Inhaltliche Datenqualität von Open Government Data

4. OGD D-A-CH-LI - Konferenz - Open X

24 Juni 2015 - Wiener Rathaus
Lichtenfelsgasse 2, 1010 Wien

Johann Höchtl, Bernhard Krabina, Jürgen Umbrich

WAS IST

1. Datenqualität

?

Datenqualität

- **ISO 9001**
  - **Gesamtheit, Richtigkeit, Konsistenz, zeitliche Nähe der Verfügbarkeit** von Daten

http://mitiq.mit.edu/IQIS/Documents/CDOIQS_200977/Papers/01_05_T2C.pdf

- **ISO 8000** – Norm für Datenqualität
  - **Syntax**
    - Each data set shall contain a **reference to the syntax** to which the data set complies... **that is publicly available**.
  - **Semantic encoding**
    - Each data element value shall **reference all concepts necessary to unambiguously define its meaning**....
  - **Conformance to requirements**
    - Each data set shall contain a reference to the data requirements statement to which the data set complies ... **shall be publicly available**.

Datenqualität - BenutzerInnensicht

What are the data quality measures for open data?

http://opendata.stackexchange.com/questions/613/what-are-the-data-quality-measures-for-open-data

5

How does a consumer know they are getting good data? Are there standard frameworks for grading the quality of an open data set? Should there be metrics published around accuracy, completeness, timeliness or validity of the data? Should there be a minimum set of controls on the part of the publisher?

releasing-data data.gov

share improve this question

2

3

I think the question, as phrased, is impossible to answer well, but I will try.

Q: "How does a consumer know they are getting good data?"

A: Let me answer with more questions. How does a consumer know they are getting a good search result from Google? How do they know when the news is of high quality? It depends. As consumers get more interested and informed about something, they do better. The most savvy and informed consumers will compare a data set against a known source. Others have to rely on some degree of trust.

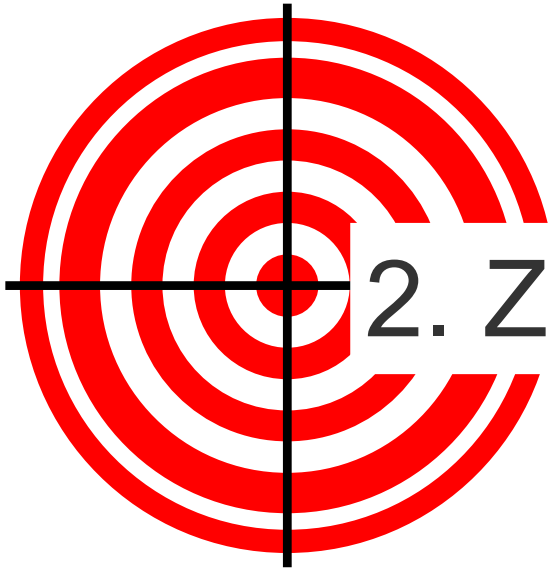
Q: "Are there standard frameworks for grading the quality of an open data set?"

A: In practice, there are defacto standards for metadata. For example, data.gov uses Dublin Core along with additional attributes. CKAN has many of the same attributes.

Also, for each type of data (or subfield) there are often industry standards or at least conventions. Good luck enumerating those!

A post from the Sunlight Foundation, Government Data Sets - Managing Expectations is a high-level gloss; it breaks down "dataset quality" into provenance, data quality, responsibility, maintenance, and documentation.

The above article is somewhat naive; the quality of a data set is not an independent thing. As Wikipedia - Data Quality points out, the quality of a data set depends on the question asked of it. There is no "one" measure of data quality. Rather, there is a subjective 'appropriateness' for each question you might ask of a data set. You can't ignore the subjective nature of data quality.



2. Zielsetzung

Technische und organisatorische Maßnahmen verbessern die **Datenqualität** aktuell verfügbarer Datensätze und **unterstützen** den **Veröffentlichungsprozess**, um in Zukunft **höhere Qualitätsniveaus**, und somit **erhöhte Nutzbarkeit und Nachhaltigkeit** von offenen Daten zu erreichen.

https://www.data.gv.at/wp-content/uploads/2012/03/Mission-Statement-AG-Qualitaetssicherung-OpenData-Portale.pdf

Grundprinzipien

1. **Datenqualität** ist eine **inhärente Eigenschaft der Verwendbarkeit**

2. **Verwendbarkeit** führt zu **Vertrauen**

Institutionalisierung von Datenqualität in Österreich

Ziele

- **Sammlung**, Auswertung und Bereitstellung von **nationalen und internationalen Erfahrungswerten**
  - Messmetriken, Vorgehensmodelle, Initiativen
- Evaluierung, Konzeption und Implementierung von **technischen Werkzeugen**
  - Werkzeuge entlang des gesamten Veröffentlichungsprozesses verbessern die Qualität von Daten
- Konkrete **Handlungsempfehlungen und Schulungen**

