execution of scientific workflows [C]// Proceedings, 16th International Conference on Scientific and Statistical Database Management. IEEE, 2004: 423-424.

[8] myExperiment [EB/OL].[2013-12-29]. http://www.myexperiment.org/.

[9] DEELMAN E, GANNON D, SHIELDS M. Workflows for e-Science [M]. Springer-Verlag London Limited, 2007.

[10] The fourth paradigm: data-intensive scientific discovery [J]. 2009.

作者简介

洪娜（1980-），中国医学科学院医学信息研究所副研究员，研究方向：语义Web、关联数据、大数据。E-mail: hong.na@imicams.ac.cn
钱庆（1970-），中国医学科学院医学信息研究所研究员，研究方向：知识组织、知识发现、大数据。
方安（1976-），中国医学科学院医学信息研究所副研究员，研究方向：知识组织，工具集成。
吴思竹（1981-），中国医学科学院医学信息研究所副研究员，研究方向：文本挖掘，知识组织。
杨林（1984-），中国医学科学院医学信息研究所副研究员，研究方向：科学数据，数字图书馆。

Data Driven Scientific Workflow and Its Application in Biomedicine

Hong Na, Qian Qing, Fang An, Wu Sizhu, Yang Lin / Institute of Medical Information of Chinese Academy of Medical Sciences, Beijing, 100020

Abstract: E-Science focuses on the scientific activities under digital environment. However, with the burst of biomedical big data, data intensive scientific research brings the new challenges to e-Science. Scientific workflow supports formalizing the flow of scientific computing, automatically coordinates multi-task and multi-steps process under a special program environment for reducing cost and promoting efficiency. In this paper, we explicit the definition of scientific workflow and then analyze the main current scientific workflow systems, besides, we propose how to deal with intensive biomedical data within scientific workflow, and some experiments have been done by using Taverna workbench. Finally, we conclude state of art of current research and some problems in this field.

Keywords: Scientific workflow, e-Science, Data intensive, Workflow share, Taverna

（收稿日期：2014-02-14）

书 讯

《国家学位论文资源管理与共享系统研究》

我国学位论文法定收藏工作已正式确立30周年。为进一步推进学位论文资源建设与共享服务，中国科学技术信息研究所所长贺德方研究员牵头组织赵嘉朱、姜爱蓉、陈传夫、张建勇、曾建勋五位专家学者对近十年来相关联合研究课题的成果进行归纳、凝练、整理和补充，形成了《国家学位论文资源管理与共享系统研究》这部学术专著。

本书总结了我国学位论文的管理历程，凝练了我国学位论文的管理建议，归纳了学位论文资源的整合方案，提出了学位论文资源的共享模式，展望了学位论文的深层次服务目标。对于全社会整体把握学位论文收藏格局，系统构建学位论文管理体系有指导作用；对于高校和研究机构的学位管理人员强化学位论文质量管理和过程监控，做好学位论文存交，推进学位论文管理和机构知识库建设有参考作用；对于图书信息机构做好学位论文管理和收藏，推进学位论文资源组织和保障利用有引导作用；也可作为高校信息管理等专业研究生学习的参考书。

《国家学位论文资源管理与共享系统研究》于2014年3月由科学技术文献出版社出版，全书27万字，定价68.00元。

数据驱动的科学工作流及其在生物医学中的应用实践

作者： 洪娜， 钱庆， 方安， 吴思竹， 杨林， Hong Na， Qian Qing， Fang An， Wu Sizhu， Yang Lin
作者单位： 中国医学科学院医学信息研究所 北京 100020

刊名： 数字图书馆论坛ISTIC
英文刊名： Digital Library Forum
年，卷(期)： 2014(3)