MoBiFlow – ein Web-2.0 basiertes Workflowsystem für die Mikrobiologie

Wolfgang Küchlin, Markus Held

Symbolisches Rechnen
Wilhelm Schickard-Institut für Informatik
Universität Tübingen
72076 Tübingen
Kuechlin@informatik.uni-tuebingen.de
www-sr.informatik.uni-tuebingen.de

Abstract: Wir diskutieren den Einsatz von Service Orientierten Architekturen und Workflows in der Wissenschaft am Beispiel der Mikrobiologie. Workflows erlauben die explizite Repräsentation und automatische Durchführung von elektronischen Experimenten. Wissenschaftliche Workflowsysteme haben aber spezifische Anforderungen. Am Beispiel von MoBiFlow, einem System für die Mikrobiologie, zeigen wir, welche Vorteile ein konsequenter Einsatz von Web-2.0 Technologien und BPEL haben kann.

1 Einleitung

Ein Geschäftsprozess *(Business Process)* ist ein Ablauf in einem Unternehmen, der in einer Folge unterscheidbarer Aktivitäten ein Resultat erzeugen soll. In diesem Sinne ist auch eine Forschungseinrichtung ein Unternehmen, dessen Geschäftsprozesse aus (reproduzierbaren) Experimenten und Verfahrensabläufen bestehen, die an der Einrichtung beherrscht werden. In der Mikrobiologie werden Experimente zusehends auch auf dem Computer durchgeführt, indem Datenbanken abgefragt oder Datenmengen mit Algorithmen der Bioinformatik behandelt werden. Diese sog. *in silico* Experimente können durch Workflows repräsentiert und automatisiert ausgeführt werden.

In der Wissenschaft können aber nicht einfach Workflow-Systeme aus der Wirtschaft übernommen werden, schon weil nur ein vergleichsweise sehr bescheidener Aufwand bei der Erstellung und beim Betrieb eines Systems geleistet werden kann. Das System muss „schlank“ sein, sich leicht erweitern, leicht bedienen und auch sehr leicht administrieren lassen. Gleichzeitig muss das System hinreichend stabil und leistungsfähig sein, damit sich ein Forschungsinstitut darauf verlassen kann. Es muss zudem auf den jeweiligen Einsatzbereich maßgeschneidert sein und darf nur minimale oder gar keine Kenntnisse in Informatik und Programmierung erfordern. Mit unserem System *MoBiFlow* versuchen wir zu zeigen, welche Vorteile eine Systemarchitektur bietet, die sich auf Java, BPEL und Web 2.0 Techniken stützt.

2 SOA und Workflows in der Mikrobiologie

SOA (*Service Oriented Architecture*) bezeichnet eine Software-Architektur, durch die wiederverwendbare (Software-)Dienste im Internet bereitgestellt und immer wieder aufs Neue zu wechselnden Anwendungen verknüpft werden. Solche Dienste werden in einer SOA Architektur nicht mehr nur von Menschen über einen Web-Browser genutzt, sondern direkt von weiteren Computerprogrammen. In der Biologie handelt es sich bei den Diensten zum Beispiel um (Gen- und Protein-) Datenbanken und Auswertefunktionen der Mikrobiologie, die von den jeweiligen Forschungsgruppen weltweit betrieben und „ins Netz gestellt" werden. Forscher an beliebigen anderen Orten können darauf aufbauend höhere Auswertefunktionen und elektronische Experimente in Form von Computerprogrammen schreiben, die weitere Daten und Schlussfolgerungen aus den einzelnen Diensten herleiten [BHW09].

