Hinter den Leuchttürmen

Auf dem Weg zu einer möglichst vollständigen Erfassung von kommunalen OGD-Angeboten

Werkstattbericht aus dem ifib/OKF-Projekt „Open Data Monitor“

Herbert Kubicek, Barbara Lippa (ifib)
Daniel Dietrich, Matthew Fullerton (OKF DE)

D-A-CH-LI-Konferenz, 6./7. August 2014 in Friedrichshafen

Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

Hintergrund und Anlass

Seit langem dieselben kommunalen OGD Leuchtturmprojekte – Aber was sieht man hinter den Leuchttümen? Und: Wie vergleichbar sind sie ?

Es gibt zurzeit keinen Überblick über Daten verschiedener Kommunen zum selben Thema, für **vergleichende Analysen** obwohl diese für die relevantesten Nutzungsszenarien erforderlich sind:

- Apps für eine einzige Stadt lohnen wirtschaftlich kaum
- Journalisten und Wissenschaftlern sind vor allem an Vergleichen interessiert

Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

Derzeitige Überblicke über kommunale Angebote

- govdata.de
  - Kein Filter nach Namen von Kommunen oder nach einer Kombi-Suche Thema + kommunale Ebene
  - Ist eben Bund-Länder-Portal
  - Über API ermittelt: zurzeit Daten von 28 Städten, Gemeinden und Kreisen
- OKF neben Country- auch City Index http://de-city.census.okfn.org
  - Existenz und Merkmale von 15 ausgewählten Datensätzen, zurzeit für 13 deutsche Städte
- Beide weit weg von Vollständigkeit der Anbieter und der einzelnen Angebote
- Mit dem Ansatz, zu warten, wer was meldet, auch nicht anders zu erwarten

Städtevergleich Hauptamt Leipzig 05/13

- 31 Städte angeschrieben -> 16 Antworten
- **In Planung bis 2015 -> 6** : Bonn, Dresden, Magdeburg, München, Nürnberg, Wuppertal
- **Nicht geplant bis 2015 -> 10:** Braunschweig, Chemnitz, Düsseldorf, Duisburg, Essen, Gelsenkirchen, Kiel, Krefeld, Mannheim, Münster
- **Keine Antwort -> 15:** Aachen, Berlin, Bielefeld, Dortmund, Erfurt, Hagen, Halle, Hannover, Karlsruhe, Köln, Mönchengladbach, Rostock, Stuttgart, Wiesbaden

http://notes.leipzig.de/appl/laura/wp5/kais02.nsf/docid/B9F0508BE42227B7C1257C7F002F23F9/$FILE/V-ds-3615-anlage-2.pdf

Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

Ziel des ifib/OKF-Projekts

Von Mai bis Dezember 2014 werden verschiedene Wege zur Gewinnung eines möglichst vollständigen Überblicks über die Angebote offener Daten aller deutschen Gebiets-körperschaften erprobt und hinsichtlich des Erhebungs-aufwands und der Qualität der Ergebnisse verglichen. Auf den Erfahrungen aus diesem ersten Schritt aufbauend soll ein Konzept für ein permanentes Angebot entwickelt werden.

Dieser Prototyp des Open Data Monitors umfasst zu Demonstrationszwecken zur Zeit **13 Städte** mit insgesamt **1850 Datensätzen** (ohne Anspruch auf Vollständigkeit)

Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

Ansatz des ifib/OKF-Projekts

- Automatisierte Suche nach bestimmten Datenformaten in kommunalen Angeboten, manuelle Indexierung und Validierung für eine Stichprobe von 100 Kommunen
  - Zusammensetzung: 40 Groß-, 20 Mittel-, 20 Klein- und 10 Landstädte + 10 größte Landkreise
  - Auswahlkriterien: eigenes Datenportal, City Census der OKF, Landeshauptstädte, Vorrecherchen, Auswahl nach Einwohnerzahl
- Für Teilstichprobe von 12 Städten systematischer verschiedener Tools:
  - Vergleich der gefundenen kommunalen OGD-URLs in Google und Bing Indizes,
  - Eigener Crawler zur Suche auf den kommunalen Webseiten,
  - Auswertung von Datenkatalogen, soweit vorhanden,
  - Auswertung der Kommunen in govdata.de.
- Ergänzend: Online Umfrage nach anderen URLs und Meldungen an Portale, z.Zt. mit dem niedersächsischen Städtetag (ca 150 Mitgliedsstädte)

