ON THE NEED FOR MODEL-DRIVEN ENGINEERING FOR DATA HARVESTERS: EXPERIENCES FROM THE GERMAN GOVDATA.DE PORTAL

Nikolay Tcholtchev, nikolay.tcholtchev@fokus.fraunhofer.de
Arun Prakash, arun.prakash@fokus.fraunhofer.de

Fraunhofer FOKUS

making cities smart

OVERVIEW

- Some Open Data Activities at Fraunhofer FOKUIS
- The German Governmental Portal GovData.DE
- Harvesting Experiences
- The Need for MDE based Harvesters
- Conclusions

© Fraunhofer FOKUS

FIRST OPEN DATA PORTAL IN GERMANY

Open Data Berlin

- Concept and realization by Fraunhofer FOKUS
- Deployment of the backend system - CKAN

- Analysis of various Open Data aspects in a corresponding study
- Definition of a Meta-data Schema

- Transfer of the pilot to Berlin Online towards a sustainable Operation
- http://daten.berlin.de

Fraunhofer FOKUS

© Fraunhofer FOKUS

ENERGY OPEN DATA OF VATTENFALL

Netzdaten Berlin

- Since December 2012: Pilot/Prototype-Portal of Vattenfall Europe on Open Data regarding the Electrical Grid of Berlin
- http://www.netzdaten-berlin.de

- Strong push towards Open Data from Industry
- 93 Datasets
  - Electricity Supply
  - Balance Sheets
  - Connections with the Grid
  - Coverage Area
  - Electrical Grid Structure
  - ...

- Concepts and realization by Fraunhofer FOKUS

VATTENFALL Fraunhofer FOKUS

Fraunhofer FOKUS

© Fraunhofer FOKUS

GOVERMENTAL DATA

Official Pilot of the German Ministry of Internal Affairs GovData.de

- The Pilot/Prototype is officially online since the 9th of February 2013
- http://www.govdata.de

- Development and Improvement of the Prototype

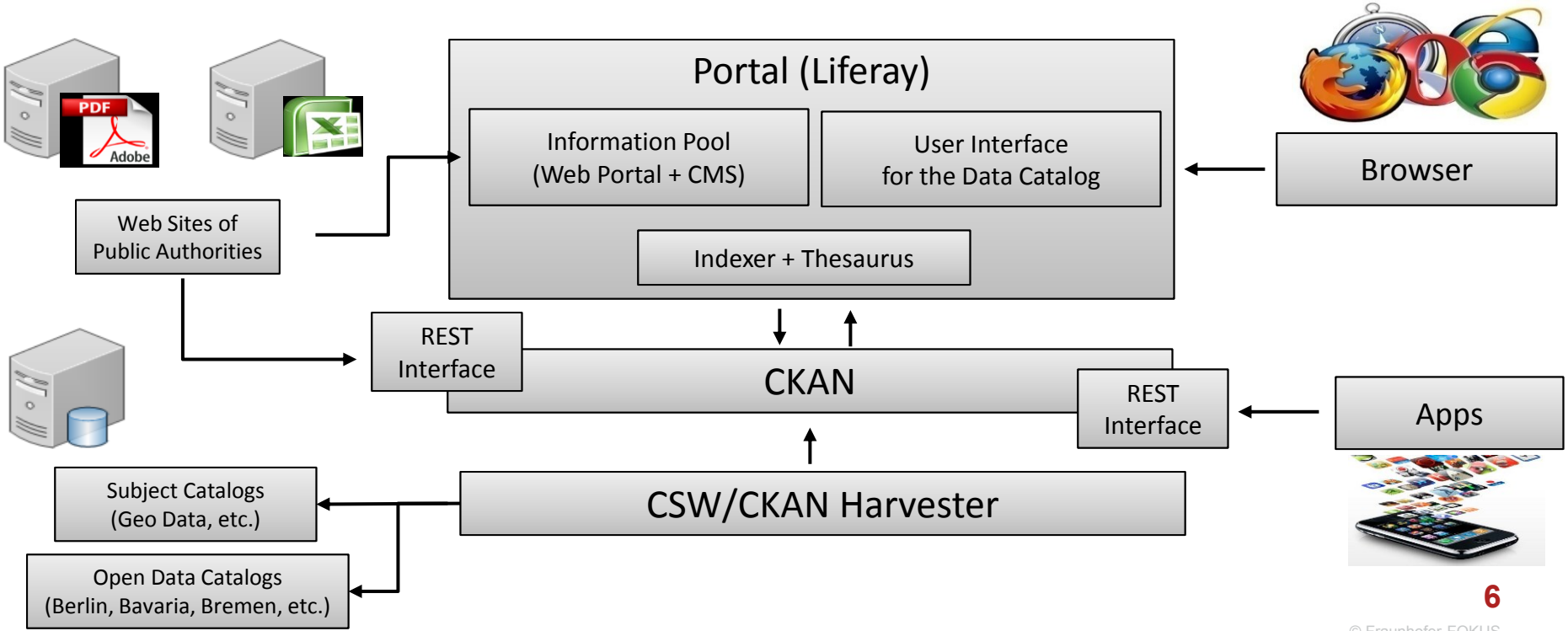
- Different Types of (Open) Data
  - Datasets
  - Documents
  - Applications