Am Beispiel der Biologie erkennt man auch den Vorteil von Workflows in der Wissenschaft. Bisher muss die Biologin jeden Dienst separat im Browser aktivieren, manuell mit Daten und Einstellungen versorgen und die Ergebnisse manuell in die Oberfläche des nächsten Dienstes kopieren *(cut-and-paste)*, wobei Formatänderungen manuell vorzunehmen sind. Wurde dagegen ein entsprechender Workflow definiert, kann dieser automatisch ablaufen und beliebig oft (mit neuen Ausgangsdaten) wiederholt werden. Da sich während der Forschung die Abfolge des Experiments häufig ändert, bis die gewünschten Ergebnisse erzielt sind, ist es von großem Vorteil, wenn die Biologin den Workflow selbst in einfacher Weise definieren und ändern kann. Danach kann der Workflow gleichzeitig als Dokumentation dafür dienen, wie die erzielten Ergebnisse zustande kamen. Bei biologischen Experimenten spricht man hier von der wertvollen *provenance information*.

BioMOBY [WL02,Bi08] ist ein quelloffener Verzeichnisdienst und Quasi-Standard für Web Services im Bereich der Bioinformatik. Die Schnittstellen der BioMOBY Dienste müssen zum Zweck der Interoperabilität vorgegebenen Konventionen genügen. Die BioMOBY Diensteveraltung (*registry*) definiert drei Ontologien für Namensräume, die bioinformatische Datenbanken, Datentypen und Analysetypen repräsentieren. Die Dienste werden über einen zentralen Server angeboten und verwaltet, der auch eine semantische Suche ermöglicht. Dienste können aufgrund ihrer Namen oder ihrer Eingabe- und Ausgabe-Datentypen gefunden werden. Jedes BioMoby-Objekt repräsentiert eine biologische Entität, die durch einen Datenbanknamensraum und eine ID eindeutig identifiziert ist. Dabei kann dieselbe biologische Entität wie z.B. ein Genom in verschiedenen Datenbanken mit verschiedenen IDs unterschiedlich repräsentiert sein. Es kann sinnvoll sein, aus Effizienzgründen nur den Identifikator eines Namensraums und eine ID als Repräsentation einer biologischen Identität zu versenden. Je nach Kontext können z.B. Strings, Formeldarstellungen oder sogar Base-64-kodierte Bilder als Repräsentation biologischer Entitäten dienen. BioMoby Dienste können sehr unterschiedlichen Funktionen dienen, so etwa dem Vergleich von Gensequenzen, der Suche in Publikationsdatenbanken, oder der Konversion von IDs verschiedener Namensräume.

Für BioMoby gibt es verschiedene Browser wie z.B. Gbrowse und Seahawk [Wi06,GS07], mit deren Hilfe Eingabedaten in mehreren Schritten von unterschiedlichen Diensten verarbeitet werden können. Solche Sitzungen stellen implizit lineare Arbeitsabläufe dar, die bei einigen BioMoby Browsern als Workflows für bioinformatische Workflow-Systeme (z.B. Taverna) gespeichert werden können. Für einige Workflow-Systeme existieren zudem BioMoby-Plugins.

Die Taverna Workbench [Oi04,Oi06] ist ein Workflow-Management System für die Bioinformatik aus dem myGrid Projekt. Taverna läuft als Desktop-Applikation, muss also lokal installiert werden. Taverna ermöglicht die Erstellung und Ausführung von Workflows und benutzt dazu die proprietäre XML-Sprache Scufl, die für diesen Zweck geschaffen wurde. Mit einer Erweiterung von Taverna können Benutzer mit einem Workflow in begrenztem Umfang über eine separate Web-Seite interagieren [LO08]. Es ist nicht gesichert, dass Scufl-Workflows auch in Zukunft ausgeführt werden können, da Taverna und Scufl keine kommerzielle Verwendung gefunden haben und ihre Finanzierung ausschließlich von öffentlicher Forschungsförderung abhängt.