Inhaltliche Ergebnisse: Daten-Angebote 100er Stichprobe

	Google – Dateien gefunden*)	Google - darunter relevante Datensätze +)	Bing - Dateien gefunden**)	Bing – darunter relevante Datensätze +)
100 Städte	**46**	**35**	**43**	**29**
40 Großstädte	29	23	26	19
20 Mittelstädte	9	6	10	5
20 Kleinstädte	2	2	1	1
10 Landstädte	2	2	3	3
10 Landkreise	4	2	3	1

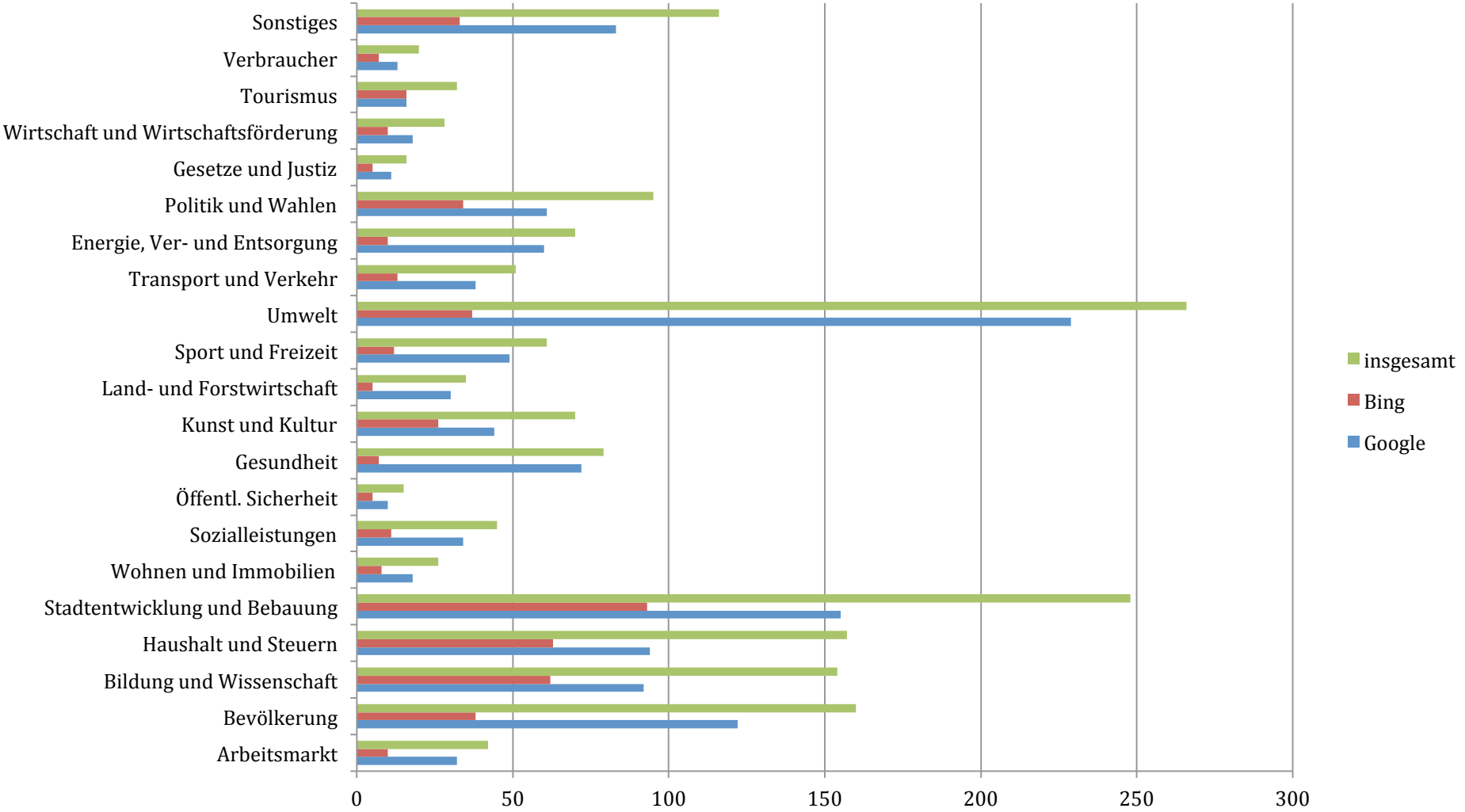
NUR TEILWEISE ÜBERSCHNEIDUNG ZWISCHEN DEN TOOLS

*) xls. xlsx, kml und kmz (andere Formate nicht indiziert oder nicht ausgewiesen)