- Focus on free Licenses
  - Datenlizenz Deutschland (de-dl, ...)
  - Creative Commons (cc-by, ...)
  - ...

© Fraunhofer FOKUS

ARCHITECTURE OF THE GOVDATA.DE PLATFORM

Implemented and operated by Fraunhofer FOKUS



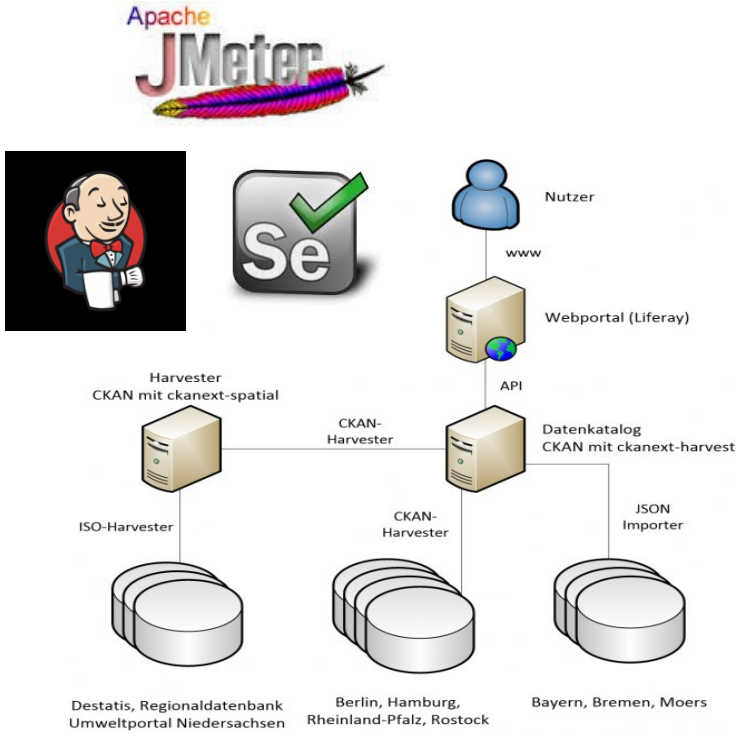
© Fraunhofer FOKUS

OPERATION OF GOVDATA.DE

Key Operational Aspects I

- Continuous Monitoring (Jenkins, Selenium GUI and Functional Tests) of the Platform‘s Operation and Availability
  - Alerting via E-Mail and SMS (using Internet based SMS Gateways)
  - 24/7 Management Support
  - Load- and Performance Testing (Apache JMeter) when required

- Redundant Deployment of key components
- Continuous Harvesting and Quality Assurance of the obtained datasets

