

Innovation with open data: Essential elements of open data ecosystems

Anneke Zuiderwijk*, Marijn Janssen and Chris Davis

Faculty of Technology, Policy and Management, Delft University of Technology, Delft, The Netherlands

Abstract. Open data ecosystems are expected to bring many advantages, such as stimulating citizen participation and innovation. However, scant attention has been given to what constitutes an open data ecosystem. The objective of this paper is to provide an overview of essential elements of open data ecosystems for enabling easy publication and use of open data. To achieve this objective, the literature has been reviewed and a scenario about the publication and use of open data has been analyzed. It was found that various applications, tools and portals are available which together can form an ecosystem. The best functionalities of this ecosystem can be selected and utilized by open data providers and users. To create an open data ecosystem at least four key elements should be captured, namely, 1) releasing and publishing open data on the internet, 2) searching, finding, evaluating and viewing data and their related licenses, 3) cleansing, analyzing, enriching, combining, linking and visualizing data and 4) interpreting and discussing data and providing feedback to the data provider and other stakeholders. Furthermore, to integrate the ecosystem elements and to let them act as an integrated whole, there should be three additional elements 5) user pathways showing directions for how open data can be used, 6) a quality management system and 7) different types of metadata to be able to connect the elements.

Keywords: Open data ecosystem, open data, open government, ecosystem, architecture, infrastructure, innovation

1. Introduction

Open data is expected to bring many advantages, such as stimulating citizen participation and innovation [1,2], stimulating transparency [3,4] and stimulating economic growth [5,6]. In this way, a more open government can be encouraged. Various open data portals and infrastructures have been developed in the last years to explore the potential of open data [7], such as national open data portals, the European open data infrastructure [8], the ENGAGE open data infrastructure [9], the Junar open data infrastructure [10] and the infrastructures of statistics agencies (e.g. Eurostat [11]). These digital government infrastructures are typically used by many different users, evolve over time and need a large user base, as they offer value only when a certain critical mass of users has been reached [12]. In addition, there are many tools and programs available which can help open data users, such as Open Refine [13] and ScraperWiki [14].

While these infrastructures and tools are important, they often provide only one part of the puzzle for open data users which often deploy more than one infrastructure or tool to process their open data. As such we plea for viewing them as part of a wider open data ecosystem in which each instrument and tool can add value. Ubaldi [15] argues that “building an ecosystem that responds to specific demands asking

*Corresponding author: Anneke Zuiderwijk, Faculty of Technology, Policy and Management, Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands. Tel.: +31 15 27 86471; E-mail: a.m.g.zuiderwijk-vaneijk@tudelft.nl.

for the provision of open government data are a necessity to create value out of OGD [added: open government data]” [p. 27]. Open government ecosystems can be used as a means for decision-making and planning [16], such as decision-making by policy-makers. They typically contain “interdependent social systems of actors, organizations, material infrastructures, and symbolic resources that can be created in technology-enabled, information-intensive social systems” [p. 900].

Up until now only few scholars have paid attention to open data ecosystems. Only little research has been performed on the essential components of open data ecosystems, while these ecosystems are important, as they may help in realizing the advantages of open data and creating value [15]. The idea of open ecosystems has been given attention in blogs [17], working papers [15], short papers [18] and reports [19,20]. There is some literature on open data ecosystems, but this literature is often focused on certain types of open data (e.g. bioscience data [21] or aid data [22]), or does not provide a clear description of what the components of an open data ecosystem should be [23]. There is hardly any literature viewing open data as an ecosystem and we did not find any scientific literature identifying the various key elements of open data ecosystems. This paper aims to fill this gap. The objective of this paper is to provide an overview of essential elements of open government data ecosystems for enabling easy publication and use of open data. For convenience’s sake we use the terms ‘open government data’ and ‘open data’ interchangeably in the remainder of this paper, yet this manuscript is focused on data opened by governmental organizations.

This paper is organized as follows. In the following section the research approach is described, followed by an investigation of the literature about ecosystems, open government ecosystems and open data ecosystems, which results in an overview of ecosystem elements that need to be investigated in the scenario. Section four presents a scenario about publishing and using open data in which elements of open data ecosystems are investigated. The results of the literature and scenario analysis are discussed in section five, leading to an overview of the essential elements of open data ecosystems. Finally, section six provides the conclusions drawn and the recommendations derived from this research.

2. Research approach

The objective of this paper is to provide an overview of essential elements of open data ecosystems for enabling publication and use of open data. We attain this objective by making use of two methods. First, we conduct a literature review which results in an overview of the most important elements of ecosystems derived from the literature. Second, we develop a scenario for the publication and use of open data, and use this scenario to identify whether additional elements were required for open data ecosystems which were not found in the literature. We report on these two methods in the following sections.

2.1. Literature review

A literature overview about open data ecosystems was created by seeking for journal and conference articles, books, governmental and non-governmental reports and other information in various databases, including Science Direct, Scopus, Google Scholar, Emerald Insight and Springer. The keywords open data ecosystem, government ecosystem and ecosystem were used for searching these databases. In addition, we used the cited references in the articles that we identified to further enrich and enlarge the literature bases. The found documents were filtered by searching for ecosystems and open data. In total, 26 publications were selected that were relevant for this research, and from these a literature overview

was created (these 26 documents can be found in Table 1 in Section 3.4). The relevance of the documents was determined by scanning the titles and abstracts of the retrieved documents. The literature review resulted in an overview of elements that are relevant to investigate in our scenario about the publication and use of open data.

2.2. Scenario

The second method used in this research is the development of a scenario to derive additional and more detailed elements of an open data ecosystem. Carroll [24] writes that scenarios can be defined as stories about people and their activities. In line with this, scenarios are narrative descriptions of interactions between users and proposed systems [25]. More specifically, “scenarios highlight goals suggested by the appearance and behavior of the system, what people try to do with the system, what procedures are adopted, not adopted, carried out successfully or erroneously, and what interpretations people make of what happens to them” [24, p. 2]. Scenarios can be used for various purposes in the interactive system development processes, and can be written from many perspectives and at multiple levels [24,26]. For instance, scenarios can evoke reflection about design issues, as they provide descriptions of end-user experiences. Furthermore, scenarios can be abstracted and categorized to create knowledge to discuss problems. Additionally, scenarios can demonstrate concretely an interpretation and a solution, while remaining open-ended and easy to revise [24]. These characteristics make scenario-based design an appropriate approach for this research, as our aim is to identify various elements and the roles of these elements for the design of desired open data ecosystems.

