Published in: GIS Report 2012/13; Harzer, Bernhard (Ed.);
ISBN 978-3-9811899-4-0; pages 15-20; 2012.

Wissenschaftliche Umweltdaten in Geodateninfrastrukturen

Stephan Mäs, Lars Bernard, Matthias Müller, Christin Henzen

Technische Universität Dresden
Professur für Geoinformationssysteme,
Helmholtzstraße 10, 01069 Dresden, Germany
{Stephan.Maes, Lars.Bernard,Matthias_Mueller, Christin.Henzen}@tu-dresden.de

Einleitung

Vor etwas mehr als 10 Jahren prägte der damalige US Vizepräsident Al Gore [Gore 1999] den vielzitierten Begriff „Digital Earth“. Seine Vorstellung von virtuellen digitalen Globen, welche den nahtlosen Zugang zu raumzeitlichen global verteilten Informationen unterschiedlichster Auflösung und Skalierung sowie die Beschreibung der Umwelt mit ihren Abhängigkeiten und Veränderungen ermöglichen, ist heute zu einem großen Teil Realität geworden. Der enorme technologische Fortschritt der letzten Jahre im Bereich der Geodatenakquisition, Rechenkapazitäten, Internet Protokolle, Bandbreite und Geodatenprozessierung erlaubt die Nutzung derartiger Anwendungen nicht nur auf Desktop PCs, sondern auch in Form von „Apps“ auf mobilen Endgeräten. Infrastrukturen für den Geodatenaustausch zwischen räumlich verteilten Organisationen bilden die Basis solcher digitalen Globen. Die Entwicklungen sind dabei sehr unterschiedlich. Produkte kommerzieller Anbieter wie Google Earth, Microsoft Bing Maps oder auch ESRIs ArcGIS Online verstärkten die öffentliche Wahrnehmung von Geoinformationen und setzen Maßstäbe im Bereich der intuitiven Bedienung, Benutzerfreundlichkeit und Performanz. Als Gegenstück dazu bieten administrative Geodateninfrastrukturen (GDI), wie sie beispielsweise im Rahmen von INSPIRE [EC 2007] aufgebaut werden, offene standardisierte Schnittstellen und verpflichten sich gegenüber den Nutzern zur Einhaltung festgelegter Parameter für die Qualität der Daten und Dienste. Wissenschaftliche Umweltdaten sind mit wenigen Ausnahmen[1] bisher kaum Teil solcher Infrastrukturen. [Craglia et al. 2012] definieren daher die Vision einer “Digital Earth” für das Jahr 2020 neu und sehen dabei folgende Herausforderungen für die Bereitstellung wissenschaftlicher Umweltdaten:

- die Verlinkung multidisziplinärer Modelle für die Vorhersage und die Bewertung globaler Veränderungen,
- die Integration von Echtzeit- (nahen) Beobachtungen aus Sensornetzwerken und sozialen Netzen,
- die Berücksichtigung verschiedener Szenarien für politische Veränderungen und deren Auswirkungen,
- die Beschreibung wissenschaftlicher Erkenntnisse zu globalen Veränderungen, die Unsicherheiten mit denen diese behaftet sind sowie die vorgeschlagenen Maßnahmen und

[1] Beispielsweise Daten des IPCC (Intergovernmental Panel on Climate Change): www.ipcc-data.org/maps/

Reaktionen in den Bereichen der Wissenschaft, von (politischen) Entscheidungsträgern und der Öffentlichkeit.

Dieser Artikel fasst die Erfahrungen und Ergebnisse beim Aufbau einer GDI für wissenschaftliche Umweltdaten (insbesondere Ergebnisse von Simulationen und Modellrechnungen) im Rahmen eines BMBF Forschungsprogrammes zusammen. Im folgenden Kapitel wird auf die konkreten Rahmenbedingungen und die Projektziele eingegangen. Danach werden einzelne technische und organisatorische Aspekte und Besonderheiten bei der Veröffentlichung von wissenschaftlichen Umweltdaten in einer GDI diskutiert.

Projekthintergrund

In dem interdisziplinär angelegten Projekt "GLUES" (Global Assessment of Land Use Dynamics on Greenhouse Gas Emissions and Ecosystem Services)[2] werden globale Landnutzungsänderungen und deren Auswirkungen auf ökosystemare Dienstleistungen und Treibhausgasemissionen analysiert [Eppink et al. 2012]. Das Projekt GLUES dient darüber hinaus der wissenschaftlichen Koordination und Synthese der Verbundforschungsvorhaben im Teil A des BMBF - Programms "Nachhaltiges Landmanagement".