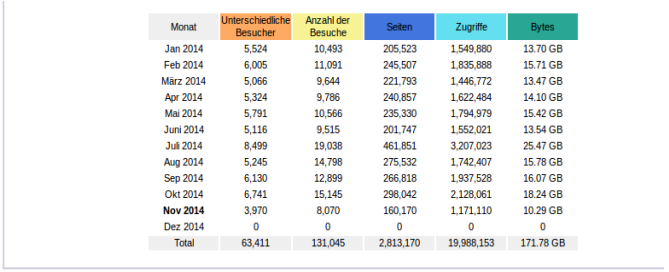


© Fraunhofer FOKUS

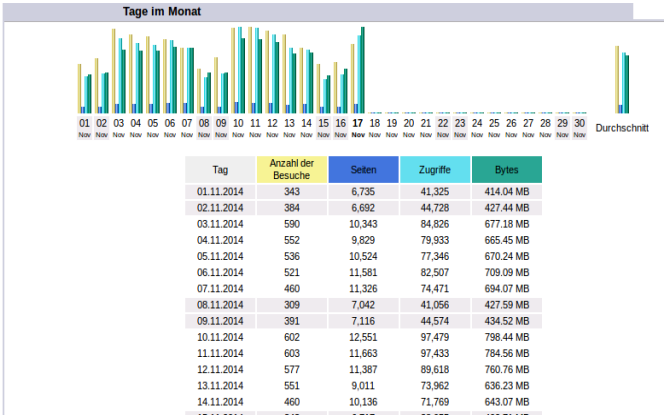
OPERATION OF GOVDATA.DE

Key Operational Aspects II

- Provisioning of advanced statistics reflecting:
  - the operation of the Platform
  - the User Interactions with GovData.DE



Monat	Unterschiedliche Besucher	Anzahl der Besuche	Seiten	Zugriffe	Bytes
Jan 2014	5,524	10,493	205,523	1,549,880	13.70 GB
Feb 2014	6,005	11,091	245,507	1,835,888	15.71 GB
März 2014	5,066	9,644	221,793	1,446,772	13.47 GB
Apr 2014	5,324	9,786	240,857	1,622,484	14.10 GB
Mai 2014	5,791	10,566	235,330	1,794,979	15.42 GB
Juni 2014	5,116	9,515	201,747	1,552,021	13.54 GB
Juli 2014	8,499	19,038	461,851	3,207,023	25.47 GB
Aug 2014	5,245	14,798	275,532	1,742,407	15.78 GB
Sep 2014	6,130	12,899	266,818	1,937,528	16.07 GB
Okt 2014	6,741	15,145	298,042	2,128,061	18.24 GB
Nov 2014	3,970	8,070	160,170	1,171,110	10.29 GB
Dez 2014	0	0	0	0	0
Total	63,411	131,045	2,813,170	19,988,153	171.78 GB



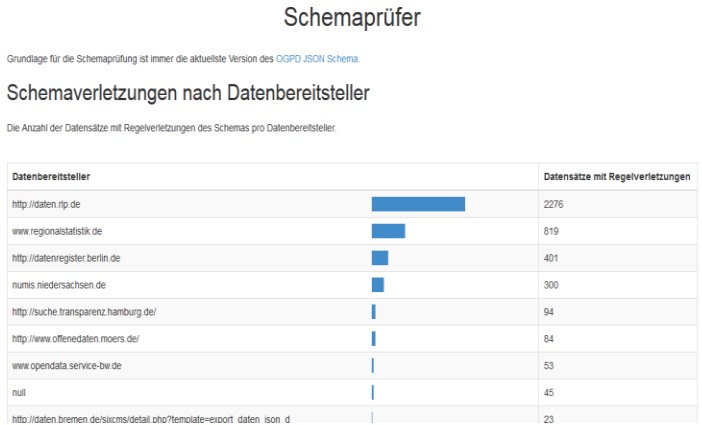
Tag	Anzahl der Besuche	Seiten	Zugriffe	Bytes
01.11.2014	343	6,735	41,325	414.04 MB
02.11.2014	384	6,692	44,728	427.44 MB
03.11.2014	590	10,343	84,826	677.18 MB
04.11.2014	552	9,829	79,933	665.45 MB
05.11.2014	536	10,524	77,346	670.24 MB
06.11.2014	521	11,581	82,507	709.09 MB
07.11.2014	460	11,326	74,471	694.07 MB
08.11.2014	309	7,042	41,056	427.59 MB
09.11.2014	391	7,116	44,574	434.52 MB
10.11.2014	602	12,551	97,479	798.44 MB
11.11.2014	603	11,663	97,433	784.56 MB
12.11.2014	577	11,387	89,618	760.76 MB
13.11.2014	551	9,011	73,962	636.23 MB
14.11.2014	460	10,136	71,769	643.07 MB

Fraunhofer FOKUS

OPERATION OF GOVDATA.DE

Key Operational Aspects III

- Established Production Level Process for Quality Assurance within the Harvesting procedures
  - Initial harvesting into a Test Environment
  - Automated Quality Assurance and Problem Reporting regarding the harvested datasets
  - Approval of the meta-data by the Dataproviders
  - Import of the harvested meta-data in the Production Level System



Schemaprüfer

Grundlage für die Schemaprüfung ist immer die aktuellste Version des OGPD JSON Schema.