Carroll [24] states that scenarios are characterized by the description of a setting and the description of agents or actors. In addition, a description is included of the goals or objectives of each agent or actor. Moreover, scenarios include sequences of actions and events, which means that they reveal what things the actors do, what happens to them and which changes occur in the circumstances of the setting. Finally, scenarios typically have a plot [24]. These elements will be described for the scenario that is presented in this manuscript.

One comprehensive scenario was developed in this research. The aim of the scenario was to identify a wide range of activities that can be performed in the so-called open data process, and to identify the elements of open data ecosystems that can be used to enable and support these activities. This means that the scenario needed to cover various activities that are performed in the processes in which open data are created, published, found, used and discussed. To get an impression of which activities are needed to be covered exactly, we searched through the open data literature and we investigated various open data infrastructures, such as national open data portals, the European open data infrastructure [8], the ENGAGE open data infrastructure [9], the Junar open data infrastructure [10] and the infrastructure of Eurostat [11] and Statistics Netherlands [27]. The requirement to create diversity in the activities of the open data scenario was formulated because diversity is expected to lead to the identification of more elements.

Furthermore, we used steps of data and information workflows and data lifecycle management described in various articles to identify which activities could be conducted in an open data scenario. A lifecycle of data on the Semantic Web includes actions such as data creation, publication, exportation, importation, use, transformation and reuse [28]. These activities can be conducted by different parties and for different purposes [ibid]. From an open data perspective, such a lifecycle could consist of the steps creating data, opening data, finding data, using data and providing feedback on and discussing data [29]. Ding et al. [30] identify the steps of initializing an open government dataset, retrieving data,

cleaning up and converting and enhancing datasets and publishing datasets after having conducted these steps. An attempt to describe or even standardize the processes involved in information flows concerns the CRISP-DM methodology for data mining, which involves multiple processing steps of data, including collecting data, preparing data for analysis, fitting data to a particular model and evaluating the results [31]. Another attempt has been done by Fry [32], who argues that the process of creating interactive data visualizations involves connecting a chain of tools that have been developed by different disciplines such as Computer Science, Statistics, Graphic Design, and Human Computer Interaction. In addition, Pirolli and Card [33] describe what they term the “Sensemaking Process” where for analysts to gain insight, they are involved in a process of gathering, filtering, and structuring data.

Common among all these visions is that they describe processes that are connected mostly in a linear fashion, with feedback loops sometimes leading from later processing stages back to earlier stages. They also describe systems where information is upgraded in many ways as it passes through each of the steps. This can be thought of as a sort of supply chain where tools or people in succeeding processing steps add value to the data. The same types of processes are occurring with open data ecosystems with the caveat that there may be many more parties involved, that the system is open and available to the public instead of being closed, and instead of having a single linear process, there is more of an ecosystem composed of different tools and groups who all use data for different purposes. We used this literature about data and information workflows and data lifecycle management in the development of the scenario about open data publication and use.

3. Literature background

In this section we describe the findings from the literature related to ecosystems in general (Section 3.1), open government ecosystems (Section 3.2) and open data ecosystems (Section 3.3). Finally, we summarize the essential elements of ecosystems found in the literature review (Section 3.4).

3.1. Ecosystems

Adner and Kapoor [34] show that an ecosystem lens can be used to make interdependencies among partners in exchange networks characterized by simultaneous cooperation and competition more explicit. Open data could be seen as such an exchange network. For instance, in an open data network there could be cooperation of various actors to facilitate the use of open government data. Open data providers usually aim to stimulate the use of their data and they often need data users to enable this. Businesses may need government data to develop services for end-users. Furthermore, there could be competition between businesses using open data, for example, to obtain open data end-users as (paying) customers for services that they have developed based on open government data. There could also be competition between open data providers, since they may want to promote their organization by stating that they open larger amounts of data or more datasets than other governmental organizations do.

Ecosystems in general can be defined as “a functioning whole in a given area” [20, p. 13] and refer to the dynamic interaction between different factors in an area [20]. This definition suggests that an ecosystem needs to combine components from different domains, since they can be seen as ‘a functioning whole’. The boundaries of (parts of) ecosystems are difficult to define and typically do not correspond to traditional industry boundaries, but are “instead defined by the strength and type of organizational interactions that occur” [35, p. 39]. The effects of the health and dynamics of an ecosystem do not correspond

with traditional industry boundaries. For example, ecosystems may be characterized by the sharing of tools and technological components or by buyer/supplier interactions [ibid].

Harrison, Pardo and Cook [16] state that the metaphor ecosystem is often used “to convey a sense of the interdependent social systems of actors, organizations, material infrastructures, and symbolic resources that can be created in technology-enabled, information-intensive social systems” [p. 900]. In addition to social factors, the literature shows that the interaction of social and technical factors may influence the outcomes [36–38]. Essential elements in ecosystem management are stakeholder collaboration, interagency cooperation, integration of scientific, social, and economic information, preservation of ecological processes, and adaptive management [39]. Stakeholders involved in ecosystems are agencies and departments, suppliers and partners, internal elements, communities and individual stakeholders [40]. In many ecosystems there are also intermediaries involved, who “are positioned between the innovation and the final customer” [41, p. 5]. The elements of an ecosystem can rely directly or indirectly on the elements of infrastructures [22]. Even though ecosystems may have various benefits, there may also be certain risks and ecosystem innovation may be costly [41]. “Creating strategy that explicitly accounts for the delays and challenges that are inherent in collaborative networks is the key to succeeding in ecosystems” [41, p. 3].

Open data ecosystems could partly be seen as business ecosystems. Moore defines a business ecosystem as “an economic community supported by a foundation of interacting organizations and individuals – the organisms of the business world” [42, p. 9]. Business ecosystems are characterized by “loosely interconnected participants who depend on each other for their mutual effectiveness and survival” [35, p. 40]. The strength and nature of interactions among participants is important. Among these participants are suppliers, customers and vast numbers of complementors providing complementary goods and services [42,43] and lead producers and other stakeholders [42]. Firms and other actors may interact even if they are separated in space and time [35]. Furthermore, ecosystems are made up of interrelated components, which can be adapted to particular contexts [19]. Various elements of the ecosystem affect the healthiness and richness of the system [44], such as interrelated tools and services [22]. The elements of an ecosystem emerge, are autonomous and self-organizing. These elements are connected to each other in feedback loops and develop mainly through local specializations and adaptation rather than through top-down design [22].

Although open data ecosystems do not only involve private organizations but also public organizations, open data ecosystems may also be viewed as a kind of innovation ecosystems, which can be defined as “the collaborative arrangements through which firms combine their individual offerings into a coherent, customer-facing solution” [41, p. 2]. Innovation ecosystems could provide various benefits, such as allowing companies and citizens to create value that they could not have created alone [ibid]. Open data ecosystems can also be seen as a kind of information ecosystem. An information ecosystem is a complex system in a certain environment consisting of people, work, value and technology, with various interconnections. The emphasis in information ecosystems is on the behavior of people supported by technology [45,46]. An information ecosystem includes information systems, databases, workflows, people and an infrastructure [47]. Open data ecosystems, in addition, may involve software ecosystems, which refer to “a networked community of organizations, which base their relations to each other on a common interest in a central software technology” [48, p. 514].

