handbuch.io

- CH

Handbuch CoScience/Daten sammeln und verarbeiten

From Handbuch.io

< Handbuch CoScience

DOI: 10.2314/coscv2.3

Autoren: Christian Hauschke

Kontributoren: Mareike König, Ulrich Kleinwechter

Kollaborative Erfassung, Verarbeitung und Visualisierung von Daten bietet Möglichkeiten, die weit über die eines Individuums hinausgehen. Dabei gibt es unterschiedliche Ansätze: das Sammeln von Daten durch ein Team, die Motivierung größerer Gruppen zum gemeinsamen Sammeln zu einem bestimmten Zweck, aber auch die Nachnutzung von Daten, die in einem anderen Kontext gesammelt wurden. Die hier vorgestellten Werkzeuge und Ansätze dienen oft nicht nur der Erfüllung eines dieser Ansätze, werden aber exemplarisch in jeweils nur einem Abschnitt präsentiert.

Contents

Citizen Science - Aktivieren Sie freiwillige Helferinnen

Die Kartierung von Meeresbodenbewohnern in einem großen Areal ist sehr zeitintensiv. Im Projekt Seafloor Explorer (http://www.seafloorexplorer.org/) wurde versucht, diese Aufgabe so zu fragmentieren, dass sie auch von Personen ohne wissenschafttlichen Hintergrund erledigt werden kann. Diese Art der Aktivierung von

Freiwilligen für die wissenschaftliche Arbeit nennt sich Citizen Science[1] und es haben sich schon allerlei Spielarten etabliert. Das Spektrum reicht von der Vogelbeobachtung[2] bis zur Erfassung von 'Lichtverschmutzung'.[3]

Im Projekt Photos Normandie (http://www.flickr.com/photos/photosnormandie/)[4] stehen rund 3.000 Fotos von der Landung der Alliierten in Frankreich im Juni 1944 unter einer CC-Lizenz auf Flickr bereit. Personen, Orte, Straßen und Gebäude können dort gemeinsam identifiziert werden.

Bei Projekten dieser Art besteht eine Herausforderung oft darin, diese Aufgabe für Nichtwissenschaftlerinnen und Nichtwissenschaftler interessant zu machen. Das Geo-Wiki Project (http://www.geo-wiki.org/), in dem nichtwissenschaftliche Personen Satellitenfotos globaler Landnutzung daraufhin untersuchen, ob an einer bestimmten Stelle Landwirtschaft stattfindet, geht diese Herausforderung in Form einer Spiele-App für Smartphones (http://www.geo-wiki.org/games/croplandcapture/) an. Die wissenschaftliche Arbeit wird als Computerspiel verpackt. In einem dazugehörigen Wettbewerb werden unter den besten Mitspielenden Preise verlost.

Achtung: Citizen Science organisiert sich nicht selbst! Die Aktivierung, Motivation und Steuerung einer ausreichend großen Community kann im Gegenteil sogar eine sehr anspruchsvolle Aufgabe sein.[5]

Erforschen Sie 'freie' Projekte

Freie Lizenzen sind wichtig.[6] Texte und Daten des Digital Bibliography & Library Project (DBLP) und der Wikipedia sind Ausgangsbasis zahlreicher Forschungsaktivitäten, weil sie legal und mit einfachen technischen Mitteln nachnutzbar sind. Im Falle der Wikipedia hat sich eine Community um die Erforschung der Wikipedia (https://en.wikipedia.org/wiki/Wikipedia:Research) gebildet. Die behandelten Fragen reichen von der Motivation der Wikipedianerinnen und Wikipedianer[7] bis zur statistischen Textanalyse.[8]

Ein weiteres Beispiel für ein frei lizenziertes Projekt ist der Google Books N-Gram Viewer (https://books.google.com/ngrams). Zwar sind nicht alle Inhalte aus Google Books überall zugänglich, nicht einmal zur einfachen Einsicht. Google stellt jedoch daraus generierte sogenannte N-Gramme (https://de.wikipedia.org/wiki/N-Gramm) zu Analysezwecken bereit. Zwar lassen sich die einfache Abfragen direkt über die Webseite stellen, die dahinter liegenden Rohdaten[9] stehen jedoch auch zum Download und zur Weiterverwendung bereit.

