Towards a Taxonomy of Standards in Smart Data

Alexander Lenk, Leif Bonorden, Astrid Hellmanns, Nico Roedder, Stefan Jaehnichen
Accompanying Research of the BMWi Smart Data Technology Program
FZI Research Center for Information Technology
Berlin, Germany
{lenk, bonorden, hellmanns, roedder, jaehnichen}@fzi.de

***Abstract*— The usage of large amounts of data has an immense potential for global economic growth and the competitiveness of countries with high technological standards. Vast amounts of data from different sources are collected and analyzed in order to seek economic profit and competitive advantages for companies and society in general. To gain profit from such data, it needs to be analyzed, processed, and interpreted. Thus, knowledge can be created and such generation of knowledge within the analysis and interpretation process constitutes the difference between "Big" and "Smart" Data. In this paper we present a taxonomy to develop standards in the field of Smart Data. It consists of 8 challenges that need to be addressed by standards and 13 fields of standardization.**

Keywords: Smart Data, Big Data, Standardization, Taxonomy

I. INTRODUCTION

Data is the natural resource of the 21st century. The reason for this development is due to its huge potential to support global economic growth and the competitiveness of countries with high technological standards, with Germany being one of the most prominent examples. Existing data pools and their high dimensionality, often referred to as Big Data, are most promising in respect to their economic value and fields of usage and exploitation. In the past, large amounts of data have been collected in order to gain economic profit and competitive benefits for companies and firms. To illustrate: recent studies predict a rapid increase of the global trade volume with the usage of Big Data of up to USD 50 billion in 2017 [1], with an increase of 6 billion just for Germany alone [2].

However, due to the increasing amount and complexity of data, companies are facing major difficulties. Statistics show that approximately 80 percent of the global data is unstructured [3] and of highest dimensionality. Thus, up to 95 percent cannot be analyzed automatically with the help of current state-of-the-art technologies [4]. In order to profit from this data, it needs to be analyzed, processed, interpreted, and transformed into knowledge. In our view, the generation of knowledge constitutes the main shift from "Big" to "Smart" Data.

Accordingly, the term "Smart Data" describes the development from initially unstructured mass-data to the intelligent processing of data and its transformation into knowledge. This knowledge and its intelligent usage is the basis for further technological innovations in almost all domains and industries. To support this development the German Federal Ministry for Economic Affairs and Energy (BMWi) launched the "Smart Data – Innovation from Data" program and selected thirteen projects with more than 60 participating companies that tackle important challenges in Industry, Mobility, Healthcare, and Energy. Figure 1 depicts the different areas of the program.

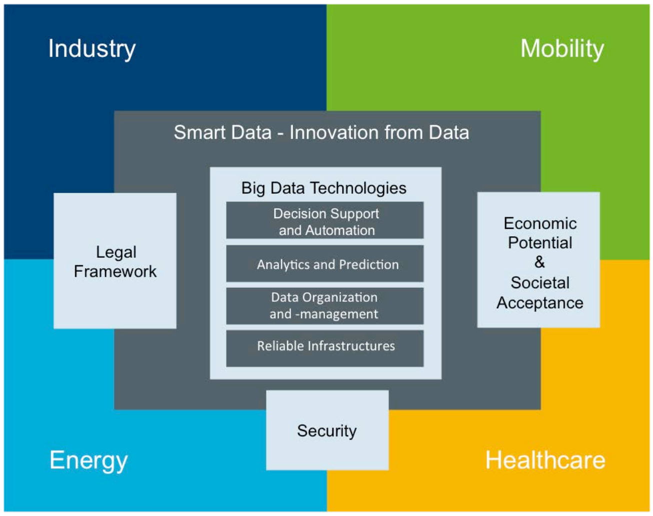

Industry
Mobility
Smart Data - Innovation from Data
Big Data Technologies
Decision Support and Automation
Analytics and Prediction
Data Organization and -management
Reliable Infrastructures
Legal Framework
Economic Potential & Societal Acceptance
Security
Energy
Healthcare

