IT-Unterstützung Translationaler Forschung im Rahmen der Clinical and Translational Science Awards

S.H.R. Wurst, G. Lamla, D. Schmelcher, F. Prasser, K.A. Kuhn

Lehrstuhl für medizinische Informatik
Klinikum rechts der Isar der TU München
sebastian.wurst@tum.de

Abstract: Im Rahmen des US-Förderprogramms Clinical and Translational Science Awards (CTSA) werden seit 2006 IT-Infrastrukturen für die translationale medizinische Forschung aufgebaut Datenbanken und Integrationsmethoden spielen eine bedeutende Rolle. Dieser Beitrag stellt die Architekturen anhand relevanter Beispiele vor.

Nach dem Erfolg des Human Genome Project entwickelte sich ein besseres Verständnis des Zusammenwirkens genetischer und umweltbedingter Faktoren bei der Entstehung von Krankheiten. Datenbanken und Integrationsmethoden spielen eine bedeutende Rolle, wenn es darum geht, genetische und molekularbiologische Daten in Beziehung zu Phänotypen und Verläufen zu setzen.

Im Aufbau oder Einsatz befindliche Integrationslösungen fokussieren auf Einrichtungen, Regionen oder internationale Strukturen. Es werden v.a. physische, aber auch virtuelle Repositories verwendet, wobei Hybridansätze vorhanden sind. Eine typische Anforderung ist die Überbrückung semantischer Heterogenität auf Typ- und Instanzebene. Data-Warehouse-Systeme und ETL-Methoden spielen neben föderierten Ansätzen eine große Rolle, wobei die Konsolidierung von Komponentenschemata und v.a. das Data Cleansing erheblichen Aufwand fordern. Der Einsatz von Ontologien soll den Integrationsaufwand beherrschbar machen. Im Rahmen der National Cancer Institute (NCI)-Initiative caBIG (Cancer Biomedical Informatics Grid) werden gemeinsame Vorgaben sowie Anwendungen und Werkzeuge für Datenmanagement und Analysen erstellt [caBIG]. Hierzu gehören ein gemeinsames Framework zur Anwendungsentwicklung, Common Data Elements (CDE) und ein Data Standards Repository (caDSR). Das Projekt caGrid realisiert in diesem Rahmen den integrierten Zugriff auf eine Sammlung von Informationsressourcen und stellt hierzu ein verteiltes Middleware - Framework zur Verfügung. Der Ansatz wurde als „Semantic Web Data Warehousing“ beschrieben [Cu09].

Wichtige Beispiele für den Einsatz des Methodenspektrums sind neben caBIG die Clinical and Translational Science Awards (CTSA) [Ze05, CTSA1]. Sie repräsentieren ein Förderprogramm des US National Institute of Health (NIH) bzw. des National Center for Research Resources (NCRR). Durch das Programm werden Zentren für klinische und translationale Forschung eingerichtet, die den Transfer von Ergebnissen zwischen

Grundlagenforschung, klinischer Forschung und Behandlung beschleunigen sollen. Zwischen 2006 und 2010 wurden 55 Institutionen in das CTSA-Programm aufgenommen, wobei die mittlere Förderung pro Institution bei über 40 Mio. USD für einen Zeitraum von 5 Jahren liegt [CTSA2]. Die biomedizinische Informatik spielt in den Konzepten der Standorte eine erhebliche Rolle.

Die Integrationslösungen der CTSA Programme erlauben einen guten Überblick über den State of the Art der Integrationslösungen für die translationale Medizin. Typische Warehouse Ansätze sind häufig. Das Mayo Clinic Life Sciences System verwendet bspw. ein Data Warehouse Konzept für die Integration von Behandlungs- mit Forschungsdaten einschließlich „Omics"-Daten. Auch das Research Warehouse der UC Davis verwendet diesen Ansatz. Das Oregon Health & Science University and Kaiser Permanente Virtual Datawarehouse verbindet zwei Data Warehouses der beiden Einrichtungen durch ein föderiertes DBMS [CTSA3]. Das Health Science Center der University of Texas in Houston baut eine Ontologie-getriebene Plattform auf, die Konzepte service-orientierter Architekturen einbezieht [Mi09].

