# PolicyCLOUD: Analytics as a Service Facilitating Efficient Data-Driven Public Policy Management

Dimosthenis Kyriazis[1](✉), Ofer Biran[2], Thanassis Bouras[3], Klaus Brisch[4], Armend Duzha[5], Rafael del Hoyo[6], Athanasios Kiourtis[1], Pavlos Kranas[7], Ilias Maglogiannis[1], George Manias[1], Marc Meerkamp[4], Konstantinos Moutselos[8], Argyro Mavrogiorgou[1], Panayiotis Michael[8], Ricard Munné[9], Giuseppe La Rocca[10], Kostas Nasias[11], Tomas Pariente Lobo[9], Vega Rodrigálvarez[6], Nikitas M. Sgouros[1], Konstantinos Theodosiou[3], and Panayiotis Tsanakas[8]

[1] University of Piraeus, Piraeus, Greece
{dimos,kiourtis,imaglo,gmanias,margy,sgouros}@unipi.gr
[2] IBM Research, Haifa, Israel
biran@il.ibm.com
[3] Ubitech, Athens, Greece
{bouras,ktheodosiou}@ubitech.eu
[4] DWF Rechtsanwaltsgesellschaft mbH, Cologne, Germany
{klaus.brisch,marc.meerkamp}@dwf.law
[5] Maggioli SpA, Santarcangelo di Romagna, Italy
armend.duzha@maggioli.it
[6] Instituto Tecnológico de Aragón, Saragossa, Spain
{rdelhoyo,vrodrigalvarez}@itainnova.es
[7] LeanXcale, Madrid, Spain
pavlos@leanxcale.com
[8] National Technical University of Athens, Athens, Greece
kmouts@gmail.com, panayiotismichael@mail.ntua.gr,
panag@cs.ntua.gr
[9] Atos Spain, Madrid, Spain
{ricard.munne,tomas.parientelobo}@atos.net
[10] EGI Advanced Computing Services for Research, Amsterdam, The Netherlands
giuseppe.larocca@egi.eu
[11] OKYS, Sofia, Bulgaria
knasias@okys.eu

**Abstract.** While several application domains are exploiting the added-value of analytics over various datasets to obtain actionable insights and drive decision making, the public policy management domain has not yet taken advantage of the full potential of the aforementioned analytics and data models. Diverse and heterogeneous datasets are being generated from various sources, which could be utilized across the complete policies lifecycle (i.e. modelling, creation, evaluation and optimization) to realize efficient policy management. To this end, in this paper we present an overall architecture of a cloud-based environment that facilitates

I. Maglogiannis et al. (Eds.): AIAI 2020, IFIP AICT 583, pp. 141–150, 2020.
https://doi.org/10.1007/978-3-030-49161-1_13

data retrieval and analytics, as well as policy modelling, creation and optimization. The environment enables data collection from heterogeneous sources, linking and aggregation, complemented with data cleaning and interoperability techniques in order to make the data ready for use. An innovative approach for analytics as a service is introduced and linked with a policy development toolkit, which is an integrated web-based environment to fulfil the requirements of the public policy ecosystem stakeholders.

**Keywords:** Data analytics as a service · Public policies · Cloud computing · Policy modelling

## 1 Introduction

The ICT advances as well as the increasing use of devices and networks, and the digitalisation of several processes is leading to the generation of vast quantities of data. These technological advances have made it possible to store, transmit and process large amounts of data more effectively than before in several domains of human activity of public interest [1]. It is undeniable that we are inundated with more data than we can possibly analyse [2]. This rich data environment affects decision and policy making: cloud environments, big data and other innovative data-driven approaches for policy making create opportunities for evidence-based policies, modernization of public sectors and assistance of local governance towards enhanced levels of trust [4]. During the traditional policy cycle, which is divided into 5 different stages (agenda setting, policy formulation, decision making, policy implementation and policy evaluation), data is a valuable tool for allowing policy choices to become more evidence-based and analytical [3]. The discussion of data-driven approaches to support policy making can be distinguished between two main types of data. The first is the use of open data (administrative - open - data and statistics about populations, economic indicators, education, etc.) that typically contain descriptive statistics, which are used more intensively and in a linked way, shared through cloud environments [3]. The second main type of data is from any source, including data related to social dynamics and behaviour that affect the engagement of citizens (e.g. online platforms, social media, crowd-sourcing, etc.). These data are analysed with novel methods such as sentiment analysis, location mapping or advanced social network mining. Furthermore, one key challenge goes beyond using and analysing big data, towards the utilization of infrastructures for shared data in the scope of ethical constraints both for the citizens and for the policy makers. These ethical constraints include "data ownership" ones, which determine the data sharing rules, as well as data localisation constraints, which may unjustifiably interfere with the "free flow of data". As for policy makers, the dilemma, to what extent big data policy making is in accordance to values elected governments promote, is created. This is a problem deriving from the fact that it is difficult to point to the scope of the consent citizens may give to big data policy analysis [5]. Furthermore, such policies provide a broad framework for how decisions should be made regarding data, meaning that they are high-level statements and need more detail before they can be operationalized [6].