Schemaverletzungen nach Datenbereitsteller

Die Anzahl der Datensätze mit Regelverletzungen des Schemas pro Datenbereitsteller.

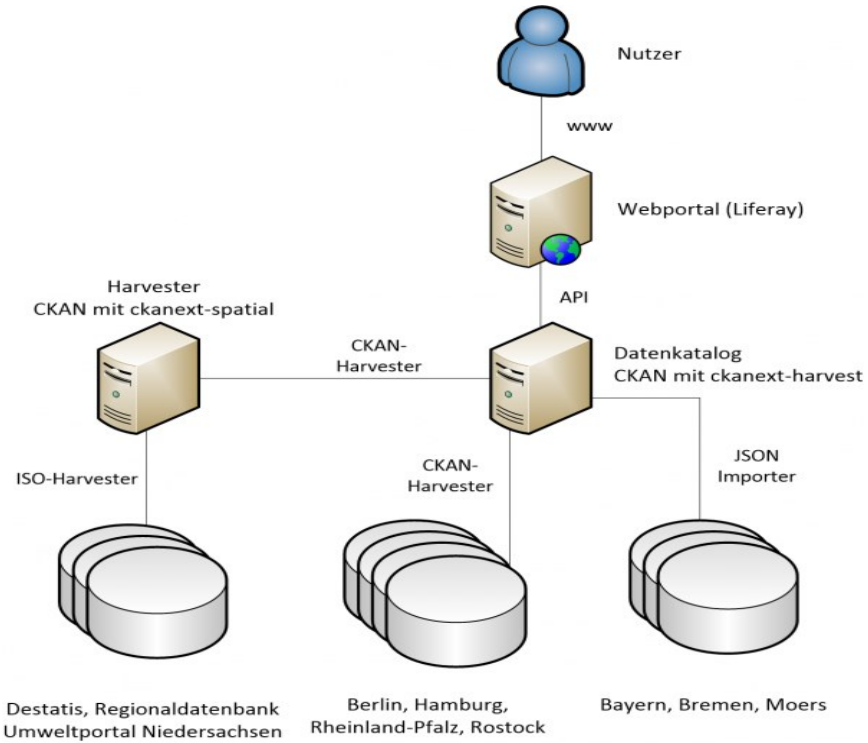
Datenbereitsteller	Datensätze mit Regelverletzungen
http://daten.rlp.de	2276
www.regionalstatistik.de	819
http://datenregister.berlin.de	401
numis.niedersachsen.de	300
http://suche.transparenz.hamburg.de/	94
http://www.offenedaten.moers.de/	84
www.opendata.service-bw.de	53
null	45
http://daten.bremen.de/sixcms/detail.php?template=export_daten_json_d	23

Fraunhofer FOKUS

© Fraunhofer FOKUS

HARVESTING EXPERIENCES

Harvesting Architecture I



- CSW (Common Service for the Web) and constitutes a REST for INSPIRE
- CKAN-2-CKAN Harvesting
- JSON Dumps

- Harvesting to the OGDD Meta-data Scheme for Germany

- Python based Harvesters implemented as CKAN extensions

© Fraunhofer FOKUS

HARVESTING EXPERIENCES

Harvesting Architecture II

- In general, the handling of the CKAN provided harvesting platform proved to be cumbersome