2 MoBiFlow: Workflow-Erstellung mit Web 2.0 Techniken

In Zusammenarbeit mit dem Tübinger Zentrum für Molekularbiologie der Pflanzen (ZMBP) haben wir prototypisch das Workflow-System „MoBiFlow“ entwickelt, das für Biologen folgende Vorteile bietet: einfache grafische Gestaltung der Workflows, Übersetzung der grafischen Workflows in BPEL als Zielsprache, kollaboratives Arbeiten innerhalb einer Browser-Oberfläche mit minimalem Installationsaufwand durch Nutzung von Web 2.0 Techniken, eingebaute Unterstützung zur Nutzung des BioMOBY Service Repositories aus der Molekularbiologie, sowie automatische Bewertung der Qualität der Workflows durch Metriken [HB08,HB09,HBW09,He10].

MoBiFlow besteht aus einem allgemeinen Teil (Hobbes) und einem Zusatzmodul (Calvin) für den Einsatz in der Molekularbiologie am ZMBP. Mit dem System sollte demonstriert werden, dass man unter Verwendung moderner Web 2.0 Softwaretechniken und unter Rückgriff auf den Sprachstandard BPEL mit vergleichsweise moderatem Aufwand ein Workflow-System mit sehr nützlichen Eigenschaften erstellen kann. Hobbes unterstützt eine grafische Darstellung und Manipulation von Workflows mit BPEL Kontrollfluss-Elementen, die nach BPEL als Zielsprache übersetzt werden. Als Internet Applikation benötigt eine Hobbes Installation nur einen einzigen zentralen Server, die Client-Oberfläche wird im Browser dargestellt und braucht nicht installiert zu werden. Dies ist z.B. für die Anwendung am ZMBP wichtig, da nur begrenzte Unterstützung durch Techniker verfügbar ist. (Die derzeitige Architektur baut auf Adobe Flex mit dem Flash plug-in auf, kann aber bei Bedarf mit begrenztem Aufwand auf Alternativen wie HTML-5 umgestellt werden.) Außerdem erlaubt diese Architektur die Unterstützung für kollaboratives Arbeiten über das Internet, etwa zwischen verteilten Arbeitsgruppen oder zwischen Heim-Anwendung und Institut.