Internationale gute Praxis

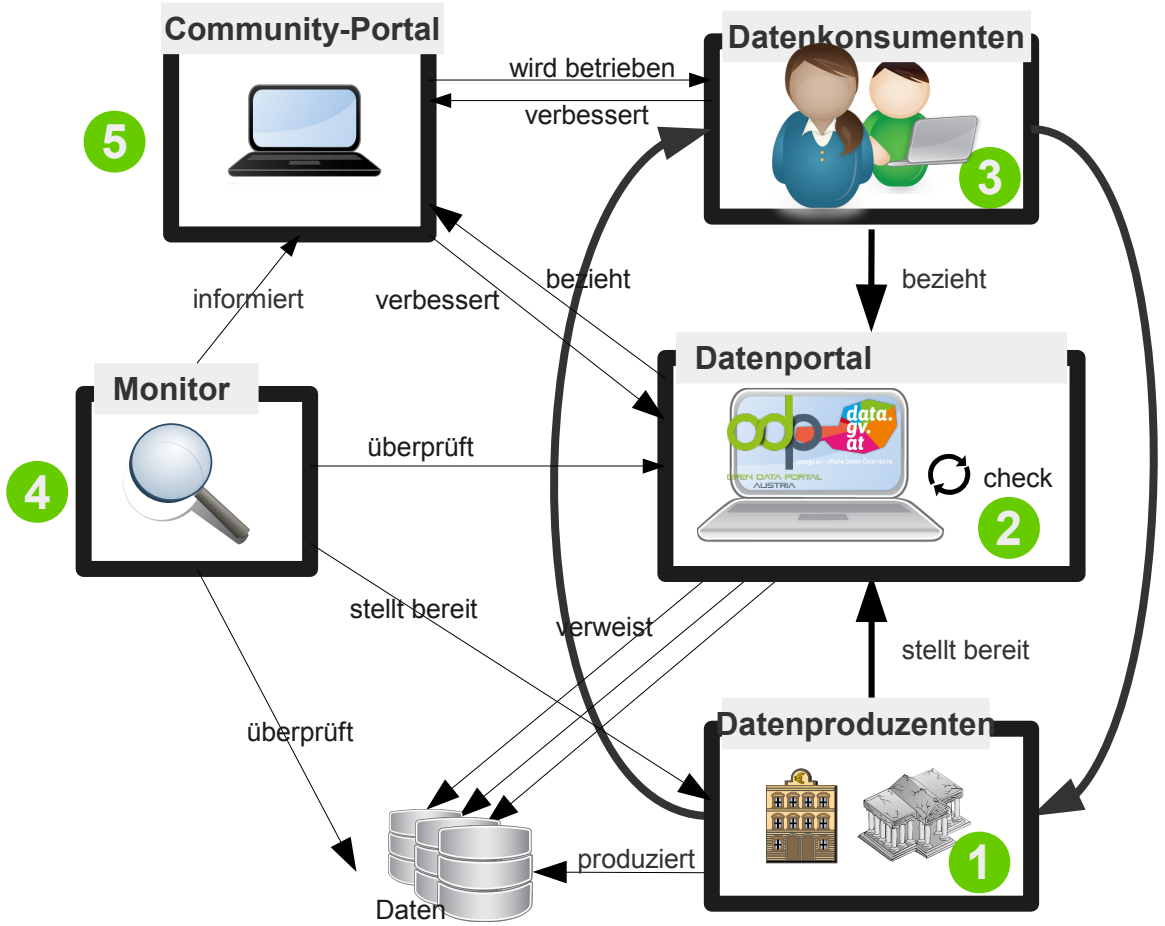
http://theodi.github.io/presentations/training-law-best-practice.html

Institutionalisierung von Datenqualität in Österreich

Nicht-Ziele

- **Erstellung von Muß-Kriterien** zur Teilnahme an Open Government Data
  - Freiwilligkeit nicht einschränken
  - Bereitstellung offener Verwaltungsdaten **nicht erschweren**
- **Generierung von Mehraufwand**
  - Werkzeugen sollen Veröffentlichungsprozess **unterstützen**
- **Neuerfindung des Rades**
  - keine Entwicklung neuer, proprietärer Ansätze
  - Integration und Erweiterung bestehender Infrastrukturen

Open Data Qualitätsframework



1. Qualitätsprozesse und **Vorgehensmodelle** auf Seite der **Datenbereitsteller**
2. **Qualitätsüberprüfungen bei** der **Bereitstellung** am Datenportal
3. Beiträge von **DatennutzerInnen**
4. **Monitoring der Datenqualität** im Zeitverlauf und Berichterstellung
5. **Datenportal der Öffentlichkeit** benutzerInnengenerierte Inhalte durchführen von Experimenten

3. Reale Probleme lösen

Reale Probleme lösen (1)

http://www.computerweekly.com/news/2240227682/Poor-data-quality-hindering-government-open-data-transparency-programme

- **Uneinheitliche Zeichenkodierungen**
  - *Microsoft Excel caused data problems even when used […] UTF-8*
  - *Data contaminated with **characters incomprehensible to UTF-8;** ill-formatted following UTF-8; flipped erratically between other character formats; used US ASCII standard, ISO-8859 standard and a similar non-ISO encoding*
- **Inkonsistente Benennung von Dateinamen und Datenfeldern**
  - *Data were regularly **formatted with commas;** changed its filename convention; **omitted or added data fields;** changed the way it formatted dates*

Reale Probleme lösen (2)

- **Kodierung**: Ist es auch wirklich UTF8?
- Überprüfung der **Qualität der Beschreibungen**
- **Verfügbarkeit von Ressourcen** („broken links“)
- **Einheitliche Formatierung** für Datumswerte, Ortsbezeichnungen
- **Eindeutige Datenkennungen** zur Korrelation
  - In der Verwaltung verwendet
  - öffentliche Bekanntheit
- **Best-Practice Guide** für Datenbereitsteller
  - Datumsformate, Uhrzeit, Geo-Kodierungen

