Open Data zwischen Anspruch und Wirklichkeit

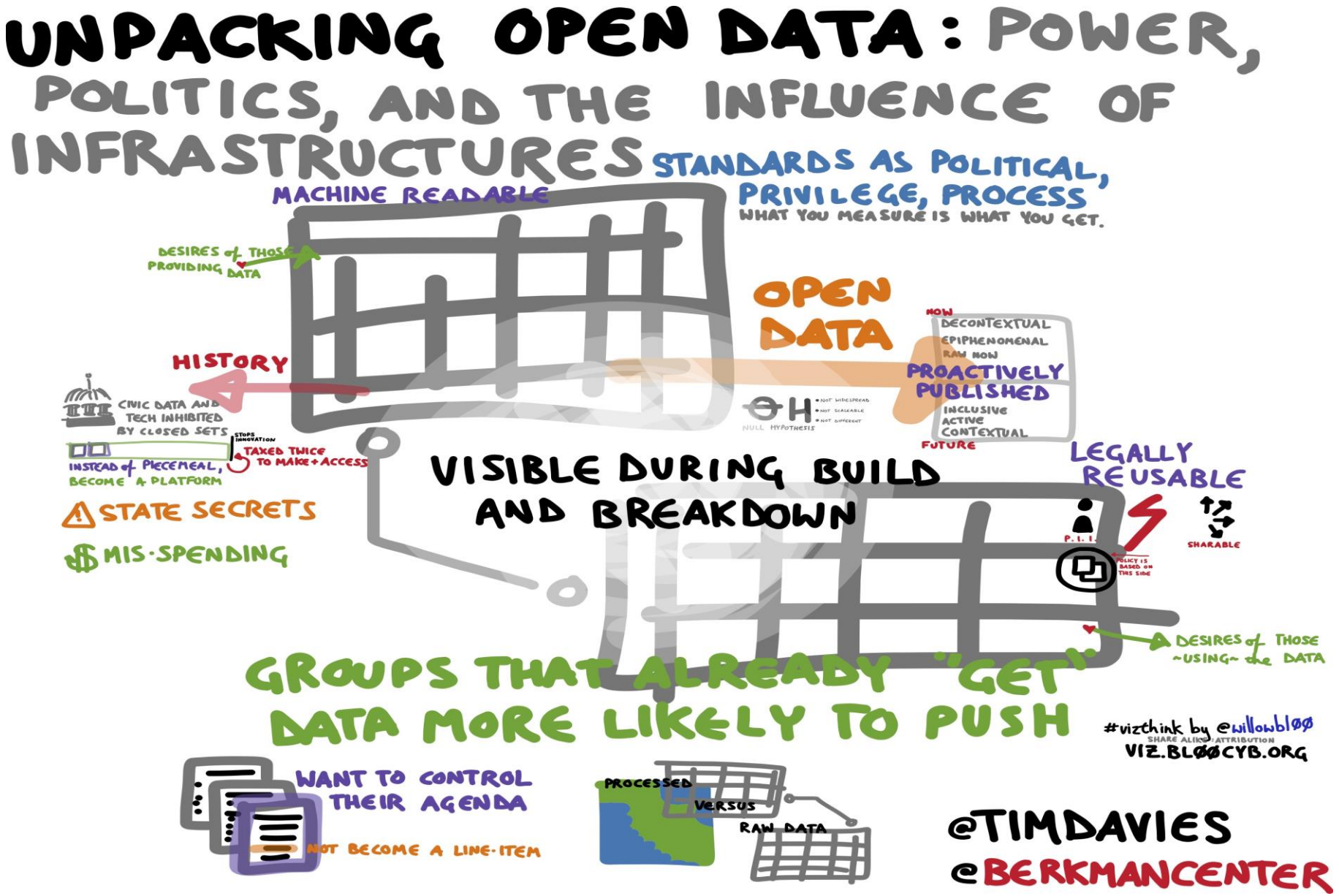
OpenDataMonitor

September 9-10, 2015
München

Collaborative Project (small or medium-scale focused research project)
FP7-ICT-2013.4.3 SME initiative on analytics

Project number: 611988

SYNYO · ODI · ΙΝΣΥ IMIS · University of Southampton · ifG.CC · City of Munich · red.es

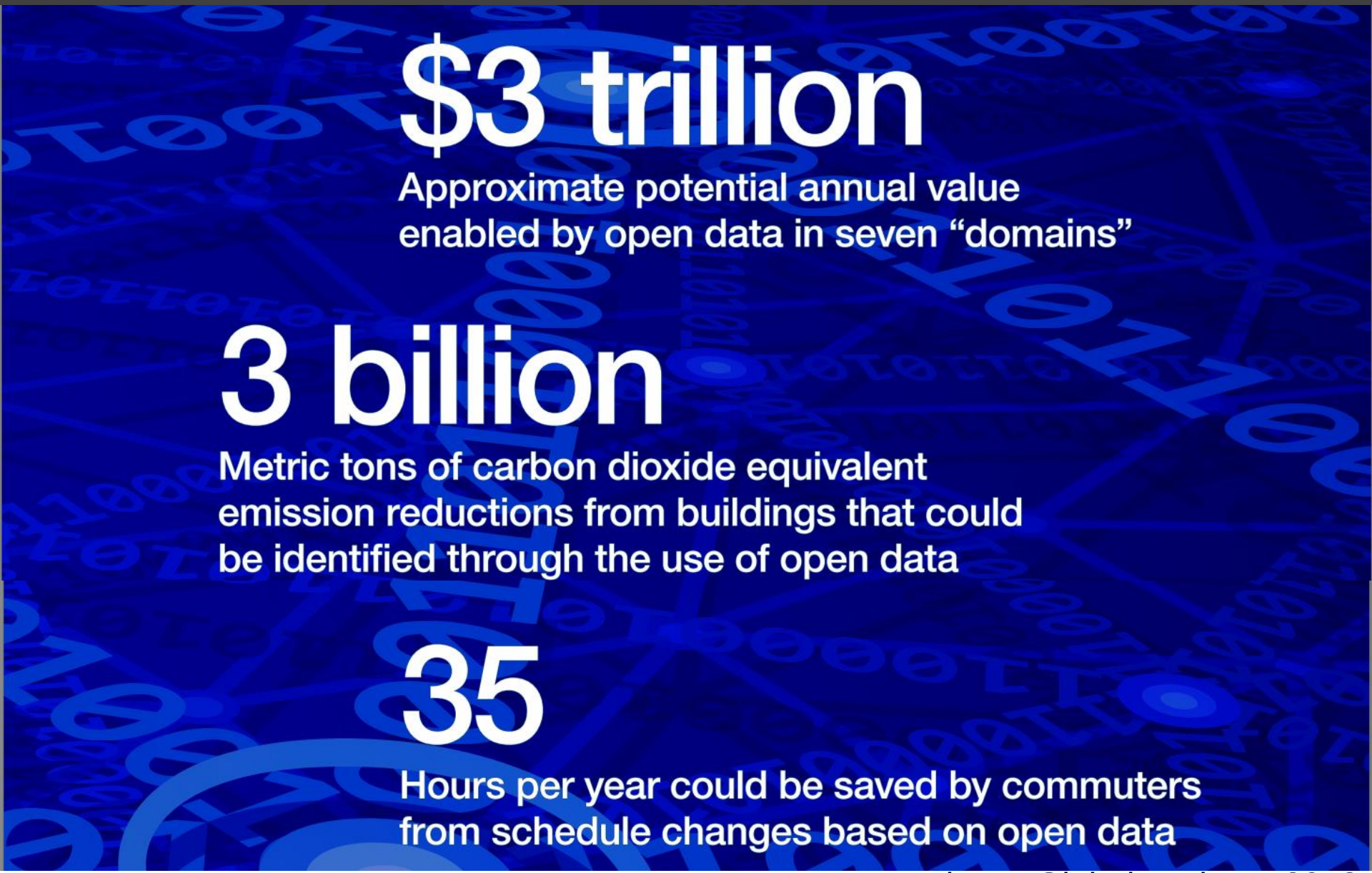

UNPACKING OPEN DATA: POWER, POLITICS, AND THE INFLUENCE OF INFRASTRUCTURES
STANDARDS AS POLITICAL, PRIVILEGE, PROCESS
WHAT YOU MEASURE IS WHAT YOU GET.
MACHINE READABLE
DESIRES of THOSE PROVIDING DATA
OPEN DATA
HISTORY
NOW
DECONTEXTUAL
EPIPHENOMENAL
RAW NOW
PROACTIVELY PUBLISHED
INCLUSIVE
ACTIVE
CONTEXTUAL
FUTURE
NULL HYPOTHESIS
CIVIC DATA AND TECH INHIBITED BY CLOSED SETS
STOPS INNOVATION
TAXED THICE TO MAKE+ACCESS
INSTEAD of PIECEMEAL, BECOME A PLATFORM
STATE SECRETS
MIS-SPENDING
VISIBLE DURING BUILD AND BREAKDOWN
LEGALLY REUSABLE
P. I. I.
POLICY IS BASED ON THIS SIDE
SHARABLE
DESIRES of THOSE ~USING~ the DATA
GROUPS THAT ALREADY "GET" DATA MORE LIKELY TO PUSH
#vizthink by @willowbloø
SHARE ALIKE ATTRIBUTION
VIZ.BLOØCYB.ORG
WANT TO CONTROL THEIR AGENDA
NOT BECOME A LINE-ITEM
PROCESSED
VERSUS
RAW DATA
@TIMDAVIES
@BERKMANCENTER

Great Expectations

Alle Bürger können die Ressourcen der Verwaltung frei nutzen

- für eine engagiertere und informiertere Bürger-/Zivilgesellschaft
- für eine transparentere und verantwortlichere Verwaltung
- für Innovation und dynamisches Wirtschaftswachstum

Große Erwartungen, hehre Ziele, fantastische Prognosen



McKinsey Global Institute, 2013

Annahmen und Voraussetzungen für Open Data

- **die Verwaltung besitzt Unmengen von Daten**
  → können vermeintlich “auf Knopfdruck” veröffentlicht werden
- **die Daten liegen in der Verwaltung qualitätsgesichert vor**
  → vermeintlich strukturiert, standardisiert und beschrieben
- **die Datenerhebung wurde mit Steuergeldern bezahlt**
  → Verwaltung darf für die Bereitstellung kein Geld verlangen
- **alle Daten sind nutzbar und nützlich**
  → ALLE Daten sollten verfügbar gemacht werden
- **Daten können genutzt werden, wenn sie maschinen-lesbar, in offenen Formaten, unter einer offenen Lizenz veröffentlicht werden**
  → Horden von Programmierern warten auf Daten, sind bereit und fähig Daten zu nutzen

Erinnern Sie sich noch an dieses “E-Government”?