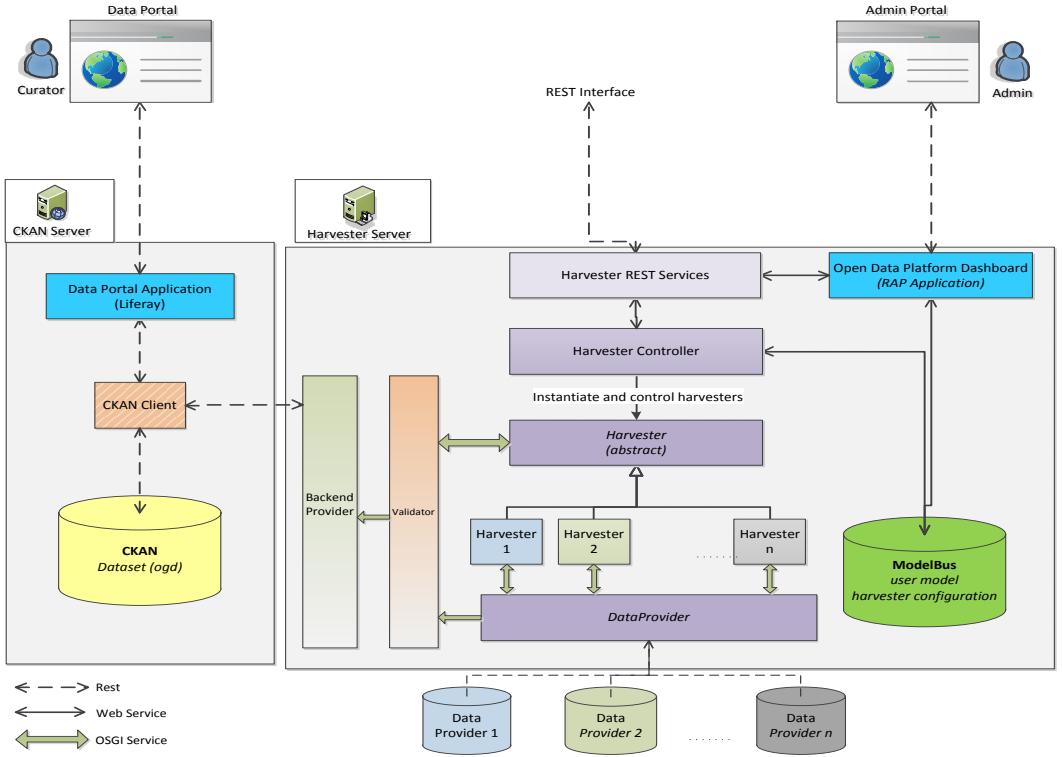
- Extensions/harvesters to be implemented in Python
  - Fine language when it comes to automating processes and hacking down tools with a specific purpose
  - Bears a large numbers of pitfalls when it is used in a complex large scale project that requires the involvement of various developers with different coding styles.
  - Aspects such as quality assurance (code review) and code styling

© Fraunhofer FOKUS

HARVESTING EXPERIENCES

Harvesting Architecture III

- OSGi based Architecture for harvesting
- First experiments show that in some cases performance deficiencies - in comparison to the Python based harvesting – should be expected
- Hence, the Java/OSGi based and Python based harvesting should be used on a case by case basis

