Linked Data basierter Explorer für krebsrelevante Ursache-Wirkungs-Beziehungen im raum-zeitlichen Kontext

von

Friedrich Müller (MSc Student, Institut für Geoinformatik, Münster),

Dorothea Lemke (Institut für Epidemiologie und Sozialmedizin, Münster)

1. Einführung

Krebs-Cluster sind ein wichtiges und sehr kontrovers diskutiertes Thema, nicht nur in Deutschland [3]. Ein Cluster wird üblicherweise als räumlich begrenztes, überzufälliges Auftreten von Krebsfällen über den Erwartungswert hinaus definiert [5]. Dabei werden häufig Anfragen aus der Bevölkerung oder von Gesundheitseinrichtungen an die epidemiologischen Krebsregister [2] gerichtet, ob es sich bei dem vermuteten Cluster um eine wirkliche räumliche und zeitliche Erhöhung der Krebsfälle handelt. Das Vorgehen sieht derzeitig so aus, das mit Hilfe von demographisch-epidemiologischen Kennzahlen ermittelt wird, ob das Risiko für diese Krebserkrankung in der Region tatsächlich statistisch-signifikant, d.h. nicht zufällig, erhöht ist. Wenn dieser Zusammenhang positiv ist, wird untersucht, ob dieses tatsächliche Cluster räumlich und zeitlich mit externen Karzinogenen (z. B. Kohlenstoffmonoxid) assoziiert werden kann. Der erste Teil der Cluster-Untersuchung wird standardisiert mit Hilfe der Daten der epidemiologischen Krebsregister durchgeführt und beantwortet. Der zweite Teil der Untersuchung, die Suche nach potentiellen Expositionsquellen, wird häufig unstrukturiert durch Zusammentragen von Informationen aus dem Internet gestaltet, was zusätzlich sehr kosten- und zeitintensiv ist. Dabei basiert die Suche nach potentiellen Karzinogenen auf den Arbeiten/Monographien der IARC (*International Agency for Research on Cancer)* [16], die aber keine direkte räumliche Verknüpfung besitzen. Es handelt sich hierbei um eine Liste unterschiedlicher chemischer Stoffe, die Krebs verursachen oder im Verdacht davon stehen. Ziel dieser vorliegenden Anwendung ist dieses Experten-Wissen der IARC mit räumlichen Ausprägungen dieser Karzinogene (z.B. CO und deren räumliche Emittenten) zu verknüpfen um eine strukturierte und fundierte Evaluation dieser Cluster-Anfragen zu ermöglichen.

Aktuell sind offene krebsrelevante, umweltbezogene Daten von verschiedenen Services verfügbar z. B. Emissionsdaten [15]. Neben der genannten Verteilung der Datenquellen sind unterschiedliche Datenformate und die Heterogenität der Daten eine Herausforderung hinsichtlich der Integration der Daten und deren Visualisierung innerhalb der Anwendung. In diesem Zusammenhang können semantische Technologien hilfreich sein um die verteilten

Daten auf eine Weise zu verknüpfen, die die einzelne Bedeutung der Ressourcen explizit darstellt und automatische Assoziationen ermöglicht [4]. Des weiteren zeigen vorangegangene Projekte aus dem Gesundheitsinformationsbereich, dass eine verbesserte und flexiblere Darstellung von Gesundheitsdaten, Analyse- und Abfragemöglichkeiten und Visualisierungen benötigt werden [6, 7].

Die hiermit präsentierte Webanwendung soll exemplarisch aufzeigen wie *Linked Data* und weitere semantischen Technologien dazu geeignet sind die Erreichbarkeit krebsrelevanter Informationen zu erhöhen. Neben der Möglichkeit sich über die krebsbezogenen Ursache-Wirkungs-Beziehungen (aus den Monographien der IARC) zu informieren, ist der Hauptnutzen der Applikation die ermöglichte Erforschung von Umweltdaten (z. B. Luftqualität, Altlasten) im Zusammenhang mit epidemiologischen Datensätzen (z. B. statistische Vergleichswerte von Krebsinzidenzen) u. a. per Geovisualisierungen für die Beispielregion Westfalen-Lippe. Der Fokus liegt hierbei auf der Verkettung: Karzinogen (z. B. CO_2) - Emissionsprozess (z. B. Verkehr) - Emissionsquelle (z. B. Auto) - Transportwege (z. B. Luft) - Exponent (z. B. Männlich/Weiblich) - Krebstyp (z. B. Lungenkrebs).

2. Realisierung

Der Workflow (vgl. Abb. 1), beginnend mit den Rohdaten über die semantische Modellierung bis hin zur Webanwendung, ist komplett auf *Open Source* Software (z. B. *Protegé* [12]*, Apache Jena* [8]*, OSM* [11] *& Leaflet* [10]) basierend, und ist in folgenden Arbeitsschritte untergliedert:

Definition des Anwendungsbereichs

Als Ausgangspunkt der Datenmodellierung wurden Elemente und Beziehungen der Ursache-Wirkungs-Kette vom Karzinogen bis hin zur Krebsinzidenz wie auch die zu betrachtenden Krebsarten und die Referenzregion für die Beispielanwendung festgelegt. Die Festlegung einer Referenzregion begrenzt den Datenfokus auf ein zu realisierendes Level. Bezüglich der Krebstypen wurde sich vorwiegend auf Typen beschränkt, die potentiell von äußeren Faktoren abhängig sind wie z. B. Lungenkrebs.