“With the help of information technology, e-government can customize services based on personal preferences and needs.” (Ho, A.T-K., PMR, 2002)

“Online technology has the potential to break down traditional barriers faced by clients.” (Department of Communications Information Technology and the Arts, 2000, p. 5)

- Transformation der Verwaltung durch/mit E-Government
- bürgerzentrierte, 24/7-One-stop-shops mit Back-Office-Daten-, Informations- und Prozessintegration
- eine effizientere, effektivere und transparentere Verwaltung

→ Technophile Mythen, die weitestgehend nicht Realität geworden sind

(Bekkers/Homburg 2007)

Voraussetzungen und Einflussfaktoren

Angebotsseitige Faktoren

- Technik: technische Readiness
- Organisation: Data Governance und Prozesse für effiziente Datenbereitstellung
- Politik: politischer Wille und Unterstützung und Budget
- Recht: Rechtsrahmen, der Weiterverwendung begünstigt; klares Lizenzregime

Nachfrageseitige Faktoren

- skalierbare, großflächige Nutzbarkeit von Daten
- umfangreiche, dynamische, detailreiche, verknüpfte Daten(sätze)
- versierte Nutzer-Community
- inhaltlich spezifische Daten von besonderem Interesse
- uvam…

Open Data zwischen Anspruch und Wirklichkeit

Angebotsseitige Realität

- Legacy-Systeme, die nie für die Freigabe von Daten gebaut wurden
- verteilte Daten in Silo-Strukturen
- Einnahmen aus dem Datenverkauf als Unabhängigkeitsquelle
- Open Data als Zusatzaufgabe, ohne unmittelbaren Nutzen für die eigene Tätigkeit
- “availability-approach” bei der Datenwahl
- Pfadabhängigkeit durch Staats-tradition und Verwaltungskultur

Nachfrageseitige Realität

- Durchschnittsbürger in keinster Weise in der Lage, einen Datensatz zu nutzen – technisch wie inhaltlich
- relativ kleine (aber laute) Gruppe von Aktivisten
- vielfach “bessere” selbsterzeugte Daten der Wirtschaft, weil personenbezogen (Stichwort: Payback, Amazon, Facebook und Google)

Inhärenter Konflikt von Open Data

Freedom of Information/ Informationsfreiheit	↯	Public Sector Information
Transparenz	↯	Innovation und Wirtschaftswachstum
putting government under scrutiny and unearthing troves of politically sensitive data	↯	kommerzielle Weiterverwendung

Inwieweit amalgamieren beide Strömungen in Open Data?

FOI/IFG vs. PSI

[We are] not considering the role of open data in transparency, but the potentiality of open data in developing of new business.[...] Now I think there is a bit of confusion between transparency and open data. This is a serious problem.” (ICT strategy unit, national level Spain)

“Basic public services: education, health, security – these kinds of data are more sensitive, because they have a political component and are really important. The other is the weather: The politicians cannot change the weather.” (open data strategy unit, Spain)

Relevant data with high potential

	Activists		Users		Public Administration	
Data category	frequency	rank	frequency	rank	frequency	rank
Politics and Elections	10.90%	1	5.78%	7	3.60%	12
Public Sector	10.26%	3	7.14%	2	6.00%	6
Finance and Budget	9.62%	4	5.78%	6	8.00%	4
Law and Justice	6.41%	5	5.10%	10	2.60%	15
Environment and Climate	3.85%	10	6.80%	5	9.60%	1
Geography and Geology	4.49%	8	4.76%	12	9.40%	2
Transportation and Traffic	5.13%	7	4.76%	11	8.40%	3
Science and Technology	3.21%	13	7.14%	4	5.60%	7

Hohe Unterschiedlichkeit zwischen Stakeholder-Gruppen:

- erkennbare Entkopplung zwischen Transparenz (activists) und Wirtschaftswachstum (users)
- Relativ eindeutige Präferenz von Verwaltung

Availability-Approach

“We have now published many data that we had in a structured way. So we didn’t have to clean so much and didn't have to prepare many data because it was data we had previously.” (open data officer of a national ministry, Spain)

"In general, it is the availability of potential data. This is still a point that data are selected based on how easy it can be made available and less based on its usefulness" (municipal level, Germany)

“Those were simply data sets that the statistical office published anyways.” (then ICT strategy unit, municipal level, Germany)

“The data we already have, what is available, is usually at the bottom of the list and not at the top.” (open data researcher, user and consultant, Spain)

Open Data zwischen Anspruch und Wirklichkeit

Datenqualität

“Data quality is also an important issue, because many times when [...] they do not want to share them, they do not want to share the data in the quality it actually has. […] They know that is incorrect […] so they cover it.” (open data researcher and consultant, Spain)

“Then we saw that some data are so old, they are not relevant for open street maps anyway.” (formerly ICT strategy unit, municipal level, Germany)

“you can have an open data catalog with thousands of data sets, it looks pretty good, lots of data are available, but if you start to work with one single data set you see that you just have rubbish.” (open data researcher and consultant, Spain)

Open Data zwischen Anspruch und Wirklichkeit

Wie sieht vor diesem Hintergrund die Wirklichkeit offener Daten in Europa aus?

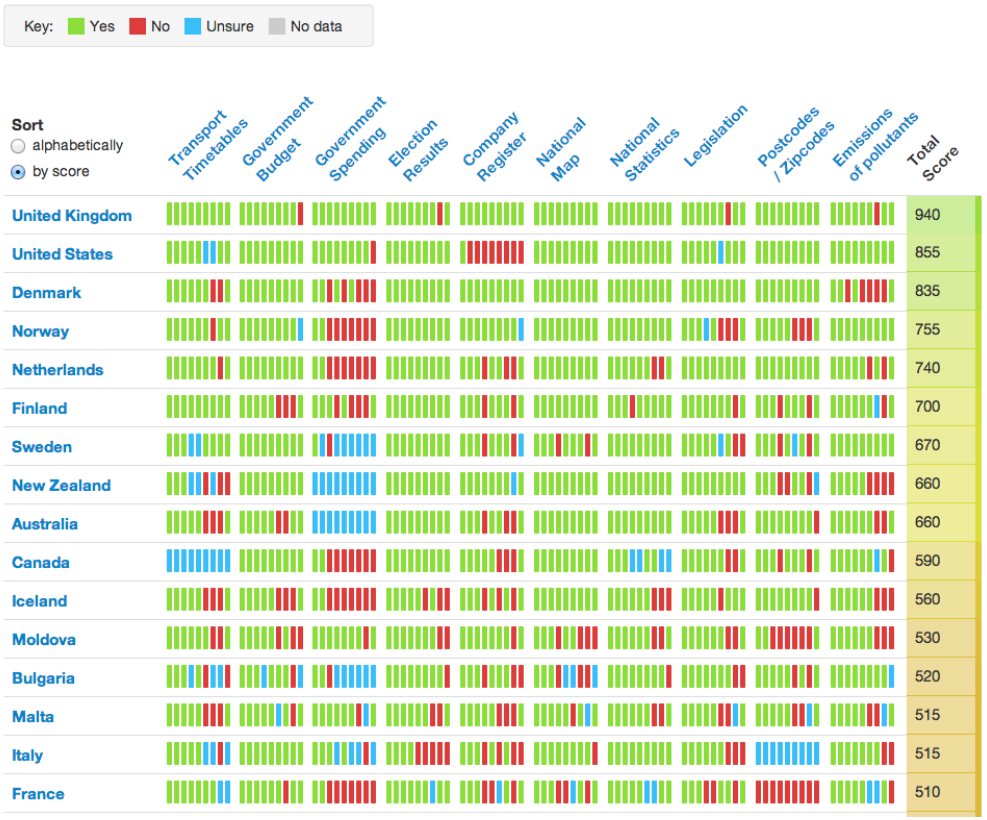
- Vielzahl von öffentlichen Organisationen veröffentlicht bereits open data oder beginnen gerade damit, jedoch **auf unterschiedliche Art und Weise**
- Open data wird fragmentiert veröffentlicht, **lokal, regional, national oder pan-Europäisch**
- **Metadaten sind oft uneinheitlich, unvollständig oder nicht akkurat**
- Situation von **open data erscheint fragmentiert und schwer zu überschauen**, was zahlreiche Fragen aufwirft:
  - Welche Kataloge und Daten sind verfügbar?
  - Gibt es (inter)nationale Trends?
  - Wo finden sich besonders vielversprechende open data Ressourcen?

