

BÜRGERINFORMATIONSSYSTEME UND SEMANTISCHE TECHNOLOGIEN

Andreas Krenmayr / Roland Traummüller

Dissertant, Universität Linz, FAW
Altenbergstraße 69, 4040 Linz-Auhof, AT
andreas.krenmayr@gmail.com

Universitätsprofessor, Johannes Kepler Universität Linz
Altenbergerstr. 69, 4040 Linz, AT
traumm@ifs.uni-linz.ac.at

Schlagworte: *Bürgerinformationssysteme, Mensch-Maschine-Kommunikation, Web of Data, Linked Open Government, Natural Language Processing*

Abstract: *Konventionelle Online-Bürgerinformationssysteme basieren auf reinem (X)HTML. Obwohl die herkömmliche Art der Benutzerinteraktion vermittelt dem Versenden von Formulardaten und dem Aufruf von vom System bereit gestellten Links mittlerweile gut verstanden ist und von Benutzern akzeptiert wird, bringt sie Nachteile mit sich. Anfragen an das System sind auf das Senden von diskreten Parametern, die von Maschinen verarbeitet werden können, beschränkt. Das Erstellen einer Antwort beschränkt sich zumeist auf die Boolesche Kombination der übermittelten Parameter. Benutzerinteraktion muss a priori parametrisiert werden und dies in möglichst eindeutiger Form. Die Kluft zwischen Alltagssprache und Fachsprache, die als bedeutende Problematik innerhalb von Bürgerinformationssystemen angesehen wird, kann auf diese Weise nicht effektiv geschlossen werden. Einen möglichen Ansatz zur Behebung dieses Defizits kann die Implementierung von semantischen Technologien aufzeigen. Dieser Beitrag will Schwachstellen konventioneller Benutzerinteraktion mit Bürgerinformationssystemen aufzeigen und einen Weg skizzieren, mithilfe semantischer Technologien diesen Schwachstellen zu begegnen. Dies bezieht die Darstellung einer Ontologie mit ein, die entwickelt wird, um ein reales Bürgerinformationssystem zu verbessern.*

1. State of the art: Maschinen-lesbare Anfragen und Menschen-lesbare Anfragen

Online-Bürgerinformationssysteme (in der Folge als OBIS bezeichnet) zielen darauf ab, Bürgern Informationen zu bestimmten Themen (etwa Fahrpläne, Wetterdaten, Börsendaten, Einkaufsempfehlungen, Informationen zu Amtswegen) online zur Verfügung zu stellen [7]. Der Einfachheit wegen wird ihr modus operandi wie folgt dargestellt: Ein *Benutzer* sendet eine *Anfrage* an das System; die Anfrage wird vom System *verarbeitet*; eine *Antwort* wird *generiert*; die Antwort wird an den Benutzer *übermittelt*. Eine erfolgreiche Antwort eines OBIS besteht aus einem oder mehreren *Contents*.

Konventionelle OBIS erlauben keine komplexeren Interaktionsmechanismen als die Benutzerelemente des HTML 4.01 Standards vorsehen. Obwohl diese Art der Mensch-Maschine-Kommunikation mittlerweile von den meisten Benutzern von OBIS gut verstanden und akzeptiert wird, bringt sie doch einige Mängel mit sich, wie gezeigt werden soll:

Konventionelle HTML-Benutzerelemente arbeiten auf zwei Arten.

[A] Die Information der Anfrage wird als diskreter Wert an den Server übermittelt. Diese Werte (Ausgangsmenge) werden vom System vorher bestimmt. Benutzerelemente dieser Art umfassen Drop-down-Listen,

Radio-Buttons und Checkboxes.

Benutzerelement	HTML Markup	Resultat
Drop-down-Liste	<pre><select> <option value="at">Austria</option> ... </select></pre>	
Radio-Button	<pre><input type="radio" value="at">Austria</pre>	
Checkbox	<pre><input type="checkbox" value="at">Austria</pre>	

Bei dieser Art von Benutzerelementen ist es offensichtlich, dass eine Anfrage an ein OBIS a priori vom Benutzer selbst parametrisiert werden muss. Die Anfrage des Benutzers muss derart bearbeitet werden, dass sie in die Ausgangsmenge passt, die das System zur Verfügung stellt. (Wenn etwa die Ausgangsmenge aus einer Anzahl europäischer Länder besteht, gibt es keine Möglichkeit, eine Anfrage einen bestimmten Wiener Gemeindebezirk betreffend zu stellen.) Der Benutzer muss seine Frage also an das System anpassen.