Aggregation der Daten

Der nächste Schritt ist das Auffinden, Sammeln und Produzieren von Datensätzen (z. B. über Vorkommen von Karzinogenen in Luft oder Boden, Emissionsquellen wie Industrieanlagen), die das Informationsfundament für die Webapplikation bilden und eine Erforschung der krebsrelevanten Daten ermöglichen. Die Rohdaten kommen von unterschiedlichen öffentlichen Services z. B. vom LANUV NRW (*Landesamt für Natur, Umwelt und Verbraucherschutz Nordrhein-Westfalen*) [17]. Als epidemiologisches Risikomaß wird mit Hilfe der Daten aus dem epidemiologischen Krebsregister und mit entsprechenden demographischen Daten das „standardisierte Inzidenzverhältnis“ (SIR) und deren 95% Konfidenzintervalle für jede Gemeinde berechnet, welches das Verhältnis zwischen beobachteten und erwarteten Krebsfällen in der Referenzregion Westfalen-Lippe wiedergibt.

Konvertierung der Datensätze ins RDF (*Resource Description Framework*) Format

Die bereinigten Rohdatensätze wurden per Skripte z. B. in Verbindung mit der *Jena Library* [8] oder per *Open Refine* mit RDF Extension [14] in RDF umgewandelt. Die RDF Strukturierung wurde anhand geeigneter Vokabulare z. B. mit der *Datacube Vocabulary* [18] erarbeitet. Die *Datacube Vocabulary,* die die Einteilung von multidimensionalen Datensätzen u. a. in Attributen, Dimensionen und Messungen erlaubt, eignet sich somit für Umweltdatensätze.

Encodierung der Datensätze als Linked Data

Der Linked Data Aspekt, die Informationen aus den erstellten Datensätzen mit anderen Ressourcen von externen RDF Sets oder Endpoints zu verlinken, ist durch die Verwendung von RDF Properties wie *rdfs:seeAlso, owl:equivalentClass* oder *owl:sameAs* bewerkstelligt. Zusätzliche Informationen, die nicht per URL verlinkt werden konnten wurden textlich per *rdfs:comment* hinzugefügt.

Modellierung der Domain Ontology

Parallel zur Erstellung der RDF Datensätze wurde die Domain Ontology, die die krebsrelevanten Ursache-Wirkungs-Beziehungen darstellt, entwickelt. Die Ontology strukturiert weitere Informationen der Ursache-Wirkungs-Kette z. B. Krebstyp, Karzinogene, Emissionsquellen, Emissionsprozesse, Transportwege, exponierte Orte, exponierte Gruppen und Geschlecht. Die Hintergrundinformationen über diese Beziehungen wurden von den Monographien der IARC herausgearbeitet. Die Erstellung der Ontology wurde mit dem Open Source Ontology Editor *Protégé* [12] durchgeführt.

Implementierung der Webapplikation

Nach der Generierung der RDF Datensätze und der Domain Ontology müssen diese zur Abfrage verfügbar gemacht werden. Dies ist zum jetzigen Zeitpunkt des Projekts unter Verwendung des Sesame RDF Frameworks [13] realisiert. Um die Anforderungen der Applikation wie Visualisierungen per Karte, Grafik oder Text webbasiert umzusetzen wurde die Anwendung u. a. per HTML, CSS, Javascript und SPARQL Technologien in Kombination mit diversen Bibliotheken wie z. B. Leaflet [10] oder jQuery [9] implementiert. Derzeit können Informationen direkt über selbstdefinierte SPARQL-Abfragen und indirekt über SPARQL eingebettete Funktionen erhalten werden.