Error creating thumbnail: File missing

Google Books N-Gram: "Collaborative Science" vs. "Coooperative Science" (https://books.google.com/ngrams/graph?content=collaborative+science%2Ccooperative+science&case_insensitive=on&year_start=1900&year_end=2014&corpus=15&smoothing=3&share=&direct_url=t4%3B%2Ccollaborative%20science%3B%2Cc0%3B%2Cs0%3B%3Bcollaborative%20science%3B%2Cc0%3B%3BCollaborative%20Science%3B%2Cc0%3B.t4%3B%2Ccooperative%20science%3B%2Cc0%3B%2Cs0%3B%3BCooperative%20Science%3B%2Cc0%3B%3Bcooperative%20science%3B%2Cc0%3B%3BCooperative%20science%3B%2Cc0)

Nutzen Sie Forschungsdaten nach

In Forschungsdatenrepositorien veröffentlichen Wissenschaftlerinnen und Wissenschaftler aus allen Disziplinen Rohdaten, also die Grundlage ihrer Forschungsaktivitäten. Im Kapitel Publikation von Forschungsdaten erfahren Sie, wie Sie selbst Daten veröffentlichen können. Da die so veröffentlichten Daten im Idealfall unter einer Lizenz stehen, die die Nachnutzung erlaubt, finden Sie dort eventuell eine Datengrundlage für Ihre Forschung (siehe auch Kapitel Freie Lizenzen und Nachnutzung).

Sie finden dort Daten jeder Art: vom Modell des weltweiten Kartoffelhandels[10] über die Mitgliederbibliotheken des Deutschen Bibliotheksverbands[11] und archäologische Artefakt-Inventare[12] bis zu den Daten rund um die Entdeckung des Higgs-Bosons.[13] Es gibt auch Repositorien für qualitative Daten, zum Beispiel das Qualitative Data Repository (https://qdr.syr.edu/) an der Syracuse University.

Eine Übersicht von Datenrepositorien finden Sie auf re3data.org (http://re3data.org) (Registry of Research Data Repositories).

Nutzen Sie Tools zur Datensammlung

Es gibt zahlreiche Tools, die Ihnen die Datensammlung erleichtern oder gar komplett abnehmen. Wenn Sie beispielsweise die Twitteraktivitäten rund um ein Hashtag analysieren wollen, können Sie die entsprechenden Tweets mittels des R-Pakets twitteR (http://cran.r-project.org/web/packages/twitteR/index.html) einsammeln. Robert Mashey stellt mit TAGS V5 (http://mashe.hawksey.info/2013/02/twitter-archive-tagsv5/) zu eben diesem Zweck eine Anwendung für Google Spreadsheets zur Verfügung, die Tweets einsammelt und einfache Auswertungen sowie eine einfache Social-Network-Analyse und eine Suchfunktion gleich mitliefert.

Anders gelagert ist [www.ushahidi.com/Ushahidi], eine Plattform, die ursprünglich entwickelt wurde, um politische Unruhen in Kenia 2007/2008 zu kartieren. Bekannt wurde Ushahidi durch den Einsatz zur Koordinierung der Katastrophenhilfen nach dem Erdbeben in Haiti 2010. Ushahidi kann Daten aus verschiedenen Quellen aggregieren und auf einer Karte darstellen.[14]

Erstellen und bearbeiten Sie Ihre Daten gemeinsam

Protegé ist ein Editor, der die Darstellung, Bearbeitung und Veröffentlichung von Ontologien ermöglicht. Seit einiger Zeit gibt es auch eine online verfügbare Variante, in der mit mehreren Personen gearbeitet werden kann. WebProtegé (http://protegewiki.stanford.edu/wiki/WebProtege) bietet explizite Kollaborationsunterstützung wie zum Beispiel durch Rechtevergabe (Wer darf was bearbeiten?), Diskussions- und Benachrichtigungsfunktionen.