Big data enabled policy making should answer modern democratic challenges, considering facts about inequality and transparency both on national and local level, involving multi-disciplinary and multi-sectoral teams.

In this context, this paper presents the main research challenges and proposes an architecture of an overall integrated cloud-based environment for data-driven policy management. The proposed environment (namely PolicyCLOUD) provides decision support to public authorities for policy modelling, implementation and simulation through identified populations, as well as for policy enforcement and adaptation. Additionally, a number of technologies are introduced that aim at optimizing policies across public sectors by utilizing the analysed inter-linked datasets and assessing the impact of policies, while considering different properties (i.e. area, regional, local, national) and population segmentations. One of the key aspects of the environment is its ability to trigger the execution of various analytics and machine learning models as a service. Thus, the implemented and integrated service can be executed over different datasets, in order to obtain the results and compile the corresponding policies.

## 2 Related Work

In terms of managing diverse data sources, the evolution of varieties of data stores (i.e. SQL and NoSQL), where each variety has different strengths and usage models, is linked with the notion of "polyglot persistence" [7]. The latter emphasizes that each application and workload may need a different type of data store, tailored for its needs (e.g. graph, time series). Moreover, the field of data warehousing addresses creating snapshots of Online Transactional Processing (OLTP) databases for the purposes of Online Analytical Processing (OLAP). This often requires copying the data and preparing it for analytics by transforming its structure (i.e. Extract Transform Load process) and building the relevant indexes for the analytical queries. This costly process is performed in order to achieve fast query times for analytical queries of interest, and in order to support data mining. However, there is an increasing trend to adopt a "just in time data warehouse" model, where data are federated on the fly according to runtime parameters and constraints [8]. As a result, data analytics frameworks increasingly strive to cater for data regardless of the underlying data store. Apache Spark [9] is an open source framework for analytics, which is designed to run in a distributed setting within a data centre. Spark provides a framework for cluster computing and memory management, and invents the notion of a Resilient Distributed Dataset (RDD) [10] that can be stored persistently using any storage framework that implements the Hadoop File System interface, including the Hadoop Distributed Filesystem (HDFS), Amazon S3 and OpenStack Swift. The Spark SQL component additionally defines a DataFrame as an RDD having a schema, and provides a SQL interface over DataFrames [11]. Built in support is provided for data sources with a JDBC interface, as well as for Hive, and the Avro, Parquet, ORC and JSON formats. Moreover, there is an external data sources API, where new data sources can be added by implementing a driver for the data source that implements the API. Many such drivers have been implemented, for example for Cassandra, MongoDB and Cloudant. These data sources can be queried, joining data across them and thus provide the ability to run batch queries across multiple data sources and formats. Spark also integrates batch processing with real time processing in the form of Spark Streaming [12] that allows real-time processing using the same underlying framework and programming paradigm as for batch computations. In Spark 2.0, streaming Spark SQL computations are also planned [13]. With the advent of IoT and the increasing capabilities available at the edge, applications may store and process data locally [14]. In the PolicyCLOUD architecture, data are managed whether in flight or at rest and are federated across multiple frameworks, data sources, locations and formats.