[B] Die Information der Anfrage wird als Freitext an den Server übermittelt. Der an den Server gesendete Wert wird nicht vom System vorgegeben. Benutzerelemente dieser Art umfassen ein- und mehrzeilige Textfelder (`<input>` und `<textarea>`). Gewöhnlich werden diese Benutzerelemente im Rahmen von Anfragen an ein OBIS ausschließlich zur Stichwortsuche herangezogen. Stichwörter werden serverseitig entweder in einem Tagging-System oder als Volltextsuche implementiert; das Prinzip bleibt letztendlich dasselbe: Ein Suchalgorithmus versucht aus einer Menge von in der Datenbank des Systems gespeicherten Contents mittels einer Trägerfunktion (von bisweilen hoher Komplexität!) die relevantesten Ergebnisse zu ermitteln.

Es ist wichtig, hier ein paar Anmerkungen anzubringen: Stichwörter zeigen an, welche Begriffe sich innerhalb eines Dokuments finden, aber sie sagen noch nichts über die Bedeutung der Begriffe und ihren intertextuellen Zusammenhang aus. Auch eine Stichwortsuche zwingt den Benutzer, ähnlich den Benutzerelementen des Typs [A], dazu, eine Anfrage zu parametrisieren. Aus diesem Grund bleibt das oben skizzierte Phänomen bestehen – wenn ein bestimmtes Stichwort nicht zur Ausgangsmenge gehört, kann es nicht zur Zielmenge beitragen.

Es lohnt sich ein Blick auf zwei tatsächliche OBIS: Die Plattform www.help.gv.at (in der Folge als HELP bezeichnet) liefert Bürgern Informationen zu Behördenwegen aller Art. Die Plattform wurde 1997 vom Bundesministerium für Finanzen lanciert und hat sich nach beständiger Weiterentwicklung zur wichtigsten Drehscheibe der österreichischen E-Administration entwickelt.

Der Informationsfindungsprozess läuft automatisch ab und die Fülle an Informationen ist beeindruckend: HELP verfolgt einen ganzheitlichen Ansatz und bietet so etwa Zugriff auf mehr als 1000 Formulare, von denen etwa 350 im Sinne des One-Stop-Prinzips online bearbeitet werden können. Hervorzuheben ist zudem die Gliederung der zur Verfügung gestellten Contents in 140 Themenbereiche, die als «Lebenslagen» bezeichnet werden.

Zusätzlich bietet HELP eine Suchfunktion. Sie wird über ein Textfeld realisiert, in welchem ein oder mehrere Suchbegriffe eingegeben werden können. Contents, die mit dem Suchparameter indiziert sind, können so gefunden werden. Die Bereitstellung einer internen Suchmaschine kann eine wertvolle Hilfe zur Auffindung von Informationen sein, hat aber ab einer gewissen Mächtigkeit der potentiellen Zielmenge auch Nachteile: So kam eine Suche nach dem generisch gewählten Suchbegriff «Jobsuche» zu 366 Ergebnissen. Die Fülle der Informationen im Hintergrund erweist sich hier als hinderlich bei der Auffindung einer spezifischen Information.

Nützlich für den User, der nicht mit der Fachsprache der Verwaltung vertraut ist, ist ein Glossar, der mehr als 600 Fachbegriffe erklärt. Was allerdings noch wünschenswerter wäre, wäre eine Form der Benutzerinteraktion, die in der Gegenrichtung funktioniert – also eine Möglichkeit der Übersetzung von Alltagssprache in Fachbegriffe, die von der Suchfunktion verstanden werden können.

Einen anderen Ansatz für die Auffindung von Informationen verfolgt das Online-Portal bestHELP.at. Best-HELP.at ist ein Sammelportal von 15 qualifizierten Berufsgruppen aus den Bereichen Therapie, Beratung und Training. Das Portal wird, anders als HELP, nicht von einer staatlichen Stelle betrieben, sondern von einem privaten Unternehmen.