Moreover, open data ecosystems can be seen as a type of digital ecosystem, which are composed of “interconnected, interrelated and interdependent digital species situated in a digital environment, that interact as a functional unit and are linked together through actions, information and transaction flows” [49, p. 7]. Digital ecosystems are considered to be robust, scalable, and self-organizing and decentralized peer-to-peer networks form an underlying tier of distributed agents [50].

3.2. *Open government ecosystems*

Open data ecosystems are also often government ecosystems, as much open data is published by the government, although elements for open data ecosystems can also be provided by the private sector. Open government ecosystems emphasize the multiple and varying interrelationships between data, open data providers, open data users, material infrastructures and institutions. The connected components are considerably interdependent [16]. Three important domains of interdependencies defined by Harrison et al. [16] are 1) government policies and practices, 2) innovators; a combination of technology, business and government and 3) users, civil society and business. Abrahams [51] writes that the key elements influencing the development of an e-government ecosystem are 1) an information, communication and networking technology infrastructure, 2) e-government, e-business and e-society, 3) human resource development and 4) a policy and regulatory environment. Obi [52] adds to this that e-government ecosystems should include Information and Communication Technologies (ICT) preparedness, ICT usage, an enterprise architecture ICT model, Information provision capability, policy marketing, resource allocation and e-participation [52]. As such, an ecosystem can stimulate user's participation in public affairs or policy making processes [44].

In summary, important characteristics of open government ecosystems are that they consist of interdependent systems and components in which various actors are involved. Taking into account these characteristics, open data can be seen as an ecosystem, as open data includes interdependent technical and social systems. An ecosystem is needed which is able to integrate various tools to enable providing and using open data. In the ecosystem various actors, such as open data providers, users and legislators play a role. In the open data ecosystem, ICT enables the discussion of open data and stimulates the participation of citizens in governmental processes of decision making and policy making.

3.3. *Open data ecosystems*

In this section various characteristics of open data ecosystems are described. Poikola et al. [20, p. 13] define an open data ecosystem as “a multi-level and multidimensional entity where raw material, as far as distribution and developing are concerned, is the target of cooperation”. Equal to open government ecosystems, open data ecosystems are characterized by the fact that the issues related to it are never entirely independent of other issues. Dependent issues in open data ecosystems are pre-existing networks of interaction, existing power and information asymmetries, levels of training and capacity, risk assessment in organizations [19]. For instance, open data users depend on the data and metadata that are published by the open data providers to be able to make use of these data. Open data providers depend on the feedback on their data that they obtain from open data users. In this way, the open data providers and open data users are involved in networks of interaction. Essential challenges that need to be addressed in open data ecosystems relate to policy, technology, financing, organization, culture, and legal frameworks [15].

Pollock [17] states that open data ecosystems contain data cycles with feedback loops, sharing of data back to publishers and sharing between so-called infomediaries. The latter are intermediate consumers of data such as builders of apps and data wranglers. Infomediaries add value to datasets by cleaning, analyzing and integrating them. Sharing these reused data means that value will be shared with many people. Pollock describes three main changes which are required to develop an open data ecosystem, namely 1) infomediaries should publish what they produce, 2) better ways to publish and share data and 3) publisher notification of patches (pull requests) with automated integration (merge) tools [17].

Ding et al. [30] refer to an ecosystem for linked open government data (LOGD) and define this as “a Linked Data-based system where stakeholders of different sizes and roles find, manage, archive, publish, reuse, integrate, mash-up, and consume open government data in connection with online tools, services and societies” [30, p. 326]. Moreover, they point out that effective LOGD ecosystems need to include many types of stakeholders such as civil servants who curate raw data, developers who build applications with open data, and citizens who view visualizations and analytical results from these data. In general, Ding et al. make a distinction between LOGD production, LOGD consumption and the LOGD community [ibid].

Ubaldi identifies three types of interacting open data ecosystems, namely 1) an ecosystem of data producers, 2) an ecosystem of infomediaries as intermediate consumers of data and 3) an ecosystem of open data users [15]. Open data ecosystems may provide data curation and sharing solutions in the form of “a common metadata tracking framework, providing tools and resources to create and manage large, heterogeneous data sets in a coherent manner, and allowing users of (parts of) data sets to ‘connect the metadata dots’” [21, p. 8]. Licensing is also an important aspect of open data ecosystems, as “a lack of licensing makes the emergence of an open-data ecosystem practically impossible, since there would be no clear rules on who can use data and how” [53, p. 53]. Sometimes linked data is seen as an open data ecosystem on itself [18,54]. For instance, Ding writes that linked open government data interconnects original data owners, service providers and data consumers [54].

3.4. Summary

The findings from the literature review on the characteristics of ecosystems are summarized in Table 1. This table shows that the research on open data ecosystems is fragmented and not uniform, as most research only focuses on one or a few activities related to the publication or use of open data, but not on the whole of activities that should be supported in an open data ecosystem. In addition, the elements of ecosystems that are derived from the literature are mainly on a high and abstract level, while a more detailed overview of elements is needed to create an open data ecosystem. We conclude that open data ecosystems are characterized by multiple interdependent socio-technical levels, dimensions, actors (including data providers, infomediaries and users), elements and components. Moreover, open data ecosystems need to address challenges related to policy, licenses, technology, financing, organization, culture, and legal frameworks and are influenced by ICT infrastructures. Open data ecosystems develop through user adaptation, feedback loops and dynamic supplier and user interactions and other interacting factors. In the development of the scenario in the following section we take these characteristics into account.

Based on Table 1, we visualized potential elements of open data ecosystems, as can be seen in Fig. 1. At the top open data suppliers are shown, whereas the bottom contains the various types of users. Open data infomediaries can participate and support both of these levels, namely by supporting data publication and by supporting data use. We use this table as a basis for the following section, and we extend it by deriving more elements for the ecosystem from the description of a scenario of open data publication and use.