Another key aspect is data interoperability given the diversity of the data sources. Among the main value propositions of the PolicyCLOUD environment and tools for policy development and management will be its ability to integrate, link and unify the datasets from diverse sources, while at the same time enabling analytics over the unified datasets. As a key prerequisite to providing this added-value, the interoperability of diverse datasets should be ensured. A wide array of data representation standards in various domains have emerged as a means of enabling data interoperability and data exchange between different systems. Prominent examples of such standards in different policy areas include: (i) the INSPIRE Data Specifications [15] for the interoperability of spatial data sets and services, which specify common data models, code lists, map layers and additional metadata on the interoperability to be used when exchanging spatial datasets, (ii) the Common European Research Information Format (CERIF) [16] for representing research information and supporting research policies, (iii) the Internet of Things ontologies and schemas, such as the W3C Semantic Sensor Networks (SSN) ontology [17] and data schemas developed by the Open Geospatial Consortium (e.g., SensorML) [18], (iv) the Common Reporting Standard (CRS) that specifies guidelines for obtaining information from financial institutions and automatically exchanging that information in an interoperable way, and (v) standards-based ontologies appropriate for describing social relationships between individuals or groups, such as the "The Friend Of A Friend" (FOAF) ontology [19] and the Socially Interconnected Online Communities (SIOC) ontology [20]. These standards provide the means for common representation of domain specific datasets, towards data interoperability (including in several cases semantic interoperability) across diverse databases and datasets. Nevertheless, these standards are insufficient for delivering what PolicyCLOUD promises for a number of reasons. Initially, there is a lack of semantic interoperability in the given domain. For example, compliance to ontologies about IoT and sensor data fails to ensure a unified modelling of the physics and mathematics, which are at the core of any sensing task. Hence, in several cases there is a need for extending existing models with capabilities for linking/relating various quantifiable and measurable (real-world) features to define, in a user understandable and machine-readable manner the processes behind single or combined tasks in the given domain. Furthermore, there is a lack of semantic interoperability across datasets from different sectors. There is not easy way to link related information elements stemming from datasets in different sectors, which typically comprise different schemas. In this context, environmental datasets and transport datasets for instance contain many related elements, which cannot however be automatically identified and processed by a system due to the lack of common semantics. Finally, one needs to consider the lack of process interoperability. PolicyCLOUD deals with data-driven policy development and management, which entails the simulation and validation of entire processes. Especially in the case of multi-sectoral considerations (e.g., interaction and trade-offs between different policies) process interoperability is required in order to assess the impact of one policy on another. PolicyCLOUD proposes a multi-layer framework for interoperability across diverse policy related datasets, which will facilitate semantic interoperability across related datasets both within a single sector and across different policy sectors. Within a specific sector of each use case, semantic interoperability will enable adhering to existing standards-based representations for the sector data and other auxiliary data (e.g. sensor data, social media data). Across different use cases, PolicyCLOUD proposes a LinkedData approach [21] to enable linking of interrelated data across different ontologies.

# 3 Main Challenges and Proposed Approach

## 3.1 Main Challenges and Objectives Addressed by PolicyCLOUD

*A Data-Driven Approach for Effective Policies Management*
The challenge is to provide a scalable, flexible and dependable methodology and environment for facilitating the needs of data-driven policy modelling, making and evaluation. The methodology should aim at applying the properties of policy modelling, co-creation and implementation across the complete data path, including data modelling, representation and interoperability, metadata management, heterogeneous datasets linking and aggregation, analytics for knowledge extraction, and contextualization. Moreover, the methodology should exploit the collective knowledge out of policy “collections”/clusters combined with the immense amounts of data from several sources (e.g. sensor readings, online platforms, etc.). These collections of policies should be analysed based on specific Key Performance Indicators (KPIs) in order to enable the correlation of these KPIs with different potential determinants of policies impact within and across different sectors (e.g. environment, radicalisation, migration, goods and services, etc.).

*Compilation, Assessment and Optimization of Multi-domain Policies*
Another challenge refers to holistic policy modelling, making and implementation in different sectors (e.g. environment, migration, goods and services, etc.), through the analysis and linking of KPIs of different policies that may be inter-dependant and inter-correlated (e.g. environment). The goal is to identify (unexpected) patterns and relationships between policies (through their KPIs) to improve policy making. Moreover, the approach should enable evaluation and adaptation of policies by dynamically extracting information from various data sources, community knowledge out of the collections of policies, and outcomes of simulations and evidence-based approaches. Policies should be evaluated to identify both their effective KPIs to be re-used in new/other policies, and the non-effective ones (including the causes for not being effective) towards their improvement. Thus, developed policies should consider the outcomes of strategies in other cases, such as policies addressing specific city conditions.