Figure 1. Smart Data Overview (c.f. [5])

Within the different projects, Smart Data technologies are used and new solutions are developed. In order for Smart Data to be successful, the projects and their results have to be utilizable for entrepreneurial efforts. As interoperability is a problem in both -Smart and Big Data-, the development and adoption of standards is of highest importance for successful exploitation. However, in current discussions, it is debatable what should be standardized in Smart Data and what could be addressed by these standards. In this paper, we present a taxonomy to classify standards in the field of Smart Data. It consists of 8 challenges and 13 different fields of standardization. In our future work, this taxonomy will serve as a framework to identify gaps in standardization that need to be addressed.

This paper is structured as follows: first, we introduce some specifics of the Smart Data technology program in section II. The current state of standardization in this area is described in section III. In section IV we present the challenges and areas of standardization. We conclude in section V with an outlook on possible future work.

II. THE SMART DATA TECHNOLOGY PROGRAM

In the four fields of interest to the "Smart Data Program", remarkable innovations are expected. In Smart Data projects use a combination of the four basic technologies *Reliable Infrastructure*, *Data Organization and Management*, *Analysis and Prediction* as well as *Decision Support and Automatization* [5] to build their solutions. This technology stack (see figure 1) can be used similar to other taxonomies as introduced e.g. in Cloud Computing [6]. Their intent is to classify technologies or at least components of these technologies.

- The category *Reliable Infrastructure* is concerned with the processing of data. Quality attributes such as performance, scalability, reliability, availability, and safety are the relevant features to be considered. Scalable on-demand infrastructures as provided by Cloud Computing environments, have already proven to be suitable and can be used for Smart Data as well.
- The challenges when dealing with unstructured and complex data are gathered in the context of *Data Organization and Management*. The application of semantic technologies has to be examined. With semantic technologies "meaning" can be encoded to enable machine-interpretability [7].
- *Analysis and Prediction* targets the analytical challenges when dealing with mass data, with support of analytical methods from the fields of statistics, machine learning, data mining, and semantics.
- *Decision Support and Automatization* provides support when based on the *Analysis and Prediction* processes such as business or productions processes have to be automatized or when decision support systems are needed for experts.

In addition to the described application areas and basic technologies of Smart Data, three fields which are of interdisciplinary interest are addressed: *Security*, *Legal Framework*, and as *Economic Potential & Societal Acceptance*.

- The topic of *Security* provides a technical basis for two other cross-section issues. In particular methods preventing unauthorized access, security solutions for the processing and storage of data, as well as the issue of anonymization are taken into account.
- The *Legal Framework* is concerned with questions about the legal requirements for the development and usability of new technologies in the context of Smart Data. In addition, necessary adaptions or changes in the regulatory frameworks need to be discussed and improvements have proposed.
- The topic of *Economic Potential & Societal Acceptance* targets the new Smart Data business models and corresponding challenges in the fields of information and network economies. The main focus is on issues concerning the social approval of the business models.

III. STANDARDIZATION IN SMART DATA

Various established Smart Data platforms are already in use by scientists and companies in order to exploit large data pools. When taking a closer look to the structure of these platforms, a distinction of six parts can be determined, as depicted in figure 2 [8]. These parts can be mapped to the different technology categories in the Smart Data Technology program. While the categories in this program define fields of standardization, this taxonomy introduces several challenges:

- *Data Storage* is manageable with the help of a variety of different technologies. Especially distributed file systems like the Hadoop File System (HDFS) [1] and relational/non-relational and In-memory databases, such as MySQL[2], Cassandra[3], or HANA[4]. It fits both the *Reliable Infrastructure* and the *Data Organization and Management categories*.
- *Data Access* deals with the authorization to grant users access to data, which is usually stored in databases in order to be efficiently processed. Methods such as streaming-procedures or query languages are compiled. Paradigms or models such as MapReduce, complex event processing (CEP), or lambda architectures can be observed as well as specific implementations of these paradigms as eg. In Hadoop MapReduce[5], or Esper[6]. This layer also fits the *Data Organization and Management category*.