Mit sehr großem Personaleinsatz und über viele Jahre hinweg wurde an der Harvard Medical School und bei Partners HealthCare eine Integrationsinfrastruktur aufgebaut, die im Rahmen der CTSA Förderung verwendet und weiterentwickelt wird. Wesentliche Komponenten sind physische Repositories für verschiedene Anwendungszwecke, die von allen assoziierten Klinika Daten empfangen. Das Partners HealthCare Clinical Data Repository setzt ein globales Schema in einem zentralen Repository um und bietet Zugriff auf Patientendaten für Krankenversorgungzwecke. Die Partners HealthCare Quality Patient Data Registry setzt einen Data Warehouse Ansatz für die Integration klinischer Daten um und wird für Abfragen im Rahmen der Qualitätssicherung verwendet. Das Partners HealthCare Research Patient Data Repository (RPDR) realisiert einen Data Warehouse Ansatz für die Integration von klinischen Daten für Forschungszwecke. Der Zugriff auf das RPDR wird durch eine einrichtungsinterne Ethikkommission geregelt, die interessierten Forschern nach einer Vorabfrage mit Aggregatwerten als Ergebnis den Zugriff auf das vollständige Ergebnis der Abfrage genehmigen kann. Das i2b2 Projekt erweitert die Funktionalität des RPDR in einer service-orientierten Architektur um Methoden der Datenverarbeitung und -analyse. Es kann eingesetzt werden, um Data Cleansing und Datenanalyse auf Daten aus RPDR Abfragen durchzuführen. Für die Abfrage über Projekte und Instanzen hinweg wurde das Abfragewerkzeug SHRINE entwickelt, das derzeit eine föderierte Abfrage über die drei i2b2 Instanzen der Harvard Klinika ermöglicht. [We09] Eine Evaluation von i2b2 für verschiedene Forschungsszenarien findet sich in [DMM09].

An der Vanderbilt University wird im Rahmen des BioVU Projekts [Ro08] eine Infrastruktur zur Integration geno- und phänotypischer Daten aufgebaut. Der dazu eingesetzte Ansatz ist mit einem Data Warehouse Ansatz vergleichbar. Zentrale Komponente ist ein physisches Repository, das aus dem EHR-System replizierte Daten verwaltet. Die Prozessschritte im Rahmen des Replikationsprozesses umfassen auch eine De-identifizierung der Daten. Zur Verknüpfung der de-identifizierten Patientendaten mit bspw. aus Biomaterialproben gewonnenen Daten wird aus dem eindeutigen EHR-Identifikator des Patienten ein Hashwert generiert und jeweils zugeordnet. Darauf

aufbauend werden Forschern verschiedene Anwendungen zur Verfügung gestellt. Mit REDCap [Ha09] lassen sich ohne großen Aufwand Formulare für die elektronische Datenerfassung entwickeln. REDCap Survey ermöglicht die einfache Erstellung von Formularen für Umfragen. Über die Portallösung StarBRITE lassen sich forschungsrelevante Informationen abrufen. Das Record Counter Werkzeug ermöglicht Abfragen nach Aggregatwerten auf den vorhandenen Datensätzen. Mit der Vanderbilt Volunteer Registry und dem Nachfolgesystem researchmatch können potentielle Probanden ihr Interesse an der Teilnahme an Studien anmelden.

Das von Intermountain HealthCare bzw. der University of Utah entwickelte FURTHeR System folgt einem leichtgewichtigen, föderierten Ansatz für die Integration von Informationssystemen auf Datenebene. Integriert werden sollen Utahs größte Patientendatenrepositories (University of Utah Healthcare, Intermountain Healthcare, Salt Lake City Veterans Administration Medical Center), Daten zum öffentlichen Gesundheitswesen des State of Utah Department of Health sowie genealogische und demographische Daten aus der Utah Population Database [Br09].