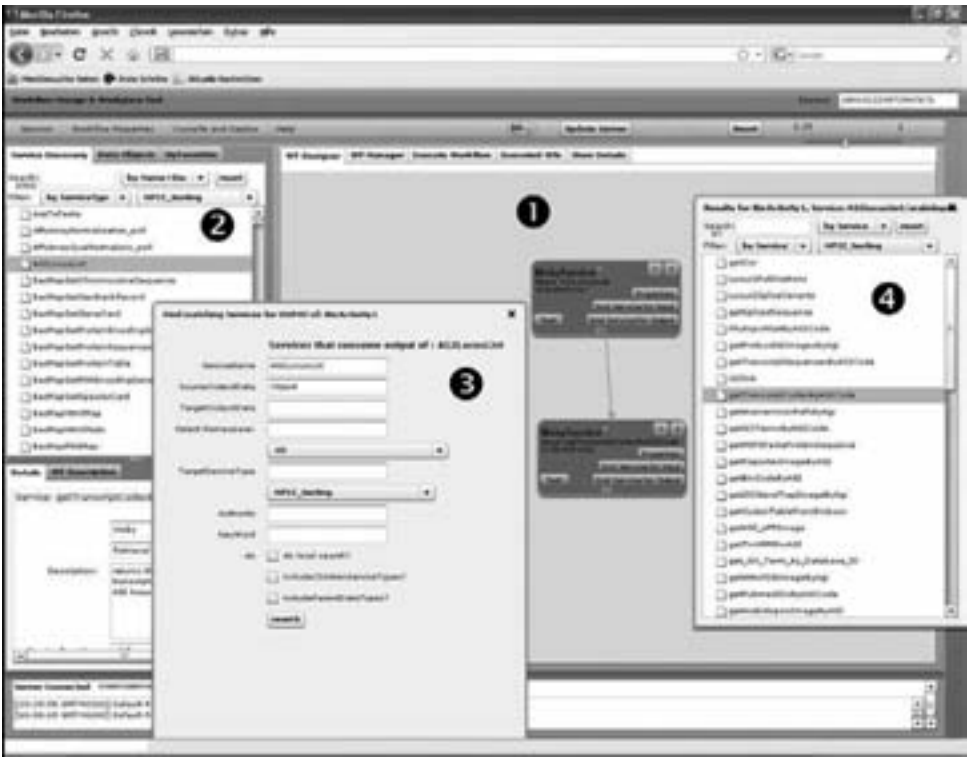

Abb. 1. The Calvin Workflow System. The workflow editing screen comprises an editing canvas (1), representing the workflow, and a list of available BioMOBY Web Services (2). Users can access a properties form for any task in the workflow (3), and search for services which consume its output data type or produce its input data type (4).

Das System Calvin kann als Zusatzmodul zu Hobbes verstanden werden, mit dem speziell die Konstruktion von Workflows unterstützt wird, die solche Web-Services aus der Molekularbiologie nutzen, die den BioMOBY Richtlinien gehorchen. Calvin kommuniziert hierzu mit dem BioMOBY Server und präsentiert dessen Dienste in einem eigenen Fenster. Für die Dienste stehen die Suchfunktionen des BioMOBY Servers zur Verfügung. Einzelne Dienste können mit der Maus in die Editierfläche gezogen werden und erhalten dadurch eine grafische Repräsentation als Kästchen mit integrierten Schaltflächen, hinter denen Anfragefunktionen des BioMOBY Servers stehen. U.a. kann auf diese Weise nach weiteren Diensten gesucht werden, mit denen die Workflow-Kette passend erweitert werden kann. Die Kästchen werden mit Pfeilen (meist entlang von Datenflüssen) zu Workflows verbunden. Durch automatische Anfragen an den BioMOBY Server kann festgestellt werden, ob die Pfeile jeweils kompatible Dienste verbinden; entsprechend werden die Pfeile grün oder rot eingefärbt. Gegenüber Hobbes sind die Calvin Workflows in ihrer Struktur auf die Bedürfnisse der Anwendung weiter eingeschränkt.

Im Umfeld von Calvin und Hobbes sind weitere Entwicklungen entstanden: Ein System von Held und Günter [HG09] zum kollaborativen Design von Web-Service Interfaces und ein System von Held und Lehle [HL09] zum Anreichern von Web-Applikationen um ein Video-Konferenztool gemäß dem REST Architekturstil. Diese Arbeiten können als Beispiel dafür dienen, dass mit den Techniken, mit denen Hobbes entwickelt wurde, auch eine Weiterentwicklung zu einer umfassenden Toolbox für Internet-basierte Experimentalforschung möglich ist.

3 Ausblick

Der Einsatz von Workflow-Systemen in der Wissenschaft bietet eine Reihe neuer Möglichkeiten und Herausforderungen. Es ist gut denkbar, dass zukünftige Workflows es auch erlauben werden, technische Apparate (z.B. Datenquellen) einzubinden und zu managen, oder dass Workflows über mobile Geräte (SmartPhones) beobachtet oder gesteuert werden können.

Wissenschaftliche Workflows repräsentieren das Herkunftswissen von wissenschaftlichen Ergebnissen und sind daher ähnlich wichtig wie die Ergebnisse selbst. Sie können Gegenstand wissenschaftlicher Diskussion werden, wenn man sie mit Kommentaren und Besprechungen versehen kann. Zumal in Verbindung mit Web 2.0 Techniken ergeben sich neue Möglichkeiten zur umfassenden Repräsentation von wissenschaftlichen Ergebnissen und zum wissenschaftlichen Diskurs (vgl. MyExperiment [RGS09,RG09] zum Austausch von Taverna Workflows).