4. Generic scenario of open data publication and use

Our investigation of the literature provided various elements of open government data ecosystems, as can be seen from Fig. 1. Since these elements are at a relatively high level of abstraction, we use a

Table 1
An overview of the characteristics of ecosystems derived from the literature

Ecosystems characteristics		Source
Design	A functioning whole	Poikola, Kola and Hintikka [1]
	Multi-level and multidimensional	Poikola, Kola and Hintikka [1]
	Boundaries are difficult to define and typically do not correspond to traditional industry boundaries	Iansiti and Levien [35]
	Develop mainly through local specializations and adaptation rather than top-down	Davies [22]
	“Interdependent social systems of actors, organizations, material infrastructures, and symbolic resources that can be created in technology-enabled, information-intensive social systems” [p. 900]	Harrison, Pardo and Cook [16]
	Can rely directly or indirectly on the elements of infrastructures	Davies [22]
	Emphasis on the behavior of people supported by technology	Nardi and O’Day [45], Weizhang and Jiefang [46]
	Influenced by 1) an information, communication and networking technology infrastructure, 2) e-government, e-business and e-society, 3) human resource development and 4) a policy and regulatory environment.	Abrahams [51]
Context	Needs to address challenges related to policy, technology, financing, organization, culture, and legal frameworks	Ubaldi [15]
	Can stimulate user’s participation in public affairs or policy making processes	Obi and Iwasaki [44]
Interdependencies and interaction	Interdependent elements (e.g. networks of interaction, power and information asymmetries, levels of training and capacity, risk assessment in organizations) and interdependent components	Ubaldi [15], Hall et al. [19]
	Dynamic interaction between different factors (e.g. organizations and individuals) in a certain area	Poikola, et al. [20], Moore [42]
	Exchange networks characterized by simultaneous cooperation and competition	Adner and Kapoor [34]
	Characterized by the sharing of tools and technological components or by buyer/supplier interactions	Iansiti and Levien [4]
	Data cycles with feedback loops, sharing of data back to publishers and sharing between so-called infomediaries	Pollock [17]
	Stakeholder collaboration, interagency cooperation, integration of scientific, social, and economic information, preservation of ecological processes, and adaptive management	Koontz and Bodine [39]
	“A networked community of organizations, which base their relations to each other on a common interest in a central software technology” [p. 514].	Mhamdia [48]
	Composed of “interconnected, interrelated and interdependent digital species situated in a digital environment, that interact as a functional unit and are linked together through actions, information and transaction flows” [p. 7]	Hadzic and Sidhu [49]
	Robust, scalable, and self-organizing and decentralized peer-to-peer networks form an underlying tier of distributed agents	Briscoe and De Wilde .[50]
	Emphasize the multiple and varying interrelationships between data, open data providers, open data users, material infrastructures and institutions	Harrison et al. [16]
The connected components are considerably interdependent	Harrison et al. [16]	
Participants	Participants are suppliers, customers and complementors providing complementary goods and services	Moore [42], Pierce [43]
	Interactions between data producers, users and infomediaries as intermediate data consumers	Ubaldi [15]
	Consisting of infomediaries to publish what public organizations produce (and tools to make this really easy)	Pollock [17], Hall [19], Adner [41]

Table 1, continued

Ecosystems characteristics		Source
Data Resources and tools	Stakeholders involved in ecosystems are agencies and departments, suppliers and partners, internal elements, communities and individual stakeholders	Evans [40]
	Needs to include many types of stakeholders such as civil servants who curate raw data, developers who build applications with open data, and citizens who view visualizations and analytical results from these data.	Ding et al. [30]
	Consisting of data packaging and patching format	Pollock [17]
	Consisting of resources that work to provide a data commons	Sansone et al. [21]
	Consisting of publisher notification of patches (pull requests) with automated integration (merge) tools	Pollock [17]
	Address licensing	Halonen [53]
	Address Linked Data	Van der Sande et al. [18], Ding [54]
	May be costly and face certain risks	Adner [41]
Should include ICT preparedness, ICT usage, an enterprise architecture ICT model, Information provision capability, policy marketing, resource allocation and e-participation.	Obi [52]	
“A Linked Data-based system where stakeholders of different sizes and roles find, manage, archive, publish, reuse, integrate, mash-up, and consume open government data in connection with online tools, services and societies” [p. 326]	Ding et al. [30]	

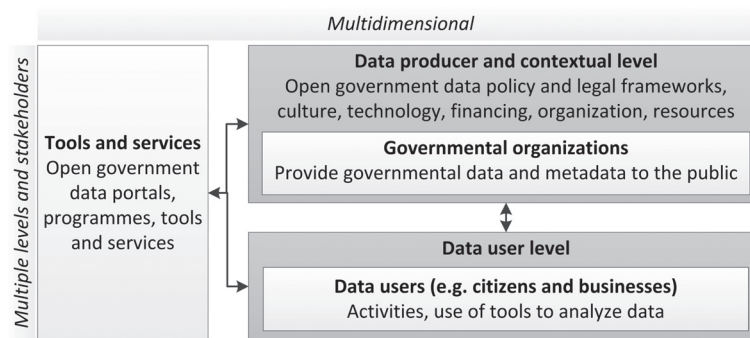


Fig. 1. Elements of an open government data ecosystem derived from the literature.

scenario to detail the elements. In this section we report on the findings of analyzing an open government data scenario, using the structure for describing scenarios that has been provided by Carrol [24]. The scenario meets our criterion that it should describe a diverse range of activities that are performed in the open data process. The scenario integrates steps of workflows for data as described in Section 2.2.

In the following, the actions and events performed in the scenario are described. After the description of each activity, a number is given which corresponds with the numbers shown in Fig. 2. Figure 2 represents the elements of an open government data ecosystem. On the various levels a number of activities are shown that can be conducted in the ecosystem, connected to the applications and tools that can be used to conduct these activities. The aim of this figure is not to provide a complete list of applications, but to provide examples of applications that the open data ecosystem can use.

The scenario resembles data lifecycle management, and in line with [28] the generation and use of data and metadata is essential. In our scenario of open data publication and use various levels are represented. On the first level, a public sector organization creates or collects data (1) and stores this data internally