Ein spezielles Service für Benutzer, die Online-Anfragen, wurde im Jahr 2005 eingeführt. Benutzer können hier in einem Freitextfeld (Typ [B]) eine Anfrage in natürlicher Sprache stellen. Zusätzlich kann u.a. spezifiziert werden, von welcher beratenden Berufsgruppe Antworten gewünscht werden – diese Angabe ist wohlgermerkt optional, wodurch ein Anfragender a priori nicht wissen muss, welche Berufsgruppe für sein Problem zuständig ist. Wissen über etwaige Fachbegriffe ist nicht notwendig.

Nach einer redaktionellen Prüfung werden die Anfragen an eine Auswahl von über 1.000 Experten aus unterschiedlichen Fachbereichen wie etwa Medizin oder Recht weitergeleitet, welche im besten Fall die Anfrage zur Zufriedenheit des Fragenden beantworten können. Nach einer erneuten redaktionellen Prüfung der Antwort wird diese für den Fragenden freigeschaltet. Das Service ist kostenlos und wird auf Wunsch anonym angeboten. Anders als bei einem automatisierten Informationsfindungsprozess erhält der Benutzer keine Antwort, die aus einer Menge von vordefinierten Contents generiert wird, sondern eine maßgeschneiderte Antwort, die genau auf Bedürfnisse des Benutzers eingeht.

Der Ansatz von bestHELP erlaubt die Verwendung von natürlicher Sprache und zwar nicht im Sinne des Surrogats einer parametrisierten Anfrage. Die Anfrage selbst muss *nicht* parametrisiert werden, daher sind sowohl Ausgangs- als auch Zielmenge offen. Der offensichtliche Nachteil dieses Ansatzes ist die Notwendigkeit, eine vermittelnde Instanz einführen zu müssen.

Zusammenfassend kann eine Tabelle erstellt werden, die die Kommunikationsmöglichkeiten von OBIS beschreibt:

<i>Benutzer sendet eine Anfrage</i>	parametrisierte Anfrage	natürliche Sprache
Anfrage wird vom System <i>verarbeitet</i>	Maschinen-lesbare Anfrage	Menschen-lesbare Anfrage
<i>Antwort wird generiert</i>	Automatisiertes Content Retrieval	Menschlicher Vermittler
Antwort wird an Benutzer <i>übermittelt</i>	Aus einer Menge vorgefertigter Contents	«Maßgeschneiderter» Content

Die nächsten beiden Kapitel beschäftigen sich mit Möglichkeiten der Verbesserung der Interaktion.

2. Linked Open Government

CHRISTIAN BIZER beschrieb das Dilemma von elektronischen Verwaltungsdaten im Jahre 2009 wie folgt: «*The great majority of public-sector data is either not accessible on the Web or accessible only in two forms: Human-readable formats [...] and proprietary data formats*». Die erste Form erlaubt keine maschinelle Auswertung der zur Verfügung gestellten Information und die zweite macht es nötig, dass die potentiellen Konsumenten der Information im Besitz der richtigen Software sind, um die benötigten Daten auslesen zu können [1]. Um diesen Defiziten zu begegnen, ist ein Modus zu schaffen, Daten des öffentlichen Sektors in Maschinen-lesbarer

Form darzustellen, in einem offenen Format und basierend auf einem gemeinsamen Vokabular – denn es sollen nicht nur verschiedenartige, abstrakte Dokumentstrukturen, sondern Contents mit einer Maschinen-lesbaren Bedeutung über das World Wide Web ausgetauscht werden können, Contents, die so reich aufbereitet sind, dass sie eine Maschinen-verständliche Beschreibung der Dinge, die sie beinhalten, bereitstellen. Dies sind die Anforderungen an Linked Open Government Data.

Man kann zweifelsfrei sagen, dass es in den letzten fünf Jahren auf diesem Gebiet Fortschritte gegeben hat, besonders im Bereich der Open Government Data Initiativen: Das US-amerikanische OGD Portal www.data.gov wurde 2009 gestartet, das britische Gegenstück www.data.gov.uk folgte 2010. Die EU fördert OGD durch die Public Sector Information Direktive (PID) 2003 und durch das Open Data Package 2011 [3]. Österreich startete sein Projekt «LOD Pilot AT» im Jahre 2014 (<http://www.lodpilot.at>) – hier wurden Datensätze von unterschiedlichen Open Data Portalen integriert, sowohl Verwaltungsdaten als auch Nicht-Verwaltungsdaten (etwa vom Open Data Portal Austria <https://www.opendataportal.at>, einer Sammlung von offenen Daten aus einem weiten Themenbereich). Diese Daten wurden in ein einheitliches Format konvertiert und können über eine Web API angesprochen werden.