Abb. 1: Schwerpunktaktivitäten von GLUES (Quelle: UFZ, www.nachhaltiges-landmanagement.de)

In diesem Förderprogramm entwickeln zwölf Regionalprojekte Beispiellösungen für global verteilte Untersuchungsgebiete. GLUES unterstützt die international fachübergreifende Zusammenarbeit in diesen Forschungsprojekten durch die Kommunikation, Koordination und Integration der Ergebnisse durch eine gemeinsame Datenplattform und durch die Entwicklung von einheitlichen Szenarien in der Landnutzung, dem Klimawandel und sozioökonomischen Veränderungen (Abbildung 1). Die Synthese integriert die Ergebnisse der regionalen Forschungsprojekte in Bezug auf die Anforderungen verschiedener Nutzer- und Stakeholdergruppen. Durch die Entwicklung von

[2] http://modul-a.nachhaltiges-landmanagement.de/en/scientific-coordination-glues

Schnittstellen zu politischen Prozessen werden die Ergebnisse potenziellen Nutzergruppen zur Verfügung gestellt und in internationale politische Prozesse transferiert.

Für den Austausch der wissenschaftlichen Daten aus Modellrechnungen und Simulationen wird im Rahmen von GLUES eine Geodateninfrastruktur (GDI) aufgebaut. Damit wird auf technischer Ebene die Zusammenarbeit innerhalb von GLUES und zwischen den regionalen Projekten des BMBF Programmes Nachhaltiges Landmanagement unterstützt und eine Analyse und Synthese von globalen und regionalen Datensätzen zu Landnutzung, Treibhausgasemissionen und ökosystemaren Dienstleistungen ermöglicht. Weiterhin stellt die GLUES GDI technische Komponenten für die Außendarstellung des Projektes bereit.

Die GLUES GDI hat dabei die folgenden wesentlichen Ziele:

1. Die involvierten Forschungsgruppen erhalten die Möglichkeit ihre Modelldaten, Analyseergebnisse und Basisszenarien über die GDI zu publizieren und auszutauschen.
2. Existierende Datenquellen können durch die GDI nahtlos miteinander verknüpft werden, beispielsweise für die Berechnung wissenschaftlicher Modelle oder Vergleichsanalysen.
3. Interessenvertreter verschiedenster Bereiche werden durch die Such- und Analysewerkzeuge der GDI dabei unterstützt, Forschungsresultate aufzufinden und diese für die eigenen Planungs- und Managementaktivitäten einzusetzen.

Ziele der Veröffentlichung wissenschaftlicher Daten

Die Veröffentlichung wissenschaftlicher Daten als zusätzlicher Output neben den wissenschaftlichen Publikationen hat viele Vorteile, einige wesentliche sind:

- die verbesserte Dokumentation, Transparenz, Vergleichbarkeit und Nachhaltigkeit der Forschungsarbeiten,
- die Stimulanz zur Wiederverwendung wissenschaftlicher Daten und damit der Kollaboration zwischen Wissenschaftlern,
- die Unterstützung von datenintensiver multidisziplinärer Forschung,
- die erhöhte Rentabilität öffentlich geförderter Forschungsinitiativen sowie
- die Bereitstellung der Daten und Forschungsergebnisse für die Öffentlichkeit und für politische Entscheidungsträger.

Die Notwendigkeit der Veröffentlichung wissenschaftlicher Daten ist daher schon seit längerer Zeit erkannt und wird auch durch entsprechende Forschungsinitiativen forciert. Hier sind insbesondere die „Initiative on Scientific Cyberinfrastructures“ der US National Science Foundation [NSF 2007] sowie im Europäischen Raum die „European roadmap for research infrastructures“ [Esfri 2008] zu nennen. Daraus resultierend gibt es vielfältige Beispiele wissenschaftlicher Dateninfrastrukturen die in der Regel auf bestimmte Domänen spezialisiert sind, wie beispielsweise das soziale Netzwerk MyExperiment[3] [Goble et al. 2010] für den Austausch von Prozessabläufen im Bereich der Bioinformatik oder die iPlant Collaborative Cyberinfrastructure[4] für Pflanzenwissenschaften. Zentrale Funktionalitäten sind in der Regel die Recherche in Katalogen, der Zugriff auf Daten und Visualisierungen, die Bereitstellung von Rechenkapazitäten sowie die Interaktion und Kollaboration zwischen Wissenschaftlern. Geodateninfrastrukturen unterstützen die meisten dieser Funktionalitäten, entsprechende Implementierungen sind aber bisher kaum vorhanden obwohl viele

[3] www.myexperiment.org
[4] www.iplantcollaborative.org

wissenschaftliche Modelle letztlich raumbezogene Phänomene beschreiben (wie z.B. Landnutzung, Biodiversität, Sozioökonomie oder Klima). Darüber hinaus würden die standardisierten Schnittstellen und Formate den interoperablen Austausch der Daten unterstützen und könnten mittelfristig die Arbeit der Wissenschaftler bei der Recherche, der Verarbeitung in unterschiedlichen Systemen und der Integration der Daten und Modelle wesentlich vereinfachen.