Bisherige Monitoring-Ansätze

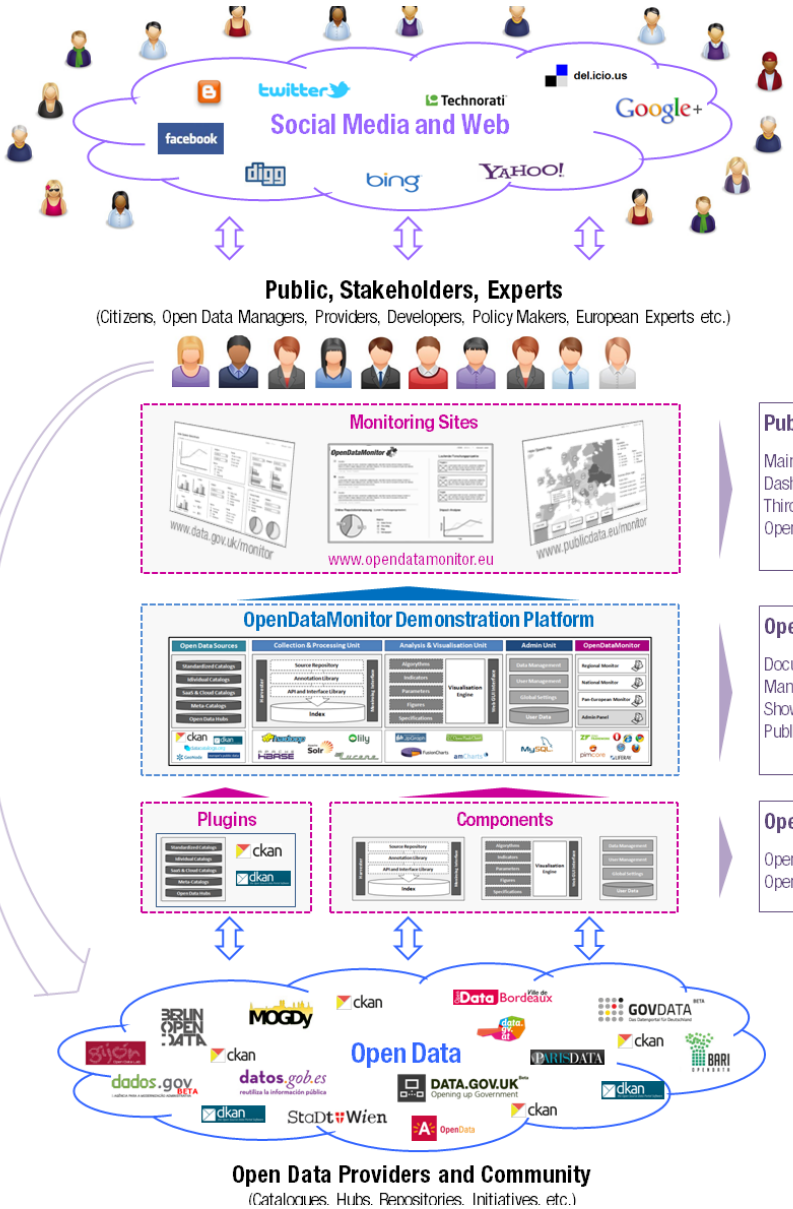
- **Open Data Barometer**
  - Methode(n): Befragung und Experteneinschätzung
  - Weltweite Datensammlung

- **Open Data Index**
  - Methode: Experteneinschätzung

- **Open Data Portal Watch**
  - Methode: automatisiertes CKAN harvesting



OpenDataMonitor Konzept



Public Access

Main Site
Dashboards and Key figures
Third party integration
Open Interfaces

Open Knowledge

Documentation
Manuals & Tutorials
Showcase
Publications

Open Source

Open Source Repositories
Open Licenses

Stakeholders can monitor and analyse open data
(using the Demonstration Site)

Presentation of the capabilities of ODM on the demonstration site

Overall architecture including data harvesting and harmonisation logic

Plugins and components that can be shared with the community

Supporting the open data community
(using particular ODM outcomes)

Aktueller Stand

238 data catalogues collected

28 countries

133 harmonised data catalogues

(from **24** countries)

3 platforms [CKAN, Socrata, HTML]

432,487 Total Distributions

915GB Total Size Distribution

2,630 Unique publishers

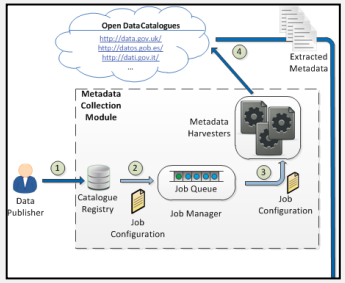
Architektur-Übersicht

Core Architecture and Processing Units

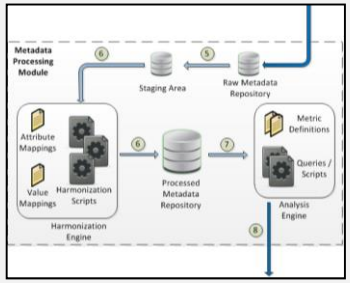
Main Architecture

Metadata Collection and Processing

Collection module



Processing module



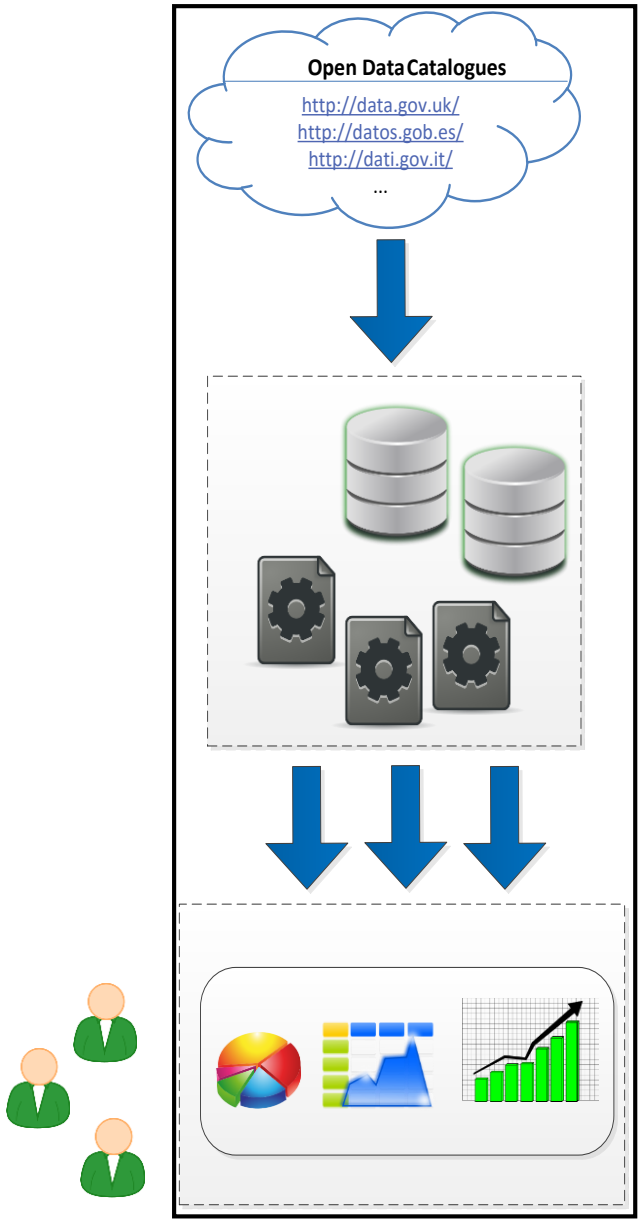
Administration panel

Future Implementations and Outlook

ODM Prototype Overview

Zwei Sub-Systeme:

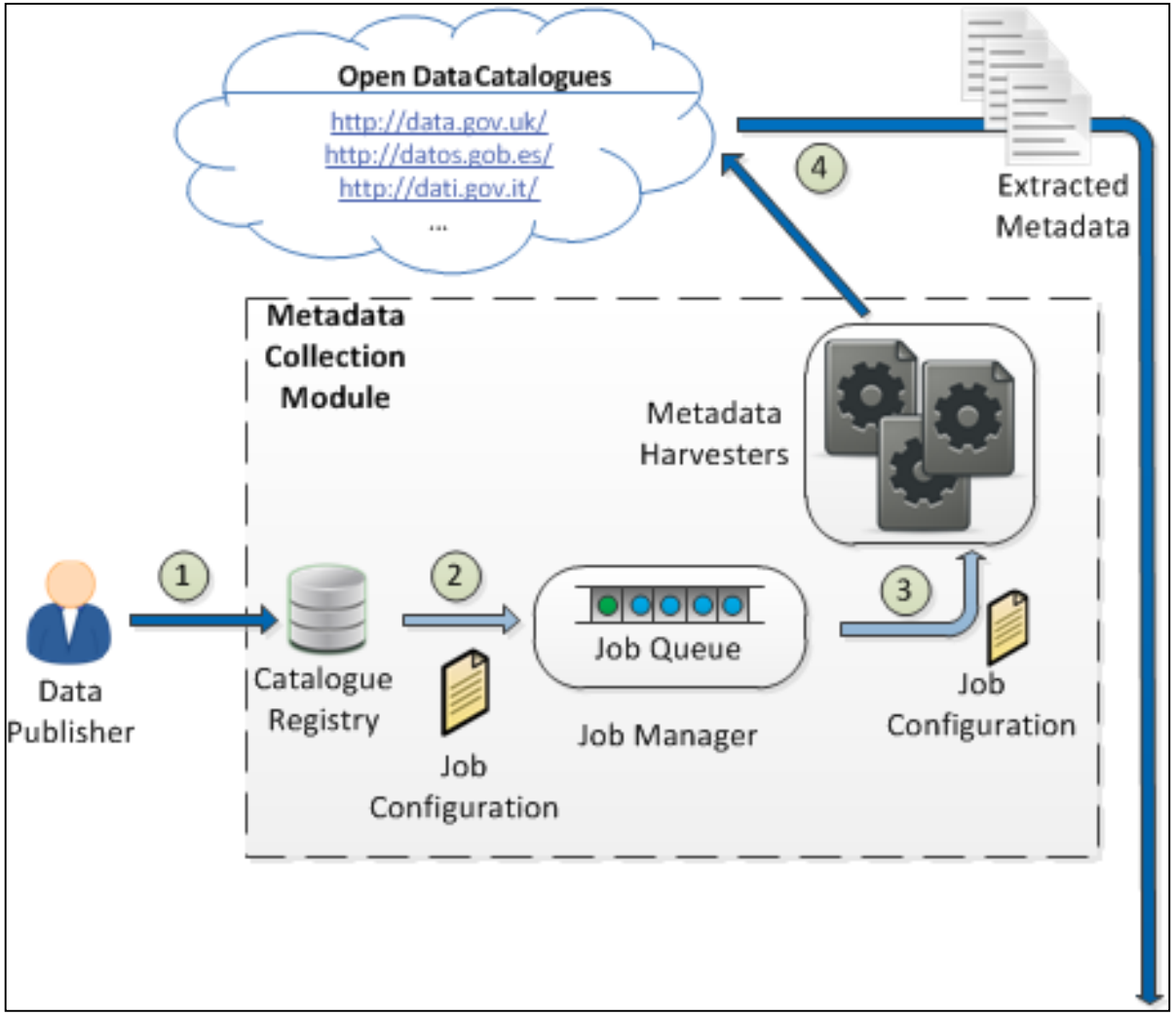
- **Metadata collection and processing:**
  - sammelt Metadaten von open data-Katalogen
  - führt Metadaten-Bereinigung und Harmonisierung aus
  - berechnet Metriken und bietet Ergebnisse via API an
- **Demonstration site:**
  - visualisiert Monitoring-Ergebnisse
  - ermöglicht dem Nutzer Benchmarking von Ländern und Katalogen sowie Berichte zu erzeugen