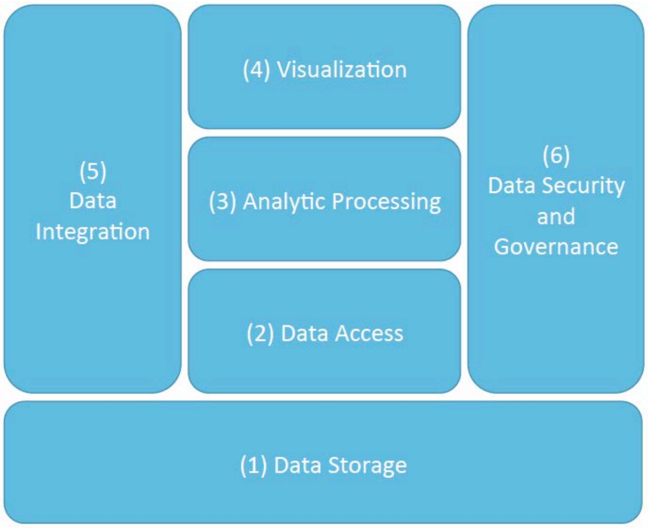


Figure 2. Taxonomy of a Smart Data Platform (c.f. [8], [9])

- In the field of *Analytic Processing,* different methods such as predictive analytics, data mining, and the

[1] http://hadoop.apache.org

[2] https://www.mysql.com

[3] http://cassandra.apache.org

[4] http://hana.sap.com

[5] http://hadoop.apache.org

[6] http://www.espertech.com/products/esper.php

application of artificial intelligence methods or machine learning technologies are of importance. For these methods, certain tools with languages as for instance SPSS[7], R[8], Matlab[9], or MLib[10] are used. *Analytic Processing* and *Visualization* can be mapped to the *Analytics and Prediction* layer of the Smart Data stack.

- The main aim of *Visualization* is to represent the relevant information in a comprehensible way and to present the derived knowledge to people as e.g. managing directors or other analysts.
- The intention of *Data Integration* is to integrate data from different sources (and probably with different formats) into a secure Smart Data environment. This implies that data from different sources and with different structures are transformed into a predefined format in order to enable the processing and unique interpretation. Whenever structure, layout, or metadata is changed, the contents remain the same.
- *Data Security and Governance* refers to the compliance and implementation of terms and regulations concerning the protection of privacy and IT security issues, which are essential for Smart Data projects [10].

As with other technological developments, the processes of standardization and unification play an important role in the context of Smart Data. Only with the use of predefined standards can interoperability between different technological solutions be guaranteed and an efficient collaboration of different agents be enabled. Furthermore, agreed standards are indispensable to ensure freedom of decision between different providers and to avoid lock-in effects. However, the levels of standardization may vary between the used technologies of the above-described application areas of Smart Data.

In the fields of *Data Storage* and *Analytic Processing* a small set of leading technologies and specific implementations are already in use. Nonetheless, interoperability between the implementations is not yet warranted.

In the fields of *Data Access* and *Data Integration* initial methods have been established, but still require time to become prevalent. Within the level of *Data Access*, there is a lack of new and more defined access methods for larger amounts of data. The problem of *Data Integration* are the different layers of data refinement as well as the different methods and solutions for *Data Access* and *Data Storage*, which have to be taken into account. This has to be established by corresponding standards in the near future.

In the fields of *Visualization* and *Data Security and Governance* the emergence of suitable standardized approaches has to be motivated, too. However, there is an urgent need to establish standards in all technological fields of Smart Data platforms.

IV. Challenges

To allow a functioning Smart Data ecosystem, the needs of several stakeholders have to be taken into account: service providers, customers, the general public, the market, and, last not least, the governing institutions. These needs are expressed as challenges, which are then grouped into eight categories. While Smart Data shares the challenges of many as-a-service models, some need to be highlighted and specially addressed. The following were identified by studying both, the Cloud Computing field [11], [12] and the current situation in Smart Data [8], [13].