Die Kooperation zwischen caBIG und CTSA-Standorten führt zu weiteren wichtigen Projekten. An der University of California, San Francisco wurde ein ontologie-basierter Ansatz entwickelt, der sich gegen ein globales Schema positioniert. Begründet wird dies mit der Volatilität der Domäne, der großen semantischen Heterogenität und der Anforderung, dass verschiedene Benutzergruppen unterschiedliche Sichten auf die integrierten Daten benötigen [Wy1]. Ein universell einsetzbarer Instance Mapper, der CTSA Health Ontology Mapper (HOM), wird gemeinsam mit Partnerinstitutionen entwickelt [Wy10]. Er soll die Abbildung lokaler (nicht standardkonform kodierter) Daten auf Standardontologien ermöglichen. Diese müssen durch das ISO/IEC 111-79 Format für Datenmodelle definiert, durch das caDSR registriert oder mittels der für caBIG geschaffenen Enterprise Vocabulary Services annotiert werden. HOM wird direkt in die i2b2 Workbench integriert. Durch die Integration von i2b2 und caGrid soll ein robustes intra- und interinstitutionelles, föderiertes Anfragesystem entstehen [Wy2].

Literaturverzeichnis

[Br09] Bradshaw RL, Matney S, Livne OE, Bray BE, Mitchell JA, Narus SP. Architecture of a Federated Query Engine for Heterogeneous Resources. AMIA Annu Symp Proc. 2009;2009:70-74.

[caBIG] https://cabig.nci.nih.gov/overview/, last access July 29, 2010

[CTSA1] http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/ Last access Mar 08, 2010.

[CTSA2] http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/#ctsa_funding last access July 27, 2010

[CTSA3] https://www.ctnbestpractices.org/networks/nih-ctsa-awardees/#ctsa Last access Mar 08, 2010

[Cu09] McCusker JP, Phillips JA, González Beltrán A, Finkelstein A, Krauthammer M. Semantic web data warehousing for caGrid. BMC Bioinformatics 2009, 10(Suppl 10):S2

[DMM09]Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. BMC Medical Research Methodology 2009, 9:70 doi:10.1186/1471-2288-9-70

[Ha09] Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics. 2009 Apr;42(2):377-381.

[Mi09] Mirhaji P, Zhu M, Vagnoni M, Bernstam EV, Zhang J, Smith JW. Ontology driven integration platform for clinical and translational research. BMC Bioinformatics 2009, **10**(Suppl 2):S2 doi:10.1186/1471-2105-10-S2-S2 BMC Medical Research Methodology 2009, 9: 70.

[Ro08] Roden DM, Pulley JM, Basford MA, Bernard GR, Clayton EW, Balser JR, u. a. Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin. Pharmacol. Ther. 2008 Sep;84(3):362-369.

[We09] Weber GM, Murphy SN, McMurry AJ, Macfadden D, Nigrin DJ, Churchill S, Kohane IS. The Shared Health Research Information Network (SHRINE): a prototype federated query tool for clinical data repositories.. J Am Med Inform Assoc. 2009;16(5):624-30. Epub 2009 Jun 30.

[Wy1] Wynden R. An Alternative Approach to Integrated Data Repository Design. http://www.ctsaweb.org/uploadedfiles/New%20Approaches%20to%20Data%20Repository%20Architecture,%20Rob%20Wynden%20%28UCSF%29.pdf last access Jul 29, 2010

[Wy2] Wynden R. Uniting i2b2.org and caGrid. http://www.ctsaweb.org/uploadedfiles/RWynden%20UCSF%20Uniting%20i2b2%20and%20caGrid.ppt last access Jul 29, 2010

[Wy10] Wynden R, Weiner MG, Sim I, Gabriel D, Casale M, Carini S, Hastings S, Ervin D, Tu S, Gennari J, Anderson N, Mobed K, Lakshminarayanan P, Massary M, Cucina RJ. Ontology Mapping and Data Discovery for the Translational Investigator. AMIA CRI Summit 2010.

[Ze05] Zerhouni EA. Translational and Clinical Science – Time for a New Vision. New England Journal of Medicine 2005; 353(15): 1621-3