Architecture Overview (I)

Metadata collection module

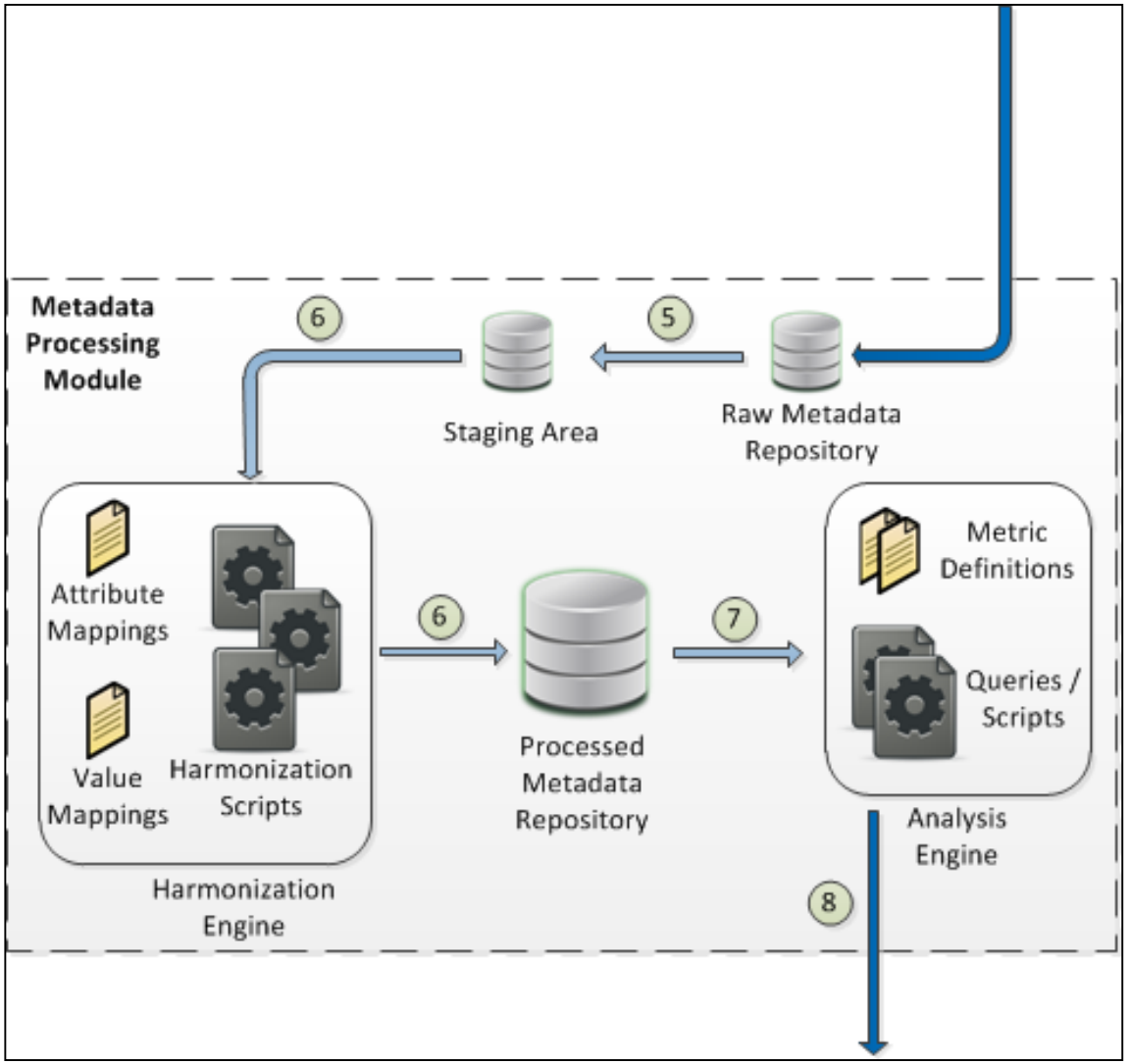
- Kataloge für das Monitoring registrieren
- periodisch Metadaten extrahieren



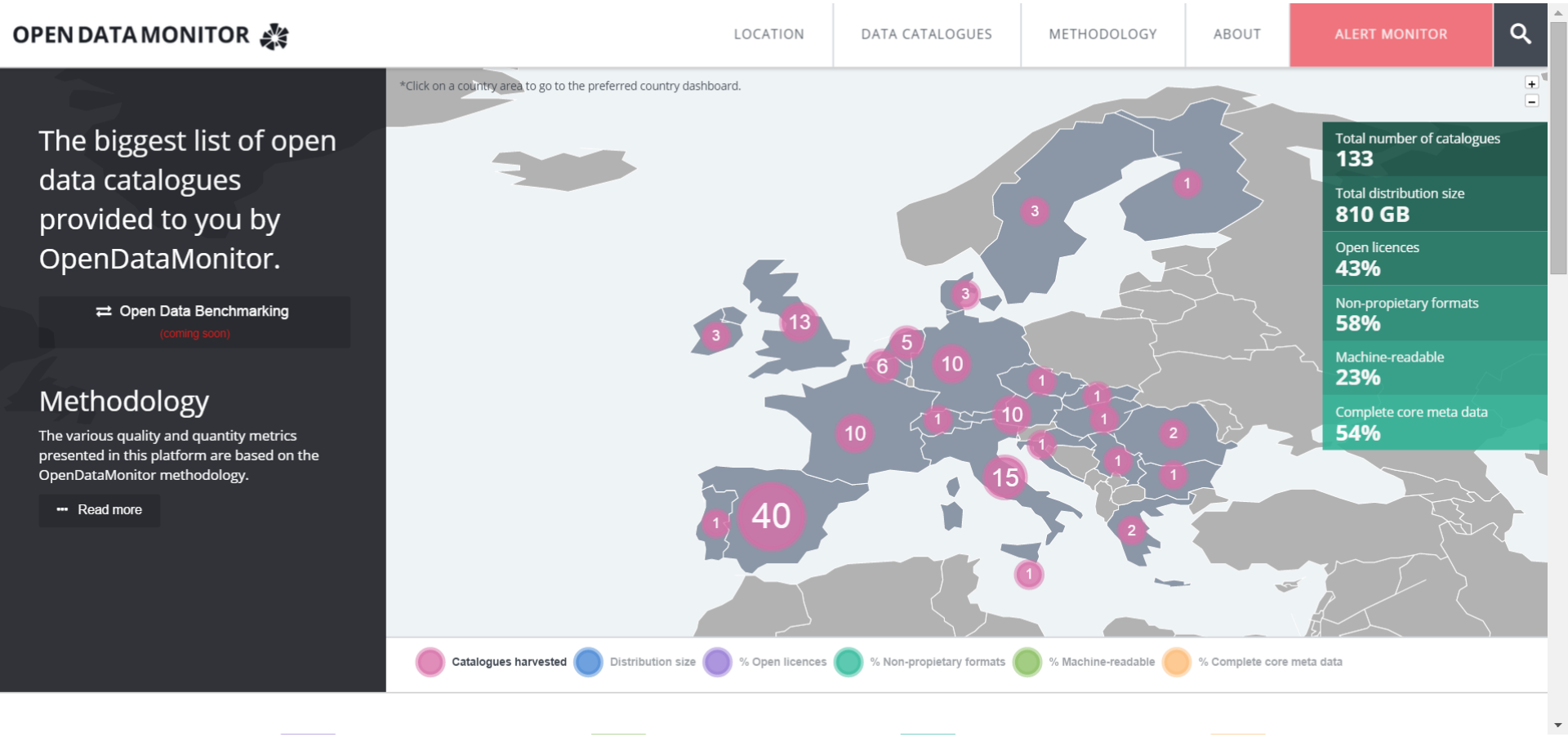
Architecture Overview (II)