A. Efficiency of Service Delivery

Service providers are interested in economically reasonable developments, and in smooth and efficient operation and configuration of services. This requires the use of efficient tools and components in the construction phase, and the easy deployment to available infrastructures – especially cloud computing – and the flexibility to adapt to new or changed requirements and resources with minimal effort [12]. The need for efficiency applies to technical, economic and human resources. So enable specific analytic tools or methods, scalable technology stacks ensure a standardized service delivery.

B. Effectiveness of Service Usage

Data services are only attractive to customers if their requirements are precisely met. General service effectiveness includes the level of agreement in basic contract design. More specific for Smart Data, flexibility and the quality of results are a major concern. Sub-Challenges are fulfillment of contracts, self-service paradigms, and governance [12].

C. Transparency of Service Delivery and Billing

Services per se, their management as well as their billing have to be comprehensible [12]. Customers have to be able to understand a provider's handling of their data: form and place of data storage and processing, and fulfillment of contracts. Providers have to supervise their own services to enable and assure quality, but in some cases this also means processing customers' data. Transparency of data location is an important challenge in the field of Smart Data since it influences technical requirements such as data protection techniques and may change the data's legal status.

D. Security

Data has to be protected against unauthorized access and loss but also against illegal usage. This includes third-party access such as hacking attempts, but also appropriate identity and rights management for parties involved [12]. Security is accompanied by trust, which is prerequisite for customers to accept an agreement or further cooperation. Thus, the providers' interest in security relates to his reputation. Security in a technical sense is discussed as a protection mechanism in the next section.

[7] http://www-01.ibm.com/software/analytics/spss/

[8] https://www.r-project.org

[9] http://www.mathworks.com/products/matlab/

[10] http://spark.apache.org/mllib/

E. *Privacy*

Data may include personal or classified information only accessible by a limited group of recipients. However, such data might also need to be processed. A standard or a technique might address the goal of processing the data, while still preserving privacy needs. This applies to raw data as well as to generated results.

F. *Interoperatbility*

Usage of most services is only possible if they are used within an ecosystem: they have to interact with other services, especially other Smart Data services, but also with other technological and management services to build a whole business process. It is also necessary that technologies are exchangeable. These lock-in effects are well-known from technologies as for instance Cloud Computing and are still a challenge in service provisioning [14]–[16].

G. *Portability*

While Interoperability is focusing on the technology these aforementioned lock-in effects can also be a problem when it comes to the data that is processed at a Smart Data provider. Being locked to a single provider or service can be the source of slowdowns in the growth process, when a provider no longer satisfies requirements. Thus, service value can be improved by interchangeability. The data itself has to remain portable i.e., source, intermediate and result data needs to be transferable from one provider to another.

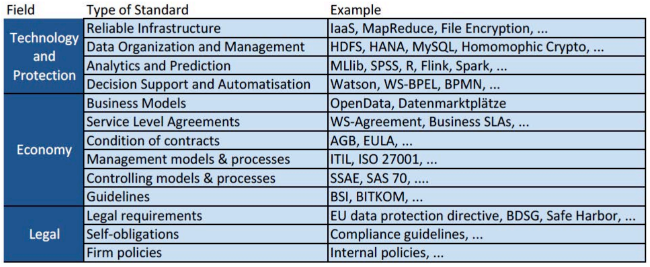
H. *Compliance*

For successful service delivery, providers must comply with different regulations and constraints: societal, market, and legal issues need to be satisfied. General acceptance needs to be aspired from society, governmental institutions may observe whether corporations comply with concepts for markets, finance and national systems, judicial institutions, and competitors and may also been checked for compliance with the law and other legal regulations. The cross-section working groups *Legal Framework* and *Economic Potential & Societal Acceptance* are occupied with this topic in the Smart Data program [10].