Error creating thumbnail: File missing

Verteilung aller Teilnehmenden der SOAP-Umfrage[15], erstellt mit Fusion Tables (Daten (http://www.google.com/fusiontables/DataSource?snapid=137611))

Ein prominenteres und breiter einsetzbares Tool ist Fusion Tables (http://www.google.com/fusiontables) von Google. Dabei handelt es sich um einen Webservice, mit dem Daten erstellt, veröffentlicht und verarbeitet werden können. Die Global Conservation Maps (http://maps.tnc.org/globalmaps/globalmaps_original.html) sind mit Fusion Tables erstellt. Weitere Beispiele finden Sie in der Fusion Tables Example Gallery (https://sites.google.com/site/fusiontablestalks/stories).

Ein Beispiel für ein Web-Tool, das es Nutzerinnen und Nutzern erlaubt, selbst komplizierte Modellrechnungen anzustellen, ist Climate Analogues (http://www.ccafs-analogues.org/). Durch die Anwendung statistischer Verfahren auf direkt in dem Tool verfügbare Klimadaten kann man herausfinden, wo auf der Erde sich klimatisch ähnliche Regionen befinden oder das Klima welcher Region heute dem Klima einer bestimmten Region in der Zukunft unter verschiedenen Klimawandelszenarien entspricht.

Visualisieren Sie Ihre Daten

Es ist nicht nur möglich, sich bei der

Error creating thumbnail: File missing

Wortwolke aus einem Zwischenstand bei der Erstellung dieses Handbuchs, hergestellt mit Many Eyes

Datensammlung helfen zu lassen. Auch für die Visualisierung von Daten sind Werkzeuge verfügbar. Diese Werkzeuge, die in manchen Fällen Operationen bis hin zu komplexen Modellrechnungen erlauben, können eine individualisierte Datenanalyse womöglich nicht ersetzen. Sie erlauben es jedoch, auf einfache Weise in einen bestimmten Datensatz einzusteigen und sich einen Überblick zu verschaffen und machen es – siehe Modellrechnungen – teilweise möglich, Analysen durchzuführen, zu denen man alleine nicht oder nur schwer in der Lage wäre.[16]

Im Bereich der Klimaforschung gibt es zahlreiche Beispiele für webgestützte und/oder kollaborative Tools im Bereich der Datenverarbeitung und -visualisierung. Ein erstes Beispiel zur Datenvisualisierung und -verarbeitung ist der KNMI Climate Explorer (http://climexp.knmi.nl). Hierbei handelt es sich um ein Werkzeug, das es erlaubt, Daten zu Klima und Klimawandel, inklusive eigener Daten, graphisch darzustellen, zu analysieren und weiterzuverarbeiten.

Mit dem Climate Wizard (http://www.climatewizard.org/) können aus den Daten globaler Klimamodelle Landkarten zum Klimawandel erzeugt werden.

RTB Maps (http://gisweb.ciat.cgiar.org/RTBMaps/) ist ein Beispiel für ein Projekt, in dem Wissenschaftlerinnen und Wissenschaftler verschiedener Forschungsinstitute ihre gemeinsame Arbeit auf einer Webplattform präsentieren. RTB Maps, das die Anwendung ArcGIS (http://www.arcgis.com/) nutzt, bündelt globale Landkarten zur Produktion verschiedener Feldfrüchte mit Karten zu Pflanzenkrankheiten und sozio-ökonomischen Variablen. Nutzerinnen und Nutzer können diese Karten nach ihren eigenen Bedürfnissen neu kombinieren und für ihre eigenen Zwecke verwenden.

Im Web finden sich darüber hinaus hunderte Werkzeuge, mit denen Sie Ihre Daten auf jede erdenkliche Art und Weise präsentieren können. [www-958.ibm.com/ Many Eyes], das von IBM gratis zur Verfügung gestellt wird, ist dabei eins der prominentesten Beispiele. Damit lassen sich auch interaktive Darstellungen von Daten mit wenigen Klicks erstellen. Sowohl die Daten als auch die Visualisierungen sind dabei immer öffentlich.

Beachten Sie dabei stets, dass Visualisierungen sehr mächtige Werkzeuge zur Verdeutlichung von Inhalten sind, es jedoch auch schnell zu Fehlinterpretationen kommen kann![17] Für einen vertieften Einstieg in die Materie sei hier stellvertretend für viele andere Werke auf die Publikationen von Edward Tufte (https://de.wikipedia.org/wiki/Edward_Tufte) verwiesen.