*Data Management Techniques Across the Complete Data Path*
Data-driven policy making highlights the need for a set of mechanisms that address the data lifecycle, including data modelling, cleaning, interoperability, aggregation, incremental data analytics, opinion mining, sentiment analysis, social dynamics and

behavioural data analytics. In order to address data heterogeneity from different sources, modelling and representation technologies should provide a "meta-interpretation" layer, enabling the semantic and syntactic capturing of data properties and their representation. Another key aspect refers to techniques for data cleaning in order to ensure data quality, coherence and consistency including the adaptive selection of information sources based on evolving volatility levels (i.e. changing availability or engagement level of information sources). Mechanisms to assess the precision and correctness of the data, correct errors and remove ambiguity beyond limitations for multidimensional processing should be incorporated in the overall environment, while taking into consideration legal, security and ethical aspects.

*Context-Aware Interoperability*
The specific challenge refers to the design and implementation of a semantic layer that will address data heterogeneity. To this end, the challenge is to research on techniques and semantic models for the interoperable use of data in different scenarios (and thus policies), locations and contexts. Techniques for interoperability (such as OSLC - Open Services for Lifecycle Collaboration) with different ontologies (as placeholders for the corresponding information) should be combined with semantic annotations. Semantic models for physical entities/devices (i.e. sensors related to different policy sectors), virtual entities (e.g. groupings of such physical entities according to intrinsic or extrinsic, permanent or temporary properties) and online platforms (e.g. social media, humans acting as providers) should be integrated in data-driven policy making environments. These models should be based on a set of transversal and domain-specific ontologies and could provide a foundation for high-level semantic interoperability and rich semantic annotations across policy sectors, online systems and platforms. These will be turned into rich metadata structures providing a paradigm shift towards content-based storage and retrieval of data instead of data-based, given that stakeholders and applications target and require such content based on different high-level concepts. Content-based networks of data objects need to be developed, allowing retrieval of semantically similar contents.

*Social Dynamics and Incentives Management*
One of the main barriers to public bodies experimenting with big data to improve evidence-based policy making is citizens' participation since they are lacking awareness of the extend they may influence policy design and the ways these will be feasible. The challenge is to raise awareness about policy consultations and enable citizens to take direct action to participate, thus ensuring higher levels of acceptance and of trust. A potential solution could be to follow a living lab approach and implement an engagement strategy based on different incentives mechanisms. Furthermore, a data-driven policy management environment should allow social dynamics and behaviour to be included in the policy lifecycle (creation, adaptation, enforcement, etc.) through the respective models and analytical tools. These will allow policy makers to obtain the relevant crowd-sourcing data and the knowledge created by the closed groups (i.e. communities evaluating proposed policies) and the engaged citizens to analyse and propose social requirements that will be turned into policy requirements. On top of this, incentives management techniques will identify, declare and manage incentives for citizens' engagement, supporting different types of incentives (e.g. social, cultural, political, etc.),

with respect to information exchange, contributions and collaboration aspirations. The environment should also provide strategies and techniques for the alignment participation incentives, as well as protocols enabling citizens to establish their participation.

*Analytics as a Service Reusable on Top of Different Datasets*
Machine and deep learning techniques, such as classification, regression, clustering and frequent pattern mining algorithms, should be realized in order to infer new data and knowledge. Sentiment analysis and opinion mining techniques should determine whether the provided contributor's input is positive or negative about a policy, thus developing a "contributor graph" for the contributors of the opinions that happen to be themselves contributors to ongoing policy making projects. In the same context, social dynamics and behavioural data analytics should provide insights regarding which data needs to be collected and aggregated in a given case (e.g. time window addressed by the policy, location of populations, expected impact, etc.), taking into consideration the requirements of the engaged citizens to model the required policies. Moreover, a main challenge refers to technologies that allow analytics tasks to be decoupled from specific datasets and thus be triggered as services and applied to various cases and datasets.