V. Fields of Standardisation

In the preceding section we defined 8 challenges for standards, now we propose a scheme for grouping the standards that should can be developed. They are generally divided into the three larger groups “Technology and Protection”, “Economy”, and “Legal”, with sub-fields.

Based on these fields, orienting knowledge, (reference) implementations or specifications, and industry standards, standards and technical standards [11] can be classified. Figure 3 shows a first approach.



Field	Type of Standard	Example
Technology and Protection	Reliable Infrastructure	IaaS, MapReduce, File Encryption, ...
	Data Organization and Management	HDFS, HANA, MySQL, Homomophic Crypto, ...
	Analytics and Prediction	MLlib, SPSS, R, Flink, Spark, ...
	Decision Support and Automatisation	Watson, WS-BPEL, BPMN, ...
Economy	Business Models	OpenData, Datenmarktplätze
	Service Level Agreements	WS-Agreement, Business SLAs, ...
	Condition of contracts	AGB, EULA, ...
	Management models & processes	ITIL, ISO 27001, ...
	Controlling models & processes	SSAE, SAS 70, ...
	Guidelines	BSI, BITKOM, ...
Legal	Legal requirements	EU data protection directive, BDSG, Safe Harbor, ...
	Self-obligations	Compliance guidelines, ...
	Firm policies	Internal policies, ...

Figure 3. First approach of a smart data standard categorization (c.f. [11])

In the following we describe the 13 different fields of standardization that are based on the fields derived from Cloud Computing [11].

A. *Technology and Protection*

The first group deals with technological fields of standardization. Technological fields can further be divided into four subgroups from the basic technologies depicted in figure 1. Protection of the technology needs to be highlighted and has to be considered in all of the subgroups. The term “protection” describes the need to improve security on a technical level, while security as one of the above-mentioned challenges describes the abstract concept. This illustrates the importance of security for the Smart Data program as a cross-section topic it is on the one hand a goal to be achieved and on the other hand a solution to achieve security, and similar goals like privacy.

1) *Reliable Infrastructure*

Data access and processing are fields that are affected by standards in many ways. Standards may apply to file formats, access protocols and to the interfaces of complete systems. The infrastructure in Smart Data is in most cases provided by a cloud environment, for which we refer to the study on the standardization of cloud computing in [11]. Especially in these virtualized environments data protection is an important issue.

2) *Data Organization and Management*

Data is the key element in Smart Data: it is generated or collected, then processed to generate beneficial results. Thus, this field of standardization includes the selection of sources and forms of anonymization, but also its annotation and structure. Data quality is addressed as it heavily influences the quality of the processes and their results. This field is closely connected to the reliable infrastructure, as it is a set of data in a certain database and structured in the databases’ format. Here, we distinguish between the infrastructure to process the data (e.g. Hadoop MapReduce) and the storage format that is built on top (e.g. HDFS).

3) *Analytics and Prediction*

Algorithms, tools, or libraries can be standardized in order to get the same results on different infrastructures or data formats. This can be tested by using benchmarks and other standardized tests. Results need to be made readable, which includes visualization, and the analytic results need to be documented, for which standardized forms can also be used.

4) *Decision support and Automatisation*

Smart Data helps to automate decisions, here two kinds of decisions can be separated: decisions on further data processing or on the following steps in a Smart Data process, and decisions taken as a result of a completed Smart Data process. Results from a Smart Data service are used to take decisions, and could be supplemented by data from other sources of information. So is especially in the field of industrial production automation and a fast reaction to changes is desired.

B. *Economy*

Economical fields of standardization cover business models with service and contract specifications, models in management and controlling, and forms of review such as audits, certifications and expert reports [11], [12]. These topics are closely linked to the cross-section issue *Economic Potential & Societal Acceptance*.

1) *Business Models*

From resource management, marketing to sales, business models are the basis of every service. Usually a business model is the reason why a service fails or succeeds. A set of standard business models can help new players on the market to be successful and grow faster.