Eine der maßgeblichen Herausforderungen von Linked Open Government Data ist, dass Datensätze in unterschiedlichen Formaten vorliegen; sie basieren auf unterschiedlichen Vokabularen und sind mit Metadaten unterschiedlicher Qualität ausgestattet [3]. Was also benötigt wird, ist ein breiter Konsens darüber, wie Daten repräsentiert werden sollen. Um einen möglichst breiten Bereich an Anwendungen zu befähigen, Web-Contents zu verarbeiten, muss man sich auf ein gemeinsames Standardformat einigen. Die Rolle dieses Standards scheint das Resource Description Framework (RDF) eingenommen zu haben [2]. RDF Graphen beschreiben Information als eine Menge von HTTP URIs, aber anders als bloße Stichwörter nicht um Inhalte zu benennen, sondern um tatsächlich existierende Phänomene (wie Autos oder Kleidung, ...) zu identifizieren. Ein Web Client, der eine derartige URI auflöst, erhält eine Beschreibung des mit der URI bezeichneten Objekts (gewöhnlich im RDF/XML oder RDFa Format). Diese Beschreibung kann wiederum Links zu anderen Objekten enthalten, die von anderen Datenquellen beschrieben werden. Ein Link hat dabei stets die Form eines RDF Triples: *Subjekt – Prädikat – Objekt*. Das Subjekt kann dabei eine URI im Namensraum eines Servers und das Objekt eine URI im Namensraum eines anderen Servers sein. Das Prädikat bestimmt den Typus des Links. Wenn ein Web Client die URI des Prädikats auflöst, erhält er eine Definition (als RDF Schema [RDFS] oder in der Web Ontology Language [OWL]) des Linktypus. Eine derartige Definition kann wiederum Links zu anderen Vokabularen enthalten. Auf diese Weise werden Vokabularen miteinander verflochten, ein Netz von *Bedeutungen* entsteht [1]. Information, die normalerweise in einer hierarchischen Struktur (als eine Form eines Baums) vorliegt, wird so in ein Netz von Relationen (und Analogien) transformiert – also in ein Konzept, das miteinander verlinkten Informationen viel besser entspricht. Information, die dergestalt zur Verfügung gestellt wird, kann dann mittels der semantischen Abfragesprache SPARQL analog zu herkömmlichen relationalen Datenbanken abgefragt werden.

Ein näherer Blick auf das Projekt «LOD Pilot AT» zeigt allerdings, dass noch viel Arbeit investiert werden muss: per Dezember 2015 waren lediglich 18 Datensätze für Entwickler als RDF Graphen verfügbar. Gleichzeitig waren am Open Data Portal Austria 350 offene Datensätze erhältlich, allerdings in nicht-vereinheitlichten Formaten. Dies zeigt: Die Standardisierung von Datensätzen in ein gemeinsames Format ist aufwändig und nicht immer unproblematisch sowohl aus technischer als auch aus rechtlicher Sicht.

3. Herausforderungen an das Natural Language Processing

Es wurde bereits festgestellt, dass konventionelle OBIS die Verarbeitung eines Anfragetextes entweder vollständig automatisieren, indem Benutzer und System mit einem Surrogat natürlicher Sprache (der parametrisierten Anfrage) arbeiten oder aber einen menschlichen Vermittler zwischen Benutzer und System etablieren,

um eine Anfrage zu verarbeiten oder weiter zu leiten.

Im Fall von OBIS wäre es wünschenswert, zumindest teilautomatisierte Systeme für die Anfrageverarbeitung einsetzen zu können, die wenigstens bis zu einem gewissen Grad fähig sind, mit natürlicher Sprache umzugehen [6]. Technologien des Natural language processing (NLP) wie *term extraction* müssen dabei nicht nur bestimmte Begriffe erkennen können, sondern auch die kontextuale Information, in der diese Begriffe auftreten, auswerten [4]. So können statt reinen Begrifflichkeiten aus einer Anfrage *Bedeutungen* abgeleitet werden.