Identifizierte Datenprobleme (AT)

https://github.com/the42/ogdat

- Nicht auflösbare Links

http://data.linz.gv.at/katalog/population/wanderung/zuzug/2009/zuzug_2009.csv liefert nicht-OK Status-Code '404' (Get)	2014-11-16 20:22:43.533842+00
http://data.linz.gv.at/katalog/population/wanderung/zuzug/2008/zuzug_2008.csv liefert nicht-OK Status-Code '404' (Get)	2014-11-16 20:22:43.533842+00
http://data.linz.gv.at/katalog/population/wanderung/zuzug/2011/zuzug_2011.csv liefert nicht-OK Status-Code '404' (Get)	2014-11-16 20:22:43.533842+00
http://data.linz.gv.at/katalog/population/wanderung/zuzug/2010/zuzug_2010.csv liefert nicht-OK Status-Code '404' (Get)	2014-11-16 20:22:43.533842+00
http://data.linz.at/katalog/stadt/wohnungen/wohnraeume/2012/twoanzrg_2012.csv liefert nicht-OK Status-Code '404' (Get)	2014-11-16 20:33:42.795532+00
http://data.linz.gv.at/ogd/katalog/politik_verwaltung/verwaltung/budget/2013/RA/ZRECHAB_2013.csv liefert nicht-OK Status-Code '404' (Get)	2014-11-17 13:32:08.374879+00

- Probleme bei Metadatenbeschreibungen

Land Tirol	Feldwert vom Typ ÖNORM ISO 8601 TM_Primitive 'YYYY-MM-DDThh:mm:ss' erwartet, Wert entspricht aber nicht diesem Typ: '2009-01-01'
Land Tirol	JSON vom Typ 'Array of String' erwartet, es wurde jedoch ein einzelner Wert geliefert
Land Tirol	kein Wert für Link angegeben (Länge 0)
Land Tirol	Beschreibung enthält weniger als 20 Zeichen (sinnvolle Beschreibung?)
Land Tirol	Zeichenkette mit Länge 0 an dieser Stelle nicht sinnvoll
Land Tirol	Zeichenkette mit Länge 0 an dieser Stelle nicht sinnvoll

Nicht verfügbare Ressourcen

Dead links on data catalogs

As I started looking at data on CKAN sites, I noticed that a lot of the datasets were links to files on other websites and that a lot of these links were dead. Then I started wondering which links were dead and how this happens.

http://thomaslevine.com/!/data-catalog-dead-links/

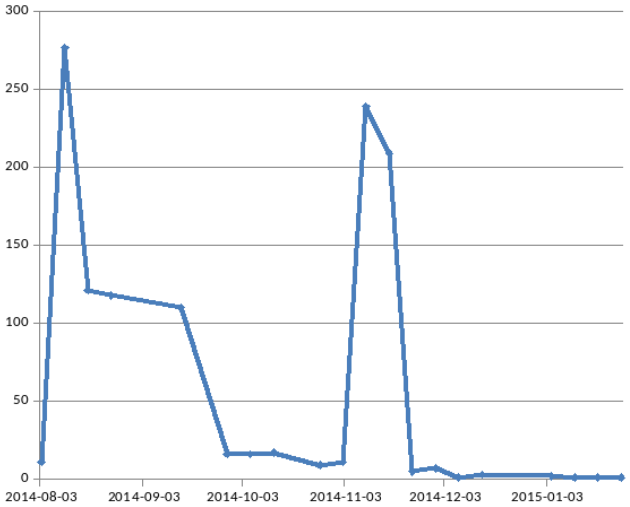
Broken links and hardly any new data on Dutch government open data portal

25/06/2014

The government of the Netherlands still lags behind in opening data to the public. The number of accessible datasets via the Dutch government open data portal is even reduced compared to a year ago.

http://openstate.eu/2014/06/nederlands-nauwelijks-nieuwe-Datasets-op-data-overheid-nl/