2) *Service Level Agreements*

Assurance of service levels can be important groundwork for individual contracts between providers and customers. This is especially useful for business-to-business transactions. Standardized contracts or key performance indicators that are part of these contracts are addressed in the following.

3) *Condition of Contracts*

Generally defined, basic components make the contract design more efficient as they contain part of the service provider's obligations.

4) *Management Models & Processes*

Management can be unitized in terms of procedure, nomenclature, and best practices. Examples are standardized processes, procedures, tasks, and checklists like ISO 20000 or ITIL [17].

5) *Controlling Models & Processes*

Billing and documentation processes have to be handled in uniform ways and therefore specified, and verified in the form of external audits or quality management processes.

6) *Guidelines*

Guidelines are a simple approach that help to achieve different goals. Guidelines are less formal than control models and processes.

C. *Legal*

Laws, directives and governmental guidelines affect Smart Data services. Decisions from industry associations and company guidelines supplement these.

Contracts and also framework contracts between service providers and customers, they may also be legal documents, but are treated as economical decisions as discussed above.

1) *Legal Requirements*

Service providers and customers are directly or indirectly influenced by government-issued regulations. In this field standards that define or used to define legal requirements are grouped.

2) *Self-Obligation*

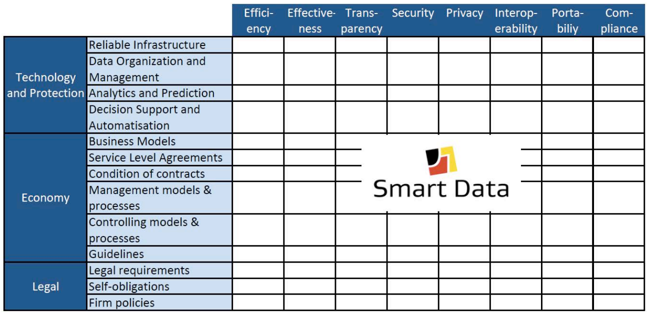
Industrial associations and other unions of corporations agree on common guidelines. So even if they are influenced by individual companies, they have the potential to influence a whole industry. In any case, associations' regulations usually only apply to the respective members.

3) *Firm policies*

The weakest form of regulations are internal guidelines, as they apply to single companies only.

VI. Conclusion and Future Work

In this paper, we introduced a taxonomy of standards for Smart Data. This taxonomy consists of 8 challenges and 13 fields of standardization addressing the broad spectrum of Smart Data. By using both, the challenges and the fields of standardization, we plan to identify already existing standards and classify the challenges they address (see figure 4). Evaluating the current state of standardization allows the identification of gaps in standardization. With this knowledge we can argue for the demand and necessity of additional standards.



		Efficiency	Effectiveness	Transparency	Security	Privacy	Interoperability	Portabiliy	Compliance
Technology and Protection	Reliable Infrastructure								
	Data Organization and Management								
	Analytics and Prediction								
	Decision Support and Automatisation								
Economy	Business Models								
	Service Level Agreements								
	Condition of contracts								
	Management models & processes								
	Controlling models & processes								
	Guidelines								
Legal	Legal requirements								
	Self-obligations								
	Firm policies								

Figure 4. Standardization Environment of Smart Data (c.f. [11])

Acknowledgment

This work is part of the Smart Data Accompanying Research that is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi).

References

[1] Statista, "Global big data market revenue segments forecast 2011-2017 | Statistic," *Statista*. [Online]. Available: http://www.statista.com/statistics/255970/global-big-data-market-forecast-by-segment/. [Accessed: 05-Oct-2015].

[2] Statista, "Big Data - Umsatz in Deutschland bis 2016," *Statista*. [Online]. Available: http://de.statista.com/statistik/daten/studie/257976/umfrage/umsatz-mit-big-data-loesungen-in-deutschland/. [Accessed: 05-Oct-2015].