Metadata processing module

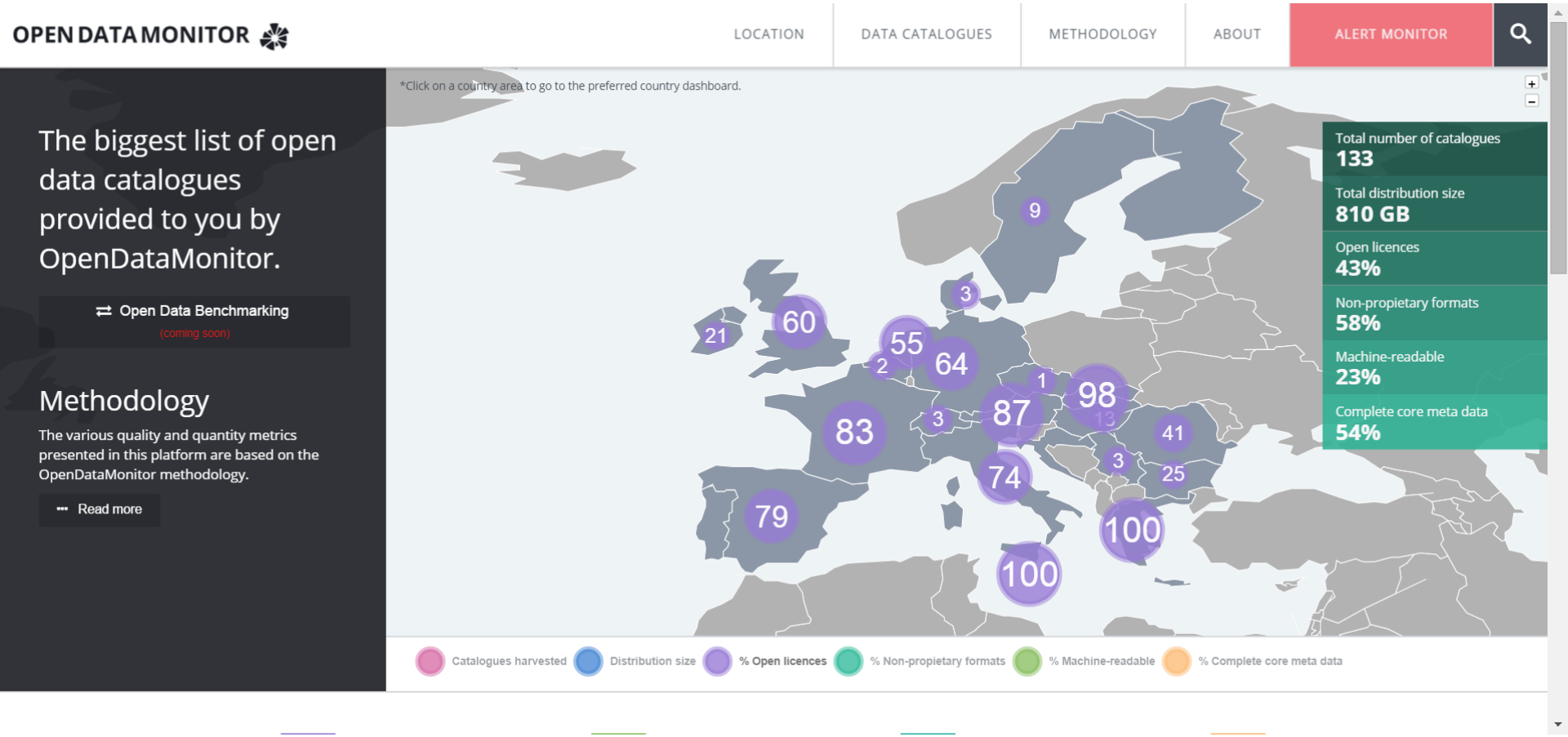
- Attribute und Werte harmonisieren
- Duplikate identifizieren
- Metriken für das Monitoring berechnen
- Ergebnisse via API verfügbar machen