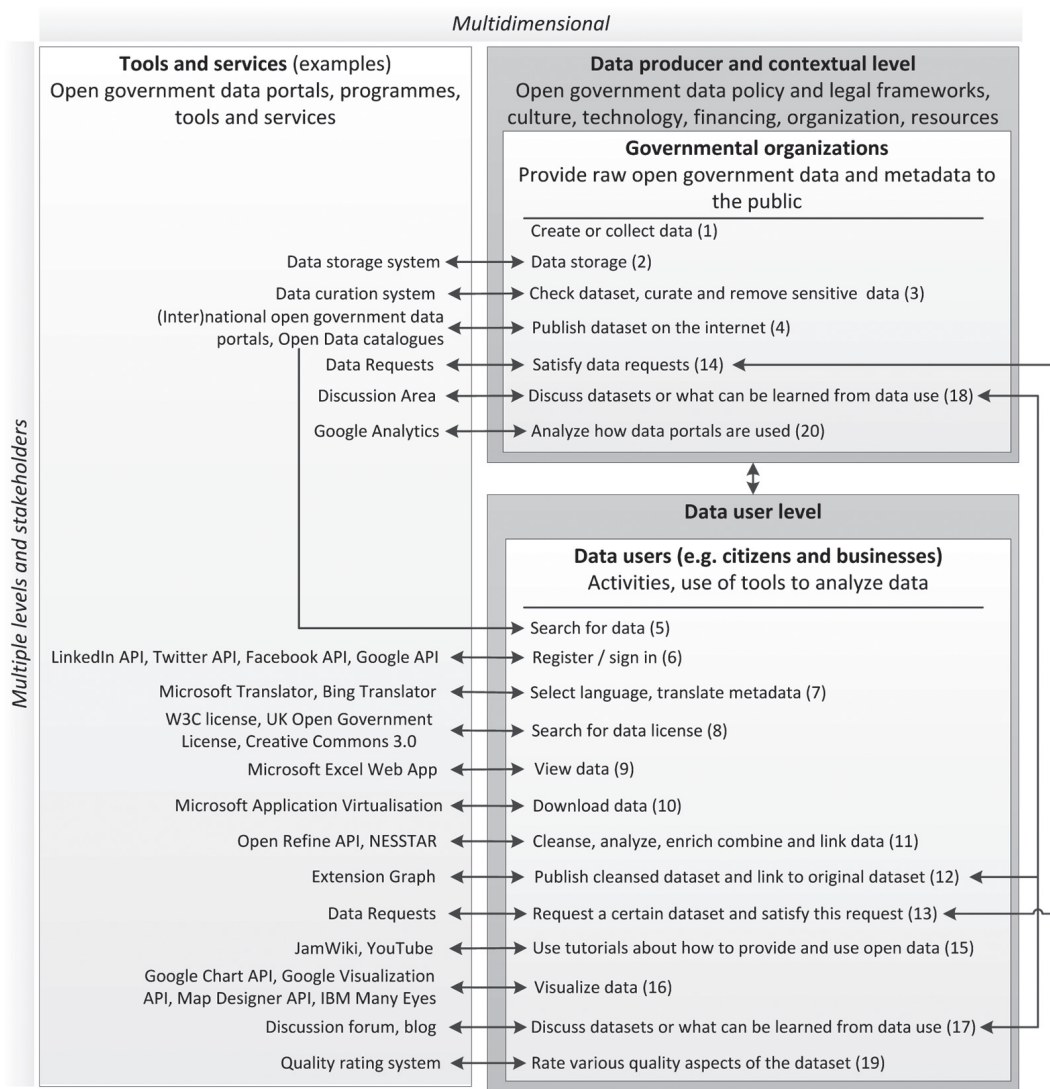


Fig. 2. Elements of an open government data ecosystem derived from the literature and the scenario.

(2). If the public sector organization decides to make this data available on the internet, the data needs to be checked, curated and sensitive data needs to be removed (3). Thereafter the public sector organization could decide to publish this data on the internet by using a data publication system and a certain open data portal (4).

After the government organization publishes its data on the internet, the published data and metadata becomes available for data end-users. On the data user level a researcher searches for open data, for example, about poverty levels and health facilities in Kenya (5). Various portals can be used to search for these open data, such as national open data portals, the European open data portal, the open data portal of the United States and open data catalogues. The data user level is connected to the data provider level, showing their interaction and interdependence. For instance, the researcher on the data user level may choose one open data portal on the data producer level and register and sign in on the portal (6).

Infomediaries can provide social media Application Programming Interfaces (APIs) to register and sign in, so that the researcher can use his user name and password of one of his social media accounts. The researcher may want to change the language that is used on the open data portal, for instance by choosing another language in the Microsoft Translator (7). The researcher wonders what the license of the dataset is and what kind of legal framework is applicable, and he searches for the license that is applicable (8). He then views a dataset in XLS format by using the Microsoft Excel Web App (9) and he decides to download the data by using the Microsoft Application Virtualization (10). The researcher can use the Open Refine API and other programs to cleanse, analyze, enrich, combine and link datasets (11). A dataset that has been extended by cleansing, analyzing, enriching, combining or linking it, can be saved so that other users can also get the value from the use of the data. This means that other users can then use the extended processed dataset, which is ideally linked to the original dataset so that it can be found easily (12). The researcher still has not gathered sufficient information and he decides to request a dataset about health facilities in Kenya using a data request system. Other users (13) or governmental organizations (14) can satisfy the request by providing the data or by providing links to the data. As the researcher wants to know more about how he can visualize the data that other users provided him with, he can use tutorials to obtain more information. Tutorials can be published, for example, using the JamWiki or YouTube (15). After having gathered sufficient information from the tutorial about how he can visualize a dataset, the researcher uses one of the visualization APIs to visualize his dataset (16). He maps both the poverty levels and the health facility levels on one map and by doing this he obtains an interesting overview of which areas have the highest need for medical aid. The researcher writes down his conclusions about Kenya's areas which are in highest need on a blog or discussion forum which is related to the dataset (17). The data provider may also be interested to know such a discussion, resulting in another interdependence of the various levels in the open data ecosystem (18). Subsequently, the researcher rates several quality aspects of the dataset using a quality rating system (19). Finally, on the data provider level the open data provider can use Google Analytics to obtain some information about how the open data portal is used (20).

The scenario described above provided an example of which activities can be performed when publishing and using open data and provided more detailed elements that can be integrated in the open government data ecosystem. Figure 2 shows how the elements of an open government data ecosystem derived from the literature are complemented by elements derived from the scenario. On the right hand side the generic elements are depicted, whereas on the left hand side examples of tools and services are shown. Note that persons and organizations in this figure can have multiple roles, and that activities can be conducted in various orders and with multiple feedback loops. For instance, a data provider could also be a data user, and after viewing or visualizing data one could find out that other (additional) data are needed and go back to the activity of search for data. Most of the activities depicted in Fig. 2 can also be conducted by infomediaries. For example, data collection, storage and visualization can be done by an infomediary. Because infomediaries can be present on many different levels and in many different dimensions in the open data ecosystem as depicted in Fig. 2, we did not create a separate infomediaries level.

5. Findings and discussion

In the previous section twenty essential elements of open data ecosystems were identified. In this section we elaborate on the findings from the scenario analysis and discuss those. The literature and the scenario showed that open data ecosystems involve multiple levels, namely a data producer and a

data user level. Within these levels there are various activities that can be conducted, such as searching, requesting, viewing, downloading, analyzing, cleansing, combining, visualizing and discussing different types of data. For instance, an open data user can search for data, then filter the results and then zoom in on the dataset and then analyze it in detail. All these elements are interdependent. For instance, it is not possible to analyze a dataset if one has not found this dataset in advance. Another example is that certain activities, such as visualization, may not be possible without consulting a tutorial and data use is not possible without data publication.