**) csv, xls. xlsx, json, shp, kml, kmz, rdf, zip u.a.m.

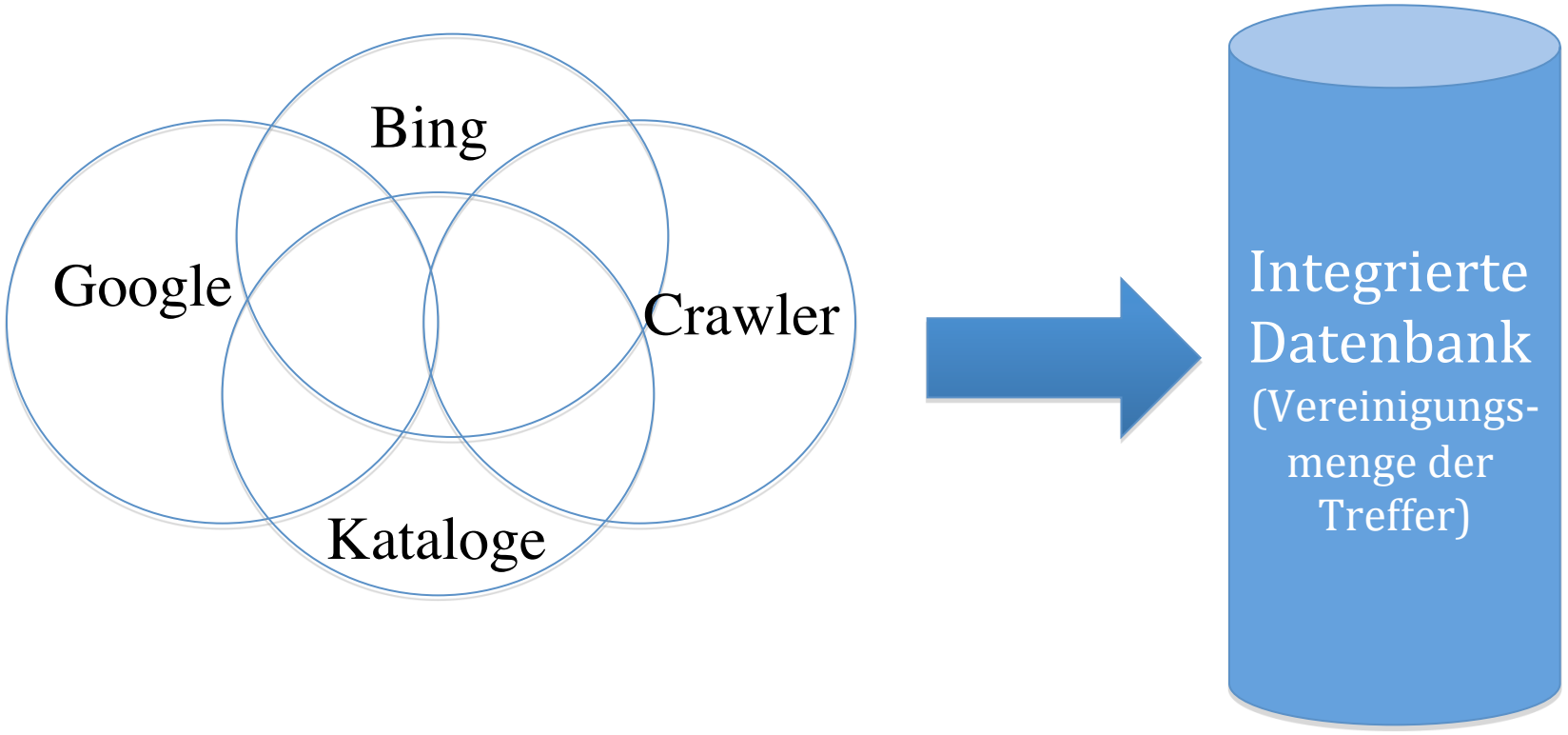
+) manuell geprüft

Inhaltliche Ergebnisse 100er Stichprobe: Treffer pro Themengebiet bei Google und BING



Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