[3] IBM, "What will we make of this moment? - 2013 IBM Annual Report." [Online]. Available: https://www.ibm.com/annualreport/2013/bin/assets/2013_ibm_annual.pdf. [Accessed: 05-Oct-2015].

[4] FZI Research Center for Information Technology, “Smart Data: A Big Data Memorandum.” [Online]. Available: http://smart-data.fzi.de/fileadmin/user_upload/smart-data-memorandum/Smart_Data_Memorandum.pdf. [Accessed: 04-Oct-2015].

[5] Bundesministerium für Wirtschaft und Energie (BMWi), “Smart Data - Innovationen aus Daten.” [Online]. Available: http://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/smart-data-innovationen-aus-daten.pdf?__blob=publicationFile&v=2. [Accessed: 28-Sep-2015].

[6] A. Lenk, M. Klems, J. Nimis, S. Tai, and T. Sandholm, “What’s inside the Cloud? An architectural map of the Cloud landscape,” in *Software Engineering Challenges of Cloud Computing, 2009. CLOUD’09. ICSE Workshop on*, 2009, pp. 23–31.

[7] R. Studer, S. Agarwal, and R. Volz, *The Semantic Web*, vol. 139. Springer Boston, 2004.

[8] Nationaler IT Gipfel, Arbeitsgruppe 2– Projektgruppe Smart Data, “Smart Data – Potenziale und Herausforderungen, Strategiepapier.” .

[9] BITKOM, “Big Data Technologien. Wissen für Entscheider.” [Online]. Available: https://www.bitkom.org/Publikationen/2014/Leitfaden/Big-Data-Technologien-Wissen-fuer-Entscheider/140228_Big_Data_Technologien_Wissen_fuer_Entscheider.pdf. [Accessed: 04-Oct-2015].

[10] Smart Data Begleitforschung, “Smart Data Fachgruppen.” [Online]. Available: http://www.digitaletechnologien.de/DT/Redaktion/DE/Dossiers/smart_data_fachgruppen.html?cms_docId=151528.

[11] Federal Ministry of Economics and Technology (BMWi), “The Standardisation Environment for Cloud Computing.” [Online]. Available: http://www.trusted-cloud.de/media/content/BMWi_Cloud_Standards_Studie_e_web.pdf. [Accessed: 28-Sep-2015].

[12] R. Bühler, “Standards in Disruptive Innovation: Assessment Method and Application to Cloud Computing,” PhD Thesis, KIT Scientific Publishing, Karlsruhe, 2015.

[13] Federal Ministry for Economics Affairs and Energy, “Smart Data Projects.” [Online]. Available: http://www.digitale-technologien.de/DT/Navigation/EN/Foerderprogramme/Smart_Data/Projekte/projekte.html. [Accessed: 28-Sep-2015].

[14] A. Lenk, G. Katsaros, M. Menzel, J. Rake-Revelant, R. Skipp, E. Castro-Leon, and Gopan V P, “TIOSA: Testing VM Interoperability at an OS and Application Level - A hypervisor testing method and interoperability survey,” in *ICE2 2014: IEEE International Conference on Cloud Engineering*, Boston, 2014.

[15] T. Kurze, M. Klems, D. Bermbach, A. Lenk, S. Tai, and M. Kunze, “Cloud federation,” presented at the CLOUD COMPUTING 2011, The Second International Conference on Cloud Computing, GRIDs, and Virtualization, 2011, pp. 32–38.

[16] N. Roedder, P. Karaenke, R. Knapper, and C. Weinhardt, “Decision-Making Based on Incident Data Analysis,” in *2014 IEEE 16th Conference on Business Informatics (CBI)*, 2014, vol. 1, pp. 46–53.

[17] S. Sahibudin, M. Sharifi, and M. Ayat, “Combining ITIL, COBIT and ISO/IEC 27002 in Order to Design a Comprehensive IT Framework in Organizations,” in *Second Asia International Conference on Modeling Simulation, 2008. AICMS 08*, 2008, pp. 749–753.