From this research we can identify four key elements that open data ecosystems should capture, namely, 1) releasing and publishing open data on the internet, 2) searching, finding, evaluating and viewing data and their related licenses, 3) cleansing, analyzing, enriching, combining, linking and visualizing data and 4) interpreting and discussing data and providing feedback to the data provider and other stakeholders. These elements will be discussed in the following. First, an open data ecosystem should capture the release and publication of data on the internet. Governmental organizations need to release their datasets to facilitate their reuse, and data users are dependent on this step.

Second, an open data ecosystem should facilitate searching, finding, evaluating and viewing data and obtaining information about the licenses related to the data. There can be many different ways in which people may search for the same type of data which is often a non-linear process containing many feedback loops. For instance, an open data user may search for all data about a certain city, which may be a difficult task as these data are tagged as having to do with many different topics, such as employment, housing, roads, finance, and other topics. For this reason the open data user should evaluate each dataset that he finds, decide whether the dataset concerns his scope and search topic, decide whether the format and quality of the data are sufficient, and then decide whether he should search on for another dataset or use this dataset. This process could be repeated many times.

Data can also be browsed, searched or obtained by requesting them. The scenario description showed that notifications of requests for data can be used to request datasets, so that open data users can be supported with obtaining this data that they want to use. Both raw data and processed data can be found in open data ecosystems. While some open data users, or sometimes called intermediaries or infomediaries, process the raw data, other users may prefer to make use of the processed data or data services and tools that are derived from these raw data. Thus, the interaction between infomediaries as intermediate consumers of data and open data users should be facilitated. Furthermore, there need to be interactions with data providers, as users can request data from data providers and provide them with feedback after they have used the data. Information about the licenses that are connected to the use of certain datasets is also important in open data ecosystems, as open data users need to know whether the license allows them to use the data in the way that they want to use them.

A third key element of open data ecosystems is cleansing, analyzing, enriching, combining, linking and visualizing data. Data cleansing refers to the detection and correction of records in a dataset [55]. Data cleansing could be a goal on itself, but is often performed to make it easier to use the dataset in another way, for instance by analyzing it. Analyzing a dataset could merely mean reading (i.e. looking at) the dataset and deriving useful information from this activity. It could also mean performing a thorough statistical analysis by using software (e.g. SPSS Statistics). An analysis of a dataset should lead to new insights and understanding of the data, possibly by analyzing data in a way that was not done before. Datasets can also be enriched in several ways. For instance, a user could annotate a dataset by describing what he or she experienced when using the data or which information other users should take into account when using the data. A dataset can also be enriched by adding information that was derived from the statistical analysis or visualization. Visualizations often provide much insight into a dataset.

Another important way of using open data is by combining data or linking them to other data, as this reveals relationships between data [56].

Fourth, an open data ecosystem needs to enable interpreting and discussing data and providing feedback to the data provider and other stakeholders. One way to enable this is by allowing for the discussion of datasets, so that data providers and other data users can discuss what can be learned from the use of the data. The results derived from the use of open data can also be shared with organizations and persons to whom this might be relevant, for instance, by using social media. A person who has used a dataset could also leave a message on an open data platform to inform other users of the dataset about certain particularities of the dataset or post a question about the data. Moreover, this can be implemented by enabling users to send each other messages about datasets, and by connecting examples of the data use (e.g. applications, visualizations and publications) to the dataset.

The scenario revealed that various tools and applications can be integrated to make data use easier and to support the four key elements of open data ecosystems. Various APIs, such as Twitter, LinkedIn, Google and Facebook, can be combined and integrated in open data ecosystems. The scenario also showed that tools such as the Microsoft Excel Web App, Open Refine, Microsoft Translator, Map Designer, JamWiki and other tools can be used to make the use of open data easier. In an open data ecosystem, these tools and applications can be combined. Open data portals can complement each other by providing open interfaces, in this way stimulating the provision and use of open data and the realization of their advantages. As a consequence, an open ecosystem is created in which the best functionalities can be utilized by open data users.

Furthermore, to integrate the ecosystem elements and to let them act as an integrated whole, there should be three additional elements, namely 5) user pathways showing directions for how open data can be used, 6) a quality management system and 7) different types of metadata for being able to connect the elements. First, user pathways can show directions for how open data can be used. Pathways make it easier to know in advance how data can be used. Second, a quality management system enables assessing the quality of the data. This system could help open data users in determining whether the quality of a dataset is satisfactory to fulfill their objectives for using the open dataset. Third, the integration of the ecosystem elements requires different types of metadata for ensuring interoperability and data processing. Discovery, contextual and detailed types of metadata need to be combined to improve the provision and use of open data [57]. Among many examples, metadata may improve storing, preservation, accessibility, visualization and multilinguality of open data. Moreover, they can improve the ability to find and interpret open data, create order within datasets, stimulate analyzing, comparing and reproducing open data, enable assessing and ranking the quality of open data, make linking data easier, avoid the unnecessary duplication of linked open data and improve visibility of and collaboration among researchers [57–59]. Interoperability in an open data ecosystem can be improved by providing metadata about shared identifiers and vocabularies and by reusing ontological elements. The best functionalities of the ecosystem can be selected and utilized by open data users.

6. Conclusions

There is only little insight in the essential elements of open data ecosystems to enable the publication and use of open data. Therefore, this paper aimed to contribute to filling this gap. Our literature review provided information about a number of essential elements of open data ecosystems. It was found that an open data ecosystem is characterized by multiple interdependent socio-technical levels, dimensions, actors (including data providers, infomediaries and users), elements and components. Moreover,

we found that open data ecosystems need to address challenges related to policy, licenses, technology, financing, organization, culture, and legal frameworks and are influenced by ICT infrastructures. Open data ecosystems develop through user adaptation, feedback loops and dynamic supplier and user interactions and other interacting factors in the field of open data. Nevertheless, the essential elements of open data ecosystems that were derived from the literature were mainly on a high and abstract level and it was concluded that a more detailed overview of elements was needed to create an open data ecosystem. In addition, the literature review showed that research on open data ecosystems is fragmented and not uniform, as most research only focuses on one or a few activities related to the publication or use of open data, but not on the whole of activities that should be supported in an open data ecosystem.

Therefore, we used a scenario to further detail the open data ecosystem elements that were found in the literature. The scenario showed that the use of open data is often a complicated process and that many different instruments and tools can be used and are available. The open data ecosystem that we developed in this paper consists of twenty activities that can be performed in the process in which open data is published and used, linked to various examples of interoperable applications that can be used on open data portals. The research showed that open data ecosystems should capture the following four key elements.