*Transferable Methods and a Unique Endpoint to Exploit Analytics in Different Cases*
A key challenge refers to an overall system, acting as an endpoint that will allow stakeholders (such as policy makers and public authorities) to trigger the execution of different models and analytical tools on their data (e.g. to identify trends, to mine opinion artefacts, to explore situational and context awareness information, to identify incentives, etc.) and obtain the results. Based on these results, the modelled policies (through their KPIs) will be realized/implemented and monitored against these KPIs. Moreover, the endpoint should allow stakeholders and public administration entities to express in a declarative way their analytical tasks and thus perform/ingest any kind of data processing. Another need is for an adaptive visualization environment, enabling policy monitoring to be visualized in different ways while the visualization can be modified on the fly. The environment should also enable the specification of the assets to be visualized: which data sources and which meta-processed information. The goal is to enable the selection of sources based on the stakeholders' needs. Incremental visualization of analytics outcomes should also be feasible enabling visualization of results as they are generated.

### 3.2 Architecture Overview

As a complete environment, the proposed architectural approach includes a set of main building blocks to realize the corresponding functionality as depicted in the following figure (Fig. 1).

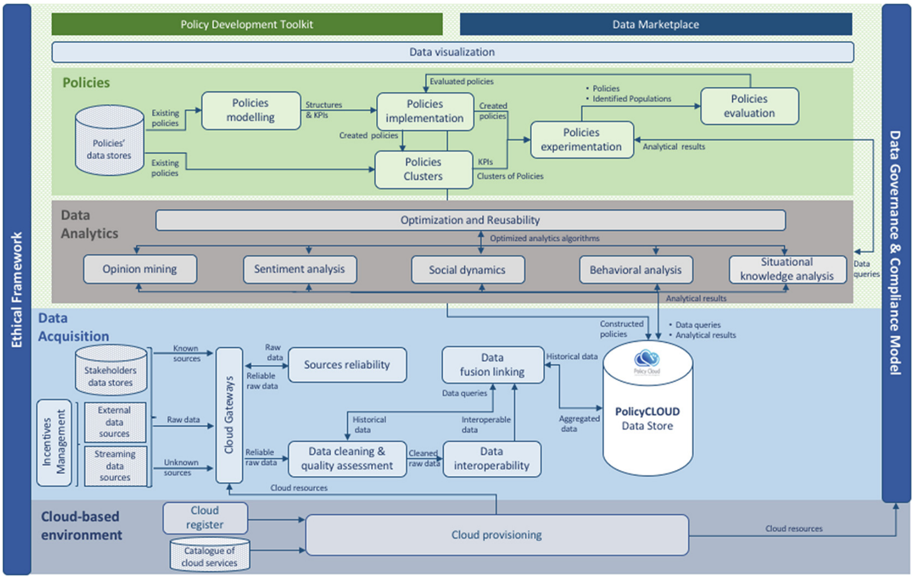

**Fig. 1.** Proposed functional architecture.

The overall flow is initiated from various data sources, as depicted in the figure through the respective *Data Acquisition* block. Data sources can be data stores from public authorities or external data sources (e.g. mobile devices, IoT sensors, etc.) that contribute data following the provision of incentives, facilitated through the *incentives management* mechanism. A set of APIs incorporated in a gateway component, enables data collection by applying techniques to identify the reliable sources exploiting the *sources reliability* component and for these sources obtain the data and perform the required *data quality assessment and cleaning*. *Semantic and syntactic interoperability* techniques are utilized over the cleaned data providing the respective interoperable datasets to the Policy CLOUD datastore following the required *data linking and aggregation* processes. The datastore is accessible from a set of machine learning models represented through the *Data Analytics* building block. Machine learning models incorporate opinion mining, sentiment and social dynamic analysis, behavioural analysis and situational/context knowledge acquisition. The data store and the analytics models are hosted and executed in a *cloud-based environment* that provides the respective services obtained from a catalogue of cloud infrastructure resources. Furthermore, all the analytics models are realized as services, thus enabling their invocation through a proposed policy development toolkit – realized in the scope of the *Policies* building block of the proposed architecture as a single point of entry into the PolicyCLOUD platform. The toolkit allows the compilation of *policies as data models*, i.e. structural representations that include key performance indicators (KPIs) as a means to set specific parameters (and their target values) and monitor the implementation of policies against these KPIs along with the list of analytical tools to be used for their computation. According to these analytics outcomes, the values of the KPIs are specified resulting to *policies implementation/creation*. It should be noted that PolicyCLOUD also introduces the concept of *policies clusters* in order to interlink different policies, and identify the KPIs and parameters that can be optimized in such policy collections. Across the complete environment, an implemented *data governance and compliance model* is enforced, ranging from the provision of cloud resources regarding the storage and analysis of data to the management of policies across their lifecycle.