Überprüfung der Ressourcenverfügbarkeit auf Data.gv.at - Wien



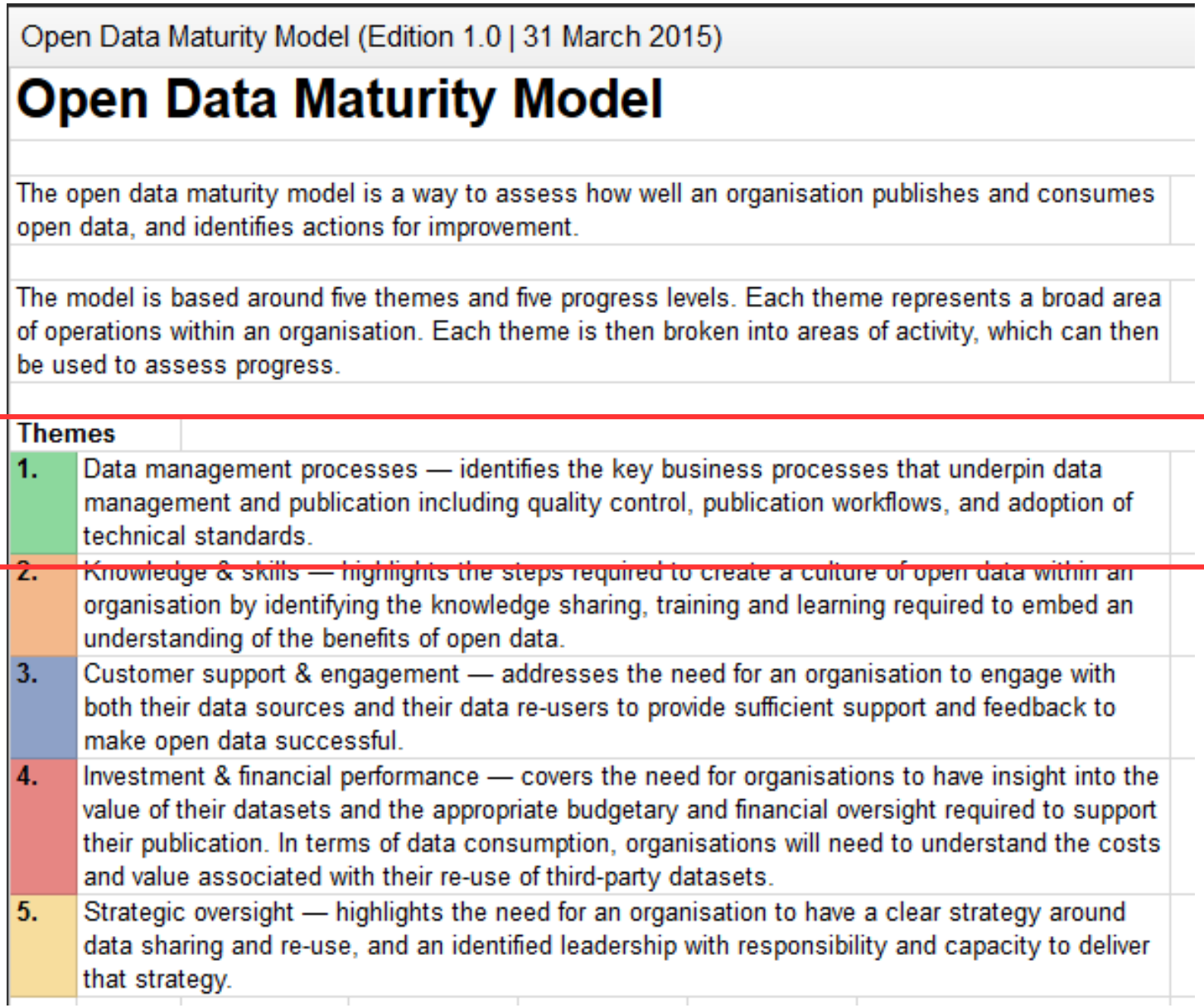
4. Maßnahmen zur Steigerung der Datenqualität

http://practicalmanagers.com/2014/05/31/fixing-bad-bosses/

Prozess-Ebene

- **Data publication** as an **integral**, well- defined and standardized **part of daily procedures** and routines
  - A. Zuiderwijk, M. Janssen, S. Choenni, and R. Meijer, “Design principles for improving the process of publishing open data,” Transforming Government: People, Process and Policy, vol. 8, no. 2, pp. 185–204, 2014.
- **Process model** towards open data as a facilitator for open government
  - G. Lee and Y. H. Kwak, “An Open Government Implementation Model: Moving to Increased Public Engagement,” IBM Center for The Business of Government, Jan. 2011 [Online]. Available: http://www.businessofgovernment.org/sites/default/files/An%20Open%20Government%20Implementation%20Model.pdf
  - B. Krabina, T. Prorok, und B. Lutz, „Open Government Vorgehensmodell“, KDZ, Vienna, Vorgehensmodell V2.0, 2012.
  - L. Dodds und A. Newman, „Open Data Maturity Model“. Open Data Institute, 2015.
- Chief Data Officer (**CDO**)
  - Y. Lee, “A cubic framework for the chief data officer : succeeding in a world of big data,” 2014.

Prozess-Ebene (2)



Open Data Maturity Model (Edition 1.0 | 31 March 2015)

Open Data Maturity Model

The open data maturity model is a way to assess how well an organisation publishes and consumes open data, and identifies actions for improvement.

The model is based around five themes and five progress levels. Each theme represents a broad area of operations within an organisation. Each theme is then broken into areas of activity, which can then be used to assess progress.