Obwohl einige wissenschaftliche Journale bereits Möglichkeiten anbieten Daten oder Software als Ergänzung zu den Artikeln zu publizieren gibt es für den einzelnen Wissenschaftler wenig Anreize dies zu tun. Die Publikation der Daten bedeutet in der Regel Mehraufwand und die Bewertung wissenschaftlicher Ergebnisse erfolgt ausschließlich über Publikationen und die Anzahl entsprechender Zitate. Solange die Publikation von Daten und deren weitere Nutzung durch andere Wissenschaftler nicht als zusätzliche Messgröße für den Output wissenschaftlicher Arbeit einbezogen werden ist es schwer Wissenschaftler zu überzeugen, dass sich dieser Mehraufwand lohnt.

Anforderungen an die Publikation wissenschaftlicher Umweltdaten

Allgemeine Metadaten

Metadaten als beschreibende Daten erfassen die wesentlichen Merkmale von Geodatenressourcen und im Idealfall unterstützen sie nicht nur das Auffinden von Daten sondern auch die Bewertung der Nutzbarkeit für eine bestimmte Anwendung und die Integration mit anderen Ressourcen. Die wohl wesentlichste Anforderung bei der Beschreibung von wissenschaftlichen Daten ist die Referenz der entsprechenden Publikationen. Die Publikationen selbst sind als Metainformation aber nicht ausreichend, da sie in der Regel auf die neuen wissenschaftlichen Erkenntnisse fokussiert sind. Sie liefern beispielsweise keine strukturierte und umfassende Beschreibung der Daten und ihrer Qualität.

INSPIRE fordert, dass jeder Datensatz mindestens durch ein Schlüsselwort aus dem GEMET (General Environmental Multilingual Thesaurus) beschrieben wird [EC 2007]. Für die wissenschaftliche Terminologie ist GEMET aber unzureichend. Derzeit gibt es keinen etablierten Thesaurus welcher geeignet wäre um die Bedeutung einzelner Terme zwischen verschiedenen wissenschaftlichen Disziplinen zu kommunizieren. Im Bereich der Umweltmodellierung trifft das bereits auf grundlegende Begriffe wie Modell, Szenario, Treiber oder Indikator zu. Obwohl oft benutzt, haben verschiedene wissenschaftliche Gruppen häufig ein unterschiedliches Verständnis dieser Begriffe. Eine detaillierte, eindeutige und formal beschriebene Terminologie, welche auch domänenübergreifend gültig ist, wird deshalb dringend benötigt.

Genauigkeit und Skalierung

Wissenschaftler vernachlässigen häufig die Beschreibung der Qualität und insbesondere der Genauigkeit der simulierten Daten. Die Elemente der ISO Metadaten Norm [ISO 2003] für die Beschreibung von Genauigkeit und Konsistenz eignen sich nicht für wissenschaftliche Simulations- oder Modellergebnisse. Sinnvoller wären Aussagen zu Abhängigkeiten zwischen den In- und Outputdaten der Modelle oder eine Einschätzung, welche der Inputdaten den größten Einfluss auf die Ergebnisqualität haben.

Außerdem bieten die ISO Metadaten Standards derzeit keine Möglichkeit um die raum-zeitliche Skala und die Auflösung von Zeitreihen- oder multidimensionalen Daten hinreichend zu beschreiben. Die räumliche Auflösung von Eingangs- und Ausgangsdaten nummerischer Modelle bezieht sich häufig auf aggregierte administrative Einheiten (z.B. Staaten) für welche die statistischen Basisdaten vorliegen. Die Bildung dieser Aggregatflächen hängt vom jeweiligen Modellziel und den erwarteten Erkenntnissen ab. Trotzdem werden diese Informationen häufig bei der Publikation der Daten nicht mit dokumentiert.