APIs und Schnittstellen

Schon 2006 fragte Declan Butler sich und seine Leserschaft, ob nun das 'Jahr der Mashups', also der Verknüpfung von Daten und Services unterschiedlicher Webseiten in einer anderen, direkt vor dem Durchbruch stünde.[18] Die große Mashup-Revolution ist bisher ausgeblieben, doch haben sich einzelne Dienste fest und von der Anwenderin/dem Anwender oft unbemerkt etabliert. So sind viele Webseiten, die auf Karten von OpenStreetMap oder Google Maps basieren, Mashups, die dadurch ermöglicht werden, dass die Kartendaten über eine Schnittstelle bereitgestellt werden.

Programmable Web (http://www.programmableweb.com/apis/directory/1?apicat=Science) listet – Stand März 2014 – fast 350 dieser Programmierschnittstellen (oder API für *application programming interface*) aus dem Bereich 'Science'. Dazu gehören das Exoplanet Archive Application Programming Interface (API) (http://exoplanetarchive.ipac.caltech.edu/docs/program_interfaces.html) der NASA oder die Schnittstellen des National Biodiversity Network (NBN) (https://data.nbn.org.uk/Documentation/Web_Services/) und des Seismic Data Portal, über das Informationen über Erdbeben abgerufen werden können.

Nutzen Sie die Computer der anderen

Die Datenmengen werden immer größer, die damit einhergehende notwendige Rechenleistung zu ihrer Bearbeitung auch. Und da nicht jeder Zugriff auf Supercomputer hat, wurden verschiedene Möglichkeiten entwickelt, wie man dennoch auf fremde Rechenleistung zugreifen kann.[19]

Ein Webservice, der sich genau dieser Aufgabe gewidmet hat, ist OpenCPU (http://opencpu.org). Der Schwerpunkt ist weniger der, möglichst große Rechenpower zur Verfügung zu haben, sondern liegt vielmehr darauf, statistische Analysen auch für Dritte nachvollziehbar und durchführbar zu machen.

Tatsächlich um Rechenleistung geht es bei BOINC (https://boinc.berkeley.edu) (Berkeley Open Infrastructure for Network Computing). BOINC ermöglicht 'verteiltes Rechnen', also das Aufteilen einer Rechenoperation auf viele Teilaufgaben, die in diesem Fall von den Computern freiwilliger Teilnehmerinnen und Teilnehmer rund um den Globus übernommen werden können. Prominente BOINC-Projekte sind SETI@Home (https://de.wikipedia.org/wiki/SETI@home), das sich der Suche nach außerirdischer Intelligenz widmete, und LHC@home (https://de.wikipedia.org/wiki/LHC@home), mit dessen Hilfe der Large Hadron Collider optimiert werden soll.

Achten Sie auf Reproduzierbarkeit und Nachhaltigkeit

Webdienste sind oft die einfachste Art, gemeinsam an Daten zu arbeiten. Beachten Sie bitte, dass im Web wenig für die Ewigkeit ist. Gibt es die von Ihnen genutzte Web-Anwendung kommendes Jahr noch? Könnten sich die Benutzungsmodalitäten für Schnittstellen ändern? Kann ich meine Daten exportieren und in anderem Zusammenhang nachnutzen?

Werkzeuge wie das oben erwähnte Fusion Tables sind sehr attraktiv. Wenn Sie es verwenden und Ihre wissenschaftliche Arbeit darauf aufbauen, sollten Sie nie aus den Augen verlieren, das Google den Dienst auch abschalten kann. Nicht nur kleine Anbieter können von einem Tag auf den anderen vor dem Aus stehen, auch große Firmen wie Google oder Yahoo sind berüchtigt dafür, Dienste wie den Google Research Datasets außer Betrieb zu nehmen.[20] Und auch die Reproduzierbarkeit Ihrer Forschung kann unter Webservices leiden, die als Black Box agieren, bei denen Sie also nicht nachvollziehen können, was tatsächlich mit Ihren Daten gemacht wird.