## 4 Conclusions

The vast amounts of data that are being generated by different sources highlight an opportunity for public authorities and stakeholders to create, analyse, evaluate and optimize policies based on the "fresh" data, the information that can be continuously collected by citizens and other sensors. To this end, what is required refers to techniques and an overall integrated environment that will facilitate not only data collection but also assessment in terms of reliability of the data sources, homogenization of the datasets in order to make them interoperable (following the heterogeneity in terms of content and formats of the data sources), cleaning of the datasets and analytics. While several analytical models and mechanisms are being developed a key challenge relates to approaches that will enable analytics to be triggered as services and thus applied and utilized in different datasets and contexts. In this paper, we have presented the aforementioned challenges and the necessary steps to address them. We also introduced a conceptual architecture that depicts a holistic cloud-based environment integrating a set of techniques across the complete data and policy management lifecycles in order to enable data-driven policy management. It is within our next steps to implement the respective mechanisms and integrate them based on the presented architecture, thus realizing the presented environment.

**Acknowledgment.** The research leading to the results presented in this paper has received funding from the European Union's funded project PolicyCLOUD under Grant Agreement no 870675.

## References

1. How can social media data be used to improve services? https://www.theguardian.com/local-government-network/2013/oct/03/social-media-improve-services-data
2. Data Growth, Business Opportunities, and the IT Imperatives. https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm
3. Anderson, J.E.: Cases in Public Policy-Making. Praeger, New York (1976)
4. Setting Data Policies, Standards, and Processes. https://www.mcpressonline.com/analytics-cognitive/business-intelligence/setting-data-policies-standards-and-processes
5. Data for Policy: A study of big data and other innovative data-driven approaches for evidence-informed policymaking. http://media.wix.com/ugd/c04ef4_20afdcc09aa14df38fb646a33e624b75.pdf
6. Big Data: Basics and Dilemmas of Big Data Use in Policy-making. http://www.policyhub.net/en/experience-and-practice/153
7. What is Polyglot Persistence? http://www.jamesserra.com/archive/2015/07/what-is-polyglot-persistence

8. Building a Just-In-Time Data Warehouse Platform with Databricks. https://databricks.com/blog/2015/11/30/building-a-just-in-time-data-warehouse-platform-with-databricks.html
9. Apache Spark Homepage. http://spark.apache.org
10. Zaharia, M., et al.: Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, Berkeley (2012)
11. Armbrust, M., et al.: Spark SQL: relational data processing in spark. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD 2015), pp. 1383–1394. ACM, New York (2015)
12. Zaharia, M., Das, T., Li, H., Hunter, T., Shenker, S., Stoica, I.: Discretized streams: fault-tolerant streaming computation at scale. In: Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, pp. 423–438. ACM, New York (2012)
13. Structuring Spark: DataFrames, DataSets and Streaming. https://databricks.com/session/structuring-spark-dataframes-datasets-and-streaming
14. Bonomi, F., Milito, R., Zhu, J., Addepalli, S.: Fog computing and its role in the internet of things. In: Proceedings of the first edition of the MCC workshop on Mobile cloud computing (MCC 2012), pp. 13–16. ACM, Helsinki (2012)
15. Data Specifications. http://inspire.ec.europa.eu/data-specifications/2892
16. Cordis. https://cordis.europa.eu/about/archives
17. Incubator. https://www.w3.org/2005/Incubator/ssn/ssnx/ssn
18. Open Geospatial Standards. https://www.ogc.org/standards/sensorml
19. FOAF Homepage. http://www.foaf-project.org
20. SIOC project. http://sioc-project.org
21. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space, 1st edn. Morgan & Claypool, San Rafael (2011)