Entstehung und Verwendung der Daten

Um die Nutzbarkeit von Daten im Hinblick auf eine bestimmte Anwendung zu bewerten sind Informationen zur Entstehungsgeschichte der Daten unabkömmlich. ISO 19115 Metadaten bieten hierfür das „Lineage“ Element, welches als Teil der Datenqualität modelliert ist. Damit ist es möglich die Entstehungsschritte eines Datensatzes von der Erfassung über alle Verarbeitungsschritte wie beispielsweise Transformationen, Aktualisierung und Korrekturen bis hin zur Generation neuer Datenprodukte zu beschreiben. Wesentliche Bestandteile der „Lineage“ sind Elemente zur Beschreibung der Quelldaten und der einzelnen Prozessschritte. Für wissenschaftliche Umweltdaten können diese Informationen Beziehungen zwischen verschiedenen Datensätzen und Modellen abbilden [Mäs et al. 2011]. Im GLUES Projekt wurde für die visuelle Illustration dieser Metainformationen ein interaktiver Webclient entwickelt (Abbildung 2), welcher an einen OGC Katalogdienst (CSW) gekoppelt ist. Damit können Wissenschaftler zum Beispiel einen schnellen Überblick darüber bekommen, welche Modelle Daten für ein bestimmtes Szenario liefern. Wenn außerdem die Verwendung der Datensätze dokumentiert ist kann gezeigt werden wo ein bestimmter Datensatz als Modellinput diente und wie groß der Einfluss („Impact“) dieser Daten auf die Arbeiten anderer Wissenschaftler ist.

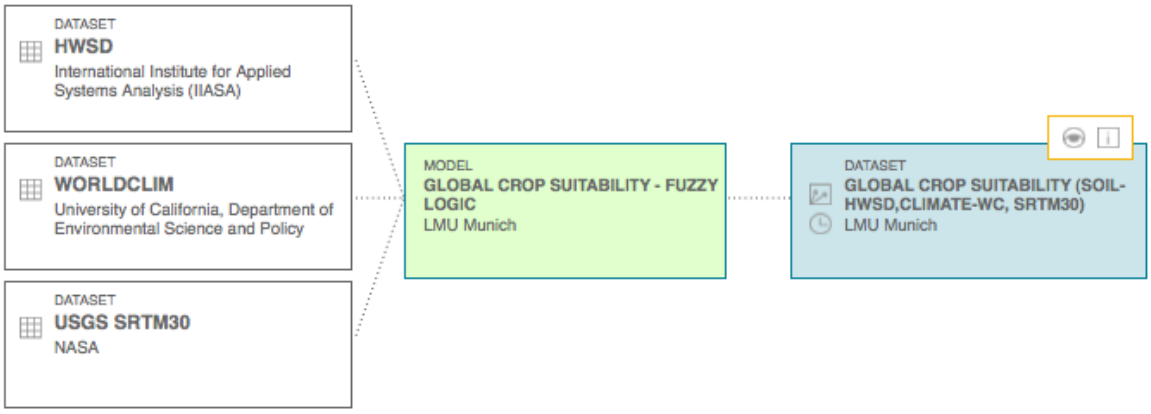

Abb. 2: Beispiel der Entstehungsgeschichte eines Datensatzes (rechts) mit drei Inputdatensätzen (links) für einen Modelldurchlauf

Eindeutige Identifikation von Datensätzen

Publizierte wissenschaftliche Daten müssen, vergleichbar zu textuellen Publikationen, persistent, eindeutig identifizierbar und verfügbar gemacht werden. DOI (Digital Object Identifier) bieten ein entsprechendes System für digitale Veröffentlichungen. Daten DOIs erlauben das Auffinden und Zitieren von Datensätzen, z.B. um innerhalb eines Artikels auf einen bestimmten Datensatz zu verweisen. DOI ist seit kurzem ein ISO Standard [ISO 2012] und wird bereits in

Publikationsplattformen wie PANGAEA (Data Publisher for Earth and Environmental Science)[5] oder dem World Data Center for Climate[6], Hamburg, eingesetzt. Die Dienste für die Registrierung einzelner DOIs werden von der Organisation Data Cite[7] bereitgestellt. Derzeitige GDIs verwenden kaum DOIs und umgekehrt, haben die mit DOI identifizierbaren Daten in der Regel keine standardisierten Metadaten bzw. Datenformate und sind nicht über standardisierte GI-Dienste verfügbar. Für die Publikation von wissenschaftlichen Umweltdaten in GDIs erscheint die Verwendung von DOIs unerlässlich. Bisher nicht gelöst sind allerdings Problemstellungen wie die Erfassung von wissenschaftlichen Zitierungen und eine entsprechende Bewertung der Datenpublikationen als wissenschaftliche Arbeiten.