Vollständige Erfassung eines Angebotes nur durch Kombination der verschiedenen Tools



Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

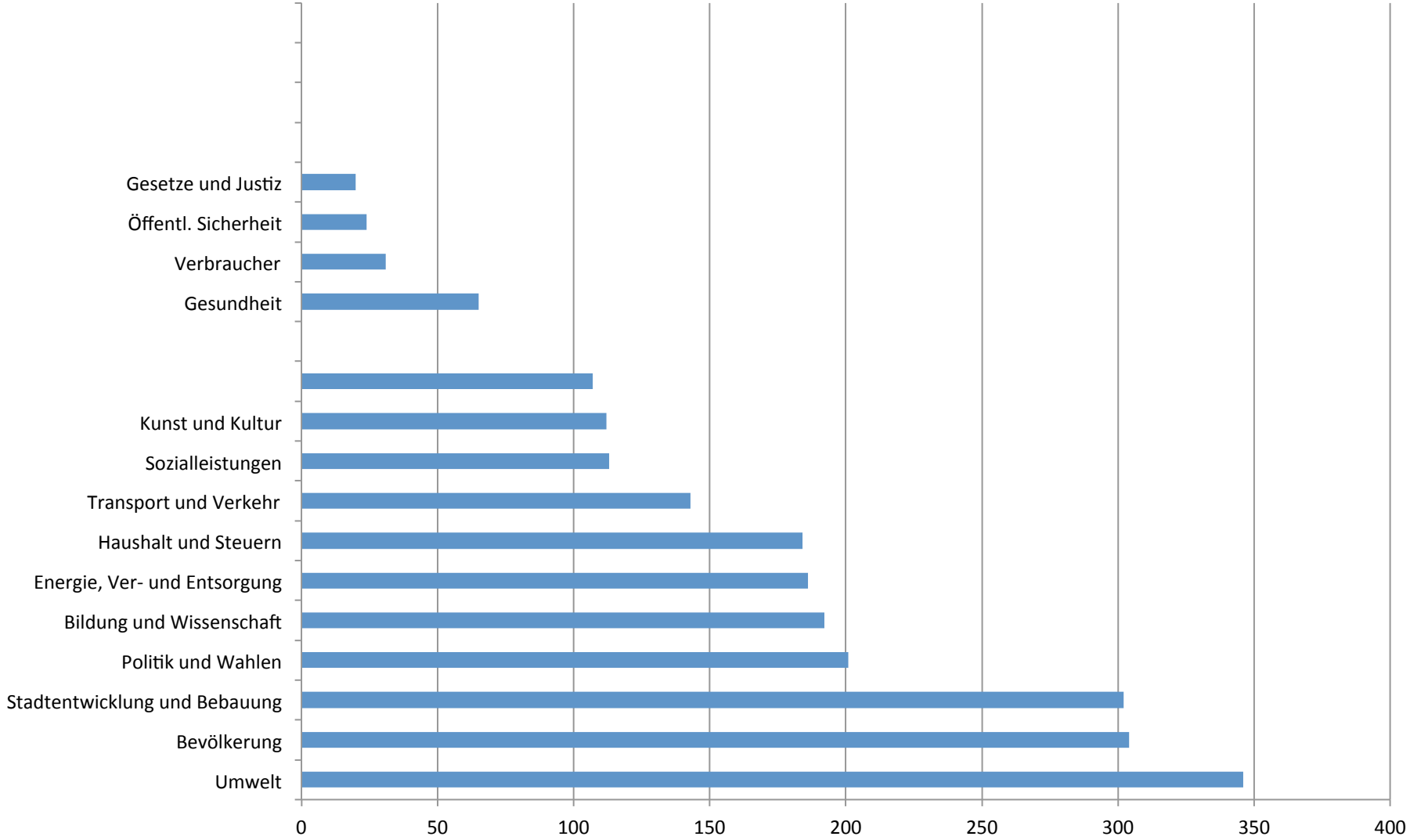
Ergebnisse: Vergleich Anzahl der Treffer