Einzelnachweise

1. Citizen Science (https://de.wikipedia.org/wiki/Citizen_Science) in Wikipedia. Abgerufen am 04.03.2014.
2. Lepczyk, Christopher A. (2005) Journal of Applied Ecology 42 (4), Seiten 672-677. DOI: http://dx.doi.org/10.1111/j.1365-2664.2005.01059.x
3. "Weißt du wie viel Sternlein stehen? Verlust der Nacht App misst Himmelshelligkeit" auf www.citizen-science-germany.de (http://www.citizen-science-germany.de/citizen_science_germany_projekte_2.html)
4. Vgl. dazu Cade, D.L.: *PhotosNormandie: An Online Archive of 3,000+ CC Photos from WWII* (http://petapixel.com/2013/04/06/photosnormandy-a-collection-of-over-3000-cc-photos-from-wwii/).
5. Siehe auch: Wiggins, Andrea; Crowston, Kevin: From Conservation to Crowdsourcing: A Typology of Citizen Science. 44th Hawaii International Conference on System Sciences (HICSS), 2011. DOI: http://dx.doi.org/10.1109/HICSS.2011.207
6. Siehe auch Kapitel Freie Lizenzen und Nachnutzung
7. Nov, Oded (2007) What motivates Wikipedians?, *Communications of the ACM*, 50 (11), Seiten 60-64. DOI: http:dx.doi.org/10.1145/1297797.1297798
8. Yasseri, Taha; Kornai, András; Kertész, János (2012) A Practical Approach to Language Complexity: A Wikipedia Case Study, *PLoS ONE*, 7(11): e48386. DOI: http://dx.doi.org/10.1371/journal.pone.0048386
9. Rohdaten des Google Books N-Gram Viewers (http://storage.googleapis.com/books/ngrams/books/datasetsv2.html)
10. Raymundo, Rubí; Kleinwechter, Ulrich; Asseng, Senthold (2014), Virtual potato crop modeling: A comparison of genetic coefficients of the DSSAT-SUBSTOR potato model with breeding goals for developing countries. ZENODO. DOI: http://dx.doi.org/10.5281/zenodo.7687
11. Hauschke, Christian (2013) Members of Deutscher Bibliotheksverband e.V. figshare. DOI: http://dx.doi.org/10.6084/m9.figshare.647329
12. MB #33 Precontact Site Artifact Inventory. Hartgen Archeological Associates, Inc. 2013 (tDAR ID: 392085) ; http://doi.org/10.6067/XCV8TD9Z8M
13. ATLAS Collaboration (2013) HepData, DOI: http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44
14. Vgl. Ruffer, Galya B. (2011) What Ushahidi can do to track displacement, *Forced Migration Review*, 38, S. 25-26. http://www.fmreview.org/technology/ruffer.html
15. Hauschke, Christian (2011) "SOAP-Daten in Google Fusion Tables". In: Infobib (http://infobib.de/blog/2011/02/18/soap-daten-in-google-fusion-tables-2/). Zuletzt abgerufen am 13.03.2014
16. Isenberg, Petra; Elmqvist, Niklas; Scholtz, Jean; Cernea, Daniel; Ma, Kwan-Liu; Hagen, Hans (2011) Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization*, 10 (4), Seiten 310-326, DOI: http://dx.doi.org/10.1177/1473871611412817
17. Vgl. Bresciani, Sabrina; Eppler, Martin J.: *The Risks of Visualization : a Classification of Disadvantages Associated with Graphic Representations of Information* (http://www.knowledge-communication.org/pdf/bresciani-eppler-risks-visualization-wpaper-08.pdf). CA Working Paper # 1/2008, February 2008.
18. Butler, Decan (2006) "Mashups mix data into global service". *Nature* (2006) 439 (7072), Seiten 6-7. DOI: http://dx.doi.org/10.1038/439006a
19. Vgl. Tsaftaris, Sotirios A. (2014) A Scientist's Guide to Cloud Computing. *Computing in Science & Engineering*, 16 (1), Seiten 70-76. DOI: http://dx.doi.org/10.1109/MCSE.2014.12
20. Mullard, Asher; Butler, Decan (2008) "Google pulls out of science data project". In: Nature News Blog (http://blogs.nature.com/news/2008/12/google_pulls_out_of_science_da_1.html). Zuletzt aufgerufen am 13.03.2014.