- 1) Releasing and publishing open data on the internet. Governmental organizations first need to provide the data.
- 2) Searching, finding, evaluating and viewing data and their related licenses. In an open data ecosystem, open data users should be able to search for data or obtain it by request. The ecosystem should make it possible to find and view both raw data and processed data. The use of open data should be made easier by providing information about the licenses that are connected to the use of the data.
- 3) Cleansing, analyzing, enriching, combining, linking and visualizing data. The ecosystem should make it possible to detect and correct records in a dataset, to (statistically) analyze data and to enrich datasets by adding useful information or data about the dataset or the use of the dataset. Moreover, visualization tools and linkage tools in the ecosystem should support the interpretation of data and reveal relationships between datasets.
- 4) Interpreting and discussing data and providing feedback to the data provider and other stakeholders. An open data ecosystem needs to enable the provision of feedback and the discussion of datasets, so that data providers and other data users can learn from this and obtain the value of data use.

Furthermore, to integrate the ecosystem elements and to let them act as an integrated whole, there should be three additional elements.

- 5) User pathways showing directions for how open data can be used.
- 6) A quality management system which enables assessing the quality of the data. This system could help open data users in determining whether the quality of a dataset is satisfactory to fulfill their objectives for using the open dataset.
- 7) Different types of metadata for ensuring interoperability and processing and for connecting the elements. Metadata can be exchanged among various open data systems. This may help improving, among others, finding, accessing and preserving data, as different open data infrastructures can be connected. This paper showed that metadata may also provide many other advantages.

An ecosystem makes it possible for open data users to select and utilize the best functionalities of this ecosystem. The various open data instruments and tools complement each other (elements 1–4) and the three additional elements (elements 5–7) ensure that the single components can be integrated, in this way enabling interoperability. We recommend to develop these elements in further research to ensure

that single elements can be integrated. By integrating the elements it should become easier to use the single tools and instruments and this should lead to more user adoption of open data and increased citizen engagement, which is crucial to stimulate innovation in ecosystems for open data. It is recommended to investigate this in further research.

Open data case studies are hardly conducted and much work still has to be done in this area. The elements identified in this paper can be extended and further detailed and refined, as this paper is only based on a single scenario. Additionally, future research could focus on scrutinizing the tools and applications that can be used in open data ecosystems. Furthermore the key to an ecosystem is the integration of the various parts which is hardly addressed to date. By doing this, a more comprehensive overview of tools and applications of open data ecosystems can be created.

Acknowledgements

This paper is related to the ENGAGE FP7 Infrastructure Project. More information can be found at www.engage-project.eu and www.engagedata.eu. The authors would like to thank their colleagues of the ENGAGE project for their input for this paper, although the views expressed are the views of the authors and not necessarily of the project.

References

- [1] J. Zhang, S. Dawes and J. Sarkis, Exploring Stakeholders' Expectations of the Benefits and Barriers of E-Government Knowledge Sharing, *Journal of Enterprise Information Management* **18** (2005), 548–567.
- [2] S. Dawes and N. Helbig, Information strategies for open government: Challenges and prospects for deriving public value from government transparency, presented at the 9th International Conference on e-government (EGOV), Lausanne, Switzerland, 2010.
- [3] J.C. Bertot, P.T. Jaeger and J.M. Grimes, Using ICTs to Create a Culture of Transparency: E-government and Social Media as Openness and Anti-Corruption Tools for Societies, *Government Information Quarterly* **27** (2010), 264–271.
- [4] C.P. Geiger and J.V. Lucke, Open Government and (Linked) (Open) (Government) (Data), *Journal of e-Democracy and Open Government* **4** (2012), 265–278.
- [5] M.T. Borzacchiello and M. Craglia, The impact on innovation of open access to spatial environmental information: A research strategy, *International Journal of Technology Management* **60** (2012), 114–129.
- [6] P. McDermott, Building Open Government, *Government Information Quarterly* **27** (2010), 401–413.
- [7] K. Braunschweig, J. Eberius, M. Thiele and W. Lehner, The State of Open Data. Limits of Current Open Data Platforms, presented at the International World Wide Web Conference, Lyon, France, 2012.
- [8] European Commission. (2012, January 8). *Data portal. The Open Data Hub of the European Union*. Available: <http://open-data.europa.eu/open-data/>.
- [9] The ENGAGE project. (2014, February 28). *ENGAGE*. Available: www.engagedata.eu.
- [10] Junar. (2009, January 8). *Junar. The open data platform*. Available: <http://www.junar.com/>.
- [11] European Commission. (2013, January 8). *Eurostat. Your key to European statistics*. Available: <http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home>.
- [12] M. Janssen, S.A. Chun and J.R. Gil-Garcia, Building the next generation of digital government infrastructures, *Government Information Quarterly* **26** (2009), 233–237.
- [13] Open Refine. (2013, October 11). *OpenRefine. A free, open source, power tool for working with messy data*. Available: <http://openrefine.org/>.
- [14] ScraperWiki. (2013, October 11). *ScraperWiki*. Available: <https://scraperwiki.com/>.
- [15] B. Ubaldi, Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives, OECD2013.
- [16] T.M. Harrison, T.A. Pardo and M. Cook, Creating Open Government Ecosystems: A Research and Development Agenda, *Future Internet* **4** (2012), 900–928.
- [17] R. Pollock. (2011, September 13). *Building the (Open) Data Ecosystem*. Available: <http://blog.okfn.org/2011/03/31/building-the-open-data-ecosystem/>.