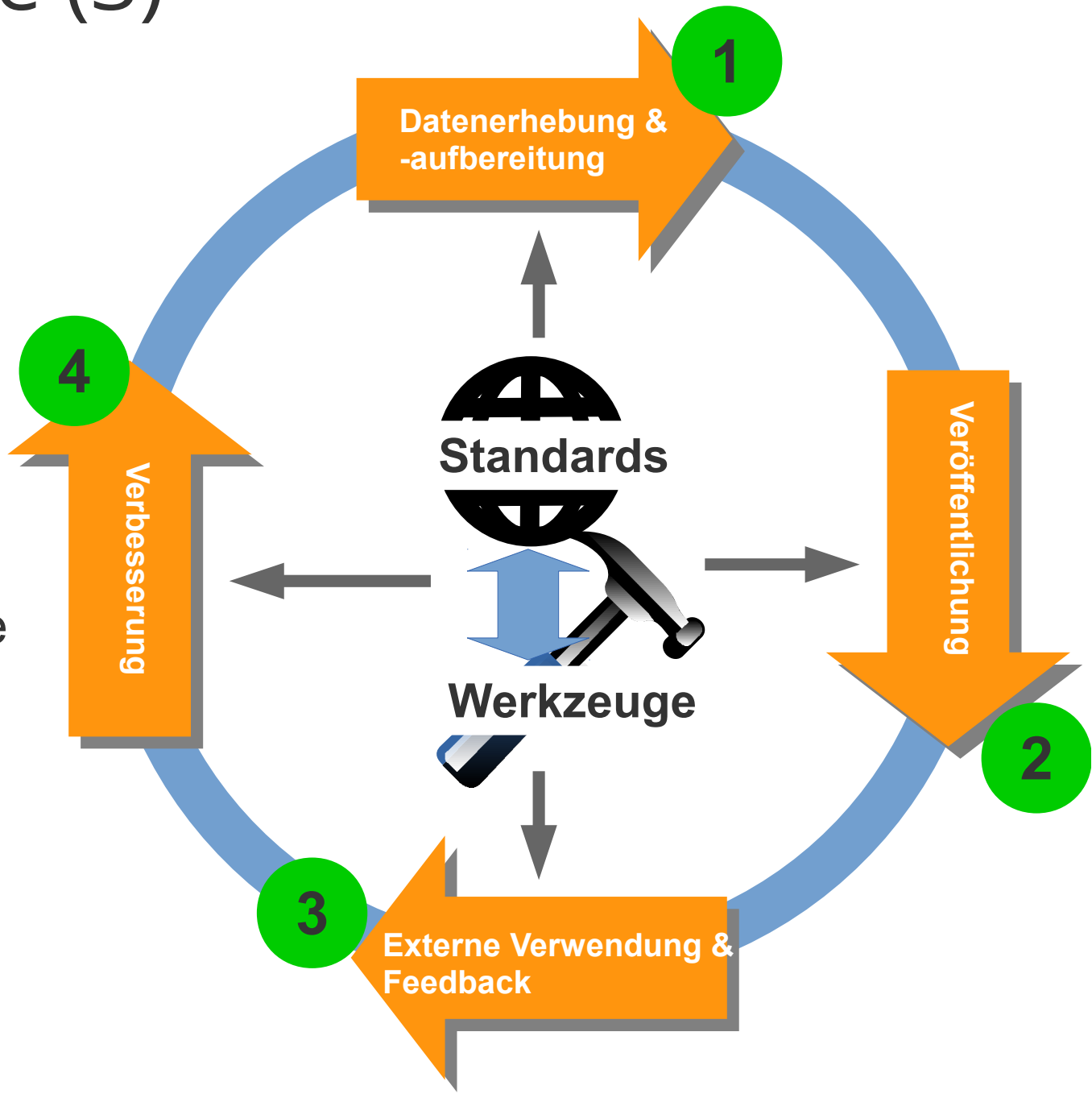
Themes

1.	Data management processes — identifies the key business processes that underpin data management and publication including quality control, publication workflows, and adoption of technical standards.
2.	Knowledge & skills — highlights the steps required to create a culture of open data within an organisation by identifying the knowledge sharing, training and learning required to embed an understanding of the benefits of open data.
3.	Customer support & engagement — addresses the need for an organisation to engage with both their data sources and their data re-users to provide sufficient support and feedback to make open data successful.
4.	Investment & financial performance — covers the need for organisations to have insight into the value of their datasets and the appropriate budgetary and financial oversight required to support their publication. In terms of data consumption, organisations will need to understand the costs and value associated with their re-use of third-party datasets.
5.	Strategic oversight — highlights the need for an organisation to have a clear strategy around data sharing and re-use, and an identified leadership with responsibility and capacity to deliver that strategy.

http://theodi.org/guides/maturity-model

Prozess-Ebene (3)

1) Automatisierte Abläufe und einheitliche Darstellungen von z.B. Datum, Uhrzeit, Geodaten
2) Encoding, Dateigröße, Formate
3) Feedback-mechanismen
4) Daten verbessern und zurückmelden

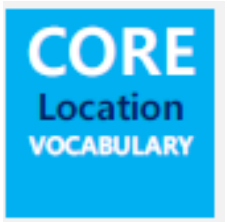


Standards (1)

- **Data on the Web**
  - Data on the Web Best Practices Working Group Charter http://www.w3.org/2013/05/odbp-charter.html
  - Encodings: UTF8
- **Dateiformate**
  - CSV: CSV on the Web Working Group http://www.w3.org/2013/csvw/wiki/Main_Page
  - Frictionless open Data: CSV Files (OKFN guidance document) http://data.okfn.org/doc/csv
- **Daten-Entitäten**
  - Geo-Data: Spatial Data on the Web Working Group Charter http://www.w3.org/2015/spatial/charter
  - Date & Time: ISO 8601 http://www.w3.org/TR/NOTE-datetime
  - Core-Vokabulare: simplified, reusable and extensible data models https://joinup.ec.europa.eu/asset/core_vocabularies

Standards (2) – Core Vokabulare

- characteristics of a legal entity, legal name, the activity, address, legal identifier, company type
- formally published on the W3C standards track as a Public Working Draft