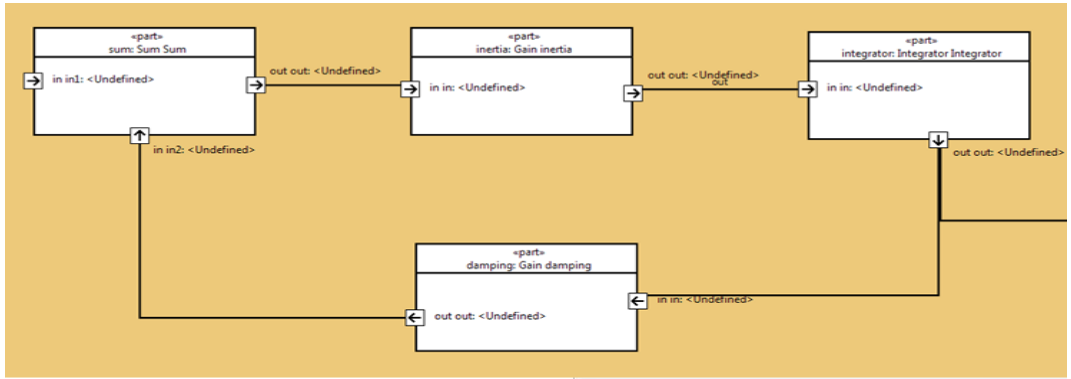


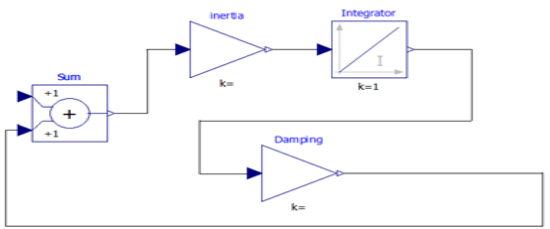
© Fraunhofer FOKUS

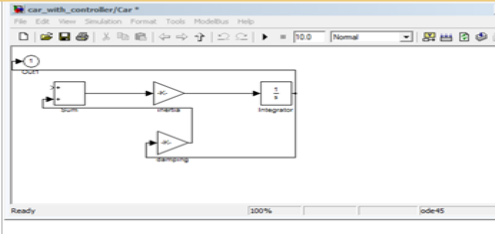
THE NEED FOR MDE BASED HARVESTERS

MDE based Harvesters

- The experiences with the OSGi/Java and the CKAN/Python based harvesting lead to the need for a platform independent model (PIM) specification for the harvesters
  - Generation of platform specific models (PSM), i.e. code, based on the use of transformations - Model-2-Code (Python or Java)







© Fraunhofer FOKUS

THE NEED FOR MDE BASED HARVESTERS

Benefits for the Involved Stakeholders

- The fact that the quality of the data – provided by the Open Data providers – would drastically increase, given the improved harvesting processes due to the use of MDE

- The time for the development of new harvesters will be drastically reduced since model based harvester engineering would allow a higher level abstraction and the involvement and collaborative harvester development by a larger set of collaborators – including people who are not pure developers and are more into the (governmental) data, its semantics and formatting

- The above aspects would increase the quality of the overall set of data provided by data platforms and will facilitate and encourage the usage of (Open) Data by companies, since the provided (meta-)data would be more timely and from higher quality (data quality and trustworthiness is one of the key topics in the light of Open Data)

© Fraunhofer FOKUS

THE NEED FOR MDE BASED HARVESTERS

Benefits for the Involved Stakeholders

- The latter (increased data quality) would lead to higher competitiveness of companies and industry using (Open) Data and would for instance allow them to pay additional taxes

- It is possible to come up with (industrial and public) fora and organizations, which would support the quality of the harvesting solutions thereby endorsing approaches such as MDE based harvesting towards establishing high quality Open Data provisioning

- These fora might also bear a financial aspect and would be responsible for financing the MDE tool providers, e.g. by paying for licenses for the MDE tools and making these tools available to the (Open) Data providers, e.g. public institutions or non-governmental organizations

© Fraunhofer FOKUS

CONCLUSIONS

- Critical aspects of our meta-data harvesting experiences around the German governmental data portal (GovData.De)
- The need for a model-driven approach for the continuous design of harvesters was derived
- Model-Driven Engineering would provide tool vendors with the possibility to commercialize their tools and let them benefit from the eco-systems emerging around (Open) Data providers
- Improved quality and timeliness of available datasets
- The possibility for commercial developers to rely on high quality data – which would in turn encourage the use of Open Data for commercial developments
- Possibility for MDE tool providers to benefit from the emerging eco-system around the topic of Open Data

© Fraunhofer FOKUS

CONTACT

Fraunhofer FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin, Germany
www.fokus.fraunhofer.de

Nikolay Tcholtchev
Senior Researcher
nikolay.tcholtchev@fokus.fraunhofer.de
Phone +49 (0)30 3463-7175

Arun Prakash
Senior Researcher
arun.prakash@fokus.fraunhofer.de
Phone +49 (0)30 3463-7358

Fraunhofer FOKUS

© Fraunhofer FOKUS

OUR SERVICES

Regarding (Open) Data Platforms Fraunhofer FOKUS offers:

- Development of guidelines for the selection and publishing of data according to their economic, security and privacy aspects with respect to governmental regulations
- Elaboration of recommendations for the provisioning and the management of (Open) Data
- Design and development of the portal and the server infrastructure for an Data Platform (e.g. an Open Data Platform)
- Design of Open Data catalogues and operation of federated catalogues and portals
- Concepts of interaction for users of (Open) Data offers
- Contributions to the standardization of formats, meta-data and licenses
- Trainings and workshops for partners regarding the realization of Open Data solutions
- Concepts, methods and tools for the analysis of data
- Development of programming interfaces for an efficient use of (Open) Data
- Design of search mechanisms, e.g. crawling

© Fraunhofer FOKUS