Lizenzierung

Bei der Veröffentlichung von Daten haben viele Wissenschaftler grundlegende Bedenken bezüglich der Lizenzierung, der Sicherung des geistigen Eigentums und der Haftbarkeit gegenüber den Nutzern der Daten. Das Creative Commons Lizenzmodell bietet hierfür eine Auswahl von vordefinierten und einfachen Lizenzbausteinen welche kombiniert werden können. Creative Commons Lizenzen werden bereits in vielen wissenschaftlichen Dateninfrastrukturen verwendet. In den meisten Fällen bleiben die Rechte an den Daten bei den Urhebern und es werden Anforderungen bezüglich Zitation oder Nennung in Publikationen der Datennutzer definiert. Eine GDI für wissenschaftliche Daten muss auch die Vertraulichkeit publizierter Daten garantieren, um beispielsweise den Wettbewerb zwischen konkurrierenden Forscherteams zu sichern. Gängige Praxis ist es zum Beispiel die Daten erst freizugeben, wenn die entsprechenden Publikationen veröffentlicht sind und vorher den Zugriff auf die Daten einzuschränken.

Fazit

Für die Publikation von wissenschaftlichen Umweltdaten in GDIs gibt es noch eine Vielzahl von Fragen und Problemen von denen hier nur einige skizziert werden konnten. In Zukunft könnten wissenschaftliche GDIs auch den Austausch und die Integration wissenschaftlicher Modelle oder Modellkomponenten [Müller et al. 2010] unterstützen. Gerade im Hinblick auf Aspekte der Organisation der als Querschnittsaufgabe angelegten GDIs hängt viel von den konkreten Rahmenbedingungen wie beispielsweise der Finanzierung bzw. Nachhaltigkeit ab[8].

Literatur

Craglia, M.; De Bie, K.; Jackson, D.; Pesaresi, M.; Remetey-Fülöpp, G.; Wang, C.; Annoni, A.; Ling, B.; Campbell, F.; Ehlers, M.; Genderen, J.V.; Goodchild, M.; Guo, H.; Lewis, A.; Simpson, R.; Skidmore, A. & Woodgate, P.; 2012. Digital Earth 2020: towards the vision for the next decade. International Journal of Digital Earth, 5 (1), 4-21.

EC (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE) 2007.

[5] http://www.pangaea.de/
[6] http://cera-www.dkrz.de/CERA/
[7] http://datacite.org/
[8] http://www.wissenschaftsrat.de/arbeitsbereiche-arbeitsprogramm/forschungsinfrastrukturen

Eppink, F.; Werntze, A.; Mäs, S.; Popp, A. & Seppelt, R.; 2012. Land Management and Ecosystem Services: How Collaborative Research Programmes Can Support Better Policies. GAIA - Ecological Perspectives for Science and Society, 21 (1), 55-63.

Esfri, 2008. European roadmap for research infrastructures [online]. European Strategy Forum on Research Infrastructures. ftp://ftp.cordis.europa.eu/pub/esfri/docs/esfri_roadmap_update_2008.pdf (Zugegriffen am 11.6.2012)

Goble, C.A.; Bhagat, J.; Aleksejevs, S.; Cruickshank, D.; Michaelides, D.; Newman, D.; Borkum, M.; Bechhofer, S.; Roos, M.; Li, P. & De Roure, D.; 2010. myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Research, 38, W677-W682.

Gore, A.; 1999. The Digital Earth: Understanding Our Planet in the 2lst Century (Speech held on January 31, 1998 in the California Science Center). Photogrammetric Engineering & Remote Sensing, 65 (5), 528-530.

ISO 19115 International Standard on Geographic information - Metadata 2003.

ISO, 2012. ISO 26324:2012 - Information and documentation - Digital object identifier system.

Mäs, S.; Müller, M.; Henzen, C.; Bernard, L.; 2011. Linking the Outcomes of Scientific Research: Requirements from the Perspective of Geosciences. Proceedings of the First International Workshop on Linked Science 2011 (LISC2011), CEUR Workshop Proceedings, Volume 783, Bonn, Germany, October 24, 2011, verfügbar unter: http://www.ceur-ws.org/Vol-783 .

Müller, M.; Bernard, L.; Brauner, J.; 2010. Moving Code in Spatial Data Infrastructures: Web Service Based Deployment of Geoprocessing Algorithms. In: Transactions in GIS 14 (2010), Nr. S1, S. 101-118.

NSF, 2007. Cyberinfrastructure vision for 21st century discovery [online]. US National Science Foundation Cyberinfrastructure Council. http://www.nsf.gov/od/oci/CI_Vision_March07.pdf (Zugegriffen am 11.6.2012)