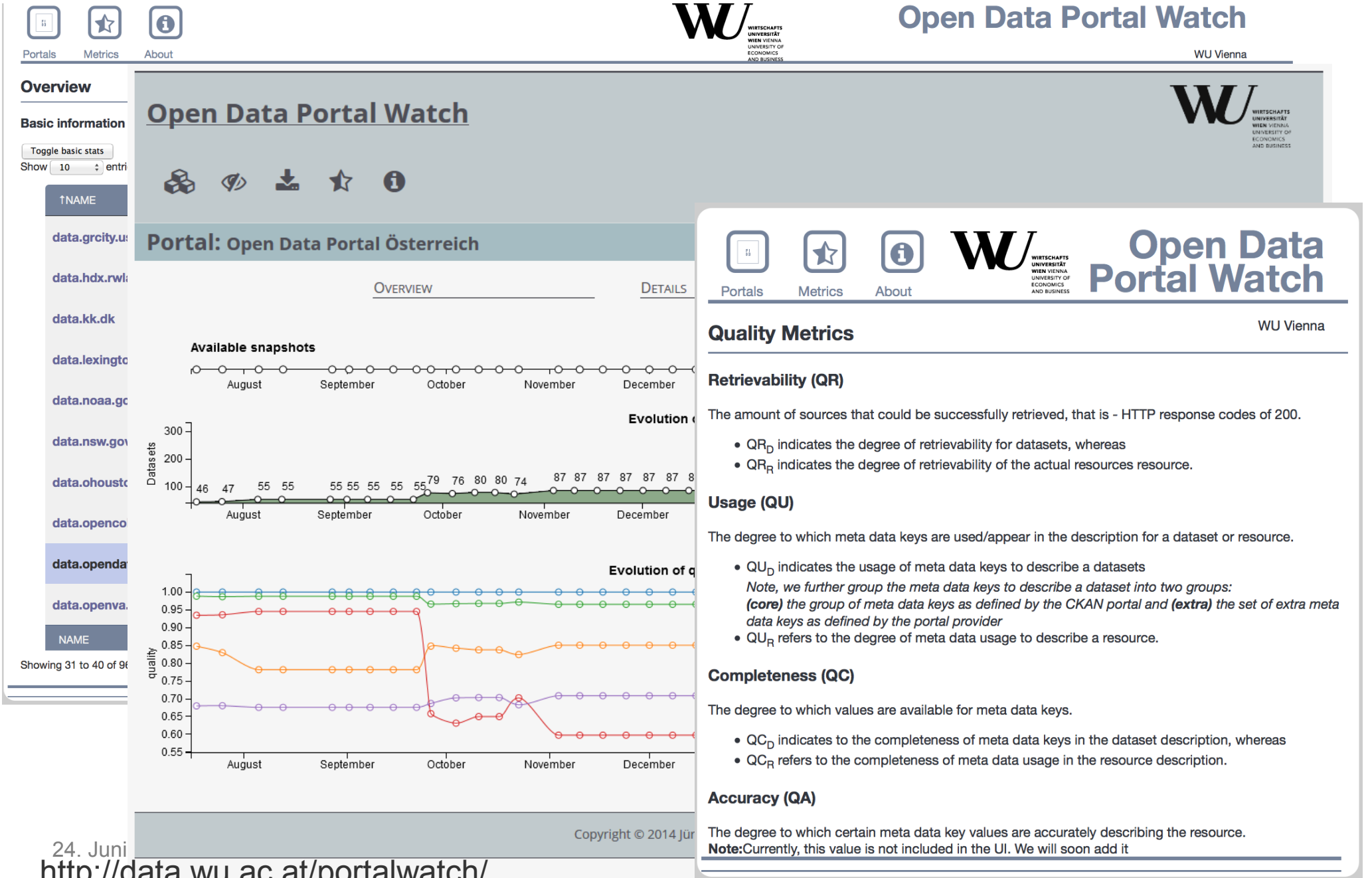
- fundamental characteristics of a location, represented as an address, a geographic name, or geometry
- aligned with the INSPIRE data specifications

- name, gender, date of birth

- fundamental characteristics of a service offered by public administration
- title, description, inputs, outputs, providers, locations

Werkzeuge (1)

Identifikation von Fehlern



http://data.wu.ac.at/portalwatch/

Werkzeuge (2)

Verbessern von Daten und Beheben von Fehlern

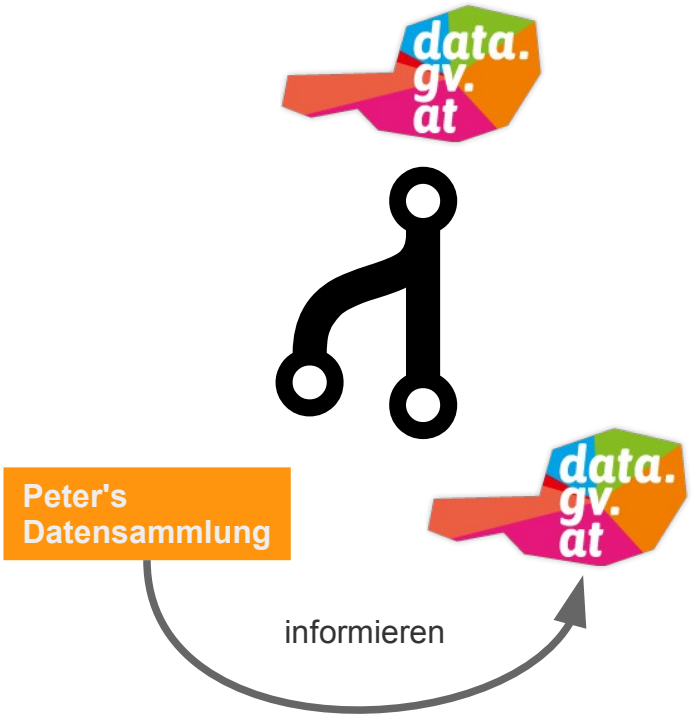
dat

Dat is a version-controlled, decentralized data sync tool designed to improve collaboration between data people and data systems.