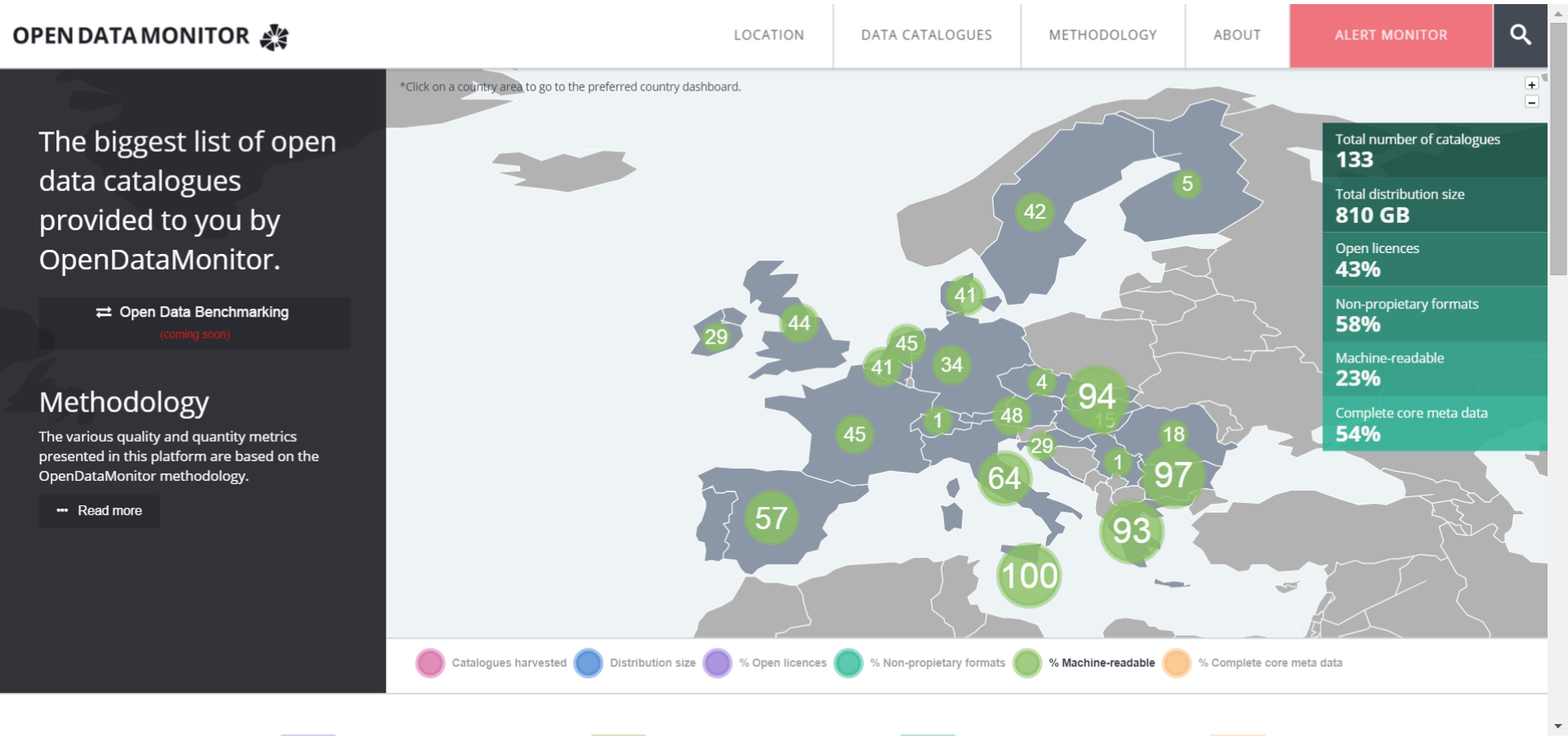
European Dashboard



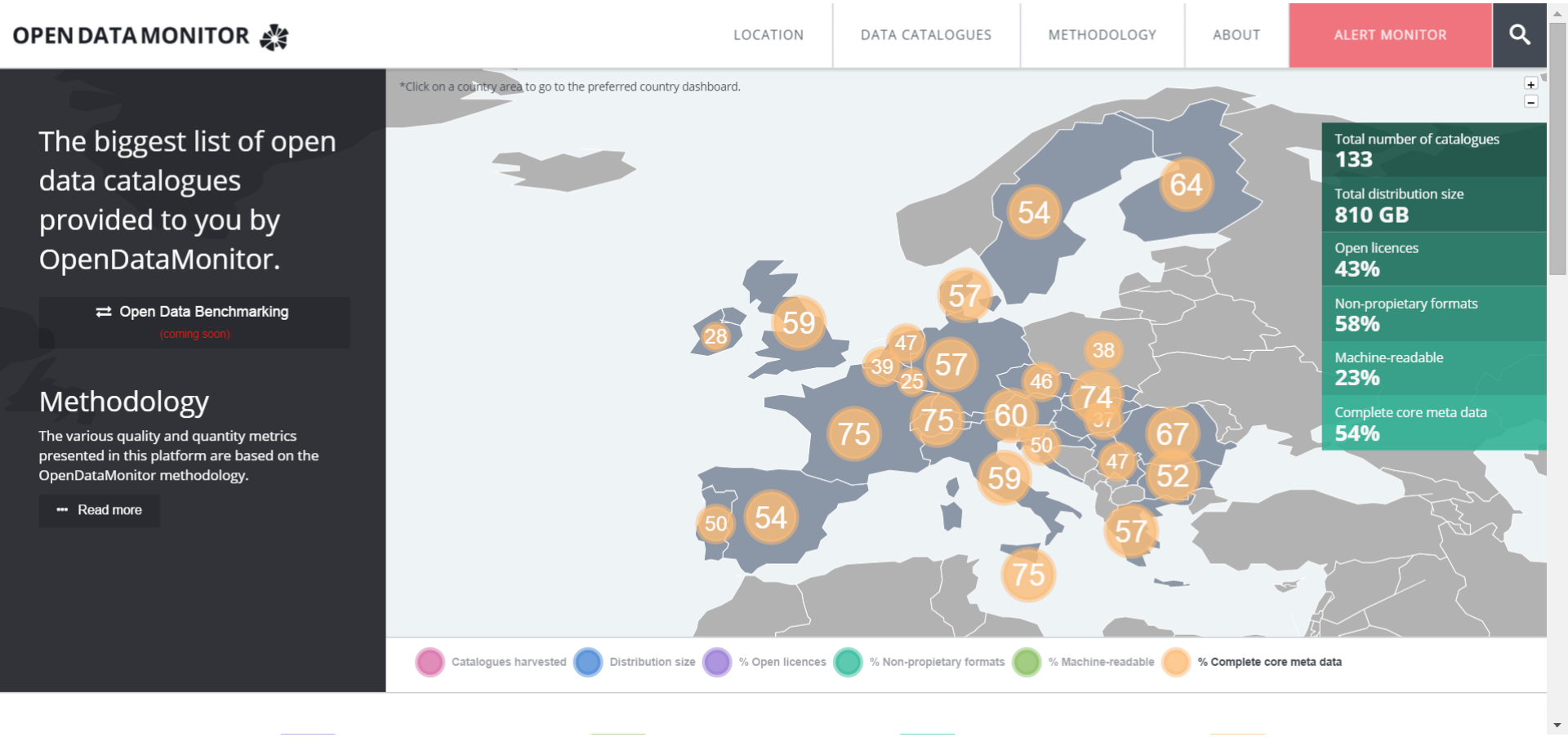
European Dashboard



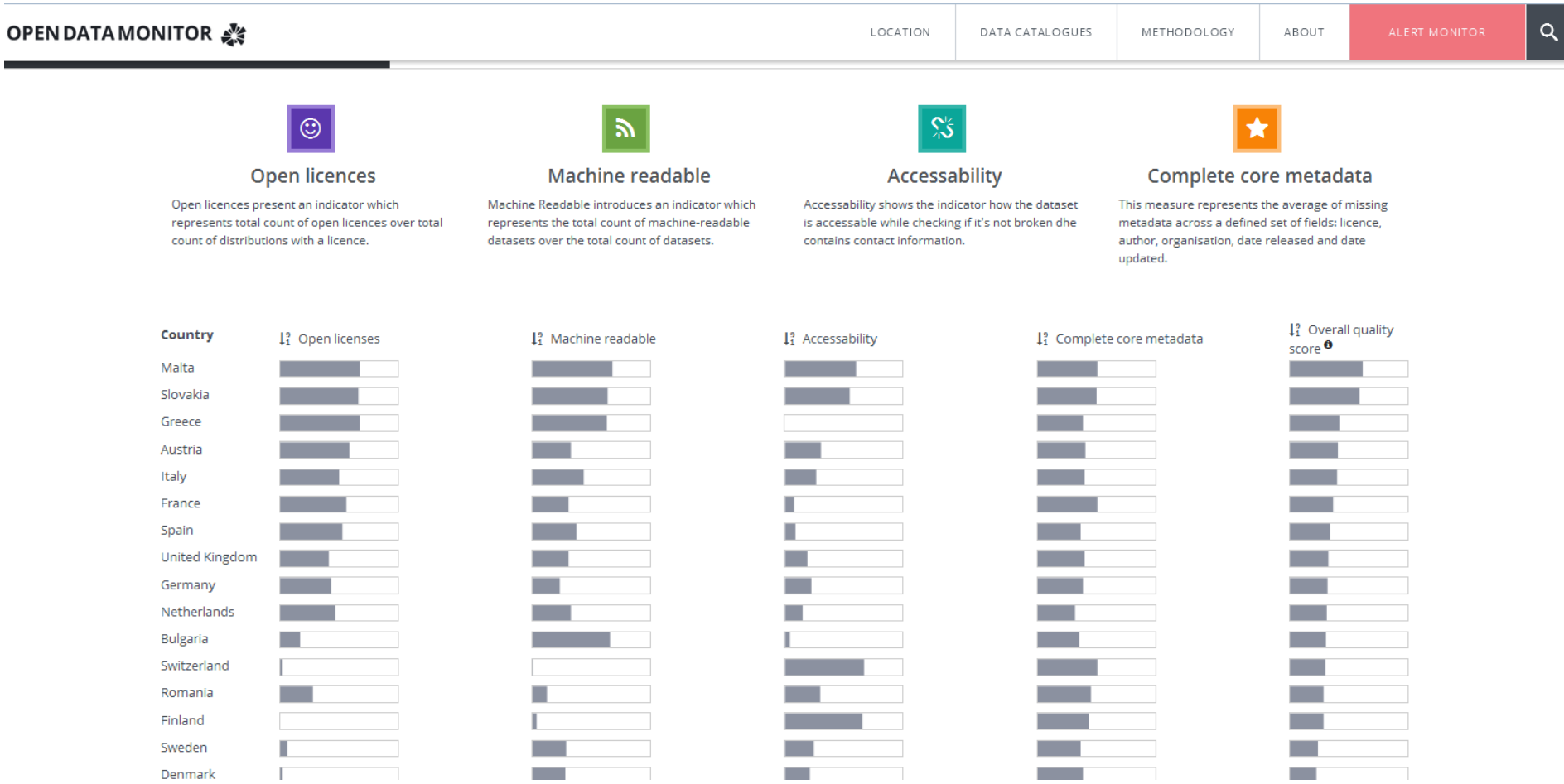
European Dashboard



European Dashboard



European Dashboard



Country Dashboard: Deutschland

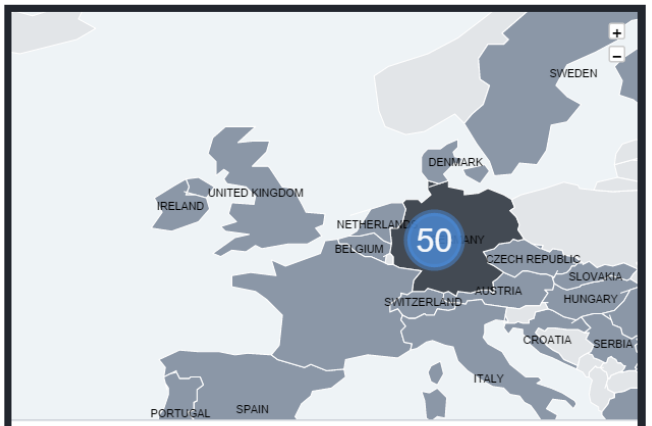
OPEN DATA MONITOR | LOCATION | DATA CATALOGUES | METHODOLOGY | ABOUT | ALERT MONITOR

YOU ARE HERE > Home > Germany

Germany

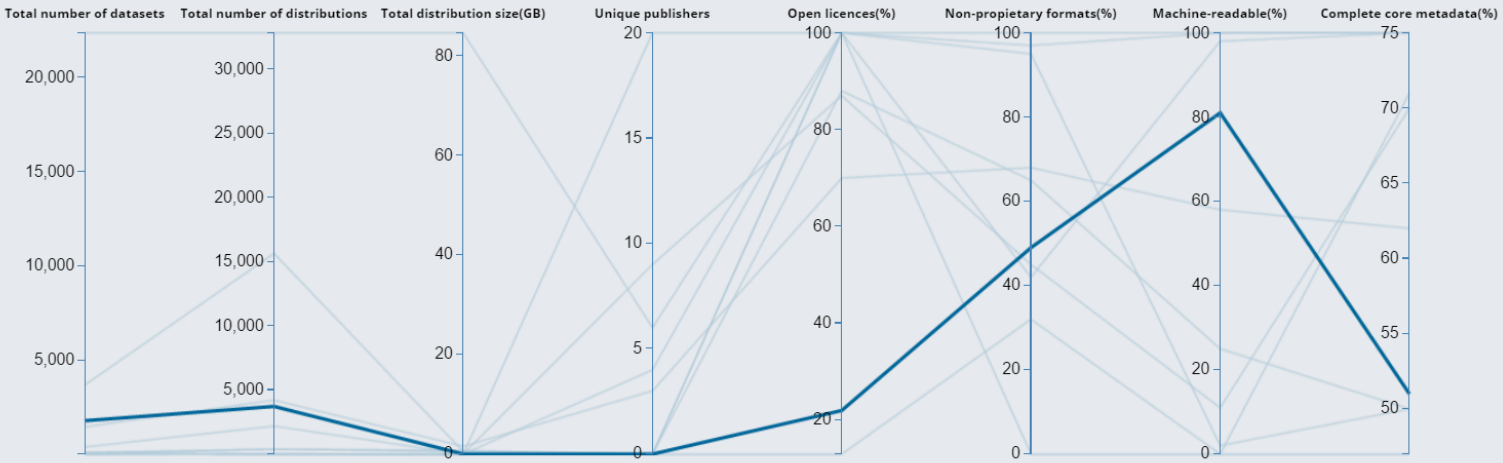
Overall quality score is 50/100 measured in 2015.