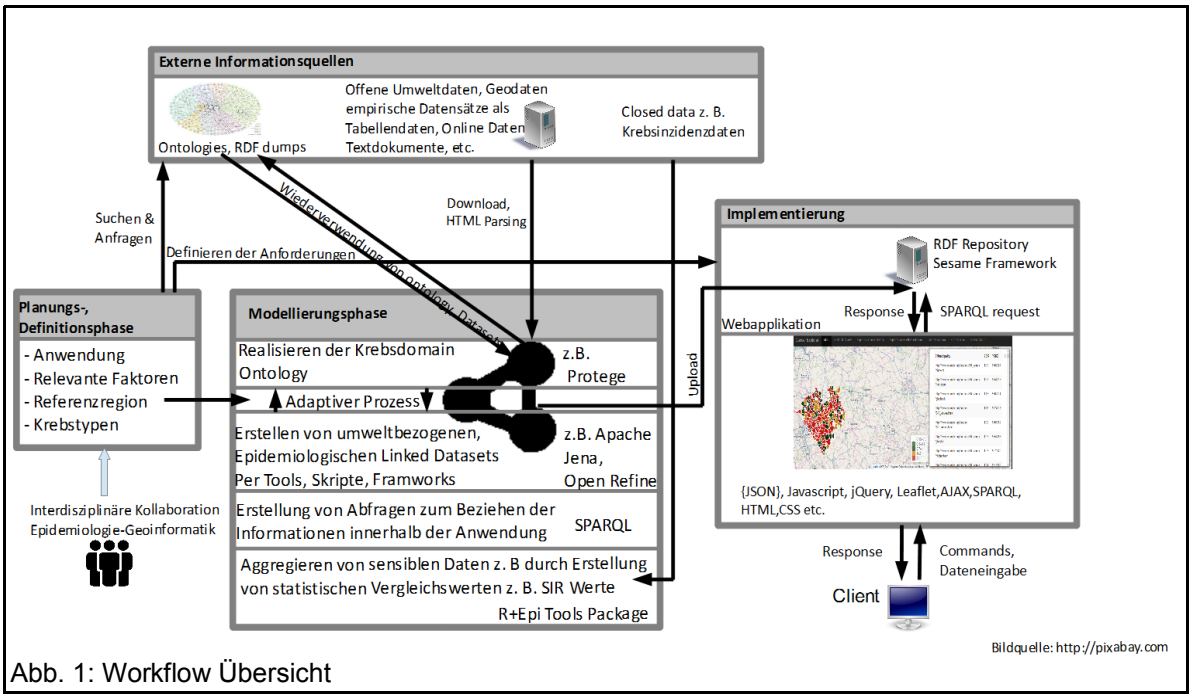

Abb. 1: Workflow Übersicht

3. Zusammenfassung

Ein erwartetes Ergebnis ist Linked Open Data, das von ursprünglich verteilten, isolierten umweltbezogenen und epidemiologischen Daten modelliert wurde. Ein weiteres Ergebnis ist die Domain Ontology, die die Ursache-Wirkungs-Kette vom Karzinogen bis hin zur Krebsinzidenz aufzeigt. Insgesamt beschäftigt sich das Projekt mit der Frage wie Linked Data dazu beitragen kann krebsrelevante Informationen, in Zusammenhang mit Krebs-Cluster Anfragen in Raum, Zeit und Semantik in geeigneter Form innerhalb einer epidemiologischen Anwendung, bereitzustellen. Die Anwendung dient als Erforschungs- und Informations-Werkzeug für krebsbezogene Ursache-Wirkungs-Beziehungen. Im größeren Rahmen trägt die Arbeit zu aktuellen Forschungsbereichen wie Geovisualisierungen im Gesundheitsbereich [1] und semantischen Technologien in Informationssystemen bei. Produkte des Projektes (z.B. Domain Ontologie) sind neben dem Code der Applikation zur Wiederverwendung auf *Github* (https://github.com/lodum/CancerExplorer) zugänglich.

Kontakt zu den Autoren:

Friedrich Müller
MSc Student, Institut für Geoinformatik, Münster
f_muel25@uni-muenster.de

Dorothea Lemke
Institut für Epidemiologie und Sozialmedizin, Münster
dorothea.lemke@uni-muenster.de

Literatur

[1] Diez, E., McIntosh B.S. 2009. A review of the factors which influence the use and usefulness of information systems. *Environmental Modelling and Software,* 24.

[2] GEKID: Bevölkerungsbezogene Krebsregister in Deutschland [http://www.gekid.de/registries.html]

[3] Kieschke J. 2010. Auswertung des EKN zur Krebshäufigkeit in der Samtgemeinde Asse, In *Book Auswertung des EKN zur Krebshäufigkeit in der Samtgemeinde Asse.*

[4] Lee T. B., Hendler J., Lassila O. 2001. The semantic web. *Scientific American* 284, 28–37.

[5] Olsen S.F., Martuzzi M., Elliott P. 1996. Cluster analysis and disease mapping - Why, when, and how? A step by step guide. *Brit Med J*, 313:863-866.

[6] Tilahun B., Kauppinen T., Keßler C., Fritz F. 2014. Design and Development of a Linked Open Data-Based Health Information Representation and Visualization System: Potentials and Preliminary Evaluation, *JMIR Med Inform,* 2(2):e31.

[7] Zhou Q., Wang C., Xiong M., Wang H., Yu Y. 2007. SPARK: adapting keyword query to semantic search, in: *ISWC*.

Internet:

[8] https://jena.apache.org/

[9] http://jquery.com/

[10] http://leafletjs.com/

[11] http://www.openstreetmap.de/

[12] http://protege.stanford.edu/

[13] http://rdf4j.org/

[14] http://openrefine.org/ + http://refine.deri.ie/

[15] http://www.gis.nrw.de/ims/ekatsmall2008/small/info.htm

[16] http://www.iarc.fr/

[17] http://www.lanuv.nrw.de/

[18] http://www.w3.org/TR/vocab-data-cube/