Beispiel: Fehlerbehebung in Datensatz zu Parkscheinautomaten in Innsbruck
https://www.data.gv.at/katalog/dataset/gisibk-psa/resource/10fd6c5e-e11e-41b2-9f76-f8df8cf279a7

A	B	C	D	E	F	G	H
OBJECTID	Nummer	Standort	Kategorie	X	Y	Lon	Lat
1	417	Reichenauer	Parkstraße w	81982,802	237769,617	11,41649	47,2736744
2	436	Langer Weg :	Parkstraße w	82842,1698	237625,452	11,4278185	47,2722703
3	438	Langer Weg :	Parkstraße w	82753,3415	237404	11,4266039	47,2702899
4	440	Langer Weg :	Parkstraße w	82643,977	237063,663	11,4250961	47,267243
5	448	Reut-Nicolus	Parkstraße w	82423,1	237446,1	11,4222483	47,27071
6	449	Andechsstra	Parkstraße w	82289,584	237211,595	11,4204411	47,2686178
7	451	Andechsstra	Parkstraße w	82139,037	237372,448	11,4184816	4 2

A	B	C	D	E	F	G	
OBJECTID	Nummer	Standort	Kategorie	X	Y	Lon	La
1	417	Reichenauer	Parkstraße w	81982,802	237769,617	11,41649	47,2736744
2	436	Langer Weg :	Parkstraße w	82842,1698	237625,452	11,4278185	47,36
3	438	Langer Weg :	Parkstraße w	82753,3415	237404	11,4266039	
4	440	Langer Weg :	Parkstraße w	82643,977	237063,663	11,4250961	47,267243
5	448	Reut-Nicolus	Parkstraße w	82423,1	237446,1	11,4222483	47,27071
6	449	Andechsstra	Parkstraße w	82289,584	237211,595	11,4204411	47,2686178
7	451	Andechsstra	Parkstraße w	82139,037	237372,448	11,4184816	47,2700832



Beispiel Finanzdaten

Transparente Finanzen laut Stabilitätspakt 2012 ?

(Stand Juni 2015)

Artikel 12

Haushaltsbeschlüsse von Ländern und Gemeinden

(1) Die Haushaltsbeschlüsse der Länder und der Gemeinden sind in rechtlich verbindlicher Form zu fassen und öffentlich kundzumachen. Bund, Länder und Gemeinden haben ihren jeweiligen Rechnungsvoranschlag und Rechnungsabschluss inklusive aller Beilagen zeitnahe an die Beschlussfassung in einer Form im Internet zur Verfügung zu stellen, die eine weitere Verwendung ermöglicht (zB downloadbar, keine Images oder PDF).

Bund: formatierte Excel-Tabellen

Länder: 1 von 9 = 11%: Wien Offener Haushalt

Burgenland: lesbares PDF (VA 2005-2015, RA 2006-2013)

Kärnten: TXT 5-steller, Beilagen (RA 2010-2013) (PDF, XLSX, VA 2005-2015, RA 2004-2013)

Niederösterreich: lesbares PDF (VA 2002-2015, RA 2000-2014)

Oberösterreich: RA 2013 CSV (VA 2015, RA 2013) 1-Steller, keine Beilagen
lesbares PDF & HTML (VA 2001-2015, RA 2001-2013)

Salzburg: TXT 3-steller, Beilagen (RA 2013), eingescanntes PDF (VA 2010-2015) und HTML

Steiermark: lesbares PDF (VA 2003-2014, RA 2005-2013)

Tirol: XLS (VA 2013-2014), eingescanntes PDF (VA 2002-2014, RA2009-2014)

Vorarlberg: Word-Dokument mit Tabellen (VA 2003-2015, RA 2004-2014)

Gemeinden: fast 30 % der Gemeinden: CSV 3-steller, RA 2001-2013 teilw 2014, VA

Offener Haushalt

Katalog

Rechnungsabschluss des Landes Oberösterreich

Budgetdaten des Landes Oberösterreich

Eindeutiger Identifikator	ooe-E90E9617F65E46A3ACA9C3E002414D1A
Datum des Metadatensatzes	2015-04-27T08:39:16.00016
Kategorie	Finanzen und Rechnungswesen
Datenverantwortliche Stelle	Land Oberösterreich, Direktion Finanzen
Lizenz	Creative Commons Namensnennung 3.0 Österreich
Zeitliche Ausdehnung (Anfang)	2011-01-01T00:00:00.00000
Veröffentlichende Stelle	Land Oberösterreich

Zusätzliche Informationen ↑ ausblenden

Bezeichnung der Metadatenstruktur	OGD Austria Metadata 2.1
Sprache des Metadatensatzes	ger
Character Set Code des Metadatensatzes	utf8
Weiterführende Metadaten	None
Attributbeschreibung	nicht erfasst

Veröffentlichende Organisation bzw. Person

Land Oberösterreich

Kategorie

Finanzen und Rechnungswesen

Schlagworte

Ausgaben · Budget · Einnahmen · Haushalt · Landeshaushalt · Rechnungsabschluss · ordentliche Ausgaben · ordentliche Einnahmen

API - Link zu allen Metadaten

/api/3/action/package_show?id=ooe-E90E9617F65E46A3ACA9C3E002414D1A

	A	B	C	D	E
1	Ausgaben - Rechnungsabschluss 2013				
2					
3	HH		Gruppe		RA 2013
4	1	Ausgaben ord. Haushalt	0	Vertretungsk�rper und allgemeine Verwaltung	627.804.487,55
5			1	�ffentliche Ordnung und Sicherheit	21.325.536,47
6			2	Unterricht, Erziehung, Sport und Wissenschaft	1.436.828.543,33
7			3	Kunst, Kultur und Kultus	185.080.932,75
8			4	Soziale Wohlfahrt und Wohnbauf�rderung	862.852.510,37
9			5	Gesundheit	742.979.658,74
10			6	Stra�en- und Wasserbau, Verkehr	401.405.064,67
11			7	Wirtschaftsf�rderung	223.526.479,02
12			8	Dienstleistungen	22.147.580,55
13			9	Finanzwirtschaft	1.041.220.606,61
14			Ergebnis		5.565.171.400,06