Kommune	Datensätze gesamt	Google Treffer	Bing Treffer	Crawler Treffer	Im Katalog
Bremen	322	14.91%	9.94%	54.66%	**47.20%**
Rostock	752	4.12%	0.13%	0.13%	**99.34%**
Wennigsen	32	3.13%	6.25%	0.00%	100.00%
Münster	4	100.00%	50.00%	50.00%	entfällt
Ulm	204	19.61%	6.86%	32.35%	**83.33%**
Köln	384	4.95%	1.56%	0.52%	**94.53%**
Bonn	57	21.05%	0.00%	0.00%	**78.95%**
Frankfurt	5	20.00%	0.00%	entfällt	entfällt
Moers	224	**0.45%**	**4.91%**	**51.34%**	**63.39%**
Berlin	701	32.95%	20.26%	entfällt	**58.63%**
Stuttgart	275	42.55%	4.36%	54.18%	entfällt
München	16	62.50%	50.00%	entfällt	entfällt

Ergebnisse: Übereinstimmung der Treffer

	Gesamt	Google-Bing	Google-Crawl	Google-Catalog	Bing-Crawl	Bing-Catalog	Crawl-Catalog
Bremen	322	14	28	7	15	6	38
Rostock	752	1	1	30	1	0	0
Wennig-sen	32	0	0	2	0	1	0
Münster	4	2	2	entfällt	2	entfällt	entfällt
Ulm	204	9	23	17	5	4	51
Köln	384	6	0	0	0	0	0
Bonn	57	0	0	0	0	0	0
Frankfurt	5	0	entfällt	entfällt	entfällt	entfällt	entfällt
Moers	224	0	0	0	6	0	35
Berlin	701	30	entfällt	8	entfällt	50	entfällt

12 Städte Vergleich: Themengebiete



Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

12 Städte Vergleich – Themengebiete pro Stadt

Stadt	*Umwelt*	*Stadt-entwick-lung und Bebau-ung*	*Bevölke rung*	*Politik und Wahlen*	*Bildung und Wissen-schaft*	*Haus-halt und Steuern*	*Trans-port und Verkehr*	*Sozial-leistun-gen*	*Kunst und Kultur*	*Wirt-schaft und Wi-förde-rung*		*Gesund-heit*	*Ver-braucher*	*Öffentl. Sicherheit*	
Bremen	74	81	79	10	97	41	53	66	23	72		22	1	5	
Rostock	14	21	10	14	8		18	6	20	1		9	8	6	
Wennigsen	4	7	2	1	1	9	2	1						1	
Münster		1											1		
Moers	2	6	54	14	28	5			2						
Ulm	2	31	19	31	15	4	6	8	4	3		3	5	5	
Köln	2	17	19	37	3	10	9	2	6	1		3		2	
Bonn		4	2	26		11			5	3					
Berlin	231	70	45	7	11	31	36	17	26	8		22	10	2	
Stuttgart	14	49	6	30	14	51	6	8	14	5		3	5	3	
München				17											
Frankfurt		2	4	4	5	2	1	3		2		2			
	343	289	240	191	182	164	131	111	100	95		64	30	24	

Gründe für die Unterschiede (1)

Bei der Suche nach bestimmten Formaten wie xls, csv, …

- Google indiziert nur kml, kmz, xls(x) -> Geringes Recall
- Eigener Crawler findet deutlich mehr, aber auch viel Irrelevantes (-> geringe Precision) wie xls-Formulare, Vorlagen, Beispielberechnungen

Gründe für die Unterschiede (2)