Workflows können zu einem integralen Bestandteil für Labor Informationssysteme (*Laboratory Information System – LIMS*) werden. Ein LIMS dient zum umfassenden Management von Labordaten gemeinsam mit den Datenquellen (Instrumenten und Workflows) sowie Meta-Daten (Publikationen, Bilder, Informationen über die beteiligten Wissenschaftler etc.). Es ist abzuwarten, ob die neue DFG Initiative „Informations-Infrastrukturen für Forschungsdaten“ sich ebenfalls auf Workflows erstrecken wird.

Literaturverzeichnis

[BHW09] Berendzen KW, Harter K, Wanke D. (2009). Analysis of plant regulatory DNA sequences by transient protoplast assays and computer aided sequence evaluation. *Methods Mol Biol*. 479:311-35.

[Bi08] BioMOBY Consortium (2008). Interoperability with Moby 1.0—It's better than sharing your toothbrush!. Briefings in Bioinformatics 2008 9(3):220-231.

[GS07] Gordon PM, Sensen CW (2007). Seahawk: moving beyond HTML in Web-based bioinformatics analysis, *BMC Bioinformatics*, vol. **8**, no. 208, 2007.

[HB08] Held M, Blochinger W (2008). Collaborative BPEL Design with a Rich Internet Application. In: *Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGrid'08)*, Lyon, France, 19-22 May 2008.

[HB09] Held M, Blochinger W (2009). Structured Collaborative Workflow Design. *Future Generation Computer Systems - The International Journal of Grid Computing and E-Science*, **25**(6):638-653, 2009. http://dx.doi.org/10.1016/j.future.2008.12.005

[HBW09] Held M, Blochinger W, Werning M (2009). E-Biology Workflows with Calvin, In: *10th International Conference on Web Information Systems Engineering (WISE'09)*, Springer LNCS 5802, Poznan, Poland, 3-7 October 2009.

[He10] Held M (2010). *Web-based Collaborative Workflow Design*. Dissertation. Fakultät für Informations- und Kognitionswissenschaften, Universität Tübingen. 2010 (eingereicht).

[HG09] Held M, Günther M (2009). Collaborative Web Service Interface Design on the Web 2.0, In: *Proc. of the 1st International Conference on Social Informatics (SocInfo 2009)*, Warsaw, Poland, 22-24 June 2009, IEEE Computer Society.

[HL09] Held M, Lehle D (2009). Augmenting Collaborative Web Applications with RESTful Video Conferencing. In: *The IASTED International Conference on Parallel and Distributed Computing and Networks*, February 16 – 18, 2009, Innsbruck, Austria.

[LO08] Lanzen A, Oinn T (2008). The Taverna Interaction Service, *Bioinformatics* **24**(8), pp. 1118–1120, 2008.

[Oi04] Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows, *Bioinformatics*, vol. 20, no. 17, pp. 3045–3054, 2004.

[Oi06] Oinn T, Greenwood M, Addis M, Alpdemir M N, Ferris J, Glover K, Goble C, Goderis A, Hull D, Marvin D, Li P, Lord P, Pocock MR, Senger M, Stevens R, Wipat A, Wroe C (2006). Taverna: lessons in creating a workflow environment for the life sciences: Research articles, *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, pp. 1067–1100, 2006.

[RG09] De Roure D, Goble C (2009). Software Design for Empowering Scientists, *IEEE Software*, Volume 26, Issue 1, Jan-Feb 2009, pages 88-95. Digital Object Identifier 10.1109/MS.2009.22

[RGS09] De Roure D, Goble, C, Stevens R (2009). The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. *Future Generation Computer Systems*, **25**(5), May 2009, Pages 561-567.

[Wi06] Wilkinson MD (2006). "Gbrowse Moby: a Web-based browser for BioMoby Services," Source *Code for Biology and Medicine*, vol. 1, no. 4, 2006.

[WL02] Wilkinson, MD, Links, M. (2002). BioMOBY: an open-source biological web services proposal. *Briefings In Bioinformatics* 3:4. 331-341.