Open licences	Machine readable	Accessability	Complete core metadata
78	47	14	62

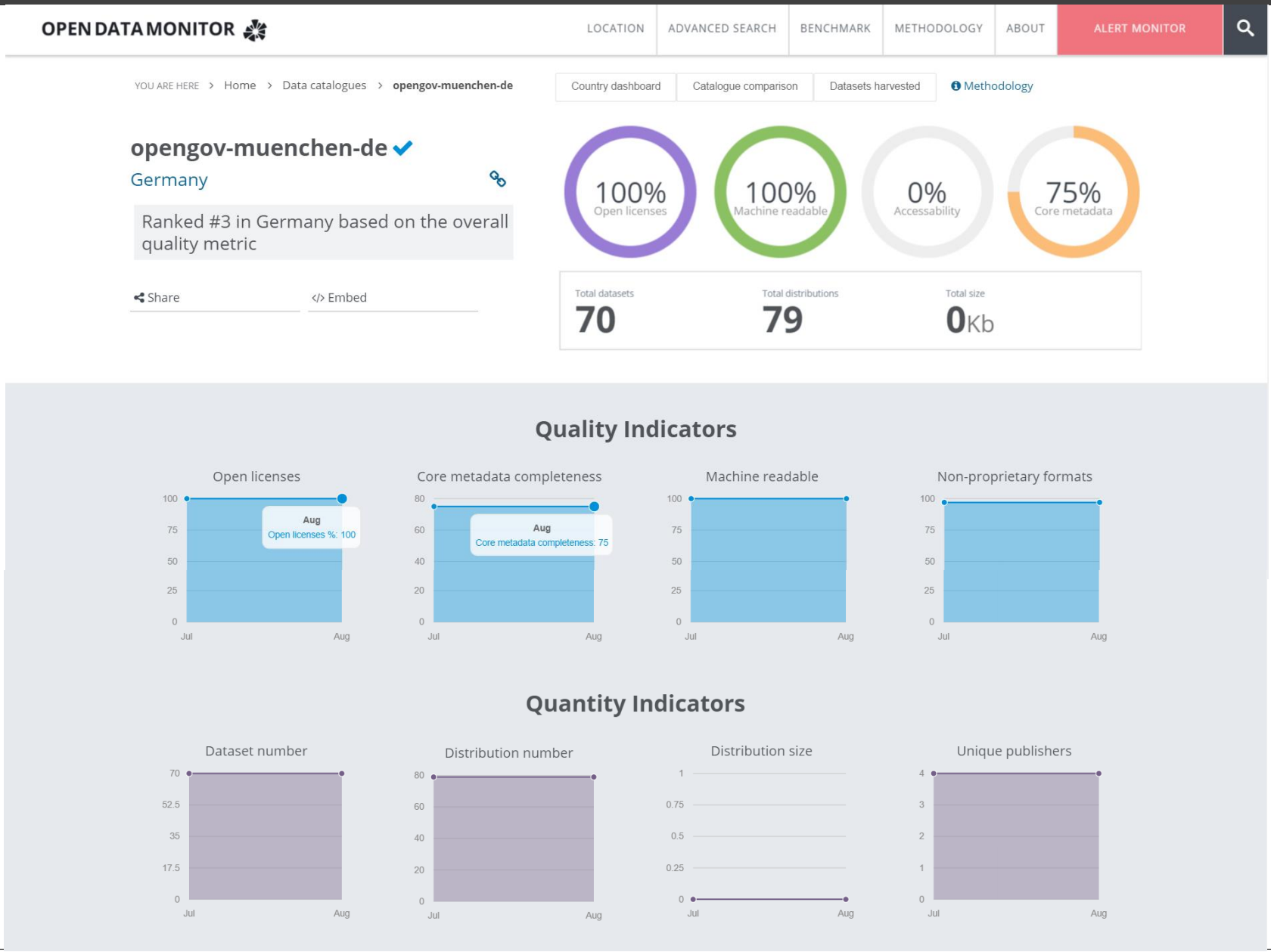


Find more about Barometer clusters

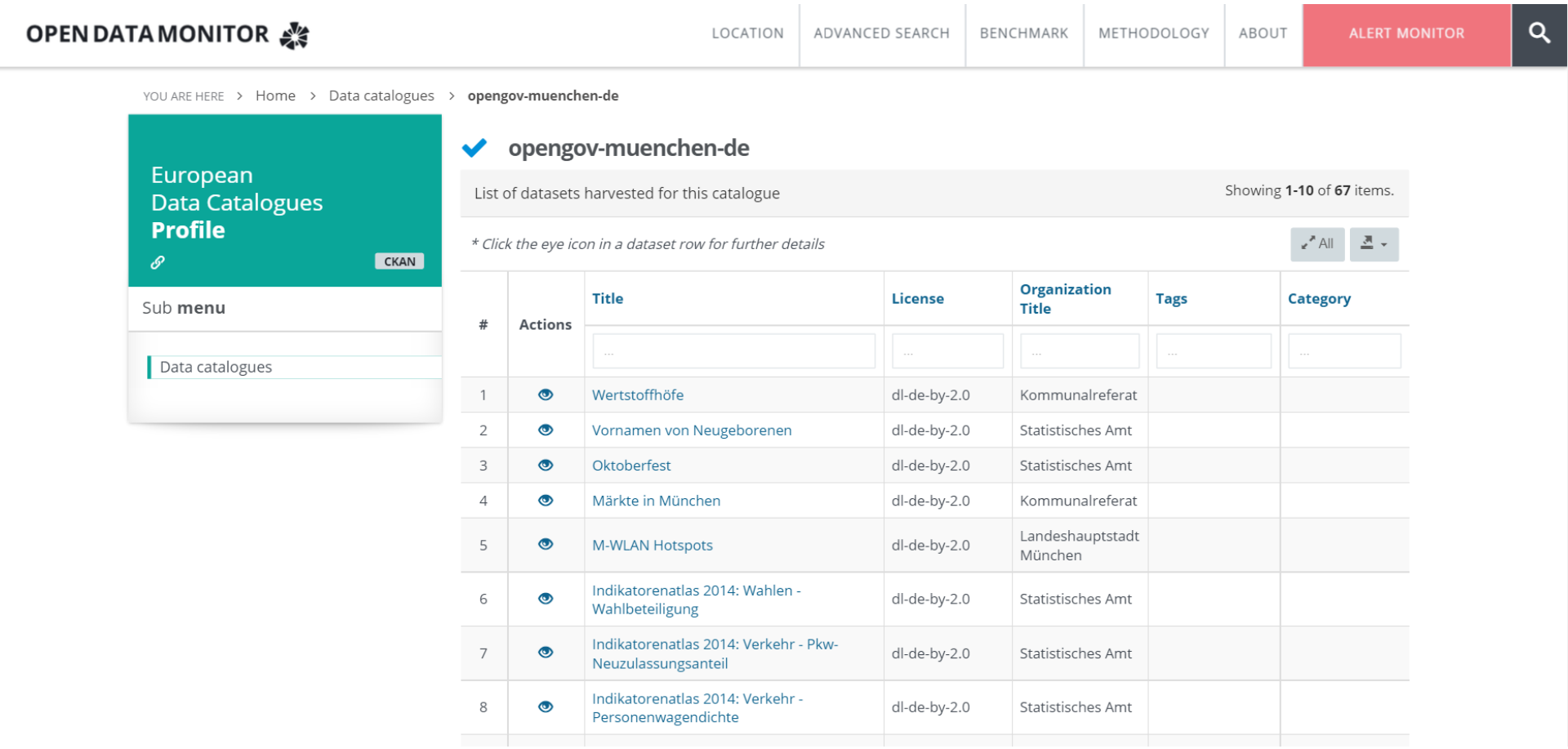
10 catalogues harvested from **Germany**



Catalogue Dashboard: München



Catalogue Dashboard: München



#	Actions	Title	License	Organization Title	Tags	Category
1		Wertstoffhöfe	dl-de-by-2.0	Kommunalreferat		
2		Vornamen von Neugeborenen	dl-de-by-2.0	Statistisches Amt		
3		Oktoberfest	dl-de-by-2.0	Statistisches Amt		
4		Märkte in München	dl-de-by-2.0	Kommunalreferat		
5		M-WLAN Hotspots	dl-de-by-2.0	Landeshauptstadt München		
6		Indikatorenatlas 2014: Wahlen - Wahlbeteiligung	dl-de-by-2.0	Statistisches Amt		
7		Indikatorenatlas 2014: Verkehr - Pkw-Neuzulassungsanteil	dl-de-by-2.0	Statistisches Amt		
8		Indikatorenatlas 2014: Verkehr - Personenwagendichte	dl-de-by-2.0	Statistisches Amt		

Empirische Ergebnisse für CKAN-Kataloge

Core metadata (licence, author, organisation, date released and date updated) meist „etwas“ angegeben, insbesondere:

- title, license, Author (email), url

Fehlende Attribute:

- Organisation info
- Temporal coverage
- Spatial coverage
- Update frequency
- Last update

Die Unordnung der Metadaten:

- Title: use random string instead of meaningful descriptions
- Licenses: same license different names
- Publisher/Maintainer: same organisation different names
- Tags: not correctly used
- Update frequency: missing or not correctly described
- And many more...
- Datasets aggregated in different ways: financial reports per year/months

Lizenzen in Deutschland: offen = offen?

- **nebeneinander von Vielzahl unterschiedlicher Lizenz-Regime**

	Berlin	Bremen	Hamburg
Anteil offener Lizenzen	88%	87%	100%
Dominantes Lizenzregime	Creative Commons	Creative Commons	Datenlizenz Deutschland
Weitere verwendete Regime	ODC, GNU	“andere” (nicht spezifiziert)	Creative Commons

- **Lizenzen im Govdata-Portal**

Lizenz	Occ.	Lizenz	Occ.
DL-DE-BY-2.0	17.762	CC-BY-4.0	5
DL-DE-BY-1.0	3.034	CC-BY	5.687
DL-DE-BY	8.338	CC-BY-SA	62
DL-DE-BY-NC	4.547	CC-0	55
DL-DE-zero	52	CC-BY-NC	9.506
GeoNutzV-Bund	70	andere (offen)	841
GeoNutz-Berlin	50	andere (nicht offen)	1.552

“Without licenses you don't have anything, [...] you have just data.” (open data researcher, user and consultant, Spain)

Diskussion

Open Data scheint in weiten Teilen (noch) nicht “ready for prime time”

- skalierbare Nutzung (Daten aus Madrid, Mailand, München, Maribor usw.) beschränkt durch
  - uneinheitliche Metadaten (auffinden und verstehen),
  - nebeneinander potenziell nicht kompatibler Lizenz-Regime,
  - unterschiedliche Datenstrukturen, Messverfahren, Bezeichnungen uvam.