Datenkataloge sind nicht vollständig

	Katalog 'Not Found'
Bremen	170
Rostock	5
Wennigsen	0
Ulm	34
Köln	21
Bonn	12
Moers	82
Berlin	290

Bremen, z.B. nicht im Katalog Bremer-Philharmoniker-Konzerteinnahmen-2005-2013.csv, Studierendenzahlen-Hochschulen-2012-2013.csv, Messebeteiligungen Stand 09-03.xls und viele Adresslisten als XML-Datei

Bonn: Wahlergebnisse 2004 – 2013, Teilweise werden gleiche oder ähnliche Daten zweimal unter verschiedenen Dateinamen veröffentlicht

Datenkataloge sind sehr unterschiedlich (wie die Leuchttürme)

Nur Köln und Bonn sind gleich

Insgesamt: sixcms/JSON Dump, CKAN API 2, DKAN, RSS + Scraping möglich, HTML Auflistung + ISO19139 XML

und oft nicht Teil der Homepage der Kommune

Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

Qualitative Kennzeichnung der Angebote kaum möglich

Hauptziel der Karte: Zuordnung zu Themengebieten (wichtigster Filter):

Bisher automatisch nur bei Datenkatalogen (10 von 100) teilweise möglich, hoher manueller Aufwand erforderlich

Wichtig für die Relevanz und Vergleichbarkeit ist ausserdem die Aktualität und der Bezugszeitraum:

Erstellungs- bzw. Aktualisierungsdatum werden in den Datenkatalogen häufiger angegeben, Aktualisierungszyklus seltener, obwohl als Referenz ganz wichtig

Ein Projekt von ifib Open Knowledge Foundation Mit finanzieller Unterstützung der bpb: Bundeszentrale für politische Bildung

Erste Schlussfolgerungen (1): Auffindbarkeit der Angebote

- Eine auch nur annähernd vollständige Erfassung der Anbieter und Angebote ist nur mit einer Kombination der verschiedenen Tools und mit sehr hohem Aufwand zu erreichen
- Erstaunlich ist die teilweise enorme Differenz zwischen den Treffern der verschiedenen Tools, insbesondere das geringe Recall von Google und teilweise auch der Datenkataloge
- Die für Nutzer interessante thematische Zuordnung jenseits der (wenigen) Datenkataloge ist wegen der unterschiedlichen Bezeichnungen und fehlender Metadaten bzw. fehlendem Zugang zu Metadaten mit noch größerem Aufwand verbunden, ohne Hoffnung auf kurzfristige technische Unterstützung.

Konsequenzen für Anbieter:

Portale und Kataloge sind notwendig, damit die Angebote gefunden werden. Auf Google alleine sollte man sich nicht verlassen

Portalbetreiber sollten selbst die kommunalen Webseite crawlen und die „wilden“ Angebote in ihrem Katalog registrieren.

Das Projekt stellt den entwickelt5en Crawler dazu zur Verfügung

Die Unterschiede zwischen den Katalogen verdeutlichen den Standardisierungsbedarf von Katalogen und Metadaten

Erste Schlussfolgerungen (2)
Erklärung der Anzahl und Verteilung der Angebote

Die 100-er Stichprobe zeigt:

Die OGD-Angebote hängen deutlich, wenn auch nicht ausschließlich von der **Größe der Kommune** ab.

Die Angebote weisen große Unterschiede zwischen den **Themengebieten** auf

Beides passt nicht zu der These von einer allgemeinen Verwaltungskultur des Amtsgeheimnisses als Hauptgrund, keine Angebote zu machen.

Gegenthese:

Es gibt in Bezug auf Transparenz und Offenlegung von Daten sehr unterschiedliche **Bereichskulturen.** Diese ergeben sich aus der **unterschiedlichen Bedeutung, die Veröffentlichungen für die Erreichung der jeweiligen Ziele in den Handlungsfeldern haben** und hängen auch von gesetzlichen Vorgaben und mit über lange Zeit etablierten Strukturen und Prozessen zusammen

„Transparenz folgt der Kultur“ und „Die Kultur folgt der Struktur“
Oder „*Das Sein bestimmt das Bewusstsein*“!

Mehr: http://open-data-map.de/