In weiterer Folge könnten Systeme, die sich NLP Technologien bedienen, von der reinen Begriffsextraktion zu dem breiteren Feld der Informationsextraktion erweitert werden. Auf diese Weise könnten aufwändige Tätigkeiten wie Ontologie-Design und das Erstellen von Vokabularien schrittweise automatisiert werden.

4. Zu einem besseren OBIS mit Hilfe semantischer Technologien?

Das österreichische Online Portal <http://www.lehre.at> setzt sich zum Ziel, Informationen über Lehrberufe zur Verfügung zu stellen. Informationen sollen so vollständig wie möglich und so einfach zugänglich wie möglich präsentiert werden. Das Portal integriert dabei Informationen aus verschiedenen Datenquellen, wie etwa dem Online Business Intelligence Tool <https://www.diebestenlehrbetriebe.at>, bei welchem Firmen die Qualität ihrer Lehrlingsausbildung evaluieren lassen können.

Um beurteilen zu können, wie das System (und zukünftige Systeme auf ähnlicher Basis) vom Einsatz semantischer Technologien profitieren könnten, wird eine Ontologie in der Domäne Lehre entwickelt. Um den Umfang der Ontologie bestimmen zu können, wird eine Liste von Fragen aufgestellt, die das System beantworten können soll (sog. *competency questions* [5]). Die Liste beinhaltet so u.a.:

- Wo kann ich in der Nähe meiner Heimatstadt eine Lehrstelle finden, die meinen Fähigkeiten entspricht?
- Welche Lehrstellen gibt es, die eine Berufsschule in der Nähe meines Wohnorts anbieten?
- Welche Fähigkeiten werden benötigt, um Augenoptiker zu werden?
- Welcher Lehrberuf wird in Wien am besten entlohnt?
- Welche ausbildungsrechtlichen Probleme könnten sich ergeben, wenn ich von Lehrberuf A zu Lehrberuf B wechsle?
- Welche Firmen im Gesundheitswesen bieten eine Ausbildung an, die von <https://www.diebestenlehrebetriebe.at> am besten bewertet wurde?

Man kann sehen, dass es sich um einen Mix aus ausbildungnerischen, geografischen, arbeitsmarkttechnischen und berufsrechtlichen Fragestellungen handelt. Auch sollen Evaluierungen, die auf dem Portal <https://diebestenlehrbetriebe.at> erstellt wurden, in Betracht gezogen werden. Die Wiederverwendung und Weiterentwicklung bestehender Ontologien wird diskutiert; allerdings existiert zum jetzigen Zeitpunkt keine verwertbare Ontologie, die die spezielle Situation der Lehrausbildung in Österreich wiederzugeben vermag.

5. Literatur

- [1] BIZER, CHRISTIAN, The Emerging Web of Linked Data, IEEE Intelligent Systems, vol. 24, no. 5, pp. 87–92, September–October 2009
- [2] HEATH, TOM/BIZER CHRISTIAN, Evolving the Web into a Global Data Space, Morgan & Claypool, 2011
- [3] DING, LI/PERISTERAS, VASSILIOS/HAUSENBLAS MICHAEL, Linked Open Government Data, IEEE Intelligent Systems, vol. 27, no. 3, pp. 11–15, May–June 2012

- [4] MAYNARD, DIANA/LI, YAORYANG/PETERS, WIM, NLP Techniques for Term Extraction and Ontology Population, <https://gate.ac.uk/sale/olp-book/main.pdf> (as per 6 January 2016)
- [5] NOY, NATALYA F./MCGUINNESS, DEBORAH, ONTOLOGY Development 101: A Guide to Creating Your First Ontology, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001, <http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf> (as per 6 January 2016)
- [6] RUSSELL, STUART/NORVIG, PETER, Artificial Intelligence. A Modern Approach, Third Edition, Pearson Education Inc., 2010
- [7] TRAUNMÜLLER, ROLAND/KRENMAYR, ANDREAS, Bürgerinformationssysteme – Neue Vorstellungen, In: Schweighofer, Erich, Kummer, Franz, Hötzendorfer (Hrsg.), Kooperation. Tagungsband des 18. Internationalen Rechtsinformatik Symposions. IRIS 2015, Wien 2015.