Ausgaben Budget Einnahmen Haushalt Landeshaushalt Rechnungsabschluss ordentliche Ausgaben ordentliche Einnahmen

Zusätzliche Informationen ↑ ausblenden

Bezeichnung der Metadatenstruktur	OGD Austria Metadata 2.1
Sprache des Metadatensatzes	ger
Character Set Code des Metadatensatzes	utf8
Weiterführende Metadaten	None
Attributbeschreibung	nicht erfasst

API - Link zu allen Metadaten

/api/3/action
/package_show?id=ooe-
F90F9617F65F46A3ACA9C3F002414D1A

Inhaltliche Datenqualität

- Es geht nicht um Metadaten, um Beschreibungen, Auffindbarkeit oder Datenformate
- Es geht um die Fragen
  - Sind die Daten genau genug (Exaktheit)
  - Sind die Daten genügend „breit“ (Anzahl der Spalten?)
  - Sind die Daten genügend „tief“ (Anzahl der Zeilen?)
  - Sind die Daten vollständig
  - Sind die Daten aktuell
- Das ist von der jeweiligen Wissensdomäne abhängig
- Aber nicht unbedingt vom Verwendungszweck!

1. ***Vollständigkeit***: Von der Verwaltung veröffentlichte Datensätze sind **so vollständig wie möglich**, **sie bilden den ganzen Umfang dessen ab, was zu einem bestimmten Thema dokumentiert ist**.

2. ***Primärquelle***: Die Daten werden von der Verwaltung an ihrem Ursprung gesammelt und veröffentlicht. Dies geschieht **mit dem höchstmöglichen Feinheitsgrad, nicht in aggregierten oder sonst wie modifizierten Formaten**.

3. ***Zeitnahe Zurverfügungstellung***: Von der Verwaltung veröffentlichten Datensätze stehen der Öffentlichkeit innerhalb eines angemessenen Zeitraums möglichst **aktuell zur Verfügung. Sie werden veröffentlicht, sobald sie erhoben und zusammengestellt wurden.**

Dabei wäre es so einfach...

- Private Haushalte
- Unternehmen
- Land- und Forstwirtschaft
- Öffentliche Einrichtungen
- Bildungseinrichtungen
- Gesundheitseinrichtungen
- Gebarung öffentlicher Sektor
  - Meldung neuer Einheiten
  - Meldung kontrollierter Einheiten
  - Erhebung Länder
  - Erhebung Gemeinden
  - Erhebung Gemeindeverbände
  - Erhebung staatlicher Einheiten
  - Erhebung Haftungen
- Registerzählung
- Abgestimmte Erwerbsstatistik
- Erhebungen A-Z

Länder

Gemäß Gebarungsstatistikverordnung, BGBl. II Nr. 345/2013 sind die Länder verpflichtet, Daten zum Rechnungsabschluss (Jahres- , Quartals- und Monatsdaten) zu melden.

Jahresdaten

Die Länder haben jährlich bis zum 31. Mai die Gebarungsdaten entsprechend der Datenschnittstelle LHD an Statistik Austria zu übermitteln.

Quartalsdaten

Daten zum Quartal sind jeweils bis zum 25. des Folgemonats, betreffend das 4. Quartal innerhalb von 5 Wochen nach Quartalsende an Statistik Austria entsprechend der Datenschnittstelle LHD (SA01 bis SA04) zu liefern.

Monatsdaten

Monatsdaten sind jeweils bis zum 25. des Folgemonats an Statistik Austria entsprechend der Datenschnittstelle LHD (SA01 und SA02) zu liefern.

Kontrolltabelle

Die Kontrolltabelle dient der inhaltlichen Prüfung der gemeldeten Jahresdaten und der Erfassung der PPP-Modelle (Jahresmeldung).

Die Datenschnittstelle LHD-V37 ist erstmals anzuwenden für:

- **Rechnungsabschluss 2014** Lieferung bis 31. Mai 2015
- **1. Quartal 2015** Lieferung bis 25. April 2015
- **Jänner 2015** Lieferung bis Februar 2015

Satzaufbau für die Lieferung der Haushaltsdaten der Länder, Version LHD-V3.7 - **Stand 29. August 2014**	(170 KB)
Handbuch LHD, Erläuterungen zur Datenschnittstelle **Stand 29. August 2014**	(360 KB)
Kontrolltabelle	(30 KB)

Inhaltliche Datenqualität

- Je nach Domäne unterschiedlich
  - Finanzdaten
  - Statistikdaten
  - Umweltmessdaten
  - GIS-Daten…
- Bei Finanzdaten
  - Es sind standardisierte, maschinenlesbare Finanzdaten aller Gemeinden und Länder schon seit Jahren verfügbar. Die Gemeinden können diese auf www.offenerhaushalt.at hochladen, die Länder könnten sie einfach veröffentlichen - siehe Kärnten.
- Qualitätssiegel, Mindeststandards?

Johann Höchtl
Center for E-Governance
Johann.hoechtl@donau-uni.ac.at

@myprivate42

at.linkedin.com/in/johannhoechtl

Bernhard Krabina
KDZ – Zentrum für Verwaltungsforschung
krabina@kdz.or.at

@krabina

at.linkedin.com/in/krabina

24.06.2015