- [18] M. Vander Sande, A. Dimou, P. Colpaert, E. Mannens and R. Van de Walle, Linked Data as enabler for Open Data Ecosystems, presented at the Open Data Week 2013 (ODW13), 2013.
- [19] W. Hall, N. Shadbolt, T. Tiropanis, K. O'Hara and T. Davies, Open data and charities. A state of the art review, Nominet Trust, Oxford 2012.
- [20] A. Poikola, P. Kola and K.A. Hintikka, Public data. An introduction to opening information sources, Ministry of Transport and Communications, Helsinki, Finland 2011.
- [21] S.-A. Sansone, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann et al., Toward interoperable bioscience data, *Nature Genetics* **44** (2012), 121–126.
- [22] T. Davies, Open Data: Infrastructures and ecosystems, presented at the ACM Web Science Conference 2012, Evanston, Illinois, 2012.
- [23] T. Davies and D. Edwards, Emerging Implications of Open and Linked Data for Knowledge Sharing in Development, *IDS Bulletin* **43** (2012), 117–127.
- [24] J.M. Carroll, Five Reasons for Scenario-Based Design, presented at the 32nd Hawaii International Conference on System Sciences (HICSS), Wailea Maui, Hawaii, 1999.
- [25] C. Potts, Using schematic scenarios to understand user needs, presented at the Proceedings of the 1st conference on Designing interactive systems: processes, practices, methods, & techniques, Ann Arbor, Michigan, USA, 1995.
- [26] Y.-K. Lim and K. Sato, Scenarios for Usability Evaluation: Using Design Information Framework and a Task Analysis Approach, in *The Proceedings of the International Ergonomics Association 15th Technical Congress*, 2003.
- [27] Statistics Netherlands (CBS). (2013, September 13). *StatLine Database*. Available: <http://www.cbs.nl/en-GB/menu/cijfers/statline/zelf-tabellen-maken/default.htm?Languageswitch=on>.
- [28] K. Möller, Lifecycle models of data-centric systems and domains. The abstract data lifecycle model, *Semantic Web* **4** (2013), 67–88.
- [29] A. Zuiderwijk, M. Janssen, S. Choenni, R. Meijer and R. Sheikh Alibaks, Socio-technical impediments of open data, *Electronic Journal of eGovernment* **10** (2012), 156–172.
- [30] L. Ding, T. Lebo, J.S. Erickson, D. DiFranzo, G.T. Williams, X. Li et al., TWC LOGD: A portal for linked open government data ecosystems, *Web Semantics: Science, Services and Agents on the World Wide Web* **9** (2011), 325–333.
- [31] R. Wirth and J. Hipp, CRISP-DM: Towards a standard process model for data mining, in *4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 2000, pp. 29–39.
- [32] B.J. Fry, (2004, March 16). *Computational Information Design*. Available: <http://dspace.mit.edu/handle/1721.1/26913>.
- [33] P. Pirolli and S. Card, The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis, in *International Conference on Intelligence Analysis*, 2005, pp. 2–4.
- [34] R. Adner and R. Kapoor, Value Creation in Innovation Ecosystems: How the Structure of Technological Interdependence Affects Firm Performance in New Technology Generations, *Strategic Management Journal* **31** (2010), 306–333.
- [35] M. Iansiti and R. Levien, *The Keystone Advantage: What the New Dynamics of Business Ecosystems Mean for Strategy, Innovation and Sustainability*. Boston: Harvard Business School Press, 2004
- [36] R.P. Bostrom and J.S. Heinen, MIS problems and failures: A socio-technical perspective. Part I: The causes, *MIS Quarterly* **1** (1977), 17–32.
- [37] G. Vonk and A. Ligtenberg, Socio-technical PSS development to improve functionality and usability – Sketch planning using a Maptable, *Landscape and Urban Planning* **94** (2010), 166–174.
- [38] A. Zuiderwijk and M. Janssen, Barriers and development directions for the publication and usage of open data: A socio-technical view, in: *Open Government. Opportunities and Challenges for Public Governance*, M. Gascó-Hernández, Ed., ed New York: Springer, 2014, pp. 115–135.
- [39] T.M. Koontz and J. Bodine, Implementing Ecosystem Management in Public Agencies: Lessons from the U.S. Bureau of Land Management and the Forest Service, *Conservation Biology* **22** (2008), 60–69.
- [40] G. Evans, *Implementing E-government: An Executive Report for Civil Servants and Their Advisors*. Burlington, U.S.A.: Gower Publishing Limited, 2003.
- [41] R. Adner, Match your innovation strategy to your innovation ecosystem, *Harvard Business Review* **84** (2006), 98–107.
- [42] J.F. Moore, *The Death of Competition – Leadership and Strategy in the Age of Business Ecosystems*, New York: Harper Business, 1996.
- [43] L. Pierce, Big losses in ecosystem niches: How core firm decisions drive complementary product shakeouts, *Strategic Management Journal* **30** (2009), 323–347.
- [44] T. Obi and N. Iwasaki, Electronic governance benchmarking: Waseda University e-gov ranking, presented at the Proceedings of the 4th International Conference on Theory and Practice of Electronic Governance, Beijing, China, 2010.
- [45] B.A. Nardi and V.L. O'Day, *Information Ecologies: Using Technology with Heart*, Cambridge, MA, USA: MIT Press, 1999.
- [46] L. Weizhang and Z. Jiefang, Development of E-Government from Information Ecology View: A Literature Review, in *E-Business and E-Government (ICEE), 2010 International Conference on*, 2010, pp. 476–479.

- [47] M.L. Brodie, Data Integration at Scale: From Relational Data Integration to Information Ecosystems, presented at the 24th IEEE International Conference on Advanced Information Networking and Applications, 2010.
- [48] A.B.H.S. Mhamdia, Performance measurement practices in software ecosystem, *International Journal of Productivity and Performance Management* **62** (2013), 514–533.
- [49] M. Hadzic and A. Sidhu, Digital health ecosystems, in *2nd IEEE International Conference on Digital Ecosystems and Technologies*, 2008.
- [50] G. Briscoe and P. De Wilde, Digital ecosystems: evolving service-orientated architectures, presented at the First international conference on Bio inspired models of network, information and computing systems, 2006.
- [51] L. Abrahams, Innovation in monitoring and evaluation for e-development and transformational government.
- [52] T. Obi, E-Government Developments, *Journal of e-Government Policy and Regulation* **32** (2009), 136–187.
- [53] A. Halonen, Being Open About Data. Analysis of the UK open data policies and applicability of open data, The Finnish Institute, London, United Kingdom 2012.
- [54] L. Ding, V. Peristeras and M. Hausenblas, Linked Open Government Data [Guest editors' introduction], *Intelligent Systems, IEEE* **27** (2012), 11–15.
- [55] A. Yaeli, H. Ship, C. Alexopoulos, V. Pterneas and N. Gaberlo, Deliverable D7.7.3. Data and knowledge annotation and linking methods, 2012.
- [56] T. Berners-Lee. (2009, October 11). *Linked data*. Available: <http://www.w3.org/DesignIssues/LinkedData.html>.
- [57] A. Zuiderwijk, K. Jeffery and M. Janssen, The potential of metadata for linked open data and its value for users and publishers, *Journal of e-Democracy and Open Government* **4** (2012), 222–244.
- [58] S. Dawes, Stewardship and usefulness: Policy principles for Information-Based Transparency, *Government Information Quarterly* **27** (2010), 377–383.
- [59] K.G. Jeffery, Metadata: The future of information systems, in: *Information Systems Engineering: State of the art and research themes*, J. Brinkkemper, E. Lindencrona and A. Sølvsberg, eds, London: Springer Verlag, 2000.