- Standardisierung sehr unterschiedlich fortgeschritten
  - Datenarten, bei denen es nationale, europäische, internationale Standards gibt (Geobasisdaten, meteorologische Daten)
  - “datenaffine” Behörden (Statistikbehörden)
  - Behörden, die langjährige Erfahrungen mit der Datenweitergabe haben (oft industrienah oder aufgrund gesetzlicher Pflicht, wie im Umweltbereich)

- tatsächlich interessante Daten im Verborgenen
  - weil aktuell damit Einnahmen erzielt werden
  - weil der Wert nicht erkennbar ist: availability-approach
  - aufgrund ungeklärter/restriktiv ausgelegter Rechtsfragen (Haftung, Verwertungsrechte, Datenschutz)

Fazit

Schwerpunkte setzen: Welche Ziele sollen erreicht werden, welche Daten sind dafür von zentraler Bedeutung?

Internationale Standards übernehmen: In welcher Form (Daten und Metadaten) werden bestimmte Arten von Daten international publiziert?

Governance, Prozesse und Systeme open data-ready gestalten: Bereitstellungsaufwand minimieren, aber nicht durch availability-approach, sondern open by design


THANK
YOU
GRACIAS
ARIGATO
SHUKURIA
DANKSCHEEN
TASHAKKUR ATU
YAQHANYELAY
SUKSAMA
EKHMET
MEHRBANI
GRAZIE
MAAKE
KOMAPSUMNIDA
GOZAIMASHITA
EFCHARISTO
JUSPAXAR
PALDIES
BOLZÏN
MERCI
BÏYAN
SHUKRIA
TINGKI

Metadata Mapping

Distribution formats

https://github.com/opendatamonitor/ckanext-harmonisation/blob/master/ckanext/harmonisation/controllers/dictionaries/basic_formats_dict.py

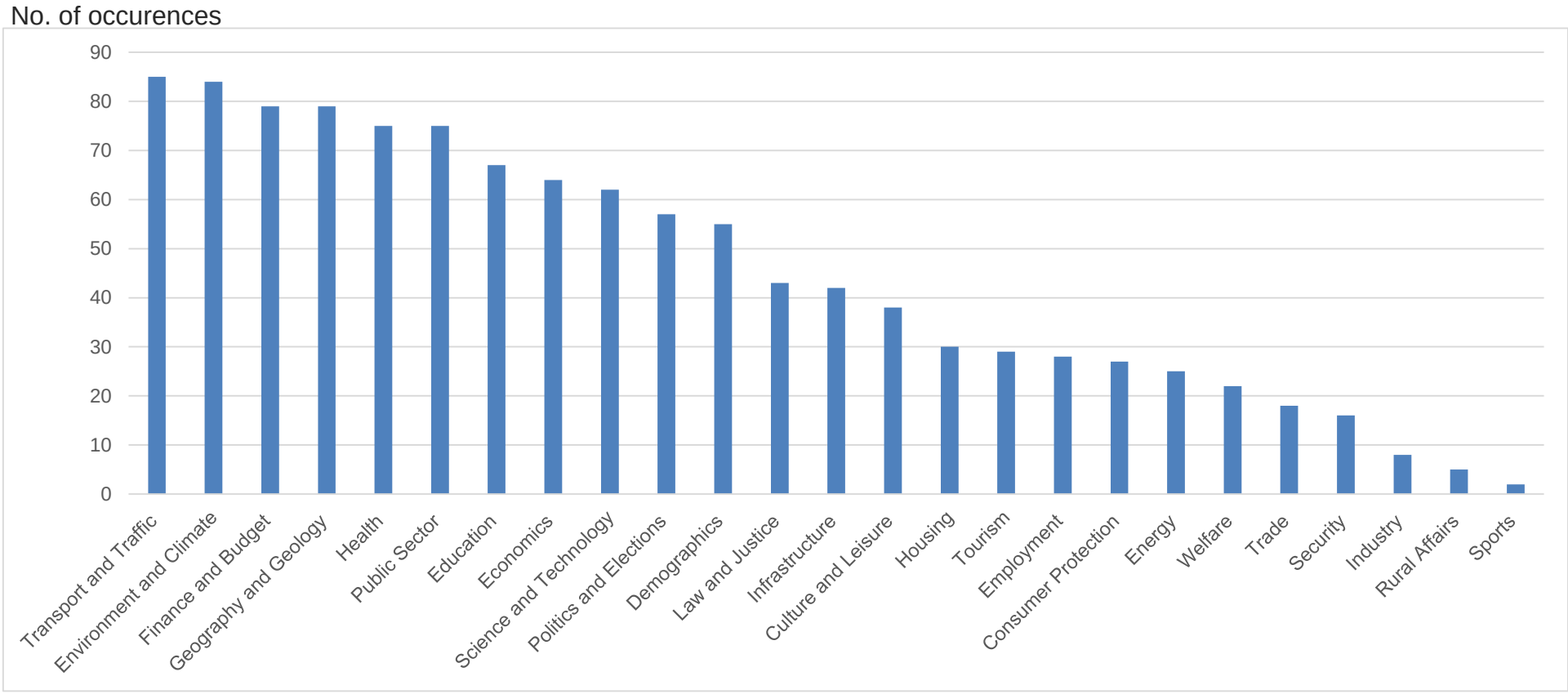
Topic categorization

https://github.com/opendatamonitor/ckanext-harmonisation/blob/master/ckanext/harmonisation/controllers/dictionaries/basic_category_values.py

Licenses

https://github.com/opendatamonitor/ckanext-harmonisation/blob/master/ckanext/harmonisation/controllers/dictionaries/basic_licenses_dict.py

Relevante Daten von hohem Interesse



Hohe Varianz zwischen den unterschiedlichen inhaltlichen Kategorien:

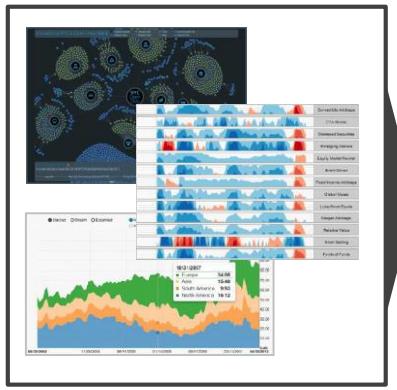
- Einige Daten hochinteressant, andere kaum relevant

→ inhaltliche Dimension scheint wichtig, obwohl Open Data nur formal technisch und rechtlich verstanden wird

Interactive Project Flow

Research studies and stakeholder analysis

Fundamental Research on Open Data and Environment



State Of The Art Analysis and Best Practices

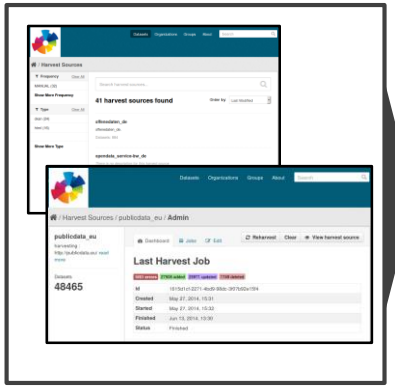
Stakeholder Surveys and Requirement Analyses

Open Data Resource Collection and Listings

Concept design and development

Framework Concept and Component Definitions

Specifications, Use Cases and Mockups



Core Implementations and Technical Development

User Interfaces and Visualisation Designs

The OpenDataMonitor Consortium

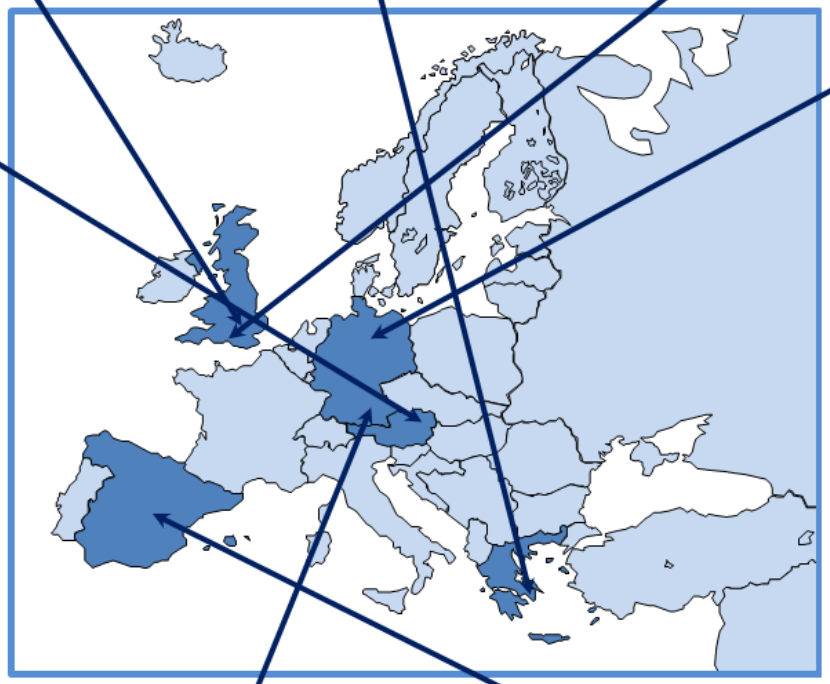
1 SYNYO GmbH
Research & Development Department
Vienna
AUSTRIA

2 Open Data Institute
Research Department
London
UK

3 Athena Research Center
Institute for the Management of Information Systems
Athens
GREECE

4 University of Southampton
Web and Internet Science Group
Southampton
UK

5 IfG.CC Potsdam eGovernment Competence Center
Research Department
Potsdam
GERMANY



6 City of Munich
Department of Labor and Economic Development
Munich
GERMANY

7 Entidad Pública Empresarial Red.es
Shared Services Department
Madrid
SPAIN